UNIVERSIDADE FEDERAL DO AMAZONAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

A HYBRID GENE SELECTION METHOD BASED ON OUTLIERS FOR BREAST CANCER CLASSIFICATION

Manaus - AM

Julho de 2023

RAYOL DE MENDONÇA NETO

A HYBRID GENE SELECTION METHOD BASED ON OUTLIERS FOR BREAST CANCER CLASSIFICATION

Thesis presented to the Graduate Program in Informatics of the Federal University of Amazonas in partial fulfillment of the requirements for the degree of Doctor in Informatics.

Advisor: Eduardo Freire Nakamura

Manaus - AM

July 2023

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Mendonça Neto, Rayol de
M539h   A hybrid gene selection method based on outliers for breast cancer classification / Rayol de Mendonça Neto . 2023
99 f.: il. color; 31 cm.

Orientador: Eduardo Freire Nakamura
Tese (Doutorado em Informática) - Universidade Federal do Amazonas.

1. Gene expression. 2. Outlier genes. 3. Breast cancer. 4. Explainable AI. I. Nakamura, Eduardo Freire. II. Universidade Federal do Amazonas III. Título

PODER EXECUTIVO
MINISTÉRIO DA EDUCAÇÃO
INSTITUTO DE COMPUTAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

UFAM

FOLHA DE APROVAÇÃO

"A hybrid gene selection method based on outliers for breast cancer classification"

RAYOL DE MENDONÇA NETO

Tese de Doutorado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Dr. Eduardo Freire Nakamura - PRESIDENTE

Prof. Dr. David Fenyö - MEMBRO EXTERNO

Claudio T. Silva

Prof. Dr. Claudio T. Silva - MEMBRO EXTERNO

Isabelle Bezerra Cordeiro

Profa. Dra.Isabelle Bezerra Cordeiro - MEMBRO EXTERNO

Prof. Dr. Eduardo James Pereira Souto- MEMBRO INTERNO

Manaus, 07 de julho de 2023.

Dedico essa tese aos meus pais e irmão por todo amor e convivência.
À minha esposa e meu filho, minha razão de viver.

Acknowledgments

Agradeço primeiramente à Deus.

Agradeço aos meus pais e minha avó que sempre me apoiaram e me incentivaram a buscar minhas vitórias por meio do estudo e trabalho.

Agradeço à todos os meu familiares, meu irmão Matheus, sogros, tios e primos, que torceram por mim ao longo desses anos.

Meus agradecimentos à minha esposa Luciana Mendonça, minha companheira e maior incentivadora, que sempre me apoiou em todas as decisões que tomei e que em momento algum deixou de caminhar lado a lado comigo.

Agradeço ao meu amado filho Thomás Mendonça, que chegou para alegrar a minha vida e tornar todos os momentos mais prazerosos.

Agradeço também o Professor Eduardo Freire Nakamura, que com toda seu conhecimento e habilidade topou entrar nessa jornada numa nova área de conhecimento e soube me guiar com maestria.

Agradeço à Professora Fábiola Nakamura por todo o acompanhamento e disponibilidade para ajudar.

Agradeço à todos os Professores do Doutorado que de alguma forma contribuíram para a minha formação.

Agradeço também aos professores David Fenyö e Claudio Silva que proporcinaram uma oportunidade de estudo única na Universidade de Nova Iorque.

Meus agradecimentos aos companheiros de PPGI, Leandro Okimoto, Hendrio Bragança, Ariel Afonso, Victória Aires, Juan Collona, Moysés Lima e tantos outros que deixei de citar, mas percorreram essa jornada junto comigo.

Agradeço também aos funcionários do ICOMP, pela disponibilidade, simpatia e gentileza.

"Comparison is the thief of joy."

(Theodore Roosevelt)

Resumo

O câncer de mama é o segundo tipo de câncer mais comum e a principal causa de mortes entre mulheres em todo o mundo. Por se tratar de uma doença heterogênea, a subtipagem do câncer de mama desempenha um papel importante na realização de um tratamento específico. Os dados de expressão gênica são uma alternativa viável para serem empregados na classificação de subtipos de câncer, pois representam o estado de uma célula em nível molecular, mas geralmente possuem um número relativamente pequeno de amostras em comparação a um grande número de genes. A seleção de genes é uma abordagem que lida com essa matriz de alta dimensão de genes contra amostras, e desempenha um papel importante na classificação eficiente de subtipos de câncer. Nesta tese, um método híbrido inovador de seleção de genes com base em *outliers* (H-OGS) é proposto para selecionar genes relevantes para classificar de forma eficiente e eficaz os subtipos de câncer de mama, e para identificar assinaturas distintas capazes de caracterizar subtipos de câncer de mama. Então, as associações aprendidas pelo classificador empregado nesse método são interpretadas localmente por SHAP Values revelando genes que são biologicamente relevantes para a classsificação de cada subtipo de câncer de mama. Em geral, nosso método seleciona apenas alguns genes altamente relevantes, acelerando a classificação e melhorando significativamente o desempenho do classificador. Experimentos mostram que nossa estratégia apresenta os melhores resultados para os subtipos Basal e Her 2, os dois subtipos de câncer de mama com os piores prognósticos, respectivamente. Nosso método também identifica três assinaturas distintas que caracterizam o subtipo basal, onde essas assinaturas possuem genes e *pathways* diretamente relacionados aos subtipos de câncer de mama. Nós também propomos um framework de avaliação que utiliza diferentes técnicas de aprendizado de máquina para uma análise mais ampla da lista PAM50 na classificação de subtipos de câncer de mama. Os experimentos mostram que o melhor método a ser utilizado na classificação dos subtipos de câncer de mama é o SVM com kernel linear.

Palavras-chave: Expressão Gênica, Genes Outlier, Câncer de Mama, eXplainable AI.

Abstract

Breast cancer is the second most common cancer type and is the leading cause of cancer-related deaths worldwide among women. Since it is a heterogeneous disease, subtyping breast cancer plays an important role in performing a specific treatment. Gene expression data is a viable alternative to be employed on cancer subtype classification, as they represent the state of a cell at the molecular level; but generally has a relatively small number of samples compared to a large number of genes. Gene selection is a promising approach to address this uneven high-dimensional matrix of genes versus samples and plays a major role in developing efficient cancer subtype classification. In this thesis, an innovative hybrid gene selection method based on *outliers* (H-OGS) is proposed to select relevant genes to efficiently and effectively classify breast cancer subtypes, and to identify distinct signatures capable of to characterize breast cancer subtypes. Then, the associations learned by the classifier employed in this method are interpreted locally by SHAP Values revealing genes that are biologically relevant for the classification of each subtype of breast cancer. In general, our method selects only a few highly relevant genes, speeding up the classification and significantly improving the classifier's performance. Experiments show that our strategy gives the best results for Basal and Her 2 subtypes, the two breast cancer subtypes with the worst prognosis, respectively. Our method also identifies three distinct signatures that characterize the basal subtype, where these signatures have genes and *pathways* directly related to breast cancer subtypes. We also propose an evaluation framework that uses different machine learning techniques for a broader analysis of the PAM50 list in the classification of breast cancer subtypes. The experiments show that the best method to classify breast cancer subtypes is the SVM with linear kernel.

Keywords: Gene Expression, Outlier Genes, Breast Cancer, eXplainable AI.

List of Figures

List of Tables

Contents

1

Introduction

Breast cancer is the second most common cancer type and is the leading cause of cancer-related deaths worldwide [Bray et al., 2018, Jemal et al., 2011]. As a highly heterogeneous disease [Miah et al., 2017], breast cancer shows distinct genetic variations, clinical outcomes, and treatment strategies between tumor subtypes [Chen et al., 2016].

Breast cancer has four major molecular subtypes [Perou et al., 2000]: Basal, Her 2, Luminal A, and Luminal B. Basal and Her 2 are the subtypes with the worst prognoses, respectively [Bertucci et al., 2012, Dwivedi et al., 2019], while Luminal A and Luminal B have better prognosis since there are effective targeted therapies for them [Dwivedi et al., 2019, Yersal and Barutca, 2014].

Cancer classification is widely used to identify cancer samples subtypes, as it can provide an efficient, accurate, and objective diagnosis for different types of cancer [Tarek et al., 2017, Tong et al., 2013]. Diagnosing cancer by biologic (or “intrinsic”) subtype adds significant prognostic and predictive information for patients with breast cancer [Ginsburg et al., 2020, Parker et al., 2009]. Therefore, classifying breast cancer into its subtypes and finding important genes related to it is crucial to properly and effectively treat patients.

Recent advances in DNA microarray technology allowed monitoring of the expression levels of thousands of genes simultaneously during important biological processes and across collections of related samples [Almugren and Alshamlan, 2019, Jiang et al., 2004], resulting in gene expression data. Therefore, presenting a viable alternative to employ on cancer classification.

One of the challenges when using gene expression data for cancer classification is that those data are represented by complex and high-dimensional matrices of genes versus samples. In general, the number of samples is much smaller than the number of genes (e.g., +10,000 genes for each sample in a set of at most a few hundred samples) [Piatetsky-Shapiro and Tamayo, 2003].

The cancer classification through gene expression data can be divided into binary classification (cancer/not cancer) [Ghorai et al., 2010, Liu and San Wong, 2017, Mostavi et al., 2020] and multiclass classification (subtypes) [Mostavi et al., 2020, Rajapakse and Mundra, 2013]. The multiclass problem is harder to solve due to the fact that the samples may contain similar histopathological appearance but can follow significantly different clinical outcomes [Dai et al., 2015].

In the literature, along with the cancer classification task, feature selection methods are also widely utilized. Whereas, by successfully removing irrelevant and redundant genes, we can reduce the dimensionality of the data, simplify the learning model, speed up the learning process, and significantly improve the classification performance [Alanni et al., 2019, Díaz-Uriarte and De Andres, 2006, Liu and San Wong, 2017].

To narrow this vast number of genes, researchers have extensively studied how these large datasets are often affected by outliers (uncommonly under-expressed or over-expressed) [Blumenberg et al., 2019, Mertins et al., 2016]. Since outliers can be genes that are functionally different from the majority of the population, they must be analyzed to find if they belong to a rare cell type, a functionally distinct group of cells, or even if they are errors in the experimental procedure [Shetta and Niranjan, 2020].

Nevertheless, despite yielding good results, most of these classification and gene selection techniques output opaque results and cannot explain how they arrive at specific decisions (which is known as the "black box" problem) [Castelvecchi, 2016]. For scientists to trust the results, it is crucial first to understand what machines do; since in many cases, it is not so much what an algorithm predicts but the relationships it establishes and how it predicts it that matters the most [Anguita-Ruiz et al., 2020].

In this sense, there is an increased need to provide machine learning (ML) models with more interpretability and explicability, which originated in what is known as eXplainable Artificial Intelligence (XAI). As one of the most employed interpretable techniques, SHapley Additive exPlanations SHAP has become a highly relevant technique within the XAI expansion, being able to generate practical knowledge understandable from the point of view of human experts [Anguita-Ruiz et al., 2020].

EXplainable Artificial Intelligence has been successfully applied to gene expression data (GED) in order to represent how the expression of one or several genes may be linked or associated with the expression of a different set of genes [Alves et al., 2010].

In this thesis, we propose the **H**ybrid **O**utlier-based **G**ene **S**election (H-OGS) method to determine the gene sets related to each breast cancer subtype. The method combines outlier detection techniques and feature elimination methods to find a small gene set capable of achieving high classification results. Our method presents promising results compared to recent approaches, with higher F_1 score when classifying the breast cancer subtypes with the worst prognoses.

Another contribution of our outlier gene selection method is the identification of gene signatures (gene sets) related to tumor biology for breast cancer subtypes. Results show that H-OGS is a viable solution in the sample subtype clustering. Results show that we were able to identify distinct signatures for the same subtype. Also, when compared to largely studied gene sets, such as the PAM50 (a well known 50-gene subtype predictor [Parker et al., 2009]), our proposed method finds signatures with less than half of the genes presented in the PAM50.

We also propose an evaluation framework that uses different machine learning techniques to classify breast cancer subtypes and investigate the features. This framework performs an analysis of the PAM50 list in the classification of breast cancer subtypes and identifies a list of genes that are important for the classification of each subtype.

Furthermore, we perform a biological analysis of the outlier genes using eXplainable AI. Our method is capable of identifying the most relevant genes for the classification of each subtype and establishing a biological association of the gene with the classified subtype.

1.1 Motivation

Cancer classification has been studied in the context of gene expression data [Yip et al., 2011], the majority of the cancer classification approaches are suited for binary classification (cancer/not cancer). To employ a feasible cancer classification method, gene selection methods are imperative to improve classification [Shukla et al., 2018], because of the high dimensionality of the data; simplify the machine learning method, speed up the classification, and significantly improve the performance of the classifier [Alanni et al., 2019, Díaz-Uriarte and De Andres, 2006, Gatto et al., 2021].

However, these methods may not be appropriate for all breast cancer types. To improve the feature selection method for breast cancer subtype classification, it is important to select the most representative genes for each of the subtypes.

Therefore, for a better gene selection method for breast cancer subtypes, it is important to select the set of genes that are able to characterize the subtypes individually, traditional approaches found in the literature usually use one set of genes for all the subtypes (e.g., PAM50).

1.2 Research Hypothesis

Gene selection algorithms generally identify genes that are most relevant for classification, ignoring those genes that are outliers. Therefore, many approaches fail to select genes that may be relevant to the classification task.

Our hypothesis is that outlier genes can be used to properly classify different subtypes of breast cancer.

Thus, our hypothesis seeks to answer the following question: *"Is it possible to characterize the breast cancer subtypes using consistently outlier genes to provide better classification quality using only a small set of genes?"*

Based on the hypothesis, we developed the OGS family of algorithms, which is compared to the 1D-CNN [Mostavi et al., 2020] algorithm, capable of classifying breast cancer subtypes using neural networks. The choice for 1DCNN is justified by the similarity of the problem addressed. We also compared it with the PAM50 gene list. A 50 genes list widely used for breast cancer subtype identification. The experiments, as well as the analysis of the proposed gene selection method, are presented in Chapters 4 and 6.

1.3 Thesis Objectives

The main objective of this work is a novel gene selection method based on outliers, which is capable of classifying subtypes of breast cancer, achieving better results when compared to state-of-art.

The specific objectives include:

1. Find relevant genes that properly characterize the different subtypes of breast cancer;

2. Demonstrate the effectiveness of the proposed method, that is, a higher quality classifier using fewer genes than more traditional methods;

3. Determine the effectiveness of different methods in the task of classifying breast cancer subtypes;

4. Determine a list of genes that are important for the classification in each subtype.

5. Improve the effectiveness of the proposed method, by enhancing the quality of the subtype classification;
6. Demonstrate the biological relevance of the outlier genes for each breast cancer subtype.

1.4 Main Contributions

The methods developed from the execution of this thesis are described and discussed in Chapters 4, 5 and 6 where gene selection methods using outliers, classifier analysis for breast cancer subtypes and gene analysis using eXplainable Ai are presented. Several experiments were carried out to evaluate and validate our methods. The results demonstrate its effectiveness. We highlight in order of conception the contributions achieved in this thesis:

- A **H**ybrid **O**utlier-based **G**ene **S**election (H-OGS) method to determine the best gene sets related to each breast cancer subtype;
- The identification of gene signatures possibly related to tumor biology for breast cancer subtypes.
- A framework to evaluate traditional machine learning methods for breast cancer subtype classification using representative genes;
- A methodology to investigate how the associations learned by the classifier are interpreted locally by eXplainable AI, revealing the biological relation of the outlier genes within each subtype.

1.5 Thesis Outline

The remainder of this thesis is organized as follows In Chapter 2, Fundamentals, we explain basic concepts, such as: gene expression and proteogenomic data, moreover, outlier detection approaches, machine learning concepts, eXplainable AI methods, and evaluation metrics. In Chapter 3, Related work, the most common approaches for cancer classification, gene selection, and works related to eXplainable AI are discussed. In Chapter 4, we describe our proposed method for outlier gene selection, we also present the methodology used, finally we detail and describe the results of our experiments. In Chapter 5, we describe our evaluation framework that uses different machine learning techniques to classify breast cancer subtypes and investigate the features, we also present the methodology and describe the results of our experiments. In Chapter 6, we explore the outlier detection techniques and feature elimination methods features and investigate how the associations learned by the classifier are interpreted locally by eXplainable AI methods, we also show the methodology and detail the results of our experiments. In Chapter 7, Project proceedings, we present the final remarks of our research, future directions, and publications that were produced during the development of this thesis.

2

Fundamentals

This chapter presents the concepts necessary for the understanding of this thesis. First, we elucidate some biological concepts such as gene expression, used as data in our method. After that, machine learning methods employed in this proposal are explained, including supervised and unsupervised approaches. Next, we demonstrate the evaluation metrics used to evaluate the classifier performance. Finally, we present a brief summary of the chapter.

2.1 Biological Background

In this proposal, we will use different terms when referring to topics related to biology, such as gene expression data, pathways, and gene sets. Therefore, it is necessary to define what each of these terms represents. Next, we will introduce some terminologies and definitions.

2.1.1 Central Dogma of Molecular Biology

The central dogma of molecular biology explains how the flow of information in the genetic code occurs.

In eukaryotic cells, the DNA located in the nucleus stores biological information in the long term and, therefore, it is able to replicate with each cell division. In order for the messages written in the DNA nucleotide alphabet to be converted into proteins, they are first transcribed into an intermediate molecule, RNA, a molecule chemically similar to DNA that uses a slightly different nucleotide code.

This RNA sends the message that was originally contained in the DNA to the cell's cytoplasm, where the translation process takes place, in which the information contained in the linear nucleotide sequence and translated into an amino acid sequence, giving rise to the protein encoded by the original message [Crick, 1970].

Due to their similar alphabets, the messages contained in DNA and RNA are easily interchangeable, however, it is evident that the information on the amino acid sequences of proteins is not inversely translated into nucleotide sequences.

In summary, the central dogma of molecular biology explains the flow of genetic information, from DNA to RNA, to make a functional product, a protein. Figure 2.1 presents the flow of the central dogma of molecular biology.

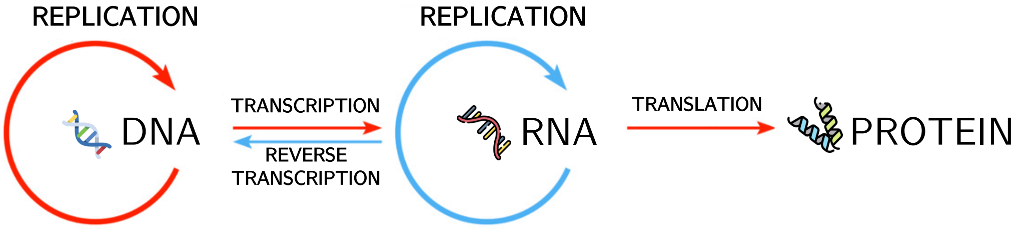


Figure 2.1: Central dogma of molecular biology flow.

2.1.2 Gene Expression Data

Gene expression is the process by which the instructions in our DNA are converted into a functional product, such as a protein. DNA microarray and RNA-Seq (which will be explained in subsection 2.1.3) are technologies that allow us to analyze, at the same time, the expression level of millions of genes. The gene expression level indicates the synthesis of different messenger RNA (mRNA) molecules in a cell. Using gene expression, it is possible to diagnose diseases, identify tumors, select the best treatment to resist illness, and detect mutations, among other processes. In order to achieve that purpose, several computational techniques, such as pattern classification approaches can be applied [Garro et al., 2016].

The Gene expression measured through microarray or RNA-Seq are stored in datasets, known as gene expression data. These datasets consists of a real-valued matrix M, with rows corresponding to gene expressions levels [Kerr et al., 2008], the columns represent the expression profiles of samples. Each cell M_{ij} is the measured expression level of gene i in sample j. The matrix is composed of n rows and m columns, which is known as a gene expression profile [Tan and Gilbert, 2003]. The rows represent genes, and the columns represent the samples as can be seen in Figure 2.2.

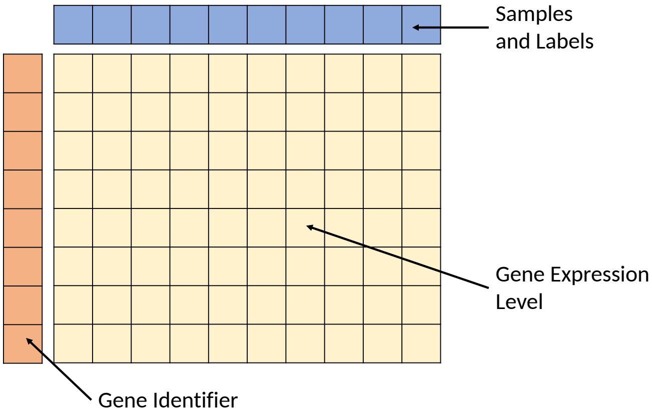


Figure 2.2: Example of Dataset, figure adapted from Ayyad et al. [2019].

Therefore, M_{ij} is a numeric value representing the gene expression level of gene i in sample j of Matrix M. Figure 2.2 shows the microarray matrix. It is worth mentioning that the scanning process to obtain the gene expression matrix presents missing value, noise, and variations resulting from each specific experimental procedure. In this step, data preprocessing is necessary to perform any experiment or analysis in the gene expression matrix. Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. It is beyond the scope of this proposal to discuss preprocessing gene expression data, but in the work of Herrero et al. [2003] this problem is addressed.

2.1.3 Current Strategies

Microarray technologies opened new ways to collect gene expression of biological material by enabling monitoring the expression of thousands of genes at the same time. According to Bhola and Tiwari [2015], in the beginning of microarray experiments, there were two common techniques: The complementary DNA (cDNA) microarray [Schena et al., 1995] and oligonucleotide arrays (oligo chip) [Lockhart et al., 1996]. Each microarray experiment have their specificities, but both of them involve three common basic procedures [Tefferi et al., 2002]:

- Chip manufacturer: A microarray is a small chip, where thousands of DNA molecules are attached in fixed grids. Each grid cell relates to a DNA sequence.
- Target preparation, labeling, and hybridization: Typically, two mRNA (Messenger RNA) samples (test and control samples) are reverse transcribed into cDNA

(Complementary DNA), labeled using either fluorescent dyes or radioactive isotopes, and then hybridized with the DNA molecules on the surface of the chip.

- The scanning process: Chips are scanned to read the signal intensity that is emitted from the labeled and hybridized targets.

The microarray technology consists of using a slide, in which the probes were immobilized in precisely defined quantities and positions (spots), to be hybridized with a pool of extracted mRNAs of biological samples (targets), which were previously marked with fluorophores. Since mRNA molecules are quite unstable when manipulated, most laboratory protocols use the reverse transcription process to convert the corresponding cDNAs during the tagging process. A microarray technology provides an indirect measure of gene expression level by quantifying the abundance of transcribed RNAs [Rosa et al., 2007].

A better [Zhao et al., 2014] and more advanced technique was created, the RNA-seq [Wang et al., 2009]. It is a methodology for RNA profiling based on next generation sequencing (NGS), and is replacing microarrays for the study of gene expression. The sequencing framework of RNA-seq enables to investigate at high resolution all the RNAs present in a sample, characterizing their sequences and quantifying their abundances at the same time. In practice, millions of short strings, called “reads”, are sequenced from random positions of the input RNAs.

The powerful features of RNA-seq, such as high resolution and broad dynamic range, have boosted an unprecedented progress of omics research [Finotello and Di Camillo, 2015]. One of the major benefits of RNA-seq is that it quantifies the expression of the over 70,000 non-coding RNAs not usually measured with microarrays.

Despite the benefits of the microarray, there are a few disadvantages of this technology. According to Jaksik et al. [2015], the biggest drawbacks of microarrays are their high cost per experiment, the abundance of probe designs based on low-specificity sequences, and the lack of control over the pool of analyzed transcripts due to the fact that the majority of widely used microarray platforms only use one set of manufacturer-designed probes. The relative lack of microarray accuracy, precision, and specificity as well as the high sensitivity of the experimental design to variations in hybridization temperature [46], the purity and rate of genetic material degradation, and the amplification process are additional flaws that could affect estimates of gene expression along with other variables.

2.1.4 Gene Set

Gene sets are exactly what the term says: a set of genes, an unordered and unstructured collection of genes [Liberzon et al., 2011]. Gene sets are defined as the collection of genes associated with a specific biological process (e.g., cell cycle) and location (e.g., on chromosome 1). It is also associated with diseases (e.g., breast cancer), or even the set of genes presented in a given pathway (e.g., the set of 128 genes involved in the Kyoto Encyclopedia of Genes and Genomes (KEGG) cell cycle pathway). Aside from containing multiple genes, there is nothing that defines a gene set. Consequently, it could be a completely arbitrary set. The Molecular Signatures Database (MSigDB) [Liberzon et al., 2011] includes over 10,000 of such gene sets defined based on many criteria, some of them seemingly arbitrary.

2.1.5 Pathway

A pathway is essentially a description of mechanisms and phenomena. A pathway is usually described by a graph that contains nodes (genes/protein) and edges (interaction between genes/proteins). There are several types of pathways such as signaling, metabolic and genetic. A pathway is meant to describe certain phenomena, interactions and dependencies. In essence, pathways are models describing the interactions of genes, proteins, or metabolites within cells, tissues, or organisms, not simple lists of genes. Well-known pathway databases include Kyoto Encyclopedia of Genes and Genomes (KEGG) [Nishimura, 2001], BioCyc [Karp et al., 2017] and reactome [Jassal et al., 2020].

In Figure 2.3, we illustrate an example of a pathway. In this Figure, RASSF1 is a protein that interacts with sets of other proteins. These proteins are related to some biological processes. As we can see, the blue proteins belong to a cell cycle process. Since all of these proteins are connected, we can identify this as a pathway.

