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

# EXPLORING EFFICIENT GENE SET ALTERNATIVES FOR IMPROVED BREAST CANCER SUBTYPE CLASSIFICATION

Manaus - AM

Janeiro de 2025

LEANDRO YOUITI SILVA OKIMOTO

# EXPLORING EFFICIENT GENE SET ALTERNATIVES FOR IMPROVED BREAST CANCER SUBTYPE CLASSIFICATION

Tese apresentada ao Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal do Amazonas, Campus Universitário Senador Arthur Virgílio Filho, como requisito parcial para a obtenção do grau de Doutor em Informática.

Orientadora: Fabíola Guerra Nakamura
Co-orientador: Eduardo Freire Nakamura

Manaus - AM

Janeiro de 2025

LEANDRO YOUITI SILVA OKIMOTO

# EXPLORING EFFICIENT GENE SET ALTERNATIVES FOR IMPROVED BREAST CANCER SUBTYPE CLASSIFICATION

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

Advisor: Fabíola Guerra Nakamura
Co-advisor: Eduardo Freire Nakamura

Manaus - AM
January 2025

Ficha Catalográfica

Elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

O41e Okimoto, Leandro Youiti Silva
Exploring Efficient Gene Set Alternatives for Improved Breast Cancer Subtype Classification / Leandro Youiti Silva Okimoto. - 2025.
82 f. : il., color. ; 31 cm.

Orientador(a): Fabíola Guerra Nakamura.
Coorientador(a): Eduardo Freire Nakamura.
Tese (doutorado) - Universidade Federal do Amazonas, Programa de Pós-Graduação em Informática, Manaus - AM, 2025.

1. Breast Cancer. 2. Gene Expression. 3. Machine Learning. 4. PAM50. 5. Subtype classification. I. Nakamura, Fabíola Guerra. II. Nakamura, Eduardo Freire. III. Universidade Federal do Amazonas. Programa de Pós-Graduação em Informática. IV. Título

Ministério da Educação
Universidade Federal do Amazonas
Coordenação do Programa de Pós-Graduação em Informática

## FOLHA DE APROVAÇÃO

## "EXPLORING EFFICIENT GENE SET ALTERNATIVES FOR IMPROVED BREAST CANCER SUBTYPE CLASSIFICATION"

**LEANDRO YOUITI SILVA OKIMOTO**

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

Profa. Dra. Fabíola Guerra Nakamura - **Presidente**

Prof. Dr. David Fenyo - **Membro Externo**

Claudio T. Silva

Prof. Dr. Claudio T. Silva - **Membro Externo**

Profa. Dra. Eulanda Miranda dos Santos - **Membro Interno**

Manaus, 16 de janeiro de 2025.

Documento assinado eletronicamente por **Fabíola Guerra Nakamura**, **Professor do Magistério Superior**, em 08/09/2025, às 17:12, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015 .

Documento assinado eletronicamente por **Eulanda Miranda dos Santos**, **Professor do Magistério Superior**, em 08/09/2025, às 20:49, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015 .

A autenticidade deste documento pode ser conferida no site https://sei.ufam.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2408083** e o código CRC **4DE299C9**.

Avenida General Rodrigo Octávio, 6200 - Bairro Coroado I Campus Universitário Senador Arthur Virgílio Filho, Setor Norte - Telefone: (92) 3305-1181 / Ramal 1193
CEP 69080-900, Manaus/AM, coordenadorppgi@icomp.ufam.edu.br

Referência: Processo nº 23105.001564/2025-53 SEI nº 2408083

*Dedico essa tese aos meus pais e irmão por todo amor e apoio que deram nessa jornada.*

*À minha esposa, que me apoiou em todas as minhas decisões.*

*"Wherever the art of medicine is loved, there is also a love of humanity."*
(Hippocrates)

# Acknowledgments

Agradeço primeiramente a Deus, pela força e direção em cada etapa desta jornada.

Aos meus pais, pelo apoio incondicional e por me incentivarem a buscar minhas conquistas por meio do estudo e do trabalho.