2.1.6 The Curse of Dimensionality

In bioinformatics, an unique challenge arises from the high dimensionality in omics (genomics, transcriptomics, proteomics, epigenomic, etc) data [Berger et al., 2013, Ma and Dai, 2011]. For gene expression data, where there are several thousands of genes and only a few hundreds of samples. This uneven number of samples and genes is due to the difficulty of collecting microarray samples. This high dimensional data may cause the "curse of dimensionality" which results in inaccurate distance metrics and impacts on classification precision [Xie et al., 2016].

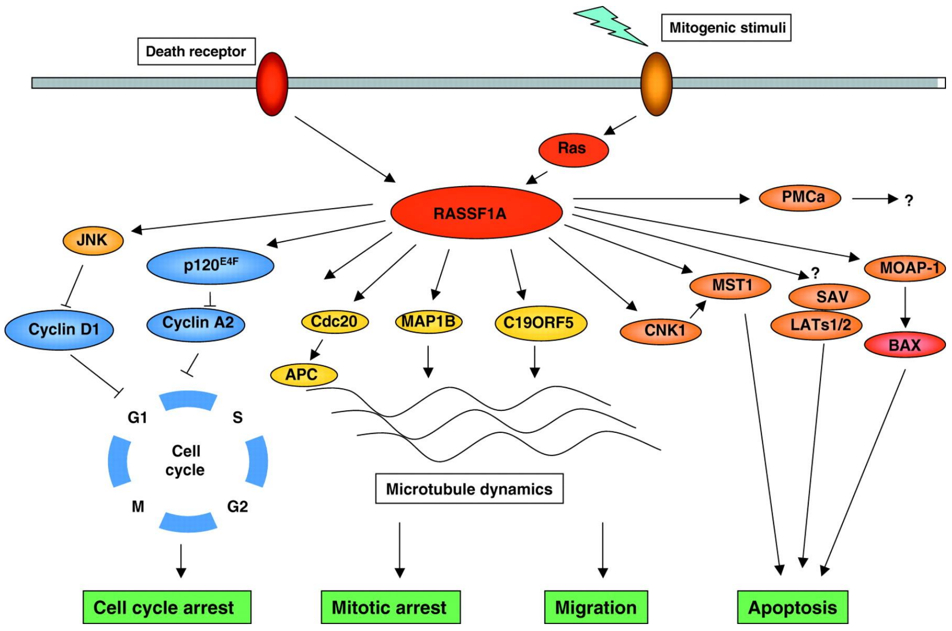


Figure 2.3: Summary of some of the known partners and pathways of RASSF1A [Jassal et al., 2020].

Having so many features and few samples creates a high probability of finding "false positives", which are due to find deferentially expressed genes in the building of predictive models [Piatetsky-Shapiro and Tamayo, 2003]. Another concerning problem is a substantial component of noise, resulting from numerous sources of variation affecting expression measurements [Mramor et al., 2005]. To solve the curse of dimensionality problem in gene expression data, the prevailing modeling approaches include gene filtering in the data preprocessing phase, and gene subset selection along with a modeling technique [Mramor et al., 2005].

2.2 Outliers

Outliers are observation (or a subset of observations) that appears to be inconsistent with the remainder of a set of data [Hodge and Austin, 2004]. An outlier often contains useful information about abnormal characteristics of the systems and entities that impact the data analysis. The recognition of such distinguished characteristics provides valuable knowledge of the analyzed data [Aggarwal, 2015].

2.2.1 Interquartile Range (IQR)

Interquartile Range (IQR) is a widely used technique that helps find outliers in continually distributed data.

It does need symmetric distributions of the data [Gelade et al., 2015]. Besides, it is not affected by outliers since it uses the middle 50% of the distribution for calculation and is computationally cheap.

The interquartile range is defined by:

$$IQR = Q3 - Q1, \tag{2.1}$$

in which $Q3$ is the third quartile and $Q1$ is the first quartile. To detect the outliers using IQR, we use the:

$$Lower\ IQR = (Q1 - 1.5 \times IQR), \tag{2.2}$$

$$Upper\ IQR = (Q3 + 1.5 \times IQR). \tag{2.3}$$

in which any data outside this range is considered an outlier.

2.2.2 Isolation Forest

Isolation forest isolates observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. This feature can be considered an outlier by the number of conditions required to separate it from the others.

Isolation Forest can scale to high-dimensional problems with a large number of irrelevant attributes; it has a linear time complexity $\mathcal{O}(n)$, and a low memory requirement [Liu et al., 2008].

It builds an ensemble of isolation trees (iTrees) for a given dataset and then defines anomalies as instances with short average path lengths on the iTrees. Isolation forest defines an Anomaly Score (AS) s, a quantitative index representing the "outlierness" degree of an observation. The anomaly score is defined for an observation x by:

$$s\left(x, \psi\right) = 2^{\left(-\frac{E(h)(x))}{c(\psi)}\right)} \in [0, 1], \tag{2.4}$$

in which $E(h(x))$ is the average path length $h(\cdot)$ over the t isolation trees and $c(\psi) = E(h(x)|\psi)$ is an adjustment factor that accounts for the cardinality of the subsampled dataset. When $E(h(x)) \to n - 1$, the AS tends to 0 meaning that x appears to be a normal instance. When $E(h(x)) \to 0$, the AS $s \to 1$, meaning that x appears to be an outlier.

2.3 Machine Learning

Machine learning is concerned with pattern recognition [Bishop, 2006] and computational learning in the context of artificial intelligence, that is, how to build programs capable of automatically improving through experience [Mitchell, 1999]. Machine learning techniques have been used in applications in a wide number of areas such as e-commerce, industry, among others [Paliouras et al., 2003]. Machine learning models can be supervised or unsupervised.

In supervised learning, the input vectors are given together with their target vectors, that is, real values that correspond to the instances, such as classes or a value to be calculated. Classifications problems are solved with supervised methods, in which the models are called classifiers. In the literature there are several techniques for classification, such as: Decision Tree (C4.5), Artificial Neural Networks (ANN), Näive Bayes (NB), kNN, SVM among others. Each of the techniques has advantages, disadvantages and the most suitable types of application. In Kotsiantis et al. [2007] a review of classification techniques is performed using metrics such as accuracy, training speed and classification speed among others.

In unsupervised learning, the data is not accompanied by its target vectors. Clustering is one of the most used unsupervised learning methods, and are widely used in biology, for exploratory and data analysis purpose. Xu and Tian [2015] reviews the most used clustering techniques, such as k-means, DBSCAN, hierarchical clustering, SOM, and others.

2.3.1 Support Vector Machine (SVM)

Support vector machines (SVM) is a supervised machine learning technique for classification, where previously identified examples are needed to build a model. In this method, the characteristics of the objects to be classified are transformed into vectors of real values, where each dimension of this vector corresponds to a feature. These feature vectors are then mapped in a space with high dimensionality through a nonlinear mapping, so that in this new space it is easier to find an optimal hyperplane that separates the mapped vectors into their different classes [Cortes and Vapnik, 1995].

This method projects each instance in a vector space, where marginal vectors are used to define the separation between classes. The idea is to find a hyperplane, represented by the vector $\vec{w}$, which besides separating the vectors of the instances of each class, maximizes the margin of separation between the classes.

According to Chen et al. [2007], given a training set of data points $G = \{(x_i y_i)\}_{i=1}^{n}, x_i \in R^m$ and $y_i \in \{+1, -1\}$. The decision function of SVM is defined by:

$$f(x) = \langle w, \phi(x) \rangle + b, \tag{2.5}$$

where $\phi(x)$ is a mapping of sample x from the input space to a high-dimensional feature space. $\langle ., . \rangle$ denotes the dot product in the feature space. The optimal values of w and b can be obtained by solving the following regularized optimization problem:

$$min\ J(w, \xi) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi i, \tag{2.6}$$

$$s.t. \begin{cases} y_i \left(\langle w, \phi(x_i) \rangle \right) + b \geqslant 1 - \xi_i, \quad i = 1, ..., n, \\ \xi i \geqslant 0, \end{cases} \tag{2.7}$$

where ξ is the ith slack variables, and C is the regularization parameter, in which determines the trade off between the maximum margin $1/\|\vec{w}\|^2$ and the minimum experience risk.

Two parameters must be considered when dealing with SVM classifiers, the kernel function and the regularization parameter. The kernel function calculates the similarity between two support vectors and is also used to map them in a space with high dimensionality. Making it possible to separate the feature vectors from the corresponding classes, which is often not possible in the original entry space. The basic types of kernel are: linear, polynomial, sgmoidal and radial based functions (RBF) [Burges, 1998]. The regularization parameter C is used to as a trade-off between training error and the flatness of the solution.

2.3.2 Neural Networks

Artificial neural networks consist of a method to solve artificial intelligence problems that are submitted by a learning process in order to store experimental knowledge and make it available for use.

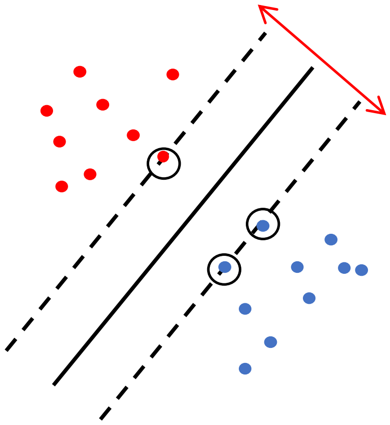


Figure 2.4: Samples in a two-dimensional space (colored dots) separated by a hyperplane supported by support vectors (circled dots), maximizing the distance between the samples closest to the threshold of the classes.

Artificial neural networks, or simply neural networks (NN), are computational models inspired by the animal central nervous system, capable of recognizing patterns and learning through data and experience [Goodfellow et al., 2016]. Since the human brain is capable of learning and making decisions based on learning, artificial neural networks must do the same.

These models follow a hierarchy, shown in 2.5, composed of several layers of connected neurons (also known as units or elements). The nodes represent the neurons and the lines represent the weights of the connections. By convention, the layer that receives the data is called the input layer, the final layer is called the output layer, and the internal ones are called hidden layers.

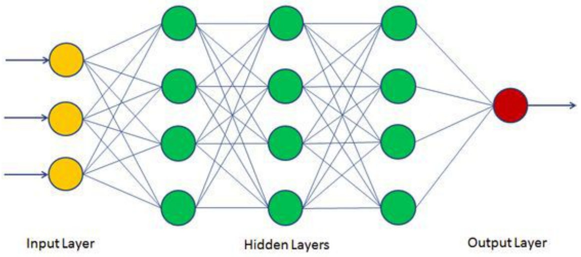


Figure 2.5: Neural network representation.

The neural network goes through a training process, acquiring the necessary system to properly execute the desired process of the data provided. Thus, the neural network is capable of extracting basic rules from real data, differing from programmed computing, where a set of rigid pre-fixed rules and algorithms is required.

Neural Network can be categorized based on the number of hidden layers and based on the way the information propagates through these layers [Shen et al., 2017]. Based on the number of hidden layers, we have:

- Shallow networks: a neural network with only a single (or a few) hidden layer.
- Deep networks: a neural network with many hidden layers.

Based on the way the information propagates, we have:

- Feed-forward networks: the information flow in one direction and there are no loops in the network. Some examples of these networks are the perceptron and the convolutional neural network.
- Feedback networks: this kind of network contains cycles or loops and therefore exhibits memorization ability and can store information, Recurrent Neural Network (RNN) is an example of this category.

2.3.3 XGBoost

XGBoost is a optimized version of the Gradient Tree Boosting (GTB) machine learning algorithm, which is an ensemble method based on decision trees [Chen and Guestrin, 2016]. It uses the boosting method in an iterative manner, where each iteration aims to correct the mistakes of the previous iteration by optimizing specific loss functions and applying regularization techniques. XGBoost, like Random Forest, is capable of detecting nonlinear relationships [Leevy et al., 2018].

2.3.4 K-means

K-means clustering is a partitioning method, this method decomposes a dataset into a set of disjoint clusters. It is commonly used to automatically partitions the input dataset into k cluster. K-Means computes the squared distances between the inputs (also called input data points) and centroids, and assigns inputs to the nearest centroid. It proceeds by selecting k initial cluster centers and then iteratively refining them as follows [Žalik, 2008]:

- Each instance d_i is assigned to its closest cluster center.
- Each cluster center C_j is updated to be the mean of its constituent instances.

The algorithm converges when there is no further change in assignment of instances to clusters

Figure 2.6 illustrates an example of the k-Means algorithm. Initially the centroids, represented by “x”, were assigned to random points. In the process of executing the algorithm, each centroid goes through the path, represented by the lines, until the groups do not change, that is, the centroid stops. In the end, we can see the formation of three clusters, represented by red, green and blue dots.

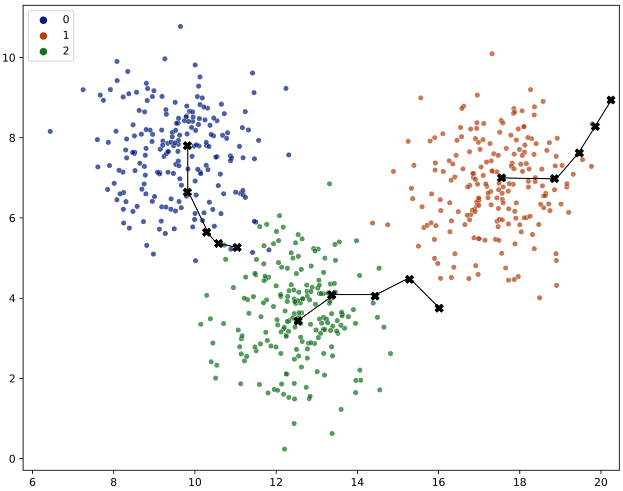


Figure 2.6: K-means algorithm.

2.3.5 k-Nearest Neighbours

The k-Nearest-Neighbours (kNN) is a non-parametric classification method, which is simple but effective in many cases. It is based on the proximity of training samples in the feature space [Guo et al., 2003].

The training process for this algorithm consists of storing the feature vector and labels (classes) of each training sample in an n-dimensional space, where n is the number of features of each sample. In the unlabeled sample classifying process, the sample is projected into space and classified according to the k closest samples.

When $k = 1$, the sample is classified according to the label of its nearest neighbor in the feature space. When k is greater than 1, the classification is given by a voting scheme, where the class with the most numerous neighboring samples is considered the sample class. For this reason, in multi-class problems, that is, with more than two possible classes, ties can occur even considering an odd number of neighbors, so a tiebreaker must be defined in the algorithm.

There are several ways to calculate the distance between two samples a and b in an n-dimensional space that can be used in kNN. Commonly, the Euclidean or Manhattan distance are used as similarity measure [Liao and Vemuri, 2002]. The following equation represents the Euclidean distance:

$$d(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \tag{2.8}$$

in which n is the number of dimensions and a and b are the samples.

2.3.6 Random Forest

Random Forest, introduced by Breiman [2001], is a machine learning method that makes predictions based on the results of multiple independent decision trees, being able to solve problems of classification, regression and other tasks. According to Treeratpituk and Giles [2009], each decision tree within the forest is assembled with a different initialization sample, extracted from the original dataset. Each tree is then built to full size without any pruning.

The selection of variables for each split in the tree is performed on a randomly selected subset of features, rather than on the full set of features, as is normally done in the traditional decision tree. Once the forest is built, ranking can be done simply by aggregating the votes of all the trees. There are only two parameters for tuning the random forest: T, the number of trees, and M, the number of features to consider when splitting each node. The error rate of a random forest depends on two factors: the correlation between the trees in the forest and the strength of each individual tree. The more correlated the trees, the higher the error rate. The stronger the individual tree (high precision), the lower the error rate.

2.3.7 Recursive Feature Elimination (RFE)

Recursive feature elimination (RFE) is a backward elimination algorithm developed by Guyon et al. [2002]. These algorithms work by iteratively removing one “worst” feature at a time until the predefined size of the final subset is reached. This method is designed to retain features that are most relevant to the classification task.

In each iteration of RFE, a linear SVM model is trained. The feature with the smallest ranking criterion is removed since it has the least effect on classification [Tang et al., 2007]. The remaining features are kept for the SVM model in the next iteration. This process is repeated until all the features have been removed. Then the features are sorted according to the order of removal. The later a feature is removed, the more important it should be.

RFE does not use the cross-validation accuracy on the training data as the selection criterion, thus is (1) less prone to overfitting; (2) able to make full use of the training data; (3) much faster, especially when there are a lot of candidate features [Yan and Zhang, 2015].

2.3.8 Forward Selection

In the forward selection method, variables are added to a model one at a time. The first predictor variable to enter the model is the variable that has the highest correlation with the response scores. At each successive step, the variables in the remaining set are considered for inclusion in the current model. The variable that is included at each step is that which produces the largest reduction in the residual sum of squares. As the model continues to improve (per that same criteria) the process continues, adding in one variable at a time and testing at each step. Forward selection continues until all variables are in the model or until a stopping rule is satisfied [Derksen and Keselman, 1992]. It is important to note that once a variable has been included in the model, it may not be removed. Because of this, one is never sure of the optimality of the variable subset [Sutter and Kalivas, 1993].

2.3.9 Multiclass problem decomposition

There are two possibilities for solving multiclass classification problems. The first is to use a classification method capable of classifying all classes simultaneously. The second possibility is to decompose the original problem into binary sub-problems. Binary decomposition has two advantages: (a) it allows the use of naturally binary classifiers, such as SVM, to solve multi-class problems, and (b) simplify decision making by simplifying the classification functions. In some problems, the decomposition managed to increase the accuracy [Fürnkranz, 2001].

There are two types of binary decomposition for multi-class problems: One-against-One (1A1) and One-against-All (1AA). Decomposition is also known as Round Robin [Fürnkranz, 2001].

The one-against-all procedure (1AA) begins by separating all samples into two sets: one set with the target class (“ +1 ”), in which only one species is included, and a second set with all the samples of the remaining classes (" -1 "). Thus, model f (.) is trained and applied to estimate the labels of the test group. If the classifier decides in favor of class 1, then the sample corresponding to this class wins a vote. Otherwise, all samples represented by class 1 win the vote.

In the second round, this procedure is repeated, but the samples in the previous round formed class 1 are now included within classes 1, and a different class is chosen to compose class 1. Again the model f (.) is trained and evaluated; all classes' votes are assigned following the same previous rule. This procedure is repeated until all classes are classified as 1.

Finally, the final decision is obtained by applying a majority vote. Although the number of iterations of this decomposition method increases linearly with the number of classes, each evaluated decision function tends to be simpler than in the multi-class case.

In our approach, we modified the 1AA model. In the first round, the samples are classified between the target class and the remaining class. In the second round, this procedure is repeated, but the samples that were classified as class 1 in the previous round are now discarded, and a different class is chosen to compose class 1. This procedure is repeated until all classes are evaluated as being 1.

This adaptation was developed because of the computational complexity of the 1AA model. While the main disadvantage of method 1A1 is its exponential time complexity $\mathcal{O}(n^2)$. Our method has a linear time complexity $\mathcal{O}(n)$, in which the total number of rounds is equivalent to the number of classes minus *1*.

2.4 Explainable AI

An advanced machine learning (ML) algorithm can produce accurate predictions, but its famous "black box" nature does not help adoption. In bioinformatics, it is crucial to have a human understanding of the decisions of a machine learning result. In order to understand feature contributions, several works explore methods to explain and visualize the importance of input variables on ML predictions [Gunning et al., 2019, Souza et al., 2022].

Our work focuses on classical ML models using SHAP Values method for explainability [Lundberg and Lee, 2017]. There are other explanation techniques associated with Deep Learning (DL) models [Gunning et al., 2019], but as they are DL-based models, it is not feasible to use in our work since it needs a large amount of data for training and explaining. Thus, we decided to use classical machine learning and SHAP values method.

2.4.1 Shap Values

The SHAP (SHapley Additive exPlanations) uses a game theory approach to explain any machine learning model, where the outcome of each feature depends on the actions of all other features [Fudenberg and Tirole, 1991]. The Shapley values method is essential in our results. It quantifies feature contributions and helps us understand the importance of outlier genes.

SHAP values can be calculated as follows [Bi et al., 2020]:

$$\phi_i = \sum_{S \subseteq F, \{i\}} \frac{|S|! \left(|F| - |S| - 1\right)!}{|F|!} \left[f_{s \cup \{i\}} \left(x_{s \cup \{i\}} \right) - f_s \left(x_s \right) \right], \tag{2.9}$$

where F represents the set of all features and S represents all feature subsets obtained from F after removing the i^{th} feature. Then, two models, $f_{s \cup \{i\}}$ and f_S, are trained again, and predictions of these models are compared to the current input $\left[f_{s \cup \{i\}} \left(x_{s \cup \{i\}} \right) - f_s \left(x_s \right) \right]$, where x_S represents the values of the input features in the set S. To estimate ϕ_i from $2^{|F|}$ differences, the SHAP approach approximates the Shapley value by either performing Shapley sampling or Shapley quantitative influence. It is interesting to note that, SHAP estimate the feature importance (magnitude of the contribution) as well as the sign (positive or negative).

2.4.1.1 Kernel explainer

The kernel explainer is a method that uses a special weighted linear regression to compute the importance of each feature [Zhang et al., 2020]. Kernel explainer provides more accurate estimates with fewer evaluations of the original model than other sampling-based estimates [Antwarg et al., 2021].

2.5 Evaluation Metrics

To assess the performance of machine learning methods, evaluation metrics are required. In this work, we are specifically interested in metrics for supervised learning, that is, classification problems. Among them, we use the metrics described below.

2.5.1 Classification Metrics

To measure the performance of our proposed method, we apply traditional classification metrics such as *precision*, *recall*, *F-score* and *accuracy.*

$$Precision = \frac{TP}{TP + FP}, \tag{2.10}$$

$$Recall = \frac{TP}{TP + FN}, \tag{2.11}$$

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}, \tag{2.12}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}, \tag{2.13}$$

where TP (true positive) is the number of times the model correctly predicts the positive class, FP (false positive) is the number of times the model incorrectly predicts the positive class, FN (false negative) is the number of times model incorrectly predicts the negative class and TN (true negative) is the number of times the model correctly predicts the negative class. With these values it is possible to assemble a contingency table (or confusion matrix), where the rows represent the actual classes and the columns the estimated classes for each classified example.

Matthews correlation coefficient and precision-recall curve are metrics suited for imbalanced datasets. While MCC is more appropriate for binary classification, the precision-recall curve is a more reliable and informative indicator of the statistical performance on multiclass problems [Chicco, 2017]. MCC is defined by:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)\,(TP + FN)\,(TN + FP)\,(TN + FN)}}. \tag{2.14}$$

The precision-recall curve is built by calculating and plotting the precision against the recall for different thresholds. It tends to maximize the correctly classified positive values and does not directly consider the correctly classified negative values [Chicco, 2017]. A high area under the precision-recall curve ($AUPRC$) represents both high recall and high precision. High precision means that the classifier is returning accurate results. High recall indicates that the classifier return most positive results. Figure 2.7 presents an example of a precision-recall curve.

We also computed *specificity* and negative predictive value (NPV), since specificity is used to check how correctly we can classify an individual in the correct cancer subtype and NPV is the probability that following a classification result, that individ-

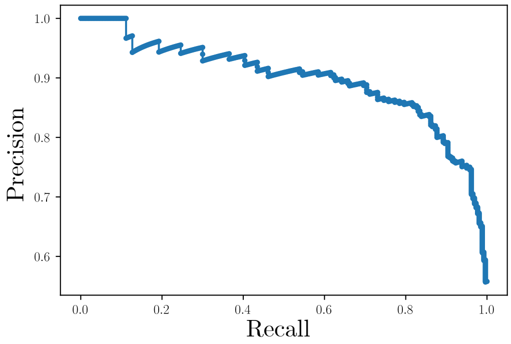


Figure 2.7: Precision Recall Curve.

ual will not have that specific cancer subtype [Parikh et al., 2008]. These metrics are defined as:

$$Specificity = \frac{TN}{TN + FP}, \tag{2.15}$$

$$NPV = \frac{TN}{TN + FN}, \tag{2.16}$$

the higher the value of these metrics, the better the result.

2.5.2 Macro metrics

Traditional metrics described in the previous section can be directly calculated from the values of the classifier confusion matrix. For example, micro accuracy is the sum of the main diagonal of the confusion matrix divided by the total of classified examples. The main disadvantage of micro metrics is that the most numerous classes, which have the highest proportion of examples for testing and training, considerably impact the results.

Macro metrics were defined to circumvent the possible bias caused in micro metrics by the imbalance number of samples per class in the [Sokolova and Lapalme, 2009] datasets. In this case, for each individual class C_i, their respective TP_i, FN_i, TN_i and FP_i are obtained, from which it is possible to calculate the Prec_i and Rec_i for each class. Since the precision and recall values are obtained by class, the macro values are obtained as the arithmetic mean of these individual values.

2.6 Chapter Remarks

In this chapter, we presented the fundamental concepts used in the development of this Thesis, including concepts related to biology and computing; we presented concepts of gene expression and pathway and machine learning fundamentals. We introduced the concept of outliers and also metrics to measure the quality of the classification. In the next chapter, related works will be presented, making use of some concepts presented in this chapter.

3

Related Work

This chapter presents a literature review on studies related to this work. We divided this section into three types of work: (i) breast cancer subtype classification, (ii) gene selection methods used to identify a small set of genes capable of improving the cancer classification task, and (iii) exploration of gene expression data using explainable AI methods.

3.1 Breast Cancer Subtype Classification

Graudenzi et al. [2017] proposed a classification framework based on Support Vector Machines (SVMs) with a feature selection strategy based on the concept of pathway activity. They identified and analyzed a list of enriched pathways in four different breast cancer subtypes, and used this information to perform the feature selection method in the classifier implementation. In terms of overall accuracy, the proposed classifier presents an accuracy around 85.00%, using 400 genes from the feature selection method.

Lee et al. [2020] used a pathway-based approach for feature selection, and applied a deep learning model with attention mechanism and network propagation for cancer classification. They used five TCGA[1] cancer datasets. The average classification accuracies of their method was 76.39% for urothelial bladder Carcinoma (BLCA), 66.91% for breast invasive carcinoma (BRCA), 71.54% for colon adenocarcinoma (COAD), 70.12% for prostate adenocarcinoma (PRAD), and 78.13% for (Stomach adenocarcinoma) STAD. They selected a total of $5,515$ genes for the classification task.

Mostavi et al. [2020] proposed three distinct convolutional neural network for cancer classification task. Regarding the prediction of breast cancer subtypes, the 1D-CNN model was employ. The authors used poor statistics methods for the feature

[1] The Cancer Genome Atlas Program.

selection step, such as standard deviation and mean. After selecting $7,091$ genes, they used their model for the classification task and achieved an average accuracy of 88.42% among five subtypes.

In the work of Li et al. [2017], the authors divided the process into two stages. First, a genetic algorithm is used as a gene selection mechanism and the KNN (k-nearest neighbors) algorithm as a classification method. The dataset contains 31 tumor types. For the sorting task using KNN, k was set to 5 with a majority voting rule. The results show that the classification accuracy was greater than 90% for 28 of the 31 types of cancer.

Lyu and Haque [2018] incorporated gene expression data into 2-D images and used a Convolutional Neural Network (CNN) to classify 33 distinct types of cancer. The authors transform cancer classification based on the gene expression problem into an image problem. The main problem is that gene expression data is highly dimensional, whereas most deep learning architectures are for 2-D imaging. As a result, the authors achieved a mean F1 across cancer types of 95.43% using $20,531$ genes.

3.2 Gene selection

Shukla et al. [2018] proposed a gene selection method composed of two-stage, in the first stage, ensemble gene selection (EGS) method using multi-layer approach and f-score approach is applied to filter the noisy and redundant genes from the dataset. In the second stage, an adaptive genetic algorithm (AGA) works as a wrapper to identify significant genes subsets from the reduced datasets, produced by EGS, that can contribute to detect cancer or tumor. On the classification step a 10-fold cross validation (10-fold CV) was employed. When using the SVM as classifier, they achieved an accuracy of 89% in the colon cancer dataset.

Rajapakse and Mundra [2013] used Pareto-front analysis (PFA) to gene selection and linear SVM for classification, they used multi-tissue, lung and leukemia datasets with 4-fold CV. The results show that PFA render minimally redundant gene subsets with improved performance in classification while compromising the stability. Using the SVM as classifier, a F-score of 95.11 was achieved using 397 genes in the mixed lineage leukemia dataset.

Huang et al. [2016] used a hybrid $L_{1/2\ +\ 2}$ regularization (HLR) function, a linear combination of $L_{1/2}$ and L_2 penalties, to select the relevant genes in the logistic regression. The HLR approach inherits sparsity characteristics from $L_{1/2}$ and grouping effect characteristics from L_2. Three different biclass datasets were used. A logistic regression classifier was used with 10-fold CV. The results show that logistic regression with the HLR approach is the promising tool for feature selection in the classification problem, since it achieved more than 89% of accuracy in all the tests.

3.3 Explainable AI

Applying eXplainable Artificial Intelligence (XAI) approaches in gene expression data presents insights into how the models are working by interpreting those "Black-box". It is capable of identifying relevant genes, which genes are essential for the classification, and even how the genes are related to each other when looking for the importance.

Yu et al. [2021] developed a tool that integrates AutoEncoder (AE) and SHapley Additive exPlanations (SHAP). The tool quantitatively evaluates the contributions of each gene to the hidden structure learned by an AE, substantially improving the expanability of AE outcomes. The tool was applied in the TCGA (The Cancer Genome Atlas), with 1,041 *samples* and $56,497$ genes. The authors were able to identify genes that are not differentially expressed and pathways in various cancer-related classes. It was also presented that Deep Auto-Encoders can be used for "small sample-sized" expression data.

In the work of Meena and Hasija [2022], they tried to find potential diagnostic biomarkers for SCC (Squamous Cell Carcinoma) by applying eXplainable Artificial Intelligence (XAI) on XGBoost machine learning models trained on binary classification datasets comprising the expression data of 40 SCC, 38 AK (Actinic Keratosis), and 46 normal healthy skin samples. After successfully incorporating SHAP values into the machine learning models, 23 significant genes were identified and were found to be associated with the progression of SCC. These identified genes may serve as diagnostic and prognostic biomarkers in patients with SCC.

Kamal et al. [2021] employed microarray gene expression data to classify the Alzheimer's disease (AD) using k-nearest neighbors (KNN), support vector classifier (SVC), and Xboost classifiers. To establish trustworthy predictive modeling, the authors introduced an explainable Artificial Intelligence method. The XAI approach they used was LIME (local interpretable model-agnostic explanations) for a simple human interpretation. This approach shows how genes were selected for a particular AD patient and the most important genes for that patient were determined from the gene expression data.

As far as our knowledge, the works available regarding eXplainable AI and gene expression data do not address the multiclass problem, trying to identify the contribution of specific genes for each subtype of a carcinoma.

3.4 Discussion

Table 3.1: Summary of related works.

Author	# of genes	Gene Selection	Classes	Classifier	Evaluation Metric
Graudenzi et al. [2017]	400	Pathway-based	Multiclass	SVM	*Precision*, recall and *accuracy*
Lee et al. [2020]	7,091	St. dev. and mean	Multiclass	1D-CNN	*Precision*, *recall* and *F-measure*
Mostavi et al. [2020]	5,015	Pathway-based	Multiclass	GCN+MAE	*Accuracy*
Rajapakse and Mundra [2013]	80	Pareto-front	Multiclass	SVM	*Precision*, *sensitivity*, *accuracy* and *F-measure*
Shukla et al. [2018]	10	EGS + AGA	Bi-class	SVM + NB + KNN + DT	*F-score*, *KWscore*, class-specific, *F-PFA* and *KW-PFA*
Huang et al. [2016]	10	Hybrid $L_{1/2+2}$ regularization	Bi-class	SVM	*Sensitivity*, *specificity* and *accuracy*

The methods presented so far have been classified according to the problem to be addressed, (i) studies that develop methods to classify breast cancer subtypes, (ii) studies that develop methods for selecting relevant genes. Table 3.1 summarizes the related work.

The works summary shows that methods that focus on a multiclass problem use dozens of genes for classification, indicating that this is a challenging problem to solve. While dealing with bi-class problems, it is much more comfortable to solve since if you can identify just a few genes that can distinguish both classes, you have a high accuracy rate. Some works even reach approximately 99% of accuracy when dealing with bi-class cancer classification problems.

However, these approaches have some limitations. In the case of multiclass problems, process a large number of genes can have a direct impact on computational cost. The works do not use two distinct dataset to address the classification of breast cancer subtypes or gene selection problem, this may have some consequences, such as bias in the results. Thus, the models need to be trained in different databases and directly compared with other methods to assess their effectiveness. Still, the idea of using gene selection for cancer classification problem is interesting to develop more efficient classification methods.

Thus, we consider as an interesting research opportunity the development of a method that is able to achieve a high accuracy using few genes, where the method is trained and tested on different bases, in order to avoid bias. From these observations, in this research we intend to develop a method for selection of genes that compares with the state-of-the-art in terms of effectiveness and efficiency and uses fewer genes.

3.5 Chapter Remarks

In this chapter, we presented a brief review of related works that address the problem of cancer classification and eXplainable AI. The works were divided according to the classification problem addressed (bi-class or multiclass). When investigating the works, we realized that there is an opportunity to develop methods with better performance than the state-of-the-art, that can achieve better results, using fewer genes. In the next chapters, we will present the methods developed in our research.

4

A Gene Selection Method Based On Outliers for Breast Cancer Subtype Classification

Cancer classification is a commonly used method for identifying cancer subtypes, offering an efficient, accurate, and objective diagnosis for different types of cancer [Tarek et al., 2017, Tong et al., 2013]. In the case of breast cancer, diagnosing it by biologic subtype provides important prognostic and predictive information for patients, making the classification and identification of related genes crucial for effective treatment [Parker et al., 2009].

DNA microarray technology has allowed simultaneous monitoring of gene expression levels during important biological processes and across related samples, resulting in gene expression data that provides a viable alternative for cancer classification [Almugren and Alshamlan, 2019]. However, using gene expression data for cancer classification poses a challenge due to the complex and high-dimensional matrices of genes versus samples. Typically, the number of genes far exceeds the number of samples, with sets containing only a few hundred samples and over 10,000 genes per sample [Piatetsky-Shapiro and Tamayo, 2003].

In this Chapter, we propose an **O**utlier-based **G**ene **S**election (OGS) method to determine the gene sets related to each breast cancer subtype. The method combines outlier detection techniques and feature elimination methods to find a small gene set capable of achieving high classification results. Compared to recent approaches, our method presents promising results, with higher F_1 score when classifying the breast cancer subtypes with the worst prognoses.

Another contribution of our method is the identification of gene signatures possibly related to tumor biology for breast cancer subtypes. To identify those signatures, we employed clustering techniques considering that is a widely used approach to identify patterns between features (genes).

4.1 Method Description

In this section, we present the outlier-based gene selection (OGS) method. In contrast to classical approaches, the outlier proposal is more efficient regarding number of important genes selected for breast cancer subtype classification. Figure 4.1 summarizes the steps that compose our approach which are discussed in the next subsections.

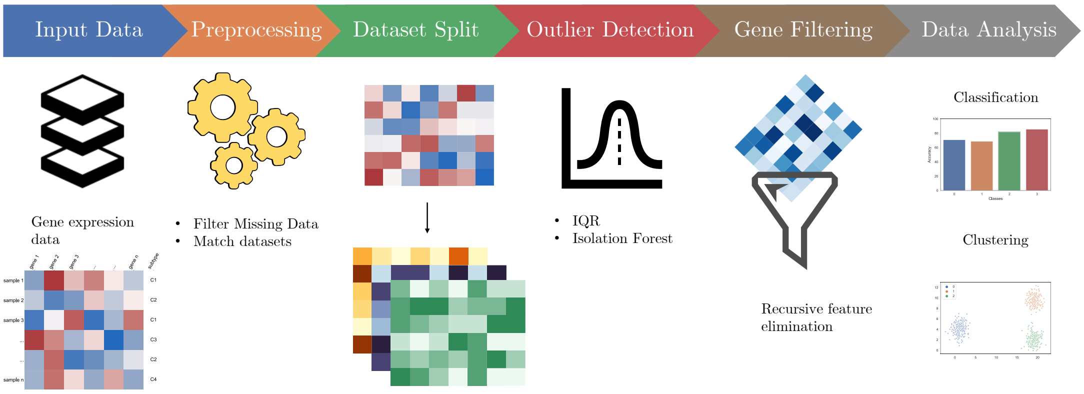


Figure 4.1: Outlier-based gene selection (OGS) method.

4.1.1 Pre-processing and subtype split

The initial step of our proposed method is pre-processing. This step is essential since the dataset presents distinct number of genes. In this step, we match the genes from the training dataset and the testing dataset, since the number of features must be the same. The genes that are presented in only one dataset are discarded. After concluding the pre-processing step, we do the subtype split, separating the dataset into smaller datasets by subtype. It is worth mentioning that all the smaller datasets have the same genes.

4.1.2 Outlier Detection

We calculate the interquartile range for outlier detection. We adopted *IQR* in our approach because it is a reasonably robust measure of variability. For each of the smaller subsets (subtype dataset), we calculated the outlier genes using the *IQR* default multiplier of 1.5. To identify if a gene can be considered as an outlier, in our approach this gene must be an outlier in at least 30% of the subtype samples.

We apply this threshold because there are some genes that are an outlier in only few samples, so applying a low threshold would result in a vast list of genes. And one of our objectives is to find a small set of genes. At the end of our computation, we have all the genes that we consider as an outlier for each of the subtypes. It is worth mentioning that this threshold was found empirically, from several tests, and may vary from dataset to dataset.

After finding the outlier genes using *IQR* we validate them using isolation forest (iForest) [Liu et al., 2008], considering that with the selected threshold, we may consider a normal gene as an outlier. We employed the isolation forest, since it can scale to high-dimensional problems with a large number of irrelevant attributes and has a linear complexity.

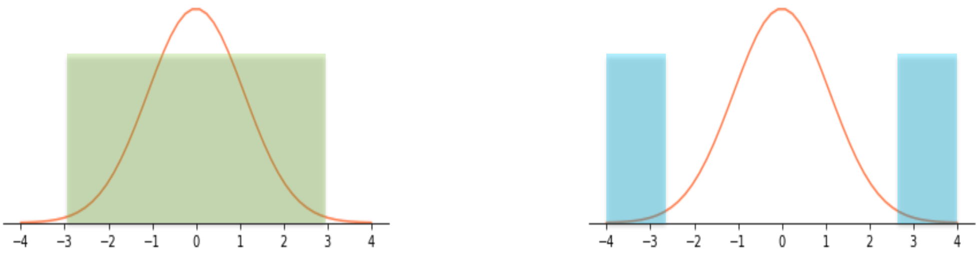
4.1.3 Gene Filtering

The gene filtering step involves two stages. In the first stage, outlier genes are selected for each of the subtypes. The second stage narrows the selected outlier genes using recursive feature elimination. This gene selection approach aims at filtering the irrelevant genes from the outlier stage, therefore improving the classification accuracy.

Outlier Approach

In the first gene filtering step, we use two strategies to find our gene set: (i) the opposite outlier and (ii) outlier vs. normal:

- **Opposite outlier** - we find the intersection of the set of genes that are outliers in one subtype and are opposite outliers in the other subtypes (Figure 4.2b). For example, suppose a gene is an underexpressed outlier in one subtype. In that case, we will only consider this gene an outlier if it is an overexpressed outlier in any of the other subtypes. Genes that are not outliers are considered normal genes. After finding the genes that are considered outliers in the outlier detection step, we find the genes that are considered normal for all the subtypes (Figure 4.2a).
- **Outlier vs. normal** - we perform the same steps of an opposite outlier. However, in this case, we compare the outlier genes from the analyzed subtype to normal genes from the other subtypes. This is necessary because some subtypes share an extensive list of outlier genes.



(a) Normal genes.

(b) Outlier genes.

Figure 4.2: Distribution of genes.

We consider a gene as normal if this gene is presented in at least 50% of the samples. Intersecting the outlier genes on the analyzed subtype and opposite outlier genes from the other subtypes is a way to find a set of genes that can distinguish the subtypes because they present different expression values among the subtypes. And when a gene expresses this difference among the subtypes (Figure 4.3), it is a premise that this gene may be useful to classify the analyzed subtype. Finally, we find our gene set after the intersection of outliers from one subtype with the opposite outlier genes or normal genes from all the other subtypes.

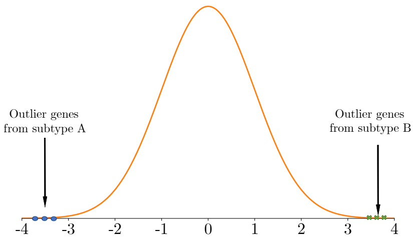


Figure 4.3: Opposite outliers

4.1.4 Narrowing Gene Set

The last stage of gene filtering is the removal of less important outlier genes from the subset found. After finding an outlier subset using opposite outlier or outlier vs. normal approaches, we implement the Recursive feature elimination (RFE) [Guyon et al., 2002] to narrow the number of genes selected.

We employed RFE because it effectively selects features in the training dataset that are more relevant in predicting the target variable; therefore being widely used to discover new biomarkers and improve the interpretation of biological data [Christin et al., 2013].

4.1.5 Data Analysis

Our analysis is divided in two parts. In the first part, we build an hierarchical classifier to validate set of genes found for each of the four breast cancer subtypes. In the second part, we focus on finding a reliable subset of genes that accurately characterizes the basal subtype of breast cancer, which has the worst prognosis [Dwivedi et al., 2019].

4.1.6 Classification

After obtaining the gene subset of each of the subtypes, we build an hierarchical classifier that is divided in three layers, as depicted in Figure 4.4, where in each layer we classify a different breast cancer subtype using the specific genes found for each subtype by our gene filtering step.

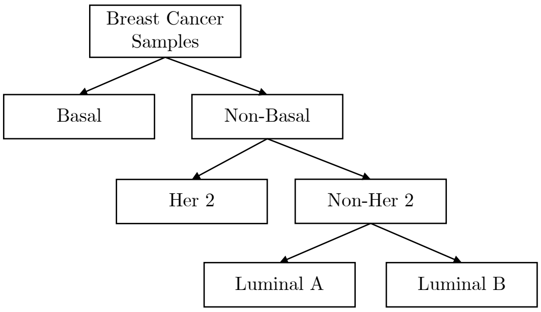


Figure 4.4: Hierarchical Classification.

It is important to emphasize that if a sample is miss classified in any of the hierarchical classifier layers, this sample will only be classified in the other layers to know which class it got mistaken.

4.1.7 Clustering

Using the subset of genes found in the previous step of OGS, we use unsupervised machine learning methods (e.g., KNN) to cluster the breast cancer subtype samples in our dataset, to identify if there are distinct signatures (gene sets) capable of characterize the basal breast cancer subtype. The assumption is that if we can cluster the basal subtype correctly, then the set of genes is relevant and deserves a biological investigation.

4.2 Evaluation Methodology

In this section, we describe the evaluation methodology used in this work. We detail the characteristics of the datasets used in the experiments. We present the parameters chosen for the machine learning methods, and also present the evaluation metrics employed.

4.2.1 Dataset

To evaluate our method, we used four datasets from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) [Edwards et al., 2015]: two RNA datasets and two protein datasets. The protein datasets provide greater analytical breadth, since they use mass spectrometry to analyze the proteomes of genome-annotated TCGA tumor samples [Mertins et al., 2016], the RNA datasets uses RNA-Seq technology for RNA sequencing. The Cptac 2C and Protein Cptac 2C datasets are used for training the models, since it has more samples, the Cptac 2D and Protein Cptac 2D are used for testing.

The datasets present gene expression and protein expression data from breast cancer patients. Those samples are divided into four principal mRNA-defined breast cancer intrinsic subtypes (See Chapter 1), normal samples were not included, since in our scope we already know that the patient has a tumor, and we are trying to identify the subtype to specify the best treatment.

Tables 4.1 and 4.2 summarize the characteristics of the datasets used in the experiments, and Figures 4.5 and 4.6 present the intersection of genes/proteins that are present on both the train and test datasets.

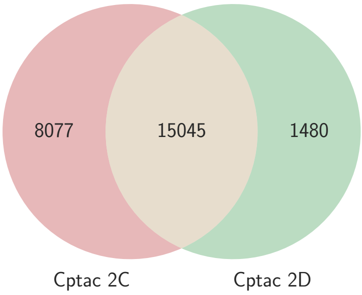


Figure 4.5: Genes intersection between Cptac 2C and Cptac 2D datasets.

Table 4.1: RNA dataset description.

Dataset	**# of genes**	**Subtypes**	**# of samples**	**Total # of samples**
Cptac 2C	23,122	Basal	29	117
		Her 2	14	
		Luminal A	57	
		Luminal B	17	
Cptac 2D	16,525	Basal	18	77
		Her 2	12	
		Luminal A	23	
		Luminal B	24	

Table 4.2: Protein dataset description.

Dataset	**# of proteins**	**Subtypes**	**# of samples**	**Total # of samples**
Protein Cptac 2C	9,771	Basal	29	117
		Her 2	14	
		Luminal A	57	
		Luminal B	17	
Protein Cptac 2D	11,311	Basal	18	77
		Her 2	12	
		Luminal A	23	
		Luminal B	24	

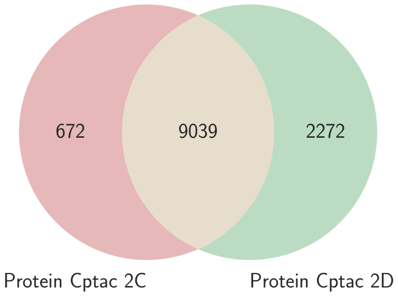


Figure 4.6: Protein intersection between Protein Cptac 2C and Protein Cptac 2D datasets.

4.2.2 Multi-level classifier

To evaluate the proposed method, we first model the bi-class problem. We employ the Support Vector Machine (SVM) to make the predictions because, as presented in the work of Mathur and Foody [2008], the SVM has a binary nature for classification. Grid search [Bergstra and Bengio, 2012] was used to optimize the parameters for each classifier model and, as a result, we chose the SVM model with Linear kernel and $C =$ 0.1 and remaining parameters set to the default of the scikit-learn[1]. Table 4.3 summarizes all the parameters used.

Table 4.3: SVM parameters

SVM Parameters	
Parameter	Value
C	0.1
break_ties	False
cache_size	200
class_weight	None
decision_function_shape	ovr
kernel	Linear
max_iter	-1
probability	False
shrinking	True
tol	0.001
verbose	False

4.2.3 Signature Identification

To validate the quality of the distinct signatures found, we employ k-means clustering. If we cluster the basal subtype correctly with a simple unsupervised learning method, then the set of genes is important. The k-means parameters are also set to the default of the scikit-learn library.

4.2.4 Evaluation Metrics

To measure the performance of our proposed method, we apply traditional metrics such as *precision*, *recall*, F_1 score and *accuracy*. Since biological data often have a sparse dataset [Chicco, 2017], we also measured the performance of OGS using the Matthews

[1] `https://scikit-learn.org/stable/`

correlation coefficient (*MCC*) [Baldi et al., 2000] and precision-recall curve. Both metrics were selected because they are suited for imbalanced datasets. While *MCC* is more appropriate for binary classification, the precision-recall curve is a more reliable and informative indicator of the statistical performance on multiclass problems [Chicco, 2017].

We also computed *specificity* and negative predictive value (*NPV*), since specificity is used to check how correctly we can classify an individual in the correct cancer subtype [Parikh et al., 2008] and *NPV* is the probability that following a classification result, that individual will not have that specific cancer subtype [Parikh et al., 2008]. All The evaluation metrics are described and detailed in Chapter 2.

We compared the OGS method with the method proposed by Mostavi et al. [2020] described in Chapter 3, since it presents state of the art neural network approach for the breast cancer subtype classification problem. This method uses an 1D-CNN (Convoluntional Neural Network) approach to classify breast cancer subtypes.

To perform the experiments, we implemented the method using the code shared by the authors[2] to model and classify our datasets. The performance was evaluated applying the same pre-processing techniques and steps presented in the authors work. We also implemented a grid-search to test their model using the best parameters and validated using *precision*, *recall*, F_1, *accuracy*, *MCC*, *precision-recall* curve, *specificity* and *NPV* as performance measures.

This baseline was selected, because they focus on breast cancer subtype classification. We also choose this baseline because it is a recent work using a state of art classification model. In addition, this baseline was designed to work with multiclass problems as well. Thus, we can determine if our method can outperform a more sophisticated method by applying a traditional classifier and a smaller set of genes.

We also compared the relevance of the subset of genes found by OGS to the then PAM50 gene list, since we understand that a single set of genes to study all the subtypes may not suffice to understand each subtype's particularities. In these experiments, we applied grid-search to find the best parameters for the classifier model and validated the results using *Precision*, *Recall*, F_1, *Accuracy*, *MCC*, *precision-recall* curve, *Specificity* and *NPV* as performance measures.

[2] `https://github.com/chenlabgccri/CancerTypePrediction`

4.3 Experiments and Results

This section describes the results obtained when applying the outlier method on a gene expression dataset. We compared the proposed method with different methods found in the literature regarding breast cancer subtype classification. Additionally, we presented that our outlier-based method is also suited for subtype signature identification.

4.3.1 Breast Cancer Subtype Classification

In the classification task, besides the 1D-CNN method proposed by our baseline, we implemented three distinct methods based on the PAM50 gene list:

- **PAM50** - We applied our hierarchical classifier.
- **RFE-PAM50** - We used the recursive feature elimination and the hierarchical classifier.
- **FLAT-PAM50** - A multiclass SVM was employed without any changes on the genes.

We also developed two hybrid methods (H-OGS and PH-OGS) that combines the classification of OGS with the RFE-PAM50 method and the genes and proteins from OGS. All of these methods were also validated using the protein datasets. For better a understanding, we identified those methods using a P identifier (P-OGS, P-1D-CNN, P-PAM50, P-RFE PAM50, P-FLAT PAM50).

In the first experiment, we compare the results obtained by all the methods in the subtype classification task. Figure 4.7 illustrates the performances in terms of macro *precision*, macro *recall*, and macro F_1. Comparing the performances obtained by all the methods, we see that PH-OGS outperformed all the other methods in *precision*, *recall* and F_1.

By using the outlier-based method, we found a small gene subset capable of achieving high classification results. Tables 4.4 and 4.5 show that we managed to select ten genes and achieved an F_1 score of 0.97, an $AUPRC$ of 0.96 and MCC of 0.96 when classifying the Basal samples.

For the Her 2 samples, we found eight genes and achieved an F_1 score of 0.86 and AUPRC of 0.79. We filtered to only 40 genes to classify both Luminals and achieved an F_1 score of 0.81 and AUPRC of 0.71 for Luminal A. For Luminal B, we achieved an F_1 score of 0.75 and an AUPRC of 0.62. Table 4.7 presents the subsets found by our proposed method for each of the subtypes.

Table 4.4: Classification results using *precision*, *recall*, F_1, and $AUPRC$ micro metrics. Best results in bold.