A todos os meus familiares, com um agradecimento especial ao meu irmão André, a quem desejo que continue seus estudos e realize todos os seus sonhos.

À minha esposa, Roberta Kathleen, minha companheira e maior incentivadora, que esteve ao meu lado em todas as decisões e nunca deixou de caminhar junto comigo. Sua presença e apoio foram fundamentais para que eu chegasse até aqui.

Agradeço à Professora Fabíola Guerra Nakamura, que, com seu conhecimento e habilidade, aceitou o desafio de explorar uma nova área de conhecimento e me guiou com maestria ao longo dessa jornada.

Ao Professor Eduardo Freire Nakamura, pela orientação, disponibilidade e constante contribuição, sempre disposto a ajudar e compartilhar seu conhecimento.

Aos professores do Doutorado, que de diferentes formas contribuíram para a minha formação acadêmica e pessoal.

Aos professores David Fenyö e Claudio Silva, pela oportunidade ímpar de aprendizado e colaboração na Universidade de Nova Iorque, experiência que marcou profundamente minha trajetória.

Aos colegas do PPGI e de trabalho, Gustavo Aquino, Hendrio Bragança, Paulo Henrique, Rayol Neto, Ariel Afonso, e tantos outros que não mencionei, mas que caminharam comigo durante essa jornada, meu sincero agradecimento.

Por fim, agradeço aos funcionários do ICOMP, pela disponibilidade, simpatia e gentileza, que tornaram o dia a dia mais leve e acolhedor.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Este trabalho foi parcialmente financiado pela Fundação de Amparo à Pesquisa do Estado do Amazonas – FAPEAM – por meio do projeto POSGRAD.

# Resumo

Nos últimos anos, os avanços na pesquisa sobre o câncer de mama destacaram a necessidade de métodos mais eficientes para lidar com a vasta quantidade de dados genômicos associados aos seus diversos subtipos. A assinatura gênica PAM50, embora bem-sucedida como uma ferramenta prognóstica, apresenta desafios devido à sua dependência de um grande número de genes, impactando tanto o custo quanto a complexidade. Esta tese tem como objetivo explorar conjuntos de genes alternativos para a classificação dos subtipos de câncer de mama que mantenham ou melhorem a precisão da assinatura PAM50. Nossa pesquisa começa com o método "Few-Shot Genes Selection", que foi especificamente desenvolvido para confrontar a assinatura gênica PAM50. Ao identificar e avaliar subconjuntos menores dentro da própria seleção de genes do PAM50, demonstramos que, mesmo com um número reduzido dos mesmos genes, é possível alcançar um desempenho de classificação igual ou superior. Essa descoberta desafia a noção de que um conjunto maior de genes é inerentemente melhor, mostrando que subconjuntos mais enxutos podem ser igualmente eficazes. Com base nisso, exploramos uma nova abordagem com o segundo foco de pesquisa, "Using Multilayer Classification to Improve Worst Prognosis Breast Cancer Subtypes Outcomes". Esse método busca identificar conjuntos de genes alternativos, distintos dos presentes no PAM50, mas com ênfase específica em melhorar a precisão da classificação para os subtipos de pior prognóstico, como Basal e Her2. Através de uma estrutura de classificação hierárquica, buscamos aprimorar os resultados para esses subtipos selecionando genes que sejam mais preditivos de casos de pior prognóstico. Em conclusão, esta tese demonstra que, refinando assinaturas gênicas existentes ou descobrindo novos conjuntos de genes, podemos desenvolver ferramentas mais eficientes para a classificação do câncer de mama. Esta pesquisa contribui para o esforço contínuo de melhorar a precisão diagnóstica e os resultados do tratamento em ambientes clínicos, particularmente para aqueles subtipos com os prognósticos mais desafiadores.

# Abstract

In recent years, advances in breast cancer research have underscored the need for more efficient methods to handle the vast genomic data associated with its diverse subtypes. The signature of the PAM50 gene, although successful as a prognostic tool, poses challenges due to its reliance on a large number of genes, impacting both cost and complexity. This thesis aims to explore alternative gene sets for breast cancer subtype classification that maintain or improve the accuracy of the PAM50 signature. Our research begins with the "Few-Shot Genes Selection" method, which was specifically designed to confront the signature of the PAM50 gene. By identifying and evaluating smaller subsets within the PAM50's own gene selection, we show that even with a reduced number of the same genes, it is possible to achieve equal or superior classification performance. This finding challenges the notion that a larger gene set is inherently better, showing that streamlined subsets can be just as effective. Building on this, we explore a new approach with the second research focus, "Using Multi-layer Classification to Improve Worst Prognosis Breast Cancer Subtypes Outcomes". This method seeks to identify alternative gene sets, distinct from those in PAM50, but with a specific emphasis on improving classification accuracy for the worst prognosis subtypes, such as Basal and Her2. Through a layered classification framework, our objective is to improve the outcomes of these subtypes by selecting genes that are more predictive of poor prognosis cases. In conclusion, this thesis demonstrates that by refining existing gene signatures or discovering new gene sets focused on specific subtypes, we can develop more efficient and cost-effective tools for breast cancer classification. This research contributes to the ongoing effort to improve diagnostic precision and treatment outcomes in clinical settings, particularly for those subtypes with the most challenging prognoses.

# List of Figures

# List of Tables

# Contents

1

# Introduction

Breast cancer ranks the second most prevalent cancer and the primary cause of cancer-related fatalities worldwide [Bray et al., 2018, Jemal et al., 2011]. Its highly heterogeneous nature [Miah et al., 2017] gives rise to distinct genetic variations, clinical outcomes, and treatment approaches across different tumor subtypes [Chen et al., 2016].

Despite this heterogeneity, researchers have found ways to address different treatments based on subtypes, there are four primary molecular subtypes of breast cancer [Perou et al., 2000]: basal, HER2, luminal A, and luminal B. This classification allowed people who have a poor prognosis to be separated from those who have a good prognosis, for example, the basal and HER2 subtypes are associated with the poorest prognoses [Bertucci et al., 2012, Dwivedi et al., 2019], while the luminal A and luminal B subtypes have more favorable outcomes due to the availability of effective targeted therapies [Dwivedi et al., 2019, Yersal and Barutca, 2014], as shown in Figure 1.1.

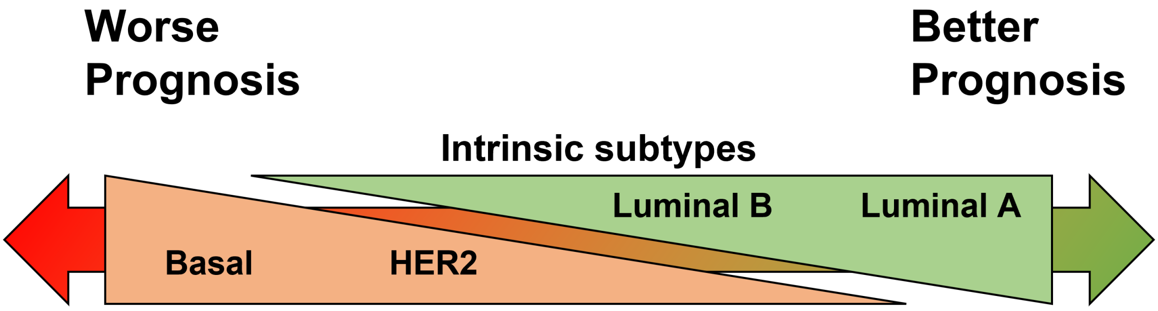


**Figure 1.1.** Breast Cancer Intrinsic Subtypes, figure adapted from Dai et al. [2015].

Cancer classification is crucial in identifying subtypes of cancer samples, providing efficient, accurate and objective diagnoses for various types of cancer [Tarek et al., 2017, Tong et al., 2013]. In addition, biological (or "intrinsic") subtype diagnosis adds significant prognostic and predictive value for patients with breast cancer [Parker et al., 2009]. Thus, accurate classification of breast cancer subtypes and identification of key genes are essential for appropriate and effective patient treatment, and it is also imperative to not misclassify patients with the worst prognosis, once they would not go through the proper treatment.