Performance on RNA dataset																
	Precision				*Recall*				F1				AUPRC			
Method	Basal	Her 2	Lum A	Lum B	Basal	Her 2	Lum A	Lum B	Basal	Her 2	Lum A	Lum B	Basal	Her 2	Lum A	Lum B
OGS	**1.00**	**1.00**	0.79	0.70	0.94	0.75	0.83	0.79	0.97	**0.86**	0.81	0.75	**0.96**	**0.79**	0.71	0.62
PAM50	**1.00**	0.73	0.92	0.86	0.94	**0.92**	**0.96**	0.75	0.97	0.81	0.94	0.80	0.96	0.69	0.89	0.72
1D-CNN	**1.00**	0.67	0.87	0.60	0.89	0.17	0.87	0.88	0.94	0.27	0.87	0.71	0.91	0.24	0.80	0.56
RFE PAM50	0.95	**1.00**	0.92	0.71	**1.00**	0.25	**0.96**	**0.92**	0.97	0.40	0.94	0.80	0.95	0.37	0.89	0.68
FLAT PAM50	**1.00**	0.77	**0.96**	**0.88**	0.94	0.83	**0.96**	0.88	0.97	0.80	**0.96**	**0.88**	0.96	0.67	**0.93**	**0.80**
H-OGS	**1.00**	**1.00**	0.92	0.81	0.94	0.75	**0.96**	**0.92**	0.97	**0.86**	0.94	0.86	0.96	**0.79**	0.89	0.77
PH-OGS	**1.00**	**1.00**	0.92	0.85	**1.00**	0.75	**0.96**	**0.92**	**1.00**	**0.86**	0.94	**0.88**	**1.00**	**0.79**	0.89	**0.80**
Performance on Protein dataset																
P-OGS	0.8	0.73	0.81	0.68	0.89	0.67	0.74	0.71	0.84	0.70	**0.77**	0.69	0.74	0.54	0.68	0.57
P-1D-CNN	0.79	0.39	0.74	0.75	0.61	**0.92**	0.74	0.38	0.69	0.55	0.74	0.50	0.57	0.37	0.62	0.48
P-PAM50	0.85	**0.82**	0.77	0.71	**0.94**	0.75	0.76	0.71	0.89	**0.78**	0.76	0.71	0.82	**0.65**	0.65	0.59
P-RFE PAM50	**0.89**	**0.82**	0.77	0.72	**0.94**	0.75	0.74	**0.75**	**0.92**	**0.78**	0.76	**0.73**	**0.86**	0.65	0.65	**0.62**
P-FLAT PAM50	0.76	0.60	**0.82**	**0.79**	0.89	0.75	**0.78**	0.62	0.82	0.67	0.80	0.70	0.70	0.49	**0.71**	0.61

Table 4.5: Classification results using *MCC*, *NPV* and *specificity* micro metrics. The best results for each metric are bold.

	MCC				*NPV*				*Specificity*			
Method	Basal	Her 2	Lum A	Lum B	Basal	Her 2	Lum A	Lum B	Basal	Her 2	Lum A	Lum B
Performance on RNA dataset												
OGS	0.96	**0.85**	0.72	0.62	0.98	0.96	0.92	0.90	**1.00**	**1.00**	0.91	0.85
PAM50	0.96	0.78	0.91	0.72	0.98	**0.98**	**0.98**	0.89	**1.00**	0.94	0.96	**0.94**
1D-CNN	0.93	0.28	0.81	0.57	0.97	0.86	0.94	0.93	**1.00**	0.98	0.94	0.74
RFE PAM50	0.97	0.47	0.91	0.71	**1.00**	0.88	**0.98**	**0.96**	0.98	**1.00**	0.96	0.83
FLAT PAM50	0.96	0.76	**0.94**	**0.82**	0.98	0.97	**0.98**	0.94	**1.00**	0.95	**0.98**	**0.94**
H-OGS	0.96	**0.85**	0.91	0.80	0.98	0.96	**0.98**	**0.96**	**1.00**	**1.00**	0.96	0.91
PH-OGS	**1.00**	**0.85**	0.91	**0.82**	**1.00**	0.96	**0.98**	**0.96**	**1.00**	**1.00**	0.96	0.92
Performance on Protein dataset												
P-OGS	0.79	0.64	0.68	0.55	0.96	0.93	0.89	0.86	0.93	0.95	**0.92**	0.84
P-1D-CNN	0.61	0.49	0.62	0.40	0.88	**0.97**	0.88	0.76	0.94	0.73	0.88	0.94
P-PAM50	0.86	**0.74**	0.65	0.57	**0.98**	0.95	0.89	0.86	0.94	**0.96**	0.90	0.86
P-RFE PAM50	**0.89**	**0.74**	0.65	0.61	**0.98**	0.95	0.89	**0.88**	**0.96**	**0.96**	0.90	0.86
P-FLAT PAM50	0.76	0.60	**0.71**	0.59	0.96	0.95	**0.90**	0.84	0.91	0.90	**0.92**	**0.92**

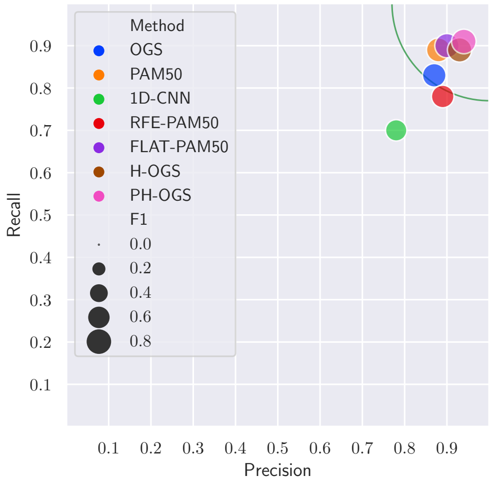


Figure 4.7: Performance obtained by all methods in terms of macro *precision*, macro *recall* and macro F_1.

For the Basal and Her 2 subtypes we used the opposite outlier approach for gene selection (Subsection 4.1.3), since those subtypes have distinct characteristics when compared to the others. However, when dealing with Luminal subtypes, we have to change our gene selection method for the outlier vs. normal approach, because both subtypes share an extensive amount of outlier genes.

Figure 4.8 shows the confusion matrices obtained by each method. Each row represents the instances of an actual class and each column represents the instances of a predicted class The baseline (1D-CNN) had trouble distinguishing the Her 2 class, performing poorly. We can observe that our method PH-OGS is the only one capable of classifying the Basal subtype correctly.

Tables 4.4 and 4.5 presents the micro performances in terms of *precision*, *recall*, F_1, $AUPRC$, *specificity* and NPV. Our method performed better when classifying the two subtypes with worst prognosis, with F_1 equal to 0.97 and AUPRC of 0.96 on the Basal subtype and F_1 equal to 0.86 and AUPRC of 0.79 on the Her 2 subtype. The *specificity* for both class were 1.00 (Table 4.5).

But even both Luminal subtypes, that tend to be more similar between them, we achieved promising results, with F_1 of 0.81 for Luminal A and 0.75 for Luminal B. Compared to the baseline, OGS outperforms in three out of four classes, achieving a macro *accuracy* score of 0.83 versus 0.77 and MCC of 0.77 versus 0.69, and also higher macro NPV of 0.94 and macro *specificity* of 0.93 (Table 4.6).

Comparing the results of our method with the three PAM50 methods in Tables 4.4 and 4.5, we see that our method still outperforms all the other methods when classifying

Table 4.6: Classification results using macro metrics. The best results for each metric are bold.

Method	F1 Macro	*ACC*	*MCC*	AVG AUPRC	AVG NPV	AVG Specificity
Performance on RNA dataset						
OGS	0.85	0.83	0.77	0.73	0.94	0.93
PAM50	0.88	0.88	0.84	0.81	0.95	0.96
1D-CNN	0.70	0.77	0.69	0.65	0.93	0.90
RFE PAM50	0.78	0.84	0.80	0.75	0.96	0.93
FLAT PAM50	0.90	0.91	0.88	0.85	0.97	**0.97**
H-OGS	0.91	0.91	0.88	0.85	0.97	0.96
PH-OGS	**0.92**	**0.92**	**0.89**	**0.87**	**0.98**	**0.97**
Performance on Protein dataset						
P-OGS	0.75	0.75	0.66	0.63	0.91	0.91
P-1D-CNN	0.62	0.62	0.53	0.48	0.87	0.9
P-PAM50	**0.79**	0.78	0.7	0.66	**0.92**	0.91
P-RFE PAM50	**0.79**	**0.8**	**0.72**	**0.68**	**0.92**	**0.92**
P-FLAT PAM50	0.75	0.75	0.67	0.63	0.91	**0.92**

the Basal and her2 subtypes, using fewer genes. As can be seen in Figures 4.8b, 4.8f and 4.8e, all of the methods have difficulty in distinguishing Her 2 from the other subtypes.

In Figure 4.8g, we can see that PH-OGS is the best method among all, where it only miss-classifies six out of 77 samples and correctly classifies all the 18 Basal samples, the subtype with the worst prognosis.

Table 4.8 presents the genes used for the subtypes classification using RFE-PAM50. The RFE-PAM50 also achieved high results using a small set of genes for the Basal classification, only five genes were necessary, while for Her 2, two genes were selected. To distinguish the Luminal samples, 48 genes were used.

We introduced two hybrid methods (H-OGS and PH-OGS) that achieved a higher F_1 score and *AUPRC* for the two subtypes with the worst prognosis. H-OGS achieved

Table 4.7: The best number of genes selected by the OGS for each subtype.

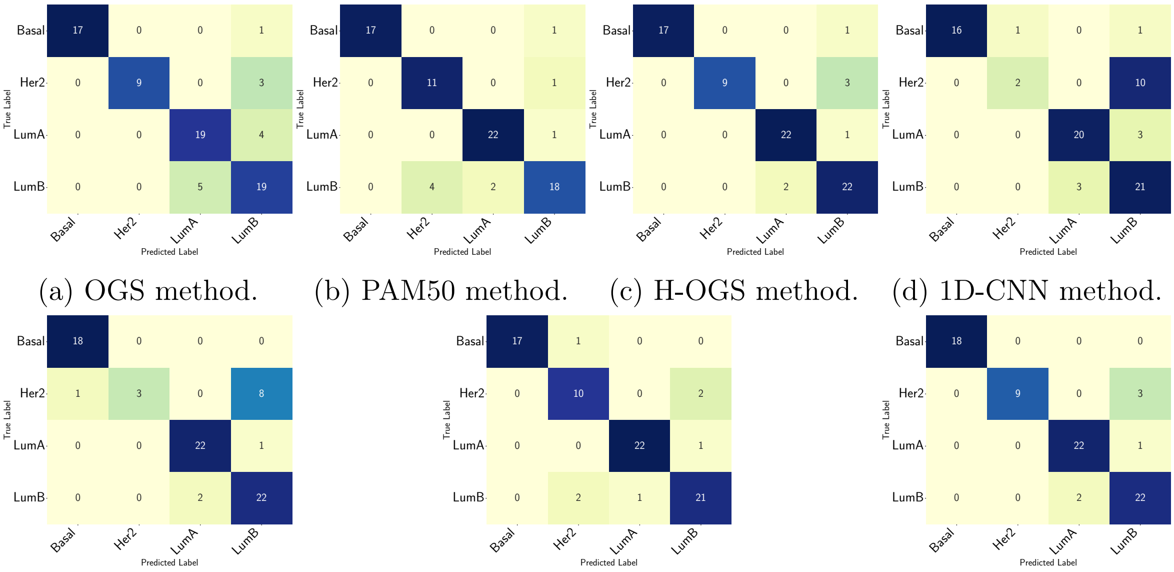
Subtype	**# of genes**	**Genes**
Basal	10	AGR2, AGR3, EN1, **FOXA1**, **FOXC1**, FZD9, KIAA1324, PRR15, SPDEF, TMC5
Her 2	8	C1orF106, CEACAM5, FBXO10, GRIK3, GRPR, MICALCL, **PGR**, TMEM145
Luminals	40	ADAM33, ADAMTS18, ATHL1, C1QTNF7, C1orf95, CACNG1, CD320, COQ3, CPA4, CX3CR1, HEPACAM2, HN1, HSPB6, KCNC3, KRT222, LGALS12, LLPH, LPCAT2, LRPAP1, **MYBL2**, PEBP4, PLIN1, PTPMT1, RBM44, RTN1, SCUBE2, SHROOM1, SLC40A1, SLC7A5, SNTB2, STAC3, STK32B, TAF6L, TEKT3, TIMP4, UCN, WTIP, YJEFN3, ZNF551, ZNF628C

Table 4.8: The best number of genes selected by the RFE-PAM50 method for each subtype.

Subtype	**# of genes**	**Genes**
Basal	5	BCL2, CEP55, FOXC1, SFRP1, TYMS
Her 2	4	BCL2, FOXC1, SFRP1, TYMS
Luminals	48	ACTR3B, ANLN, BAG1, BCL2, BIRC5, BLVRA, CCNB1, CCNE1, CDC20, CDC6, CDH3, CENPF, CEP55, CXXC5, EGFR, ERBB2, ESR1, EXO1, FGFR4, FOXA1, FOXC1, GPR160, GRB7, KIF2C, KRT14, KRT17, KRT5, MAPT, MDM2, MELK, MIA, MKI67, MLPH, MMP11, MYBL2, MYC, NAT1, NDC80, NUF2, ORC6, PGR, PHGDH, PTTG1, SFRP1, SLC39A6, TMEM45B, TYMS, UBE2C

an F_1 score of 0.97 and $AUPRC$ of 0.96 for the Basal subtype, while Her 2 presented an F_1 score of 0.86 and $AUPRC$ of 0.79 (Table 4.4). In this hybrid method, we combined the OGS with the RFE-PAM50 method. The first two subtypes are classified with the OGS genes, while to distinguish the Luminal subtypes, we used the RFE-PAM50 genes.

This hybrid method improved the Basal and Her 2 classifications because OGS does not penalize the classifier regarding Her 2 subtype (Table 4.6). H-OGS also outperforms all the other methods in macro NPV, achieving a performance of 0.97. When we look to the *specificity* micro metrics, we see that every Basal and Her 2 samples classified as Basal and Her 2 by H-OGS were correctly classified. Table 4.9 presents the genes used by H-OGS to classify each subtype.



(a) OGS method. (b) PAM50 method. (c) H-OGS method. (d) 1D-CNN method.

(e) RFE-PAM50 method. (f) FLAT-PAM50 method. (g) PH-OGS method.

Figure 4.8: Confusion matrices obtained by each method.

The other hybrid method is the PH-OGS, this method uses the genes found by OGS with the genes that direct the synthesis of the proteins found by OGS on the protein dataset. On our experiments, when we use the OGS on the protein dataset, the P-OGS presents the worst performance among all the other methods, achieving the best F_1 score of only 0.84 for the Basal subtype (Table 4.4).

On the other hand, when using the expression of the proteins that are synthesized by PAM50 genes, we achieve better results. The P-RFE PAM50 and the P-PAM50 reach an F_1 macro of 0.79 4.6, while the P-FLAT PAM50 achieves results that are similar to the P-OGS approach 4.6. When comparing the overall results of the classification methods using the expression of genes or proteins, we can see that using the expression of genes leads to better results compared to the use of the expression of proteins. Nevertheless, we wanted to verify the impact of outlier proteins on the H-OGS method.

The results show that the outlier proteins found give complementary information, therefore increasing the classification score for the Basal subtype. PH-OGS achieved an F_1 score of 1.00 and $AUPRC$ of 1.m for the Basal subtype, while Her 2 presented an F_1 score of 0.86 and $AUPRC$ of 0.79. When analysing the macro metrics (Table 4.6), PH-OGS outperforms the other methods in all the metrics, with a F_1 macro and *accuracy* of 0.92. The resulted genes for each subtype classification is presented in Table 4.10.

Therefore, our PH-OGS uses relatively fewer genes than the compared ap-

Table 4.9: The best number of genes selected by the hybrid method (H-OGS) for each subtype.

Subtype	# of genes	Genes
Basal	10	AGR2, AGR3, EN1, FOXA1, FOXC1, FZD9, KIAA1324, PRR15, SPDEF, TMC5
Her 2	8	C1orF106, CEACAM5, FBXO10, GRIK3, GRPR, MICALCL, PGR, TMEM145
Luminals	48	ACTR3B, ANLN, BAG1, BCL2, BIRC5, BLVRA, CCNB1, CCNE1, CDC20, CDC6, CDH3, CENPF, CEP55, CXXC5, EGFR, ERBB2, ESR1, EXO1, FGFR4, FOXA1, FOXC1, GPR160, GRB7, KIF2C, KRT14, KRT17, KRT5, MAPT, MDM2, MELK, MIA, MKI67, MLPH, MMP11, MYBL2, MYC, NAT1, NDC80, NUF2, ORC6, PGR, PHGDH, PTTG1, SFRP1, SLC39A6, TMEM45B, TYMS, UBE2C

Table 4.10: The best number of genes selected by the protein hybrid (PH-OGS) method for each subtype

Subtype	# of genes	Genes
Basal	18	AGR3, CEACAM6, CLDND2, CLSTN2, CYP4Z2P, GP2, MAPT, MB, MBOAT2, MLPH, NEK10, NRK, S100A8, SCGB2A2, SCUBE2, SERPINA5, TMEM45B, TTC39A
Her 2	8	C1orf106, CEACAM5, FBXO10, GRIK3, GRPR, MICALCL, PGR, TMEM145
Luminals	48	ACTR3B, ANLN, BAG1, BCL2, BIRC5, BLVRA, CCNB1, CCNE1, CDC20, CDC6, CDH3, CENPF, CEP55, CXXC5, EGFR, ERBB2, ESR1, EXO1, FGFR4, FOXA1, FOXC1, GPR160, GRB7, KIF2C, KRT14, KRT17, KRT5, MAPT, MDM2, MELK, MIA, MKI67, MLPH, MMP11, MYBL2, MYC, NAT1, NDC80, NUF2, ORC6, PGR, PHGDH, PTTG1, SFRP1, SLC39A6, TMEM45B, TYMS, UBE2C

proaches. While FLAT-PAM50 uses 50 genes for all the subtypes, we can achieve better results for Basal and her two subtypes using only 18 and 8 genes, respectively. The 1D-CNN uses $7,000$ genes and present the worst result among the methods tested. In general, given the high F_1 score for all the subtypes, it is clear that our proposed approach is viable once the fewer genes needed to identify the subtype, less computation is required for the classification task.

4.3.2 Analysis of the Basal signature

While predicting if a sample is cancerous or not using gene expression data may be relatively simple, predicting cancer subtypes, such as breast cancer, is an ongoing research topic. In addition to using the outlier based-method for classification, we also used it to identify genetic signatures that may characterize a particular breast cancer subtype.

We conducted the experiments on the Basal subtype since it is the one with the worst prognosis among all breast cancer subtypes [Bertucci et al., 2012], so it is imperative to identify a small set of genes capable of characterizing this subtype.

To do this task, we had to find outlier gene sets among the same subtype that do not intersect with each other and yet are capable of clustering all the samples from this subtype in the same group. The assumption is that if we can cluster the Basal subtype correctly, even with k-means, then the set of genes is relevant. For this experiment we used the Cptac 2D dataset (Table 4.1).

To find distinct signatures, we applied different gene filtering methods (e.g., Forward Selection, RFE) in the gene selection step of our method (Subsection 4.1.3). After finding that both gene filtering methods returned gene sets with more than 80% of intersection, we defined a new outlier threshold. When we add new genes to the dataset by changing the outlier threshold, the gene filtering methods present distinct results, since the methods are based on the whole group of features, so when you add or remove new features, the results tend to be diverse.

Regardless of finding distinct signatures using the *outlier vs. normal* approach (Subsection 4.1.2), we also applied this to the *opposite outlier* approach, whereas those two approaches present a distinct set of genes. By empirically testing several outlier thresholds and gene filtering methods on *opposite outlier* and *outlier vs. normal genes*, we were able to find three distinct signatures that correctly cluster all Basal samples in the same group. Table 4.11 presents the signatures found.

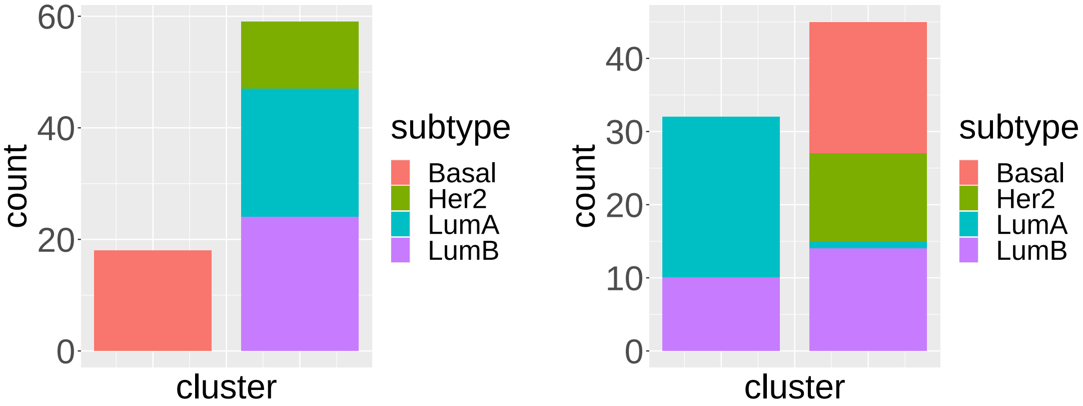
The results of our three distinct gene sets are presented in Figure 4.9a. As can be seen, regardless of the gene set used all Basal samples were clustered together. While by using the PAM50 gene list 100% of the her2, 58% of Luminal B, and 4% of Luminal A samples in the same cluster with the Basal samples (Figure 4.9b).

Interestingly, the intersection of the three gene sets found and the PAM50 includes only four genes: ESR1, FOXA1 MLPH and TMEM45B (three of them widely known to be related to breast cancer).

Since we were able to find distinct signatures for Basal breast subtype, it is necessary to do a biological investigation in those genes found that are not presented

Table 4.11: Distinct signatures for Basal subtype identification. Genes that intersect with the PAM50 are bold.

Gene filtering method	Recursive Feature Elimination		Forward Selection
Outlier Threshold	45%	55%	40%
Outlier Approach	Opposite Outlier (A)	Outlier vs. normal (B)	Opposite outlier (C)
Number of Genes	22	9	9
Genes	ACOX2, AGAP11, ART3, C1orf168, C5orf34, CMBL, COCH, **ESR1**, FERMT1, GPR98, HDC, HPX, IL20RB, MGC16142, **MLPH**, MPV17L, NAPRT1, PLA2G4F, RERG, SLCO5A1, TRPM8, TTYH1	AFF3, BCAS1, C9orf152, DNAJC12, **FOXA1**, FZD9, RET, SLC44A4, ST8SIA1	A2ML1, ABCA12, ABCC11, ADAMTS15, AGR3, ANKRD30A, ANKRD30B, CA12, **TMEM45B**



(a) Outlier-based Method and RFE-PAM50.

(b) PAM50 genes.

Figure 4.9: Outlier-based method vs. PAM50.

in the PAM50 gene list to identify which role they play in developing the associated subtype.

We also conducted the experiments using the RFE-PAM50 method, to identify if there is a subset of genes in the PAM50 gene list capable of group all the Basal samples in the same cluster, therefore identifying a PAM50 gene signature for Basal subtype.

Figure 4.9a shows that we can find a signature for the Basal subtype using the PAM50 gene list. We used only 13 genes among all the 50 genes available. Table 4.12 presents the genes used to cluster all the Basal samples.

Table 4.12: PAM50 signature for Basal subtype.

Method	**# of genes**	**Genes**
RFE-PAM50	13	BCL2, CEP55, CXXC5, ERBB2, FOXC1, MIA, MLPH, NDC80, NUF2, PGR, PTTG1, SFRP1, TYMS

We analyzed the signatures' pathway using the gene set enrichment analysis (GSEA)[3]. Signature B presents five out of nine genes capable of distinguishing between Luminal And Basal subtypes, finding genes up-regulated in Luminal Breast cancer subtype compared to the Basal one (FOXA1, BCAS1, AFF3, C9orf152, SLC44A4) and four genes down-regulated in Basal subtype (BCAS1, DNAJC12, RET, SLC44A4).

Signature C presents among the nine genes, five that are up-regulated in the Luminal Breast cancer subtype (CA12, ABCC11, ABCA12, TMEM45B, ANKRD30A) and two genes that are differentially expressed between Basal and Luminal (ABCA12 and ANKRD30A).

Signature A does not show any pathway related to breast cancer subtypes, indicating that these genes need further investigation since they are not associated with any breast cancer pathway but can distinguish the Basal subtype. However, this signature presents two genes (ESR1 and MLPH) that are in the PAM50 gene list.

Individually analyzing those genes, Supplementary Table I shows that signature A contains only three genes associated with breast cancer (ESR1, RERG, TRPM8). However, although not directly associated, other genes are known to be related to breast cancer. ACOX2 is related to a potential novel therapeutic biomarker in ER+ breast tumors [Bjørklund et al., 2015]. AGAP11 is related to treatment and prognosis targets and biomarkers for breast cancer [Wang et al., 2019]. ART3 is a critical triple negative breast cancer marker with functional significance [Tan et al., 2016].

FOXA1 and MLPH was implicated in the development of breast cancer [He et al., 2015]. FERMT1 belongs to a six-gene signature that can classify breast tumors with a higher propensity to metastasize to the lungs, independent of the molecular subtype [Driouch et al., 2009]. HPX is closely related to breast cancer and may be involved in robust detection of disease progression [Cine et al., 2014]. C5orf34 and MGC16142

[3] https://www.gsea-msigdb.org/gsea

have uncharacterized proteins associated, studies on these are necessary for better understanding.

The other genes found in signature A (ART3, C1orf168, CMBL, COCH, GPR98, HDC, IL20RB, MPV17L, NAPRT1, PLA2G4F, SLCO5A1, and TTYH) do not present documented relation with breast cancer, but are related to several distinct diseases, indicating the necessity to understand the link between these genes and breast cancer.

Signature B (Supplementary Table II) presents three genes (BCAS1, DNAJC12, FOXA1) individually associated with breast cancer. Researchers show that AFF3 over expression in breast cancer cells resulted in tamoxifen resistance[Shi et al., 2018]. RET can have an important role in breast cancer, but only in the subset of ER+ tumors, where it is found overexpressed [Nigro et al., 2019]. SLC44A4 is found high expressed in breast cancer. The C9orf152 is an uncharacterized protein. FZD9 and ST8SIA1 are associated with two distinct syndromes.

Signature C (Supplementary Table III) presents five genes related to breast cancer (ABCC11, ADAMTS15, AGR3, ANKRD30A, ANKRD30B). Besides that, ABCA12 presents potential modifiers of progression and response to the chemotherapy of breast cancer. CA12 can be seen as a new prognostic indicator and even a new target for treatment. A2ML1 is a top 10 upregulated gene in breast cancer. High TMEM45A expression is associated with poor prognostic in breast cancer patients.

Those results indicate that our gene selection method is a viable way to identify relevant distinct signatures. The genes that are not related to breast cancer need further investigation to identify potential pathways or the impact in the tumour.

4.4 Chapter Remarks

In this Chapter, we proposed a method that identifies relevant outlier genes (uncommonly over-expressed or under-expressed) that assist in gene signature identification and breast cancer subtype classification. As a differential, we highlight that our method identifies distinct subsets of genes for each of the breast cancer subtypes.

Compared to the state of the art approach, based on convolutional neural network, our method, OGS, reached a gain of at least 3% in relation to the F_1 metric (quality of classification) in two out of four classes. In Her 2 class, our method reached a gain of more than 50%.

We also presented a hybrid method (PH-OGS), where we improved the classification results for the four subtypes, achieving an F_1 score of 1.0 and 0.86 for the Basal and Her 2, respectively. The two subtypes with the worst prognosis.

Besides, considering the number of genes, the proposed method presented fewer genes than the baseline and the PAM50. The efficiency gain results from the good gene filtering method proposed, which is capable of selecting a few highly relevant genes.

By analysing the signatures, our method was able to identify three distinct signatures capable of cluster all the Basal samples in the same group. Two of the three signatures have pathways that are related to breast cancer subtypes.

5

Classification of breast cancer subtypes: A study based on representative genes

Efficient, accurate, and objective cancer diagnosis is commonly achieved through the use of classification methods. Identifying the biological subtype of breast cancer can provide patients significant prognostic and predictive information, making correct classification into subtypes critical for effective treatment [Parker et al., 2009].

Despite the abundance of available data, the genetic mapping of breast cancer and its subtypes remains incomplete. A set of fifty genes, known as PAM50 [Chia et al., 2012], has been established as a reference for characterizing breast cancer subtypes. However, further investigation is still necessary, as the PAM50 gene list alone is not precise enough to accurately distinguish between the subtypes.

This shows that although the study of gene expression is already a reality, there is still no definitive understanding of all genes related to breast cancer and, especially, a definitive understanding of the interactions between these genes. Thus, an important contribution to accurate diagnosis is identifying a subset of genes capable of characterizing the subtypes and differentiating them from each other.

In this context, we propose an evaluation framework that uses different machine learning techniques to classify breast cancer subtypes and investigate the features. Given the particularities of gene expression data, mainly caused by the sensitivity of different technologies for its acquisition, it is not a simple application of machine learning methods and packages from a computational point of view. Since depending on the technologies used (e.g., cDNA microarray [Schena et al., 1995] or oligonucleotide arrays [Lockhart et al., 1996]) to quantify the gene expression data. Results are presented differently.

Accordingly, there is a clear need to investigate existing techniques to treat these input data with different characteristics and reliability. Based on this, this Chapter presents (i) A study of different methods in the task of classifying breast cancer subtypes, (ii) an analysis of the PAM50 list in the classification of breast cancer subtypes, and (iii) a list of genes that are important for the classification in each subtype.

5.1 Evaluation Framework

The evaluation framework in this work consists of the following steps. (i) collection of databases that have gene expression; (ii) data pre-processing to select only the genes involved in the study; (iii) classification of samples among breast cancer subtypes; and (iv) analysis of the performance of classifiers with different evaluation metrics.

5.1.1 Dataset and Pre-processing

This framework starts by choosing the dataset, where the data can be extracted from genomic data repositories containing gene expression data. After the data collection phase, pre-processing is necessary to identify if the database has all 50 genes from the PAM50 list. Thus, among all the genes that exist in the chosen dataset, only the 50 genes from the PAM50 list are selected, if the 50 genes are not in the dataset, another dataset needs to be used to validate the work. In this step, we understand that employing the PAM50 genes in the classification task is already a way of feature selection since it reduces our scope from thousands of genes to only 50.

5.1.2 Classification

After selecting only the PAM50 genes for the training and testing basis, we classify them with different classifiers. This step aims to understand how different classification methods are able to distinguish breast cancer subtypes using gene expression data.

5.1.3 Data Analysis

To measure the performance of the methods, we apply traditional metrics such as *precision*, *recall*, *F-measure*, and *accuracy*. Since biological data usually have a sparse dataset [Chicco, 2017], we also measure the performance of the methods using the Matthews correlation coefficient (*MCC*) [Baldi et al., 2000] and *precision vs. recall* curve ($AUPRC$). Both metrics were selected because they are suitable for unbalanced databases. While MCC is more appropriate for binary classification, the *precision vs. recall* curve is a more reliable and informative indicator of statistical performance in multiclass problems [Chicco, 2017]. We also calculate *specificity* as this measure is used to see how correctly we can classify an individual into the correct cancer subtype [Parikh et al., 2008]. All The evaluation metrics are described and detailed in Chapter 2.

After performing the classification and evaluating the performance of each classifier, we apply the Shapley Values (SHAP) [Lundberg and Lee, 2017] to evaluate the feature importance for each of the classifiers. SHAP is a game theory based approach to describe the performance of a machine learning model. SHAP can provide explanations for local and global models, estimating feature contributions to the output of the model. To produce an interpretable model, SHAP uses an additive feature attribution method. It is interesting to note that SHAP estimates the feature importance (magnitude of the contribution) as well as the sign (positive or negative).

5.2 Evaluation of the proposed Framework

This section presents an analysis of the different classifiers used to classify breast cancer subtypes using the PAM50 gene set. Additionally, we detail the methodology used to apply the proposed approach.

5.2.1 Methodology

In this subsection, we describe the evaluation methodology used in this work. We detail the characteristics of the datasets used in the experiments. We present the parameters chosen for the machine learning methods and also explain the evaluation metrics used.

Dataset

We used two distinct gene expression datasets from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) [Edwards et al., 2015] to validate the methods. The datasets present the gene expression of patients with breast cancer. These specimens are divided into four intrinsic breast cancer subtypes (see Section 1).

To evaluate the performance of the classifiers, we employ a 10-fold cross-validation classification. We combined both datasets using ComBat [Johnson et al., 2007], a widely used tool for correcting technical biases in gene expression data. Therefore, obtaining a larger dataset with 194 samples. Then, we separated the resulting merged dataset into 70% for training and 30% for testing. The training set was used for the 10-fold cross-validation.

Table 5.1 summarizes the characteristics of the databases used in the experiments:

Table 5.1: Dataset description.

Dataset	**# of genes**	**Subtypes**	**# of samples**	**Total # of samples**
Cptac 2C	23,122	Basal	29	117
		Her 2	14	
		Luminal A	57	
		Luminal B	17	
Cptac 2D	16,525	Basal	18	77
		Her 2	12	
		Luminal A	23	
		Luminal B	24	

Classifiers

To evaluate the performance of the PAM50 list for classifying breast cancer subtypes, we employed five distinct methods. The Grid search [Bergstra and Bengio, 2012] was used to optimize the parameters for each classifier, the parameters set for the Grid Search were chosen empirically. Table 5.2 presents the chosen classifiers and parameters. The remaining parameters have been set to the scikit-learn[1] default configuration.

Table 5.2: Classifier parameters.

Method	Parameters
SVM(Linear)	$C = 0.1$
SVM(RBF)	$C = 1.1$
KNN	$p = 1$, $n\ neighbors = 5$, $weights = uniform$
Random Forest	$bootstrap = False$, $min\ samples\ split = 6$, $n\ estimators = 28$
XGBoost	$gamma = 0.04$, $learning\ rate = 0.07$

[1] https://scikit-learn.org/stable/

To calculate the evaluation metrics, we used the scikit-learn and pandas-ml[2] libraries. We compare the different classifiers to see which method performs better overall and for each subtype separately. We use the evaluation metrics presented in Section 5.1.

5.2.2 Results

Different classifiers take into account distinct ways of classifying samples into different classes. Some use spacing between classes to differentiate them (SVM), while others check which class predominates among the elements closest to the analyzed sample (KNN). Some use decision trees (Random Forest) to perform the classification, and others start from a primary hypothesis and try to improve it to reach a better result (XGboost). Therefore, we expect that there will be different results for the tested classifiers, even if they are submitted to the same test conditions.

Classification Analysis

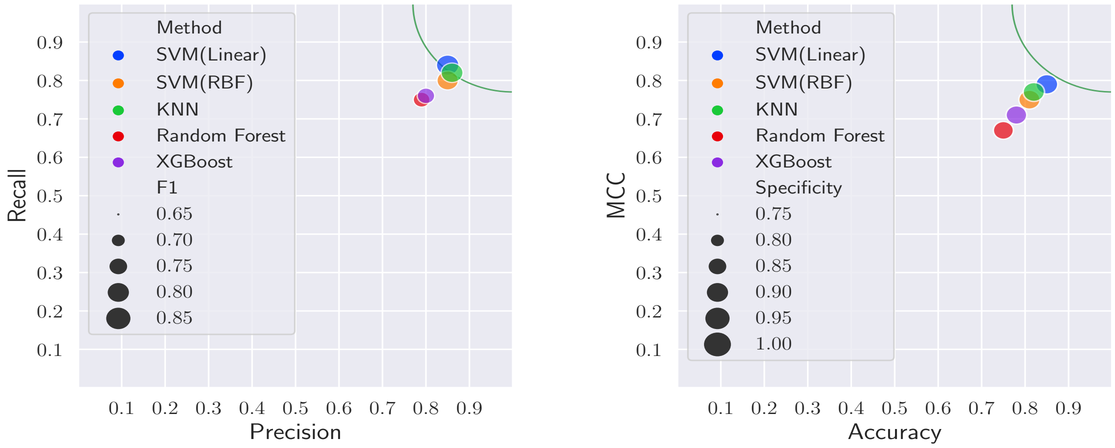
In the first experiment, we compared the results obtained by all methods to classify the four subtypes (Figure 5.1). Figure 5.1a illustrates performance in terms of precision, recall, and F_1. Figure 5.1b illustrates performance in terms of accuracy, MCC, and specificity. The X axis presents the precision (Figure 5.1a) and the accuracy (Figure 5.1b). The Y axis presents the recall (Figure 5.1a) and the specificity ((Figure 5.1b). The larger the circle, the higher the F_1 in Figure (Figure 5.1a) and the specificity in (Figure 5.1b). The color of the circle indicates the method.

Comparing the performance obtained by the methods, we see that SVM(Linear) outperformed all other methods in the six analyzed macro metrics. This performance can be explained by the fact that it is the classifier that best managed to separate the samples from Luminal A from Luminal B.

Table 5.3: Classification results using *precision*, *recall* and F_1 micro metrics. Best results in bold.

Method	Precision				Recall				F1			
	Basal	Her2	LumA	LumB	Basal	Her2	LumA	LumB	Basal	Her2	LumA	LumB
SVM (Linear)	**0.99**	0.91	**0.87**	**0.86**	0.96	**0.88**	**0.93**	**0.81**	**0.98**	0.90	**0.90**	**0.83**
SVM (RBF)	**0.99**	0.91	0.83	0.84	0.96	0.84	0.89	0.79	**0.98**	0.88	0.86	0.82
KNN	0.96	**0.97**	0.83	0.84	**0.99**	0.86	0.89	0.79	0.97	**0.91**	0.86	0.81
Random Forest	0.96	0.85	0.80	0.80	0.93	0.78	0.86	0.74	0.95	0.81	0.83	0.77
XGBoost	0.97	0.90	0.80	0.85	0.91	0.84	0.85	0.79	0.94	0.87	0.82	0.82

[2] https://pypi.org/project/pandas-ml/



(a) *Precision, recall and* F_1.

(b) *Accuracy, MCC and specificity.*

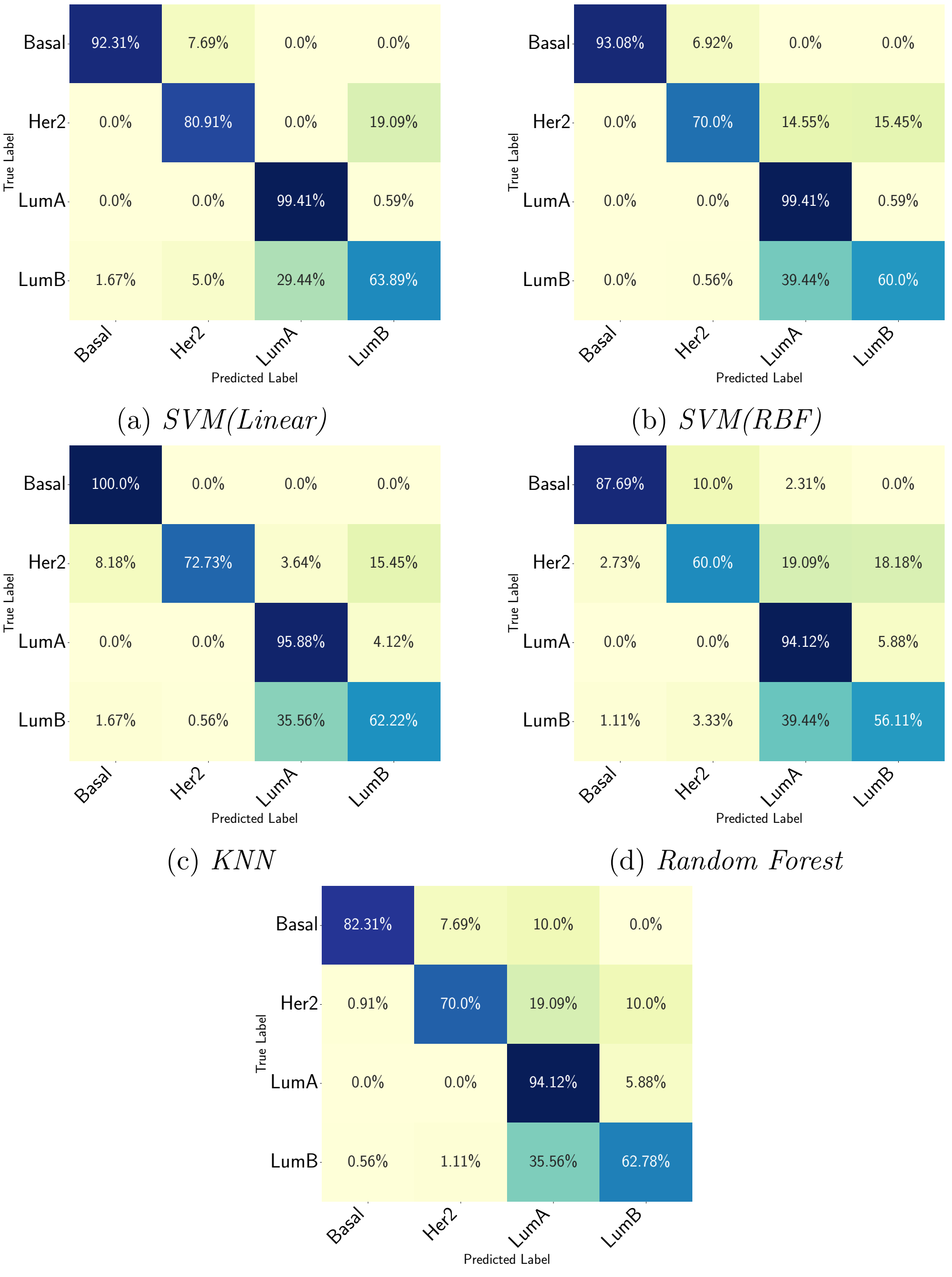
Figure 5.1: Performance obtained by the methods using macro metrics.

Table 5.4: Classification results using *MCC*, *AUPRC* and *Specificity* micro metrics. Best results in bold.

Method	*MCC*				*AUPRC*				*Specificity*			
	Basal	Her2	LumA	LumB	Basal	Her2	LumA	LumB	Basal	Her2	LumA	LumB
SVM(Linear)	**0.95**	0.79	**0.80**	**0.66**	0.96	**0.88**	**0.93**	**0.81**	**1.00**	0.97	**0.86**	**0.95**
SVM(RBF)	**0.95**	0.75	0.72	0.63	0.96	0.84	0.89	0.79	**1.00**	0.98	0.82	0.94
KNN	**0.95**	**0.82**	0.72	0.63	**0.99**	0.86	0.89	0.79	0.97	**1.00**	0.82	0.94
Random Forest	0.90	0.62	0.65	0.54	0.93	0.78	0.86	0.74	0.99	0.97	0.77	0.93
XGBoost	0.87	0.73	0.65	0.63	0.91	0.84	0.85	0.79	**1.00**	0.98	0.77	**0.95**

Figure 5.2 shows the average confusion matrices obtained by each method. Each row represents the instances of an actual class, and each column represents the instances of a predicted class. We can see that the results obtained with the *SVM(Linear)* (Figure 5.2a presented better results than the other classifiers, where 7.69% of the Basal samples were wrongly classified as Her 2, 19.09% of Her 2 samples and 0.59% of Luminal A misclassified as Luminal B. In the case of *SVM(RBF)* (Figure 5.2b, the classifier had difficulty separating the Her 2 samples, where 14.55% of the samples were incorrectly classified as Luminal A and 15.45% incorrectly classified as Luminal B.

The *KNN* classifier (Figure 5.2c) had the same classification problem as the *SVM(RBF)*, where 35.56% of the Luminal B samples were classified as Luminal A. In addition, 15.45% of the Her 2 samples were incorrectly classified as Luminal B. The KNN was the only method to correctly classify 100% of the Basal subtype. Tests with the *Random Forest* (Figure 5.2d) present that 39.44% of Luminal B were erroneously classified as Luminal A. Finally, *XGBoost* (Figure 5.2e) misclassified 27.69% of the



(a) *SVM(Linear)*

(b) *SVM(RBF)*

(c) *KNN*

(d) *Random Forest*

(e) *XGBoost*

Figure 5.2: Confusion matrices obtained by each method.

Basal samples and confused 35.56% of the Luminal B samples with Luminal A subtype.

In summary, the *SVM(Linear)* classifier provides the best performance, with the least amount of wrongly classified samples. It can also be noted that the subtype with the worst prognosis, Basal, had the few characterization problems, regardless of classifier, thus being the most characteristic subtype among the four. In contrast, the Luminal B subtype is the subtype where the classifiers have greater difficulty in

Table 5.5: Classification results using macro metrics compared with the related work. The best results for each metric are bold.

Method	# of genes	F_1 Macro	ACC	MCC
SVM(Linear)	50	**0.85**	**0.84**	**0.79**
SVM(RBF)	50	0.83	0.81	0.75
KNN	50	0.84	0.82	0.77
Random Forest	50	0.76	0.75	0.67
XGBoost	50	0.80	0.77	0.71
1D-CNN (Mostavi)	50	0.63	0.73	-

classifying the samples. It also happens that all classifiers have classified at least 10% of the Her 2 subtype as Luminal B.

Analyzing Table 5.3, containing the results obtained by the five classifiers tested, we can identify that the Basal subtype obtained a precision score of 99% in both SVM classifiers. The Luminal A subtype had the highest precision score of 87% with SVM(Linear), plus a recall above 85% in four of the five classification methods used. Observing the Her 2 subtype, it can be seen that it manages to obtain a score of 97% of precision with the KNN classifier, with a recall near 90% only with SVM(Linear) classifier. The Luminal B subtype had a maximum precision of 86% and a maximum recall of 81%.

When we analyze the F_1 score, the SVM(Linear) classifier holds three highest scores in the four subtypes, The Basal subtype had a score of 98%, while the Luminal A subtype had 90% and the Luminal B subtype had 83%. For the Her 2 subtype, the KNN classifier had a score of 91%.

Table 5.4 contains the results for the metrics *MCC*, *AUPRC* and *Specificity*. Examining the data obtained with the metric *MCC*, we notice that the classifier SVM(Linear) along with SVM(RBF) and KNN have the highest scores for the subtypes Basal with 95%. Only SVM(Linear) achieves a *MCC* of 80% in Luminal A. In contrast, the Her 2 subtype had a maximum score of 82% with the *KNN* classifier.

In the *AUPRC* metric, the SVM(Linear) classifier obtained scores of 88% for the Her 2 subtype, 93% for the Luminal A subtype, and 81% for the Luminal B subtype. The KNN classifier obtained an AUPRC of 99% for the Basal subtype. Analyzing the *Specificity*, we notice that the Basal subtype got 100% with the classifiers SVM(Linear), SVM(RBF), and XGBoost. The Her 2 subtype obtained a maximum of 100% using the KNN classifier. The Luminal A subtype obtained 86% with the SVM(Linear) classifier, and the Luminal B subtype obtained 95% of *Specificity* using the SVM(Linear) and

XGBoost.

Analyzing the data from the Table 5.5, with the results of the macro metrics, and the results of the related works. We can conclude that SVM(Linear) outperforms all of the other classifiers. The work of Mostavi et al. [2020] uses a CNN but only achieves a F_1 score of 63%. The SVM(Linear) has the best score in the following employed evaluation metrics: 85% for F_1, 84% for *Accuracy*, 79% for *MCC*, 89% for *AUPRC* and 94% for *specificity*.

In general, we notice that the Basal, the subtype with the worst prognosis, is the most characteristic among all since the classifiers obtained better results in this subtype. Concerning Her 2 (subtype with the second-worst prognosis), we noticed that it obtained the second best result among the subtypes. While Luminal B had the worst result, thus being the most difficult to be classified. Finally, the results showed that the evaluation framework combined reveals that the PAM50 gene list has good results when classifying the breast cancer subtypes, and the SVM(Linear) is the best classifier to employ.

Gene Analysis

In this step, we analyze which features (genes) are more important for each of the methods to classify the breast cancer subtypes using SHAP values (Section 5.1). Although to compute the SHAP values, we face an exponential computational complexity [Messalas et al., 2019], in the scope of our project, we were able to apply it to all samples since our merged dataset has 194 samples.

Figure 5.3 shows the features SHAP values obtained by each method. The larger the bar, the more critical the gene is for the classification of the subtype. We can see that distinct classifiers present distinct gene importance. For the SVM(Linear) (Figure 5.3a), the larger bars belong to the Basal subtype, therefore showing why this subtype has the best classification score.

For the SVM(RBF) (Figure 5.3b), the genes are also important in classifying the Luminal B subtype, explaining why these methods only looses to SVM(Linear) when classifying the Luminal B subtype. The KNN (Figure 5.3c) also presents the genes as important for the Basal subtype, explaining the performance in the classification of this subtype. For the other subtypes, this method presents similar results when compared with SVM(RBF) and Random Forest.

The Random Forest (Figure 5.3d) presents genes very important for the basal subtype classification and also for Her 2 and Luminal A. Finally, the XGBoost (Figure 5.3e) presents the worst result among all the methods, this can be explained since the XGBoost does not perform well when trained on small datasets [Rácz et al., 2021].

Summarizing the results, we can see that the ESR1 gene has almost the same importance for each classifier and is the most important gene to classify the Her 2 subtype. This is because mutations on this gene are acquired frequently in metastatic hormone receptor-positive breast cancer [Turner et al., 2020].

When we look for the PGR gene, we can see that it is more important for the Luminal A subtype in all the methods. This is because the expression of PGR is a potent prognostic indicator for evaluating the long-term prognosis of Luminal A [Kurozumi et al., 2017]. The FOXA1 and MLPH genes are the more critical genes to classify the basal subtype. Both are implicated in the development of breast cancer [He et al., 2015]. FOXA1 segregates with genes that characterize the luminal subtypes in DNA microarray analyses [Badve et al., 2007].

In our next experiment (Figure 5.4), we analyze the SHAP values individually for each class (subtype). We chose the SVM(*Linear*) for this analysis since it presented the best classification performance among all the methods tested (Subsection 5.2.2). This summary plot shows the importance of features and how their SHAP values are spread across the data. The plot uses SHAP values to show the distribution of each feature's impacts on the model output. The dots represent each sample in the test dataset.

For each subtype (Figures 5.4a, 5.4b, 5.4c and 5.4d), we can see which features are most influential in the model's output, the importance of the features are ranked in ascending order. For example, in Figure 5.4a, the ESR1 gene is the more important gene for the Basal subtype, while the CCNE1 gene is the 15th more important gene.

The horizontal location of the samples (the dots across the plot) shows whether the effect of that value is associated with a higher or lower prediction, and the color shows whether that variable is high (in red) or low (in blue) for that observation. As can be seen in Figures 5.4a, 5.4b, 5.4c and 5.4d, the more important gene for each subtype is more spread across the plots.

We can see that the importance of genes is different for each subtype. It is interesting to note that ESR1 is the most important gene for Basal, Her 2, and Luminal B, while for Luminal A is the second most important gene. We can also see that the distribution of the samples varies depending on the subtype.

For example, while for the Basal subtype (Figure 5.4a) most samples have negative SHAP values and a high correlation with the gene expression, for the Luminal B subtype (Figure 5.4d), most samples have positive SHAP values. These results complement the experiments presented in Figure 5.3a, as they demonstrate the behavior of the SHAP values in each of the samples for each of the subtypes.