Advancements in omic techniques have transformed the analysis of gene expression data, enabling better understanding of various diseases, including cancer. For example, breast cancer, a heterogeneous disease, has different subtypes with unique biological properties and treatment responses [Consortium et al., 2012, Kundaje et al., 2015, Lonsdale et al., 2013, Weinstein et al., 2013]. Although an abundance of genomic and proteomic data has become available in recent years, harnessing this information requires efficient theoretical frameworks, tools, and computational methods [Graudenzi et al., 2017, Raghu et al., 2020].

A challenge in using gene expression data for cancer classification is handling high-dimensional matrices of genes versus samples. Typically, the number of samples is considerably smaller than the number of genes (for example, more than 10,000 genes for each sample in a dataset containing only a thousand samples) [Piatetsky-Shapiro and Tamayo, 2003].

Feature selection methods are widely used in the literature to address high-dimensional classification tasks. By removing redundant and irrelevant genes, these methods can reduce the dimension of the data, streamline learning models, accelerate the learning process, and significantly improve classification performance [Alanni et al., 2019, Díaz-Uriarte and De Andres, 2006, Liu and San Wong, 2017].

In recent years, the signature of the PAM50 gene, a set of 50 genes used to classify breast cancer into four intrinsic molecular subtypes, has emerged as a valuable prognostic tool in breast cancer research, providing insight into tumor subtypes and guiding therapeutic decision making [Bastien et al., 2012]. However, the PAM50 approach has limitations, including its reliance on a relatively large number of genes, which can increase costs and complexity in both the research and clinical settings [Huang et al., 2018]. Although 50 genes may not seem excessive at first glance, in medical research and diagnostics, handling such a number can be quite a burdensome. Each gene requires specific tests, reagents, and validation processes, all of which contribute to increased time, cost, and logistical challenges. As a result, there is a growing need for more efficient methods that can accurately classify breast cancer subtypes using fewer genes, making the process both more cost-effective and accessible.

This proposal explores two complementary approaches to address this challenge: a "Few-Shot Genes Selection" method and a "Multi-Layer Classification" technique.

The primary objective of this research is to investigate the possibility of achieving an accurate classification of breast cancer subtypes using a reduced number of characteristics.

## 1.1 Motivation

Cancer classification research has extensively focused on gene expression data [Yip et al., 2011]; most cancer classification approaches are tailored for binary cancer classification (cancer / noncancer). Implementing a practical cancer classification method requires the use of gene selection techniques [Shukla et al., 2018].

Gene selection methods address the high dimensionality of data, streamline machine learning techniques, accelerate classification, and significantly enhance classifier performance [Alanni et al., 2019, Díaz-Uriarte and De Andres, 2006, Liu and San Wong, 2017].

In order to advance gene selection methods for breast cancer subtypes, it is essential to identify a minimal set of genes capable of accurately characterizing a patient's tumor. Unfortunately, traditional approaches in the literature often depend on a single set of genes for all subtypes (e.g., PAM50). Occasionally, these strategies are linked to unsupervised methodologies that have been established in the past. As a result, medical professionals may still rely on outdated strategies, which underscores the need for more refined and updated gene selection methods.

## 1.2 Research Hypothesis

Breast cancer subtype classification is an essential aspect of precision medicine, as it allows for the development of customized treatment plans for patients. Identifying key genes in these subtypes can enable a more accurate classification and improved therapeutic strategies. Current methods, such as the well-known PAM50 gene signature, have proven useful in subtype classification. However, it is necessary to explore whether more refined and reduced gene sets can achieve the same level of accuracy. This research aims to investigate the subsets derived from PAM50 and address the question **"How far can we reduce the gene set while maintaining an accurate classification?"**.

We must consider potential classification errors as we go deeper into the analysis of reduced gene sets and their impact on classification. An intriguing question is whether misclassified samples lie near the boundaries between different cancer subtypes.

If this is the case, it may be possible that such samples possess characteristics of multiple subtypes, leading to misclassification. Understanding the nature of these samples and the reasons behind their misclassification will provide valuable information on the complexity of breast cancer subtypes and inform the development of more accurate classification methods.

Based on the first hypothesis, we have developed the “Few-Shot Gene Selection” approach, which aims to investigate the potential for precise categorization of breast cancer subtypes using a reduced gene set derived from the well-established PAM50 gene signature. The “Few-Shot Gene Selection” method randomly selects smaller subsets from PAM50. Evaluate your classification performance using machine learning metrics and a linear model, specifically employing the Support Vector Machine (SVM) classifier.

## 1.3 Contributions

The contributions of this thesis are grounded in two key publications. First, the development of a novel approach for breast cancer subtype classification using reduced gene sets derived from the PAM50 signature, as presented in *Few-shot genes selection: Subset of PAM50 genes for breast cancer subtypes classification* (published in BMC Bioinformatics) [Okimoto et al., 2024]. Second, a collaborative effort to explore machine learning techniques for analyzing the PAM50 gene set, detailed in *Classification of breast cancer subtypes: A study based on representative genes* (published in the Journal of the Brazilian Computer Society) [Mendonca-Neto et al., 2022].

## 1.4 Objectives

The primary objective of this work is to evaluate the possibility of maintaining the accuracy of the classification while using a more compact gene set and simplifying the process of identifying the subtypes.

The specific objectives include:

1. Identify relevant genes derived from the well-known PAM50 Gene Signature that effectively represent the various breast cancer subtypes;

2. Demonstrate the efficiency of the proposed method, resulting in a higher quality classifier that uses fewer genes compared to traditional approaches;

3. Identify relevant genes derived in a multilayer classification process that effectively represent the various breast cancer subtypes;

4. Discover new gene sets capable of characterizing the subtypes: Basal, Her2, Luminal A, and Luminal B;
5. Identify misclassified samples occurring at the boundaries between subtypes in order to investigate the reasons for misclassification in the future.

## 1.5 Document Outline

In Chapter 2, Fundamentals, we introduce essential concepts, including gene expression, gene sets, and PAM50 Gene Signature, along with machine learning principles, evaluation metrics, and dimensionality reduction for visualization. In Chapter 3, Related work, we discuss prevalent approaches for cancer classification and studies associated with gene selection. In Chapter 4 and Chapter 5, we outline two approaches to our proposed methods for gene selection, describe the methodology employed, and provide a comprehensive account of our experimental results. Finally, in Chapter 6, Final Remarks, we highlight our research contributions, outline the limitations of this work, and future directions.

2

# Fundamentals

This chapter presents the concepts necessary for the understanding of this proposal. First, we elucidate some biological concepts, such as gene expression, used as data in our method. The machine learning methods employed in this proposal are then explained, including supervised and unsupervised approaches. Next, we demonstrate the evaluation metrics used to evaluate the performance of the classifier. 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 terminology and definitions.

### 2.1.1 Gene Expression Data

Gene expression refers to the process of converting genetic information encoded in DNA into functional products such as proteins. Technologies like DNA microarrays and RNA-Seq have made it possible to analyze the expression levels of thousands of genes simultaneously. These expression levels reflect the synthesis of mRNA molecules in cells, offering insights into various biological processes. By analyzing gene expression, researchers can diagnose diseases, identify tumors, determine optimal treatments, detect genetic mutations, and more. To achieve these objectives, a variety of computational techniques, including pattern classification methods, are often applied [Garro et al., 2016].

Gene expression data, obtained through microarray or RNA-Seq platforms, are stored in what are referred to as gene expression datasets. These datasets are organized as a matrix in which rows represent the expression levels of different genes, and columns correspond to the expression profiles of various samples [Kerr et al., 2008]. Each entry in this matrix reflects the measured expression level of a specific gene in a specific sample. The matrix has $n$ rows (genes) and $m$ columns (samples), which form what is known as a gene expression profile [Tan and Gilbert, 2003]. An example of such a matrix is shown in Figure 2.1.