5.3 Chapter Remarks

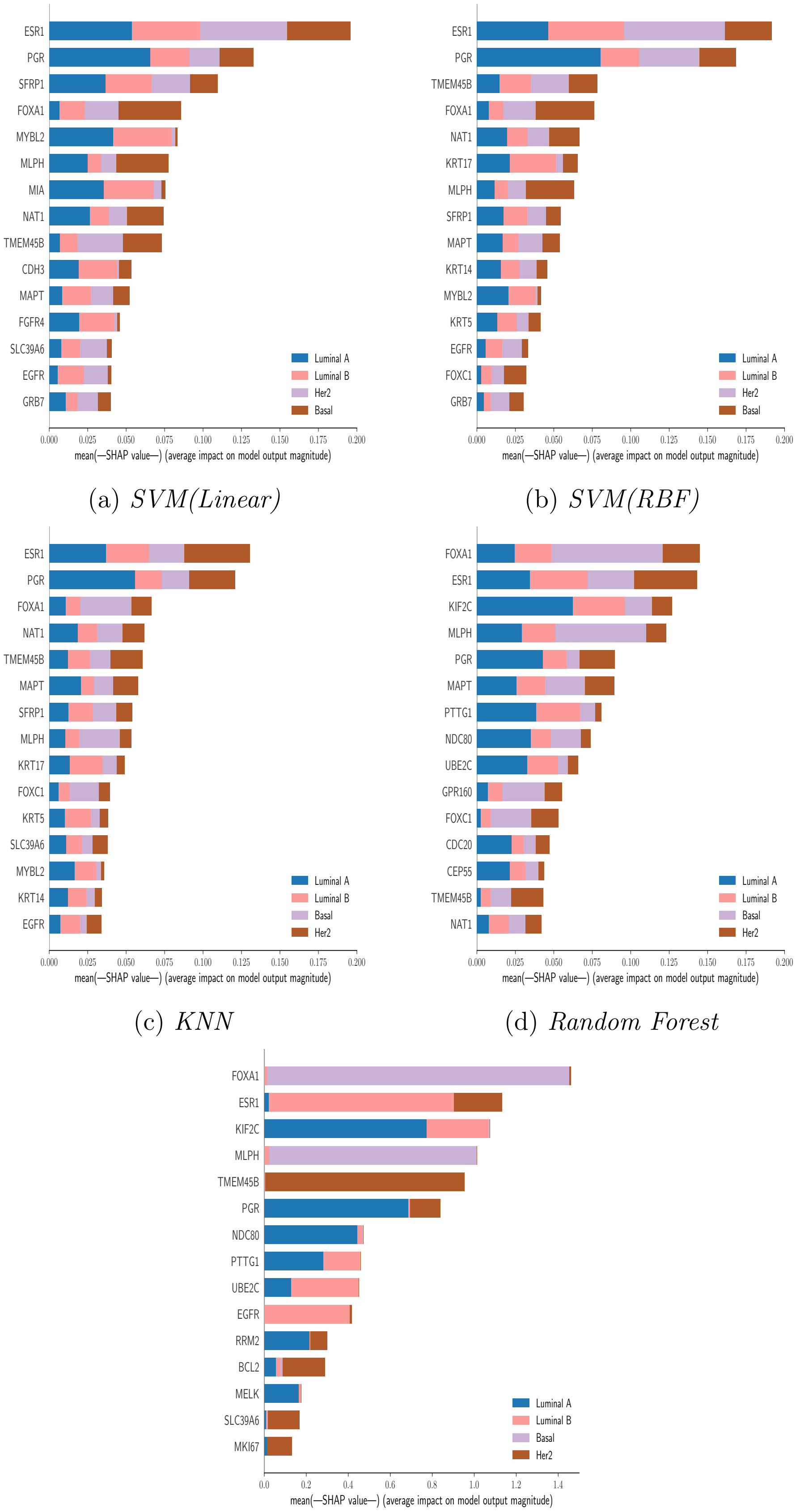
This chapter presented an evaluation framework for classifying breast cancer subtypes based on the PAM50 gene list. We employed distinct classification methods, each with different characteristics, to analyze whether there is a difference between them when classifying the breast cancer subtypes. Seven evaluation metrics were employed to evaluate the methods to get an overview of how the methods perform.

As a result, we noticed that the SVM(Linear) obtained better macro results than the others. We also verified that the Basal subtype (the one with the worst prognosis and the most characteristic), the classifier KNN outperformed all the other methods, reaching an F1 score of 100%. In addition, the other classifiers remained with a score above 80% for this subtype.

It is noticed that Her 2, the subtype with the second-worst prognosis, has the third best results in the classification. It reaches a maximum F1 score of 80%, achieving a minimum of 60% with the classifier Random Forest, in which the Her 2 samples are confused with all other subtypes.

Among the Luminal A and Luminal B subtypes, there is confusion between the samples, given that they are highly correlated. Although the PAM50 has only 50 genes, this is a good set for ranking as it scored in four of the five classifiers an F1 Macro score above 75%. At the micro-level, SVM(Linear) also managed to maintain an F1 score above 82% for all subtypes.

When analyzing the features, we can see that SHAP values identify the more important genes for the classification of each subtype and when we study those genes, we understand how they are related to the classified subtypes.



(a) *SVM(Linear)*

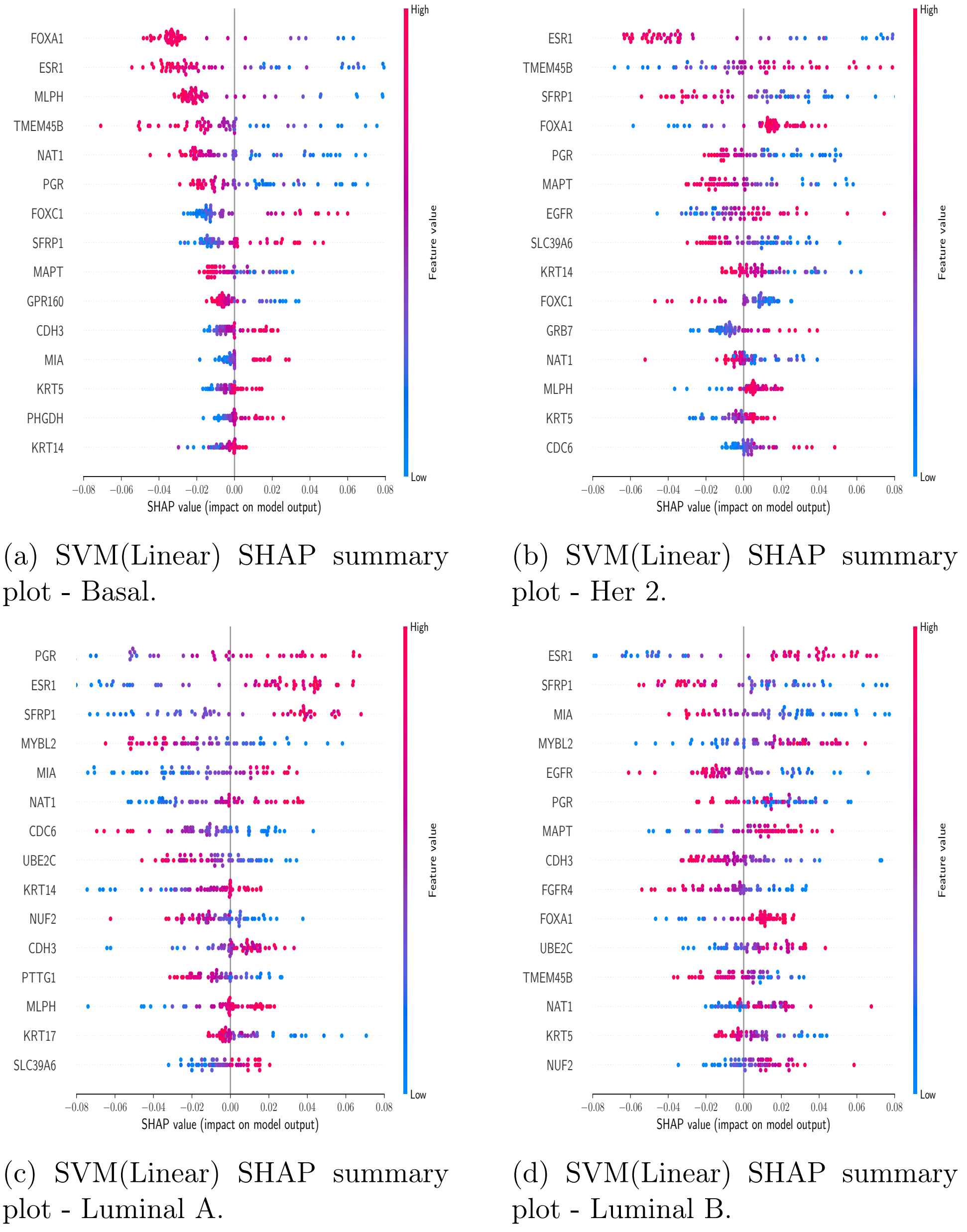
(b) *SVM(RBF)*

(c) *KNN*

(d) *Random Forest*

(e) *XGBoost*

Figure 5.3: Feature SHAP values for each method.



(a) SVM(Linear) SHAP summary plot - Basal.

(b) SVM(Linear) SHAP summary plot - Her 2.

(c) SVM(Linear) SHAP summary plot - Luminal A.

(d) SVM(Linear) SHAP summary plot - Luminal B.

Figure 5.4: SVM(Linear) SHAP summary plot for each subtype.

6

An outlier-based XAI approach to identify biologically significant genes for breast cancer subtypes

In the field of gene expression analysis, the ability to accurately predict and understand how genes behave in different biological contexts is critical for advancing our understanding of biological systems and developing effective treatments for diseases such as cancer.

Machine learning techniques have shown remarkable success in analyzing gene expression data and identifying patterns and correlations that may not be apparent to human experts. However, the complexity of these models often makes it difficult to understand how they arrive at their predictions, raising concerns about their interpretability and potential biases.

Explainable AI (XAI) methods aim to address these issues by providing insights into the decision-making processes of machine learning models. In this context, XAI can play a crucial role in enhancing the transparency and interpretability of gene expression analysis models, enabling researchers and clinicians to make more informed decisions and facilitating the development of more effective treatments. Therefore, the development and application of XAI in gene expression analysis have become a growing research focus in recent years.

In this context, we expand the **O**utlier-based **G**ene **S**election (OGS) method to investigate how the associations learned by the classifier are interpreted locally by SHAP Values, revealing genes and patterns that are important for all the subtypes.

6.1 Outlier based gene selection method with XAI

The general objective of this research is to explore, through a Machine Learning (ML) explainability technique, the importance of outlier genes for the classification of breast cancer subtypes, and how these genes are associated with the breast cancer subtype severity. For this, our method is divided into 6 steps (Figure 6.1): input data, preprocessing, dataset split, outlier detection, gene filtering, and explainable models.

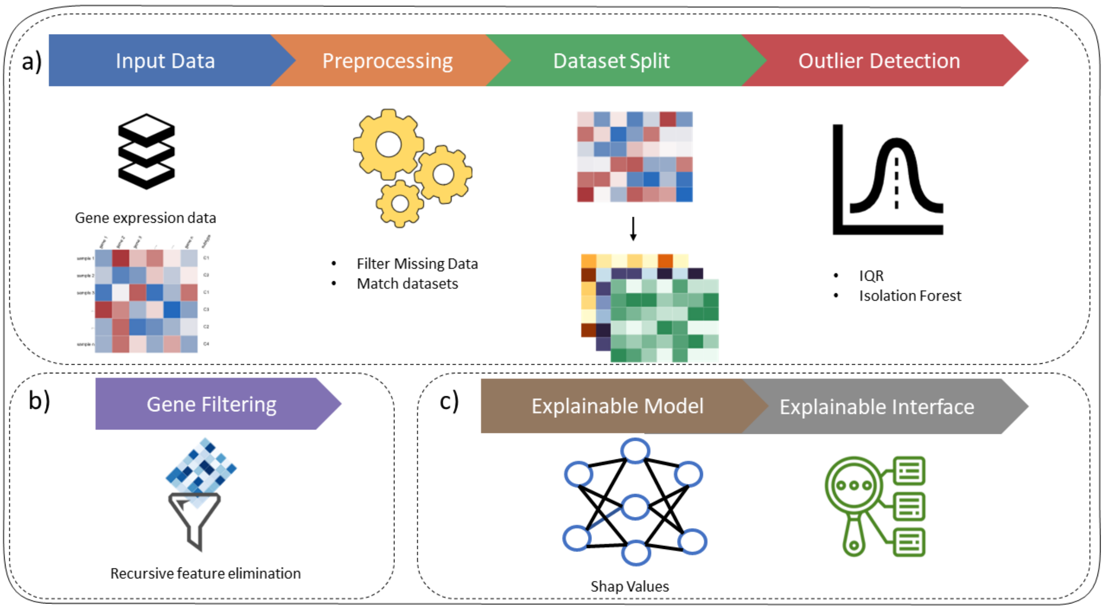


Figure 6.1: In the proposed method: (a) the method receives gene expression data, performs the filtering of genes and split the datasets into subtypes; subsequently, outlier genes are detected using IQR and validated through Isolation Forest; (b) after that, gene filtering step is performed to select the most important outlier genes; (c) finally, the classification is done and each result is interpreted by Shap Values, thus generating explanations for each gene.

The initial step of the outlier-based method is Preprocessing, in which samples with more than 60% missing data are discarded, since genes with a high number of missing data may confuse the classifier.

We also match the genes from the training dataset and the testing dataset, since the number of features must be the same. After concluding the preprocessing step, we do the subtype split, separating the dataset into smaller datasets, containing only the respective subtype samples. It is worth mentioning that all the smaller datasets have the same genes.

6.1.1 Outlier Detection

We compute the interquartile range (IQR) for outlier detection. IQR was employed because it is a reasonably robust measure of variability. It does need symmetric distributions of the data [Gelade et al., 2015]. Besides, it is not affected by outliers since it uses the middle 50% of the distribution for calculation and is computationally cheap.

For each of the smaller subsets (subtype dataset), we computed the outlier genes using the IQR default multiplier of 1.5. After finding the outlier genes using IQR we validated them by using isolation forest (iForest) [Liu et al., 2008].

6.1.2 Gene Filtering

The gene filtering step involves two stages. In the first stage, outlier genes are selected for each of the subtypes. The second stage narrows the selected outlier genes using recursive feature elimination. This gene selection approach aims filtering the irrelevant genes from the outlier stage, thereby improving the classification accuracy. When a gene exhibits a difference in expression between subtypes, it is a premise that this gene may be useful to classify the analyzed subtype

The last stage of gene filtering is the removal of less important outlier genes from the subset found. After finding an outlier subset using opposite outlier or outlier vs. normal approaches, we implement the Recursive feature elimination (RFE) [Guyon et al., 2002] to narrow the number of genes selected.

We employed Recursive Feature Elimination because it effectively selects features in the training dataset that are more relevant in predicting the target variable; therefore being widely used to discover new biomarkers and improve the interpretation of biological data [Christin et al., 2013]. The main goal of RFE is to rank the features by iteratively removing the features with the lowest weights. This method is designed to retain features that are most relevant to the classification task.

In this last stage, we divided the proposed method since one of our objectives is to compare the efficiency of employing the recursive feature elimination method in outlier genes.

6.1.3 Explainable Model

After obtaining the gene subset of each of the subtypes, we classify the samples into subtypes. In this stage, we examine the genes that are more crucial for determining the subtypes of breast cancer using the SHAP values [Lundberg et al., 2018]. After that, a biological analysis of the genes found to be important for each subtype is done.

6.2 Evaluation Methodology

In this section, we describe the evaluation methodology used in this work. We detail the characteristics of the datasets used in the experiments. We describe the eXplainability method used and also present the evaluation metrics employed.

6.2.1 Dataset

To validate the models, the experiments were carried out using distinct gene expression datasets from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) [Edwards et al., 2015]. The Cptac 2C dataset is used to train the models, as it has a more significant number of samples, while the Cptac 2D is used for testing. The datasets present the gene expression of patients with breast cancer. These samples are divided into four intrinsic breast cancer subtypes (see Chapter 1).

Table 6.1 summarizes the characteristics of the databases used in the experiments:

Table 6.1: Dataset description.

Dataset	**# of genes**	**Subtypes**	**# of samples**	**Total # of samples**
Cptac 2C	23,122	Basal	29	117
		Her 2	14	
		Luminal A	57	
		Luminal B	17	
Cptac 2D	16,525	Basal	18	77
		Her 2	12	
		Luminal A	23	
		Luminal B	24	

6.2.2 Model Explainability

An advanced machine learning (ML) algorithm can produce accurate predictions, but its famous “black box” nature does not help adoption. In bioinformatics, it is crucial to have a human understanding of the decisions of a machine learning result. In order to understand feature contributions, several works explore methods to explain and visualize the importance of input variables on ML predictions [Gunning et al., 2019, Souza et al., 2022].

Our work focuses on classical ML models using SHAP Values method for explainability [Lundberg and Lee, 2017]. There are other explanation techniques associated with Deep Learning (DL) models [Gunning et al., 2019], but as they are DL-based models, it is not feasible to use in our work since it needs a large amount of data for training and explaining. Thus, we decided to use classical machine learning and SHAP values method.

In this work, we employed SHAP (SHapley Additive exPlanations) (see Chapter 2). It uses a game theory approach to explain any machine learning model [Kwon and Lee, 2023], where the outcome of each feature depends on the actions of all other features [Fudenberg and Tirole, 1991]. The Shapley values method is essential in our results. It quantifies feature contributions and helps us understand the importance of outlier genes.

6.2.3 Evaluation Metrics

To measure the performance of our proposed method, we apply traditional metrics such as *precision*, *recall* and F_1 score. Since biological data often have a sparse dataset [Chicco, 2017], we also measured the performance of our method using Area Under Precision-Recall Curve $AUPRC$. The precision-recall curve is a more reliable and informative indicator of the statistical performance on multiclass problems [Chicco, 2017]. All The evaluation metrics are described and detailed in Chapter 2

6.3 Experiments and Results

This section describes the results obtained when applying the outlier method on a gene expression dataset with and without RFE. We compared the two distinct approaches. To measure the performance of the outlier-based method, facing the unbalanced datasets present in Table 6.1, we employed the AUPRC metric (Area Under the Precision-Recall Curve) [Davis and Goadrich, 2006] was chosen to evaluate the model, in addition to the metrics: *precision*, *recall* and F_1-score. Additionally, we present an explainable AI analysis of our experiments.

6.3.1 Breast Cancer subtype Classification

In the classification task, we implemented two distinct outlier based methods:

- **PH-OGS** - Described in Methods section, where we apply the RFE approach for feature elimination.

- **NR-OGS** - Described in Methods section, in which we do not employ any feature elimination method.

The PH-OGS presented by Mendonca-Neto et al. [2021], uses the outliers genes to classify the Basal and Her 2 subtypes. To classify the Luminals subtype, the authors used the PAM50, a well known 50-gene subtype predictor [Parker et al., 2009], since this set of genes outperformed the outliers genes. Since it is a binary classifier, we present the analysis of the genes for Luminal A and Luminal B as one, here we will call them Luminals.

Table 6.2 summarizes the resulted number of genes after the outlier-based gene selection method. After employing the RFE technique, we reduced the number of genes for the Basal and Her 2 subtypes in 85%. For the Luminals subtype, we were able to reduce only 4% (2 genes), since it is based on the PAM50 gene list, an already reduced number of genes.

Table 6.2: Number of genes used in each Method.

Summary of the genes used.			
	Number of Genes		
Subtypes	PH-OGS	NR-OGS	Difference
Basal	18	240	92.50%
Her 2	89	8	86.44%
Luminals	50	48	4.00%

In the first experiment, we compared the classification results of PH-OGS with NR-OGS. Table 6.3 presents the performance in terms of *precision* and *recall* and Table 6.4 illustrates performance in terms of F_1 and AUPRC. Comparing the performance obtained by the methods, we see that both methods have high results for all four subtypes, achieving an F_1 above 0.80 for all the subtypes. Besides that, PH-OGS outperformed the NR-OGS in terms of F_1 and AUPRC for the Basal and Her 2 subtypes. This performance can be explained by the fact that it is the method that best managed to separate the samples from Basal to Non-Basal.

Figure 6.2 shows the confusion matrices obtained by each method. Each row represents the instances of an actual class and each column represents the instances of a predicted class.

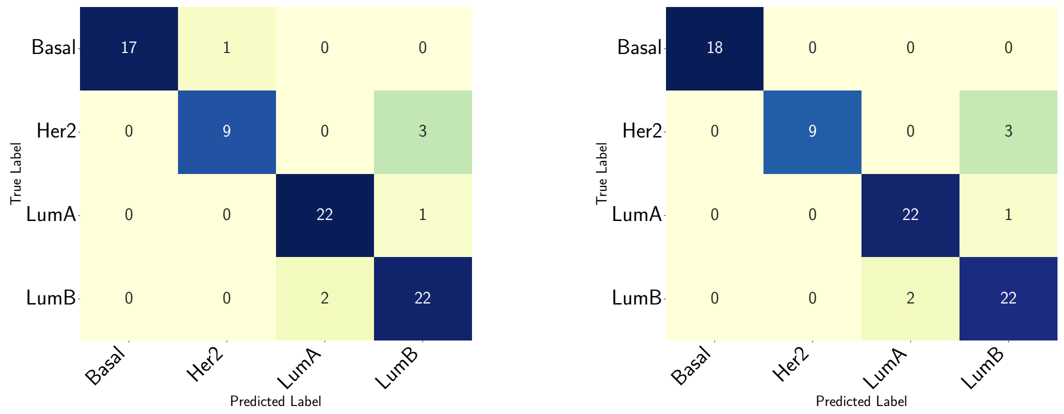
We can observe that the results obtained with the PH-OGS (Figure 6.2) presented better results than the NR-OGS method, where only two samples were wrongly

Table 6.3: Classification results using *precision* and *recall* metrics. Best results in bold.

Methods Performance								
	Precision				Recall			
Method	Basal	Her2	LumA	LumB	Basal	Her2	LumA	LumB
PH-OGS	**1.00**	**1.00**	**0.92**	**0.85**	**1.00**	**0.75**	**0.96**	**0.92**
NR-OGS	**1.00**	0.90	**0.92**	**0.85**	0.94	0.75	**0.96**	**0.92**

Table 6.4: Classification results using F_1 and AUPRC metrics. Best results in bold.

Methods Performance								
	F1				AUPRC			
Method	Basal	Her2	LumA	LumB	Basal	Her2	LumA	LumB
PH-OGS	**1.00**	**0.86**	**0.94**	**0.88**	**1.00**	**0.79**	**0.89**	**0.80**
NR-OGS	0.97	0.82	**0.94**	**0.88**	0.96	0.71	**0.89**	**0.80**



(a) Without RFE.

(b) With RFE.

Figure 6.2: Confusion matrices obtained by each method.

classified as Her 2, two samples as Luminal A, and three as Luminal B. The PH-OGS method provides the best performance, with the least amount of wrongly classified samples, six in total. We can notice that the subtype with the worst prognosis, Basal, is the one that had the least characterization problems, regardless of the method. In contrast, the Her 2 subtype is the subtype where the methods have greater difficulty classifying the samples. Regarding the Basal and Her 2, we can see that RFE plays an essential role in eliminating non-important genes, improving the classification and reducing the number of genes. For the Luminal A and Luminal B subtypes, similar results are expected for both methods since it is used almost the same amount of genes.

6.3.2 Global Explanations Using Shap Values

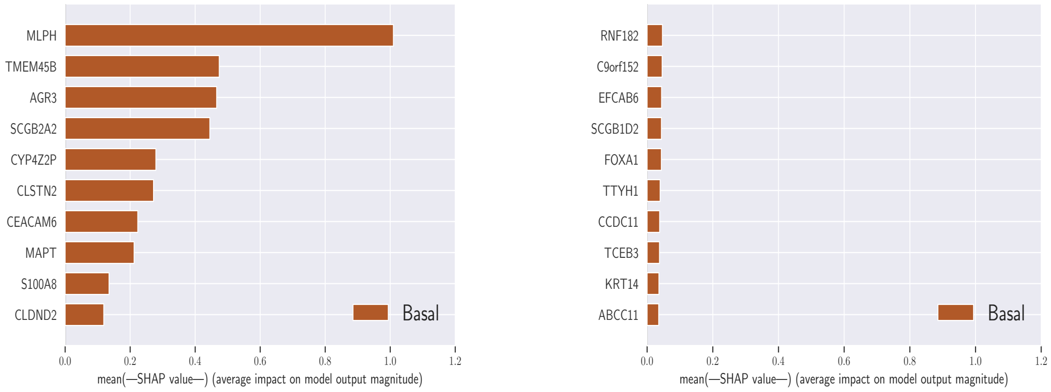
In the explainability step, we analyze which features (genes) are more important for each of the methods to classify the breast cancer subtypes using SHAP values [Lundberg et al., 2018]. Although to compute the SHAP values, we face an exponential computational complexity [Messalas et al., 2019], in the scope of our project, we were able to apply it to all samples since our test dataset has only 77 samples. Figures 6.3, 6.4 and 6.5 shows features SHAP values obtained by each method. The larger the bar, the more critical the gene is for the classification of the subtype. We can see that the distinct methods present distinct gene importance.

Figures 6.3 and 6.4 show that the number of genes used for classification impacts directly in the feature importance. The PH-OGS results (Figures 6.3a and Figure 6.4a) has a perceptive discrepancy across genes importance (Figure 6.3a). While for PHG-OGS, it is very clear which genes are the most important, for NR-OGS, all genes appear to have close importance.