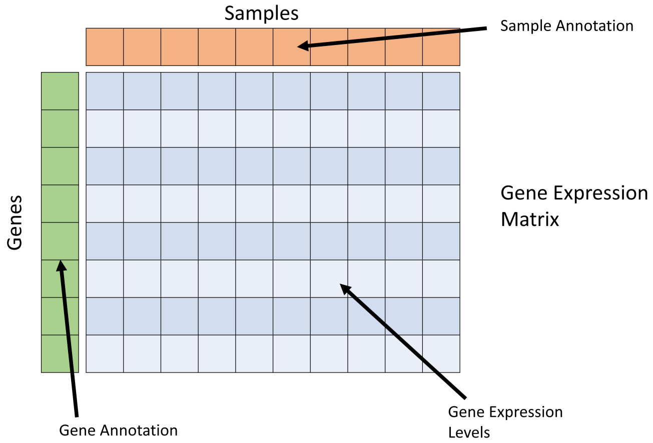


**Figure 2.1.** Dataset example, figure adapted from Ayyad et al. [2019].

Gene expression data are fundamentally a numerical representation of the expression levels of each gene across different samples. It is important to note that the process of obtaining gene expression data can introduce challenges such as missing values, noise, and variations due to experimental conditions. To address these issues, data preprocessing is a critical step in transforming raw data into a more usable format. Although the detailed steps involved in the preprocessing are beyond the scope of this discussion, works like Herrero et al. [2003] provides a detailed overview of the techniques used.

From an informatics perspective, gene expression data can be viewed as a matrix, where each row corresponds to a feature (gene) and each column represents a sample [Bhandari et al., 2022]. In this context, the expression level of a gene is analogous to the feature values typically encountered in machine learning problems. This structure allows for the application of standard machine learning algorithms, where genes are treated as features and samples are seen as instances to be classified or clustered. Using computational techniques such as feature selection, clustering, and classification, researchers can extract meaningful patterns from gene expression data.

The advantage of this perspective is that it provides the ability to treat genes as measurable characteristics, enabling the use of a wide range of data mining and machine learning methods [Dey et al., 2021]. These methods allow for more accurate predictions and a better understanding of underlying biological processes, especially when working with high-dimensional datasets like those found in gene expression studies. By focusing on gene expression data as a structured matrix, we can harness the power of computational algorithms to facilitate more robust analyses of complex biological data.

### 2.1.2 Curse of Dimensionality

The data that we use, Gene Expression Data (Subsection 2.1.1), typically consist of several thousand genes, but only a few hundred samples. This imbalance between the number of samples and the genes arises from the difficulties of collecting microarray samples. High-dimensional data can lead to the "curse of dimensionality," which refers to the exponential increase in volume associated with adding extra dimensions to data, resulting in sparse data points across this larger space. As a consequence, distance metrics become imprecise, classifiers struggle to generalize, and overall accuracy can degrade significantly [Xie et al., 2016].

In gene expression data, where thousands of genes are measured across relatively few samples, this curse is particularly evident. The high dimensionality not only amplifies computational challenges but also introduces several biological and statistical pitfalls. One key problem is the risk of overfitting, as the model may capture noise or random fluctuations in the data rather than true underlying patterns [Piatetsky-Shapiro and Tamayo, 2003]. The vast number of genes increases the likelihood of spurious correlations, where irrelevant genes appear to be significant merely by chance. Additionally, gene expression data often contain noise arising from biological variability, technical artifacts, or environmental factors, making the analysis even more difficult [Mramor et al., 2005].