Comparing the features importance (Figure 6.3) when classifying Basal versus Non-Basal samples. For PH-OGS, the top four relevant genes are MLPH, TMEM45B, AGR3, and SCGB2A2. The MLPH is more important to distinguish the two classes in PH-OGS (6.3a), and it is known to have an elevated expression levels in ERα(+) breast cancer patient [Thakkar et al., 2015]. The TMEM45B gene is in the PAM50 gene list and is found overexpressed in breast cancer Her 2 subtype [Parker et al., 2009, Zhao et al., 2016]. AGR3 presents significantly higher expression in luminal breast cancer subtypes [Garczyk et al., 2015]. SCGB2A2 gene is significantly downregulated in Her 2-enriched/Triple Negative [Prat et al., 2013], while the genes importance in NR-OGS is similar (Figure 6.3b) for all the top 10 genes.

To separate Her 2 Samples from Non-Her 2 samples, the more important genes for PH-OGS (Figure 6.4a) are PGR and GRPR. The expression of PGR, a gene included in the PAM50 gene list, is identified as a significant prognostic marker in ER-positive/Her 2-negative breast cancer patients [Kurozumi et al., 2017]. While the GRPR is found to be overexpressed in estrogen receptor (ER)(+) subtypes [Morgat et al., 2017]. For the NR-OG (Figure 6.4b), the GFRA1 is the most important gene for the classification, in which is overexpressed in the majority of breast cancers [Bosco et al., 2018].

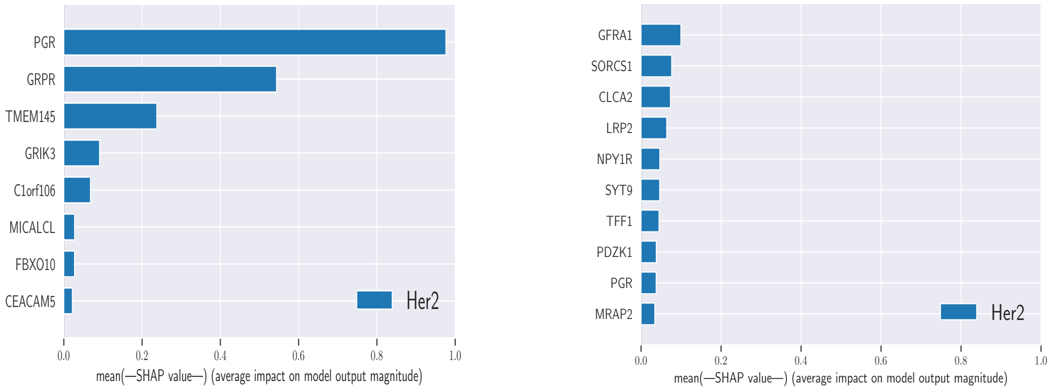
Figure 6.5 illustrates the importance of the genes in separating the Luminal A



(a) PH-OGS.

(b) NR-OGS.

Figure 6.3: Genes SHAP values obtained by each method - Basal Subtype.



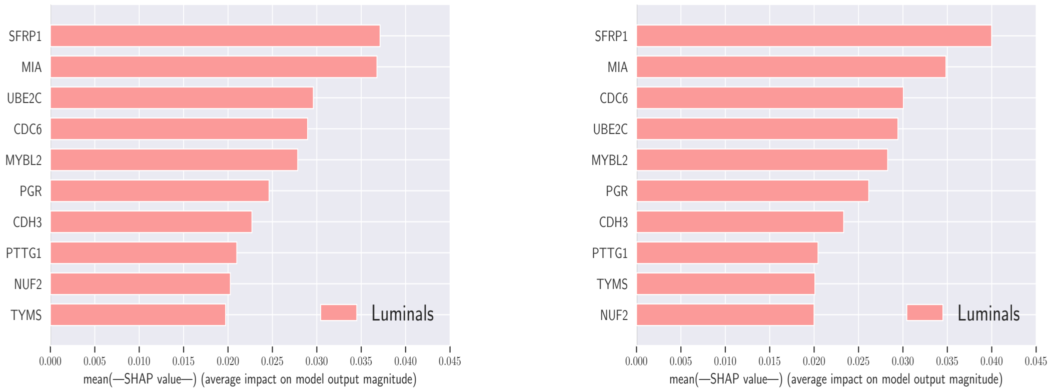
(a) PH-OGS.

(b) NR-OGS.

Figure 6.4: Genes SHAP values obtained by each method - Her 2 Subtype.

from Luminal B subtype. Since we employed the PAM50 for this classification, the results are almost similar for both methods. SFRP1 is the most important gene for both methods. This gene is under-expressed in breast cancer tissue when compared to normal tissue [Clemenceau et al., 2020]. It was found that SFRP1 is highly expressed in Basal subtype and Triple-Negative breast cancer [Baharudin et al., 2020]. The second most important gene for both methods is MIA; It is highly expressed in breast cancer. Its overexpression leads to increased metastasis of malignant melanoma cells by enhancing invasion and extravasation [El Fitori et al., 2005].

Given a feature, we also extract the importance proportion for each class Figure 6.6. By analyzing the contribution of each class for each feature using the PAM50 gene list. ESR1 gene has a greater contribution to three of the four subtypes, which is known to be frequently mutated in endocrine-resistant ER-positive (ER+) breast cancer and linked to ligand-independent growth and metastasis [Li et al., 2022]. While



(a) PH-OGS.

(b) NR-OGS.

Figure 6.5: Genes SHAP values obtained by each method - Luminals Subtype.

the most important gene for Luminal A is PGR. We tried to improve our classification results by looking to the most important for each subtype and adding to our outlier gene list. The results showed that this do not improve our classification performance.

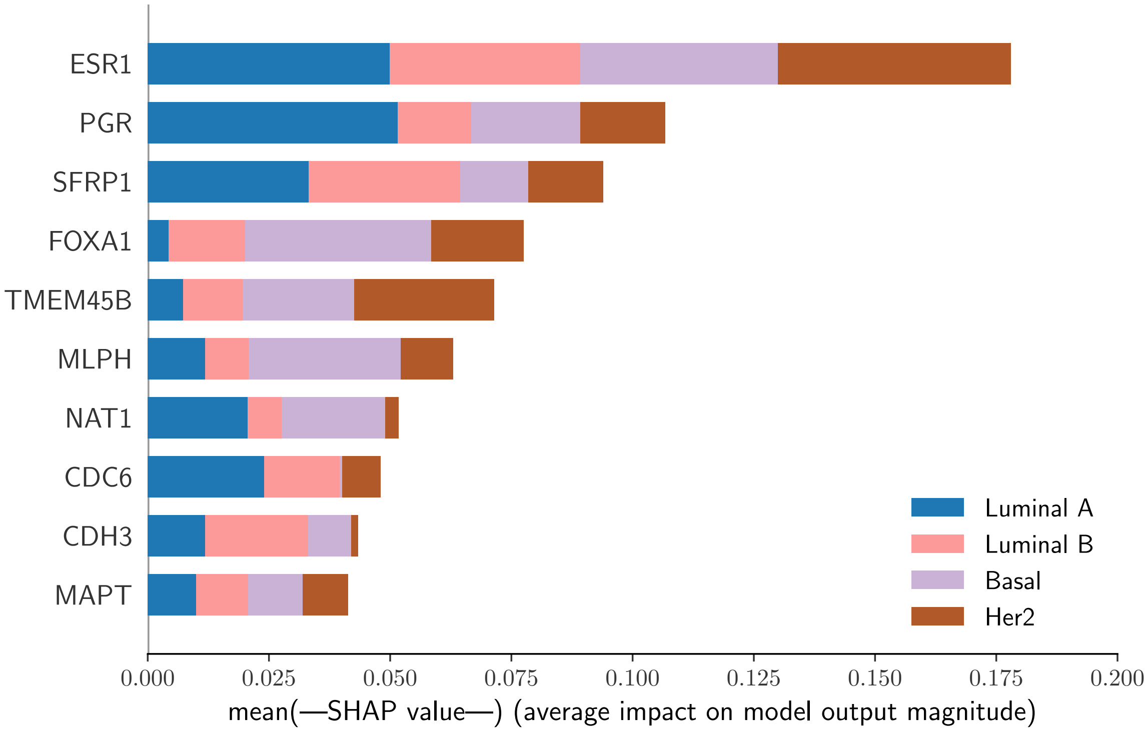


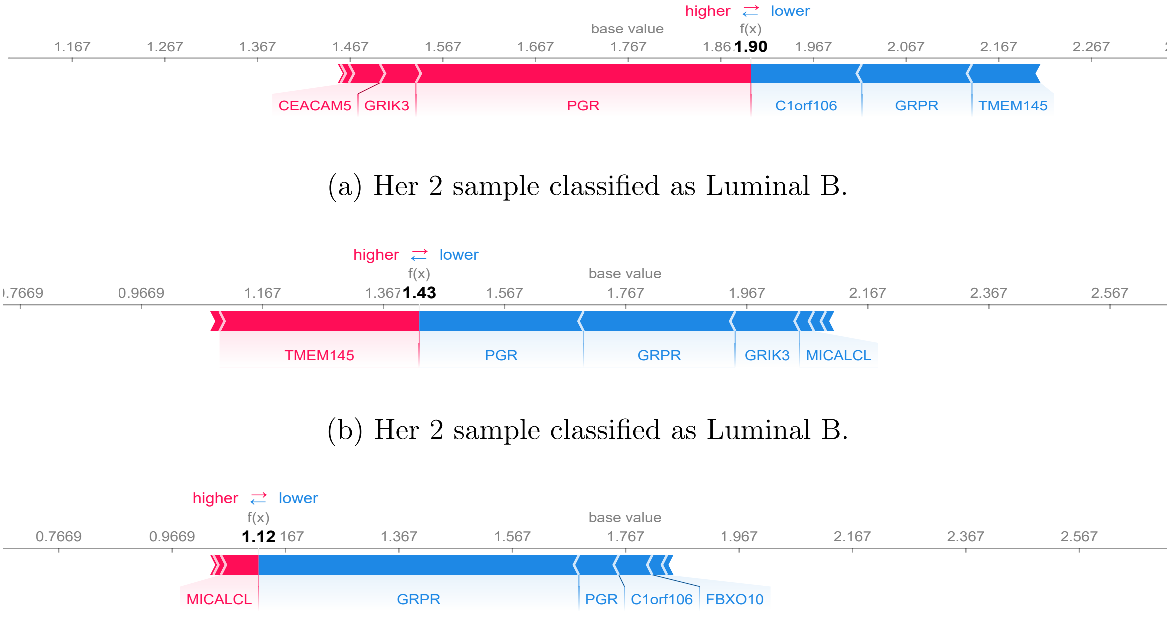
Figure 6.6: Genes importance proportion for each subtype.

6.3.3 Explaining Individual Predictions

In this section, we give a closer look at individual predictions to understand how validation influences at instance level. For this purpose, we present results based on shap force plot [Lundberg et al., 2018, Molnar, 2020]. The force plot shows shap values contributions in generating the final prediction using an additive force layout. We can visualize feature attributions as "forces".

The prediction starts from the baseline. The baseline for Shapley values is the average of all predictions [Bragança et al., 2022]. In the plot, each Shapley value is an arrow that pushes to increase with positive value (red) or decrease with negative value (blue) the prediction. These forces balance each other out at the actual prediction of the data instance.

We used the shap force plot to all the six miss classified samples of PH-OGS (Figure 6.2a). As can be seen in Figure 6.7, all the three miss classified samples have blue arrows forcing towards the negative class. While Figure 7a indicates a balanced force, making the classifier confuses both classes, Figures 7b and 7c shows that both samples are leaning towards the wrong class.



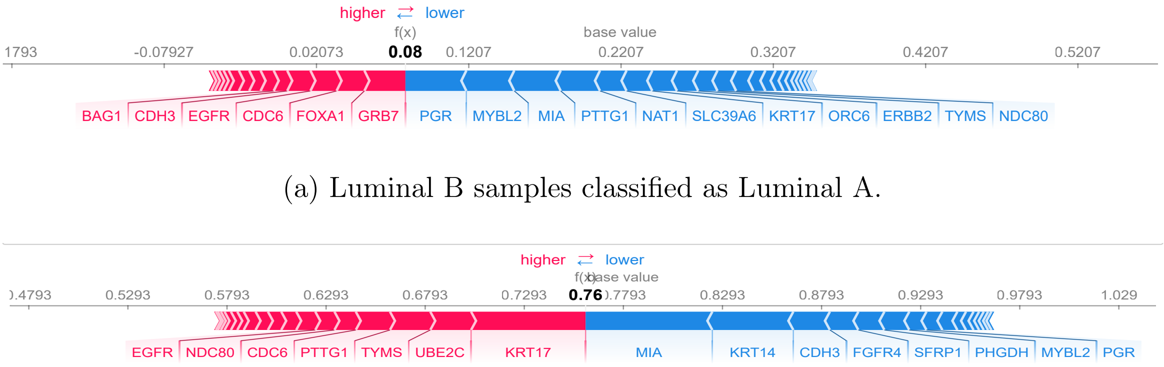
(a) Her 2 sample classified as Luminal B.

(b) Her 2 sample classified as Luminal B.

(b) Her 2 sample classified as Luminal B.

Figure 6.7: Her 2 individual analysis

Figure 6.8 illustrates the miss-classified Luminal B samples. Interestingly, both results are that none of the samples were completely confused with the other subtypes. Figure 8b indicates a totally balanced force of features. This is explained by the fact that both Luminal subtypes have similar characteristics [Hashmi et al., 2018]. Another fact is that most of the genes are forcing towards the negative class. Despite that, some genes are in opposite directions, for example, the CDH3, which is expressed in estrogen receptor (ER)(-), progesterone receptor (PgR), and Her 2-enriched/Triple Negative [Paredes et al., 2007]. Some genes impose strong force towards negative classes for one classification, and in the other not, for example, the PGR.



(a) Luminal B samples classified as Luminal A.

(b) Luminal B samples classified as Luminal A.

Figure 6.8: Luminal B individual analysis.

In Figure 6.8 it is illustrated the shap force plot of Luminal A sample classified as Luminal B. We can see that most of the genes are a forcing towards the negative class. It is the same problem faced when Luminal B samples are classified as Luminal A. Since both subtypes have similar characteristics, to performance the separation of both subtypes is a hard task.

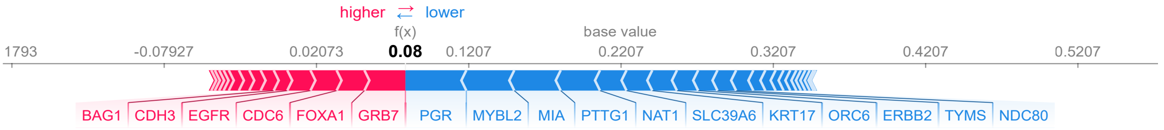


Figure 6.9: Luminal A sample classified as Luminal B.

We can observe in these results that, when analyzing the predictions individually for a given class, the classifier can change the order of importance between the features.

6.4 Chapter Remarks

In this Chapter, we described an outlier method capable of identify relevant gene for breast cancer subtype classification, present gene rank importance and how the more important genes are related to the breast cancer subtypes. In our analysis, employing eXplainable AI methods is crucial to understand the biological meaning of the genes used in the classification, since we can understand the most characteristic genes.

While the results presented are promising, in many situations, it is not trivial to find meanings in the genes importance. However, our results show that the outlier-based gene select method can significantly influence the selection of features and improve the classification performance.

Our work has presented manners that can be explored by using explainable algorithms to improve the transparency of creating machine learning models. The SHAP results reinforce that some samples, even from the subtype, are differently interpreted by machine learning methods.

7

Final Remarks

This chapter concludes and summarizes this doctoral thesis, in addition to pointing to future directions for current research. In Section 7.1, we present our final considerations on the solutions presented, in Section 7.2 we present the future development of this proposal, and in Section 7.3, we present the publications derived from this research.

7.1 Conclusion

In this thesis, we addressed the cancer subtype classification issue using gene expression data. This is one of the challenges when using gene expression for cancer classification since those data are represented by complex and high-dimensional matrices of genes versus samples. Therefore, selecting a few highly relevant genes becomes a major effort.

In this context, the contributions of this research include four main results: (i) the classification of breast cancer subtypes based on outlier genes; (ii) the identification and analysis of signatures based on outliers that characterize the basal subtype; (iii) a framework that uses different machine learning techniques to classify breast cancer subtypes and investigate the features; (iv) and how the associations learned by the classifier are interpreted locally by an eXplainable AI method, revealing the biological meaning of the genes and linking them to each subtype.

Finally, we proved through experiments that the solutions present in this thesis are capable of selecting a few highly relevant genes for classifying the breast cancer samples into their intrinsic subtypes using outlier genes, also identifying signatures that characterize the cancer subtypes with the worst prognosis. We also demonstrate a study of different methods in the task of classifying breast cancer subtypes through the PAM50 gene list. In addition, we demonstrated through eXplainable AI that the number of genes used for classification impacts directly the feature importance and that all the outlier genes found by our method are biologically connected to one or more subtypes of breast cancer.

7.2 Future Directions

Based on the partial results obtained, we intend to continue our experiments, seeking to improve the gene selection method through outliers. Regarding the methods used in our works, the following activities are still open:

In future work, we intend to identify and perform a functional analysis of the distinct gene signatures for the other breast cancer subtypes. Furthermore, we plan to expand our approach to subtypes of other carcinomas.

1. Identify and perform a functional analysis of the distinct gene signatures generated from outlier genes for the other breast cancer subtypes.
2. Expand our outlier-based gene selection method for subtypes of other heterogeneous carcinomas.
3. Expand our eXplainable AI approach to identify if the biologically relevant outlier genes are related to each other in pathways.
4. Extend the analysis of classifier to a multilevel classification, in which we will employ a hierarchical classifier to perform the classification of breast cancer subtypes.

7.3 Publications

Here we present the publications that were produced during the development of this thesis.

7.3.1 Main Publications

This work resulted in the publication of the following manuscripts:

- *Using Outliers to Find Genes Relevant for Basal Breast Cancer* as part of BIOKDD 2020: 19th International Workshop on Data Mining in Bioinformatics in San Diego, California. This abstract describes a strategy to identify relevant genes for the basal subtype based on outliers.

- *A Gene Selection Method Based On Outliers for Breast Cancer Subtype Classification* as part of the BIOKDD Special Issue in ACM/IEEE Transactions on Computational Biology and Bioinformatics. This paper describes an outlier-based gene selection method for breast cancer subtype classification.

- *Classification of breast cancer subtypes: A study based on representative genes* as part of the Journal of the Brazilian Computer Society (JBCS). This paper describes an approach that uses different machine learning techniques for a broader analysis of the PAM50 list in the classification of breast cancer subtypes.

7.3.2 Colaboration Publications

Our classification method have been successfully applied to other articles:

- *Classificação de subtipos de câncer de mama: Um estudo baseado em genes representativos* as part of the XLVIII Seminário Integrado de Software e Hardware - SEMISH. This paper describes an approach that uses different machine learning techniques for a broader analysis of the PAM50 list in the classification of breast cancer subtypes.

- *Um método de Estimação de Expressões Gênicas de Câncer de Mama com Base em Correlação* as part of the L Seminário Integrado de Software e Hardware - SEMISH. This paper describes a method to treat missing values in breast cancer gene expressions.

- *Redução dimensional de dados de expressão gênica para classificação de subtipos de câncer de mama* as part of the XVII Brazilian e-Science Workshop - BreSci. This paper describes the use of Siamese Networks as a new approach to improve the classification in breast cancer subtypes.

7.3.3 Submitted Papers

This work also resulted in the submission of the paper that is under review:

- *An outlier-based XAI approach to identify biologically significant genes for breast cancer subtypes* as part of the ACM/IEEE Transactions on Computational Biology and Bioinformatics. This article explores the classification of subtypes using outlier genes and the identification of biologically relevant genes for classification using eXplainable AI methods.

Bibliography

Aggarwal, C. C. (2015). Outlier analysis. In *Data mining*, pages 237--263. Springer.

Alanni, R., Hou, J., Azzawi, H., and Xiang, Y. (2019). Deep gene selection method to select genes from microarray datasets for cancer classification. *BMC bioinformatics*, 20(608):1--15.

Almugren, N. and Alshamlan, H. (2019). A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. *IEEE access*, 7:78533--78548.

Alves, R., Rodriguez-Baena, D. S., and Aguilar-Ruiz, J. S. (2010). Gene association analysis: a survey of frequent pattern mining from gene expression data. *Briefings in Bioinformatics*, 11(2):210--224.

Anguita-Ruiz, A., Segura-Delgado, A., Alcalá, R., Aguilera, C. M., and Alcalá-Fdez, J. (2020). explainable artificial intelligence (xai) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research. *PLoS computational biology*, 16(4):e1007792.

Antwarg, L., Miller, R. M., Shapira, B., and Rokach, L. (2021). Explaining anomalies detected by autoencoders using shapley additive explanations. *Expert Systems with Applications*, 186:115736.

Ayyad, S. M., Saleh, A. I., and Labib, L. M. (2019). Gene expression cancer classification using modified k-nearest neighbors technique. *Biosystems*, 176:41--51.

Badve, S., Turbin, D., Thorat, M. A., Morimiya, A., Nielsen, T. O., Perou, C. M., Dunn, S., Huntsman, D. G., and Nakshatri, H. (2007). Foxa1 expression in breast cancer—correlation with luminal subtype a and survival. *Clinical cancer research*, 13(15):4415--4421.

Baharudin, R., Tieng, F. Y. F., Lee, L.-H., and Ab Mutalib, N. S. (2020). Epigenetics of sfrp1: the dual roles in human cancers. *Cancers*, 12(2):445.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412--424.

Berger, B., Peng, J., and Singh, M. (2013). Computational solutions for omics data. *Nature reviews genetics*, 14(5):333.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1):281--305.

Bertucci, F., Finetti, P., and Birnbaum, D. (2012). Basal breast cancer: a complex and deadly molecular subtype. *Current molecular medicine*, 12(1):96--110.

Bhola, A. and Tiwari, A. K. (2015). Machine learning based approaches for cancer classification using gene expression data. *Machine Learning and Applications: An International Journal*, 2(3):12.

Bi, Y., Xiang, D., Ge, Z., Li, F., Jia, C., and Song, J. (2020). An interpretable prediction model for identifying n7-methylguanosine sites based on xgboost and shap. *Molecular Therapy-Nucleic Acids*, 22:362--372.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Bjørklund, S. S., Kristensen, V. N., Seiler, M., Kumar, S., Alnæs, G. I. G., Ming, Y., Kerrigan, J., Naume, B., Sachidanandam, R., Bhanot, G., et al. (2015). Expression of an estrogen-regulated variant transcript of the peroxisomal branched chain fatty acid oxidase acox2 in breast carcinomas. *BMC cancer*, 15(1):1--13.

Blumenberg, L., Kawaler, E., Cornwell, M., Smith, S., Ruggles, K., and Fenyo, D. (2019). Blacksheep: A bioconductor and bioconda package for differential extreme value analysis. *BioRxiv*, page 9.

Bosco, E. E., Christie, R. J., Carrasco, R., Sabol, D., Zha, J., DaCosta, K., Brown, L., Kennedy, M., Meekin, J., Phipps, S., et al. (2018). Preclinical evaluation of a gfra1 targeted antibody-drug conjugate in breast cancer. *Oncotarget*, 9(33):22960.

Bragança, H., Colonna, J. G., Oliveira, H. A., and Souto, E. (2022). How validation methodology influences human activity recognition mobile systems. *Sensors*, 22(6):2360.

Bray, F., Ferlay, J., Soerjomataram, I., L. Siegel, R., Torre, L., and Jemal, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries: Global cancer statistics 2018. *CA: A Cancer Journal for Clinicians*, 68:394--424.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5--32.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121--167.

Castelvecchi, D. (2016). Can we open the black box of ai? *Nature News*, 538(7623):20.

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.

Chen, X., Hu, H., He, L., Yu, X., Liu, X., Zhong, R., and Shu, M. (2016). A novel subtype classification and risk of breast cancer by histone modification profiling. *Breast cancer research and treatment*, 157(2):267--279.

Chen, Z., Li, J., and Wei, L. (2007). A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue. *Artificial Intelligence in Medicine*, 41(2):161--175.

Chia, S. K., Bramwell, V. H., Tu, D., et al. (2012). A 50-gene intrinsic subtype classifier for prognosis and prediction of benefit from adjuvant tamoxifen. *Clinical cancer research*, 18(16):4465--4472.

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData mining*, 10(1):1--17.

Christin, C., Hoefsloot, H. C., Smilde, A. K., Hoekman, B., Suits, F., Bischoff, R., and Horvatovich, P. (2013). A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. *Molecular & Cellular Proteomics*, 12(1):263--276.

Cine, N., Baykal, A. T., Sunnetci, D., Canturk, Z., Serhatli, M., and Savli, H. (2014). Identification of apoa1, hpx and potee genes by omic analysis in breast cancer. *Oncology reports*, 32(3):1078--1086.

Clemenceau, A., Diorio, C., and Durocher, F. (2020). Role of secreted frizzled-related protein 1 in early mammary gland tumorigenesis and its regulation in breast microenvironment. *Cells*, 9(1):208.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273--297.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561--563.

Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., and Shi, B. (2015). Breast cancer intrinsic subtype classification, clinical use and future trends. *American journal of cancer research*, 5(10):2929--2943.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 233–240, New York, NY, USA. Association for Computing Machinery.

Derksen, S. and Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 45(2):265--282.

Díaz-Uriarte, R. and De Andres, S. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(3):13.

Driouch, K., Bonin, F., Sin, S., Clairac, G., and Lidereau, R. (2009). Confounding effects in "a six-gene signature predicting breast cancer lung metastasis": reply. *Cancer Research*, 69(24):9507--9511.

Dwivedi, S., Purohit, P., Misra, R., Lingeswaran, M., Vishnoi, J. R., Pareek, P., Sharma, P., and Misra", S. (2019). Application of single-cell omics in breast cancer. In *Single-Cell Omics*, volume 2, pages 69--103. Academic Press.

Edwards, N. J., Oberti, M., Thangudu, R. R., Cai, S., McGarvey, P. B., Jacob, S., Madhavan, S., and Ketchum, K. A. (2015). The cptac data portal: a resource for cancer proteomics research. *Journal of proteome research*, 14(6):2707--2713.

El Fitori, J., Kleeff, J., Giese, N. A., Guweidhi, A., Bosserhoff, A. K., Büchler, M. W., and Friess, H. (2005). Melanoma inhibitory activity (mia) increases the invasiveness of pancreatic cancer cells. *Cancer cell international*, 5(1):1--8.

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.

Fudenberg, D. and Tirole, J. (1991). *Game theory*. MIT press.

Fürnkranz, J. (2001). Round robin rule learning. In *Proceedings of the 18th International Conference on Machine Learning (ICML-01): 146–153*. Citeseer.

Garczyk, S., von Stillfried, S., Antonopoulos, W., Hartmann, A., Schrauder, M. G., Fasching, P. A., Anzeneder, T., Tannapfel, A., Ergönenc, Y., Knüchel, R., et al. (2015). Agr3 in breast cancer: prognostic impact and suitable serum-based biomarker for early cancer detection. *PloS one*, 10(4):e0122106.

Garro, B. A., Rodríguez, K., and Vázquez, R. A. (2016). Classification of dna microarrays using artificial neural networks and abc algorithm. *Applied Soft Computing*, 38:548--560.

Gatto, B. B., Santos, E. M. d., Koerich, A. L., Fukui, K., and Junior, W. S. (2021). Tensor analysis with n-mode generalized difference subspace. *Expert Systems with Applications*, 171:1--11.

Gelade, W., Verardi, V., and Vermandele, C. (2015). Time-efficient algorithms for robust estimators of location, scale, symmetry, and tail heaviness. *The Stata Journal*, 15(1):77--94.

Ghorai, S., Mukherjee, A., Sengupta, S., and Dutta, P. K. (2010). Cancer classification from gene expression data by nppc ensemble. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(3):659--671.

Ginsburg, O., Yip, C.-H., Brooks, A., Cabanes, A., Caleffi, M., Dunstan Yataco, J. A., Gyawali, B., McCormack, V., McLaughlin de Anderson, M., Mehrotra, R., et al. (2020). Breast cancer early detection: A phased approach to implementation. *Cancer*, 126:2379--2393.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

Graudenzi, A., Cava, C., Bertoli, G., Fromm, B., Flatmark, K., Mauri, G., and Castiglioni, I. (2017). Pathway-based classification of breast cancer subtypes. *Front Biosci*, 22(10):1697--1712.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., and Yang, G.-Z. (2019). Xai—explainable artificial intelligence. *Science Robotics*, 4(37):eaay7120.

Guo, G., Wang, H., Bell, D., Bi, Y., and Greer, K. (2003). Knn model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS,* *DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*, pages 986--996. Springer.

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.

Hashmi, A. A., Aijaz, S., Khan, S. M., Mahboob, R., Irfan, M., Zafar, N. I., Nisar, M., Siddiqui, M., Edhi, M. M., Faridi, N., et al. (2018). Prognostic parameters of luminal a and luminal b intrinsic breast cancer subtypes of pakistani patients. *World journal of surgical oncology*, 16(1):1--6.

He, J., Yang, J., Chen, W., Wu, H., Yuan, Z., Wang, K., Li, G., Sun, J., and Yu, L. (2015). Molecular features of triple negative breast cancer: microarray evidence and further integrated analysis. *PloS one*, 10(6):e0129842.

Herrero, J., Díaz-Uriarte, R., and Dopazo, J. (2003). Gene expression data preprocessing. *Bioinformatics*, 19(5):655--656.

Hodge, V. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2):85--126.

Huang, H.-H., Liu, X.-Y., and Liang, Y. (2016). Feature selection and cancer classification via sparse logistic regression with the hybrid l1/2+ 2 regularization. *PloS one*, 11(5):1--15.

Jaksik, R., Iwanaszko, M., Rzeszowska-Wolny, J., and Kimmel, M. (2015). Microarray experiments and factors which affect their reliability. *Biology direct*, 10:1--14.

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F., May, B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Weiser, J., Wu, G., Stein, L., Hermjakob, H., and D'Eustachio, P. (2020). The reactome pathway knowledgebase. *Nucleic acids research*, 48(D1):D498–D503. Provided by Reactome. Citation Accessed on Wed Feb 10 2021.

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., and Forman, D. (2011). Global cancer statistics. *CA: a cancer journal for clinicians*, 61(2):69--90.

Jiang, D., Tang, C., and Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge & Data Engineering*, 11:1370--1386.

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118--127.

Kamal, M. S., Northcote, A., Chowdhury, L., Dey, N., Crespo, R. G., and Herrera-Viedma, E. (2021). Alzheimer's patient analysis using image and gene expression data and explainable-ai to present associated genes. *IEEE Transactions on Instrumentation and Measurement*, 70:1--7.

Karp, P. D., Billington, R., Caspi, R., Fulcher, C. A., Latendresse, M., Kothari, A., Keseler, I. M., Krummenacker, M., Midford, P. E., Ong, Q., Ong, W. K., Paley, S. M., and Subhraveti, P. (2017). The BioCyc collection of microbial genomes and metabolic pathways. *Briefings in Bioinformatics*, 20(4):1085–1093.

Kerr, G., Ruskin, H. J., Crane, M., and Doolan, P. (2008). Techniques for clustering gene expression data. *Computers in biology and medicine*, 38(3):283--293.

Kotsiantis, S. B., Zaharakis, I., and Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1):3--24.

Kurozumi, S., Matsumoto, H., Hayashi, Y., Tozuka, K., Inoue, K., Horiguchi, J., Takeyoshi, I., Oyama, T., and Kurosumi, M. (2017). Power of pgr expression as a prognostic factor for er-positive/her2-negative breast cancer patients at intermediate risk classified by the ki67 labeling index. *BMC cancer*, 17(1):1--9.

Kwon, S. and Lee, Y. (2023). Explainability-based mix-up approach for text data augmentation. *ACM transactions on knowledge discovery from data*, 17(1):1--14.

Lee, S., Lim, S., Lee, T., Sung, I., and Kim, S. (2020). Cancer subtype classification and modeling by pathway attention and propagation. *Bioinformatics*, 36(12):3818--3824.

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., and Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal of Big Data*, 5(1):1--30.

Li, Y., Kang, K., Krahn, J. M., Croutwater, N., et al. (2017). A comprehensive genomic pan-cancer classification using the cancer genome atlas gene expression data. *BMC genomics*, 18(1):508.

Li, Z., McGinn, O., Wu, Y., Bahreini, A., Priedigkeit, N. M., Ding, K., Onkar, S., Lampenfeld, C., Sartorius, C. A., Miller, L., et al. (2022). Esr1 mutant breast cancers show elevated basal cytokeratins and immune activation. *Nature Communications*, 13(1):2011.

Liao, Y. and Vemuri, V. R. (2002). Use of k-nearest neighbor classifier for intrusion detection. *Computers & security*, 21(5):439--448.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740.

Liu, C. and San Wong, H. (2017). Structured penalized logistic regression for gene selection in gene expression data analysis. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(1):312--321.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. *2008 Eighth IEEE International Conference on Data Mining*, pages 413--422.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Norton, H., et al. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature biotechnology*, 14(13):1675.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J., et al. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10):749--760.

Lyu, B. and Haque, A. (2018). Deep learning based tumor type classification using gene expression data. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pages 89--96.

Ma, S. and Dai, Y. (2011). Principal component analysis based methods in bioinformatics studies. *Briefings in bioinformatics*, 12(6):714--722.

Mathur, A. and Foody, G. M. (2008). Multiclass and binary svm classification: Implications for training and classification users. *IEEE Geoscience and remote sensing letters*, 5(2):241--245.

Meena, J. and Hasija, Y. (2022). Application of explainable artificial intelligence in the identification of squamous cell carcinoma biomarkers. *Computers in Biology and Medicine*, 146:105505.

Mendonca-Neto, R., Li, Z., Fenyö, D., Silva, C. T., Nakamura, F. G., and Nakamura, E. F. (2021). A gene selection method based on outliers for breast cancer subtype classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(5):2547--2559.

Mertins, P., Mani, D., Ruggles, K. V., Gillette, M. A., Clauser, K. R., Wang, P., Wang, X., Qiao, J. W., Cao, S., Petralia, F., et al. (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*, 534(7605):55--62.

Messalas, A., Kanellopoulos, Y., and Makris, C. (2019). Model-agnostic interpretability with shapley values. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1--7. IEEE.

Miah, S., Banks, C. A., Adams, M. K., Florens, L., Lukong, K. E., and Washburn, M. P. (2017). Advancement of mass spectrometry-based proteomics technologies to explore triple negative breast cancer. *Molecular BioSystems*, 13(1):42--55.

Mitchell, T. M. (1999). Machine learning and data mining. *Communications of the ACM*, 42(11):30--36.

Molnar, C. (2020). *Interpretable machine learning: A Guide for Making Black Box Models Explainable.* Lulu. com.

Morgat, C., MacGrogan, G., Brouste, V., Vélasco, V., Sevenet, N., Bonnefoi, H., Fernandez, P., Debled, M., and Hindie, E. (2017). Expression of gastrin-releasing peptide receptor in breast cancer and its association with pathologic, biologic, and clinical parameters: a study of 1,432 primary tumors. *Journal of Nuclear Medicine*, 58(9):1401--1407.

Mostavi, M., Chiu, Y.-C., Huang, Y., and Chen, Y. (2020). Convolutional neural network models for cancer type prediction based on gene expression. *BMC Medical Genomics*, 13(44):1--13.

Mramor, M., Leban, G., Demšar, J., and Zupan, B. (2005). Conquering the curse of dimensionality in gene expression cancer diagnosis: tough problem, simple models. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 514--523. Springer.

Nigro, C. L., Rusmini, M., and Ceccherini, I. (2019). Ret in breast cancer: pathogenic implications and mechanisms of drug resistance. *Cancer Drug Resist*, 2:1136--52.

Nishimura, D. (2001). Biocarta. *Biotech Software & Internet Report: The Computer Software Journal for Scient*, 2(3):117–120.

Paliouras, G., Karkaletsis, V., and Spyropoulos, C. D. (2003). *Machine learning and its applications: advanced lectures*, volume 2049. Springer.

Paredes, J., Correia, A. L., Ribeiro, A. S., Albergaria, A., Milanezi, F., and Schmitt, F. C. (2007). P-cadherin expression in breast cancer: a review. *Breast Cancer Research*, 9:1--12.

Parikh, R., Mathai, A., Parikh, S., Sekhar, G. C., and Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian journal of ophthalmology*, 56(1):45--50.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8):1160--1167.

Perou, C. M., Sørlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., et al. (2000). Molecular portraits of human breast tumours. *Nature*, 406(6797):747--752.

Piatetsky-Shapiro, G. and Tamayo, P. (2003). Microarray data mining: Facing the challenges. *SIGKDD Explorations*, 5:1–5.

Prat, A., Adamo, B., Cheang, M. C., Anders, C. K., Carey, L. A., and Perou, C. M. (2013). Molecular characterization of basal-like and non-basal-like triple-negative breast cancer. *The oncologist*, 18(2):123--133.

Rácz, A., Bajusz, D., and Héberger, K. (2021). Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification. *Molecules*, 26(4):1111.

Rajapakse, J. C. and Mundra, P. A. (2013). Multiclass gene selection using pareto-fronts. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(1):87--97.

Rosa, G. J. d. M., Rocha, L. B. d., and Furlan, L. R. (2007). Estudos de expressão gênica utilizando-se microarrays: delineamento, análise, e aplicações na pesquisa zootécnica. *Revista Brasileira de Zootecnia*, pages 186--209.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467--470.

Shen, D., Wu, G., and Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221--248.

Shetta, O. and Niranjan, M. (2020). Robust subspace methods for outlier detection in genomic data circumvents the curse of dimensionality. *Royal Society open science*, 7(2):14.

Shi, Y., Zhao, Y., Zhang, Y., AiErken, N., Shao, N., Ye, R., Lin, Y., and Wang, S. (2018). Aff3 upregulation mediates tamoxifen resistance in breast cancers. *Journal of Experimental & Clinical Cancer Research*, 37(1):254.

Shukla, A. K., Singh, P., and Vardhan, M. (2018). A hybrid gene selection method for microarray recognition. *Biocybernetics and Biomedical Engineering*, 38(4):975--991.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427--437.

Souza, L. R., Colonna, J. G., Comodaro, J. M., and Naveca, F. G. (2022). Using amino acids co-occurrence matrices and explainability model to investigate patterns in dengue virus proteins. *BMC bioinformatics*, 23(1):1--19.

Sutter, J. M. and Kalivas, J. H. (1993). Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection. *Microchemical journal*, 47(1-2):60--66.

Tan, A. C. and Gilbert, D. (2003). Ensemble machine learning on gene expression data for cancer classification. *Applied Bioinformatics*, 2.

Tan, L., Song, X., Sun, X., Wang, N., Qu, Y., and Sun, Z. (2016). Art3 regulates triple-negative breast cancer cell function via activation of akt and erk pathways. *Oncotarget*, 7(29):46589.

Tang, Y., Zhang, Y.-Q., and Huang, Z. (2007). Development of two-stage svm-rfe gene selection strategy for microarray expression data analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(3):365--381.

Tarek, S., Elwahab, R. A., and Shoman, M. (2017). Gene expression based cancer classification. *Egyptian Informatics Journal*, 18(3):151--159.

Tefferi, A., Bolander, M. E., Ansell, S. M., Wieben, E. D., and Spelsberg, T. C. (2002). Primer on medical genomics part iii: microarray experiments and data analysis. In *Mayo Clinic Proceedings*, volume 77, pages 927--940. Elsevier.

Thakkar, A., Raj, H., Ravishankar, Muthuvelan, B., Balakrishnan, A., and Padigaru, M. (2015). High expression of three-gene signature improves prediction of relapse-free survival in estrogen receptor-positive and node-positive breast tumors. *Biomarker insights*, 10:BMI--S30559.

Tong, M., Liu, K.-H., Xu, C., and Ju, W. (2013). An ensemble of svm classifiers based on gene pairs. *Computers in biology and medicine*, 43(6):729--737.

Treeratpituk, P. and Giles, C. L. (2009). Disambiguating authors in academic publications using random forests. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 39--48.

Turner, N. C., Swift, C., Kilburn, L., Fribbens, C., Beaney, M., Garcia-Murillas, I., Budzar, A. U., Robertson, J. F., Gradishar, W., Piccart, M., et al. (2020). Esr1 mutations and overall survival on fulvestrant versus exemestane in advanced hormone receptor–positive breast cancer: A combined analysis of the phase iii sofea and efect trials. *Clinical Cancer Research*, 26(19):5172--5177.

Wang, X., Wan, J., Xu, Z., Jiang, S., Ji, L., Liu, Y., Zhai, S., and Cui, R. (2019). Identification of competitive endogenous rnas network in breast cancer. *Cancer medicine*, 8(5):2392--2403.

Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57.

Xie, H., Li, J., Zhang, Q., and Wang, Y. (2016). Comparison among dimensionality reduction techniques based on random projection for cancer classification. *Computational biology and chemistry*, 65:165--172.

Xu, D. and Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165--193.

Yan, K. and Zhang, D. (2015). Feature selection and analysis on correlated gas sensor data with recursive feature elimination. *Sensors and Actuators B: Chemical*, 212:353--363.

Yersal, O. and Barutca, S. (2014). Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World journal of clinical oncology*, 5(3):412--424.

Yip, W.-K., Amin, S. B., and Li, C. (2011). A survey of classification techniques for microarray data analysis. In *Handbook of Statistical Bioinformatics*, pages 193--223. Springer.

Yu, Y., Kossinna, P., Liao, W., and Zhang, Q. (2021). Explainable autoencoder-based representation learning for gene expression data. *bioRxiv*.

Žalik, K. R. (2008). An efficient k-means clustering algorithm. *Pattern Recognition Letters*, 29(9):1385--1391.

Zhang, K., Xu, P., and Zhang, J. (2020). Explainable ai in deep reinforcement learning models: A shap method applied in power system emergency control. In *2020 IEEE 4th Conference on Energy Internet and Energy System Integration (EI2)*, pages 711--716. IEEE.

Zhao, L.-c., Shen, B.-y., Deng, X.-x., Chen, H., Zhu, Z.-g., and Peng, C.-h. (2016). Tmem45b promotes proliferation, invasion and migration and inhibits apoptosis in pancreatic cancer cells. *Molecular BioSystems*, 12(6):1860--1870.

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of rna-seq and microarray in transcriptome profiling of activated t cells. *PloS one*, 9(1).

A

Genes functional description

Supplementary Table I: Signature A genes functional description.

Gene	Description
ACOX2	Encodes the branched-chain acyl-CoA oxidase. Deficiency of this enzyme results in the accumulation of branched fatty acids and bile acid intermediates, and may lead to Zellweger syndrome, severe cognitive disability, and death in children.
AGAP11	Putative GTPase-activating protein. Among its related pathways are Endocytosis.
ART3	Encodes an arginine-specific ADP-ribosyltransferase. The encoded protein catalyzes a reversible reaction which modifies proteins by the addition or removal of ADP-ribose to an arginine residue to regulate the function of the modified protein. It is associated with Esophageal Candidiasis.
C1orf168	Adapter protein that plays a role in T-cell receptor (TCR)-mediated activation of signaling pathways. It is associated with Hypertropia and Syringomyelia.
C5orf34	C5orf34 is a Protein Coding gene. It is an uncharacterized protein.
CMBL	Is a cysteine hydrolase of the dienelactone hydrolase family that is highly expressed in liver cytosol. Diseases associated with CMBL include Waardenburg Syndrome Type 3 and Type 1.
COCH	Plays a role in the control of cell shape and motility in the trabecular meshwork. Hybridization to this gene was detected in spindle-shaped cells located along nerve fibers between the auditory ganglion and sensory epithelium. It is associated with Deafness and Autosomal Dominant 9.

(*continues on next page*)

(*continued from previous page*)

ESR1	This gene encodes an estrogen receptor, a ligand- activated transcription factor composed of several domains important for hormone binding, DNA binding, and activation of transcription. Breast cancer is associated with ESR1.
FERMT1	Encodes a member of the fermitin family, and contains a FERM domain and a pleckstrin homology domain. The encoded protein is involved in integrin signaling and linkage of the actin cytoskeleton to the extracellular matrix. Mutations in this gene have been linked to Kindler syndrome.
GPR98	Encodes a member of the G-protein coupled receptor superfamily. The encoded protein contains a 7-transmembrane receptor domain, binds calcium and is expressed in the central nervous system. Mutation in this gene are associated with Usher syndrome 2 and familial febrile seizures.
HDC	Encodes a member of the group II decarboxylase family and forms a homodimer that converts L-histidine to histamine in a pyridoxal phosphate dependent manner. Histamine regulates several physiologic processes, including neurotransmission, gastric acid secretion,inflamation, and smooth muscle tone. It is associated with Gilles De La Tourette Syndrome and Mastocytosis, Cutaneous.
HPX	Binds heme and transports it to the liver for breakdown and iron recovery, after which the free hemopexin returns to the circulation. The encoded protein is an acute phase protein that transports heme from the plasma to the liver and may be involved in protecting cells from oxidative stress. It is associated with Hepatitis E and Blackwater Fever.
IL20RB	Forms a heterodimeric receptor for interleukin-20. Among its related pathways are AKT Siganaling Pathway and PEDF Induced Signaling.
MGC16142	MGC16142 is an RNA gene, and is affiliated with the lncRNA class. It is an uncharacterized protein.
LPH	Rab effector protein involved in melanosome transport. A mutation in this gene results in Griscelli syndrome type 3, which is characterized by a silver-gray hair color and abnormal pigment distribution in the hair shaft.
MPV17L	Isoform 1 participates in reactive oxygen species metabolism by up- or down-regulation of the genes of antioxidant enzymes. Among its related pathways are Peroxisome.
NAPRT1	Catalyzes the first step in the biosynthesis of NAD from nicotinic acid, the ATP-dependent synthesis of beta-nicotinate D-ribonucleotide from nicotinate and 5-phospho-D-ribose 1-phosphate. It is associated with Pellagra.
PLA2G4F	Calcium-dependent phospholipase A2 that selectively hydrolyzes glycerophospholipids in the sn-2 position. It is associated with Ileocolitis and Inflammatory Bowel Disease.

(*continues on next page*)

(*continued from previous page*)

RERG	Binds GDP/GTP and possesses intrinsic GTPase activity. Has higher affinity for GDP than for GTP. Breast cancer is associated with RERG.
SLCO5A1	Encodes a 12 transmembrane domain protein that is a member of the solute carrier organic anion transporter superfamily. It is associated with Mesomelia-Synostoses Syndrome and Mesomelia.
TRPM8	Receptor-activated non-selective cation channel involved in detection of sensations such as coolness. It is expressed in breast tumours.
TTYH	Encodes a member of the tweety family of proteins. It is associated with Pediatric Infratentorial Ependymoma and Neuropathy, Congenital Hypomyelinating, and Autosomal Recessive.

Supplementary Table II: Signature B genes functional description.

Gene	Description
AFF3	Putative transcription activator that may function in lymphoid development and oncogenesis. It is associated with Childhood T-Cell Acute Lymphoblastic Leukemia and Eaf.
BCAS1	Resides in a region at 20q13 which is amplified in a variety of tumor types and associated with more aggressive tumor phenotypes. Found to be highly expressed in three amplified breast cancer cell lines and in one breast tumor without amplification at 20q13.2.
C9orf152	C9orf152 is a Protein Coding gene. It is an uncharacterized protein. Therefore, it needs further studies.
DNAJC12	This gene encodes a member of a subclass of the HSP40/DnaJ protein family. Is associated with complex assembly, protein folding, and export. It is asssociated with proliferative and non-proliferative type fibrocystic change of breast.
FOXA1	Encodes a member of the forkhead class of DNA- binding proteins. These hepatocyte nuclear factors are transcriptional activators for liver-specific transcripts such as albumin and transthyretin, and they also interact with chromatin Diseases associated with FOXA1 include Estrogen-Receptor Negative breast cancer and Estrogen-Receptor Positive breast cancer.

(*continues on next page*)

(*continued from previous page*)

FZD9	Encode 7-transmembrane domain proteins that are receptors for Wnt signaling proteins. The deletion of the FZD9 gene may contribute to the Williams syndrome phenotype. It is associated with Williams-Beuren Syndrome and Exudative Vitreoretinopathy.
RET	Involved in numerous cellular mechanisms including cell proliferation, neuronal navigation, cell migration, and cell differentiation upon binding with glial cell derived neurotrophic factor family ligands. It is associated with Multiple Endocrine Neoplasia, Type Iia and Thyroid Carcinoma.
SLC44A4	Plays a role in the choline-acetylcholine system and is required to the efferent innervation of hair cells in the olivocochlear bundle for the maintenance of physiological function of outer hair cells and the protection of hair cells from acoustic injury. It is associated with Deafness and Autosomal Dominant 72.
ST8SIA1	Involved in the production of gangliosides GD3 and GT3 from GM3; gangliosides are a subfamily of complex glycosphinglolipds that contain one or more residues of sialic acid. It is associated with Miller Fisher Syndrome.

Supplementary Table III: Signature C genes functional description.

Gene	Description
A2ML1	Encodes a member of the alpha-macroglobulin superfamily. In a few cases, presents mutations of the protein in breast cancer. Is able to inhibit all four classes of proteinases by a unique "trapping" mechanism. Is associated with Otitis Media and Noonan Syndrome 1.
ABCA12	Probable transporter involved in lipid homeostasis. Belongs to the ABC transporter superfamily. ABC proteins transport various molecules across extra- and intracellular membranes. Associated with Ichthyosis, Congenital and Autosomal Recessive 4B.
ABCC11	Is a member of the superfamily of ABC transporters. The product of this gene participates in physiological processes involving bile acids, conjugated steroids, and cyclic nucleotides. This gene is overexpressed in breast tissue.
ADAMTS15	Encodes a member of the ADAMTS protein family. Is associated with Spondyloepimetaphyseal Dysplasia, Missouri Type and Jacobsen Syndrome. This gene may function as a tumor suppressor in colorectal and breast cancers.
AGR3	Encodes a member of the disulfide isomerase (PDI) family of endoplasmic reticulum (ER) proteins that catalyze protein folding and thiol-disulfide interchange reactions. This gene is overexpressed in breast cancer.

(*continues on next page*)

(*continued from previous page*)

ANKRD30A	Encodes a DNA-binding transcription factor that is uniquely expressed in mammary epithelium and the testis. Altered expression levels have been associated with breast cancer progression. It is associated with breast cancer and Vestibulocochlear Nerve Disease.
ANKRD30B	Is a Protein Coding gene. An important paralog of this gene is ANKRD30A. It is also expressed in breast cancer.
CA12	Belongs to the alpha-carbonic anhydrase family. Participate in a variety of biological processes, including respiration, calcification, acid-base balance, bone resorption, and the formation of aqueous humor, cerebrospinal fluid, saliva, and gastric acid. It is highly expressed in normal tissues, such as kidney, colon and pancreas, and has been found to be overexpressed in 10% of clear cell renal carcinomas.
TMEM45B	Belongs to the TMEM45 family. It promotes proliferation, invasion and migration and inhibits apoptosis in pancreatic cancer cells.