Addressing the curse of dimensionality in gene expression data requires careful consideration, particularly when the goal is to retain interpretability for clinical applications. Traditional dimensionality reduction techniques, such as Principal Component Analysis (PCA) [Ringnér, 2008], are often inappropriate in this context. While PCA and similar methods can reduce the dimensionality by merging genes into composite features, these approaches obscure the identity of individual genes. In the field of oncology, where personalized medicine and targeted therapies are paramount, it is crucial to maintain a clear understanding of which specific genes are driving a patient's cancer profile. As such, gene filtering and selection methods that preserve the biological relevance of individual genes are more appropriate for breast cancer subtype classification.

To address these challenges, gene filtering and selection approaches are commonly employed. These methods focus on reducing the number of genes used in the analysis, aiming to isolate those that are most informative while discarding irrelevant or redundant features. Key strategies include:

- **Gene Filtering:** During the data preprocessing stage, filtering techniques are applied to exclude genes that show little variation across samples or that are unlikely to contribute to the classification task. Methods such as variance filtering or statistical tests for differential expression can be used to narrow down the gene set to those most likely to be relevant to the subtype [Mramor et al., 2005].

- **Feature Selection Algorithms:** Machine learning algorithms such as Recursive Feature Elimination (RFE), Lasso Regression, and others are utilized to select subsets of genes that contribute most to the model's performance. These techniques enable us to identify a small set of informative genes without merging or transforming them into abstract components, preserving their individual identities for clinical interpretation [Huang et al., 2018].

- **Biological Pathway-Based Selection:** Another approach is to filter genes based on their known involvement in specific biological pathways related to cancer. By leveraging prior biological knowledge, we can target genes that are likely to be of high relevance to the disease, further reducing dimensionality while maintaining biological interpretability [Berger et al., 2013].

- **Hybrid Methods:** Some approaches combine statistical filtering with biological insight. For instance, initial statistical filtering can remove the bulk of irrelevant genes, and then further refinement can be based on pathway involvement or known gene functions to ensure that the most biologically relevant genes are retained [Ma and Dai, 2011].

By focusing on reducing the number of genes rather than combining them into latent factors, we maintain the interpretability of our models, allowing clinicians to understand which genes are driving subtype classification. This interpretability is crucial for translating genomic insights into effective treatments, as it enables more targeted therapeutic strategies.

In summary, addressing the curse of dimensionality in gene expression data for breast cancer subtype classification requires a balance between reducing dimensionality and maintaining biological interpretability. Gene filtering and feature selection methods allow us to focus on a smaller, more informative set of genes, improving the accuracy and robustness of our models without sacrificing their potential clinical utility. These methods are crucial for reducing the complexity of gene expression data while still allowing for precise and personalized treatment decisions.

### 2.1.3 Gene Set

A gene set is a collection of genes, an unstructured and unequal group associated with specific biological processes, locations, diseases, or pathways [Liberzon et al., 2011]. Gene sets are typically defined based on a shared biological function (e.g., cell cycle regulation), a specific location in the genome (e.g., genes located on chromosome 1), a disease association (e.g., genes linked to breast cancer), or their involvement in a biological pathway (e.g., the group of 128 genes that make up the KEGG cell cycle pathway). Although gene sets can be organized based on these functional or biological criteria, there are no strict rules governing their structure, which allows for a degree of arbitrariness in how they are compiled. For example, the Molecular Signatures Database (MSigDB) contains more than 10,000 gene sets, curated according to various criteria, some of which may appear subjective or context-dependent [Liberzon et al., 2011].

In the context of informatics, gene sets play a pivotal role as selected features within machine learning models. In our work, we consider genes as features in the data, and gene sets serve as a method for filtering and selecting relevant features to analyze. By focusing on specific gene sets, we reduce the dimensionality of the data, allowing the model to focus on the most biologically relevant subsets, rather than processing thousands of genes at once. This filtering of genes provides a balance between retaining meaningful biological information and improving the computational efficiency and accuracy of our models.

Gene sets, when used in predictive modeling, allow us to target groups of genes that have been pre-identified as relevant to the problem domain, in our case, breast cancer subtypes [Bastien et al., 2012]. This approach ensures that we are working with biologically informed features, enhancing the interpretability of the models and supporting the discovery of potential therapeutic targets. By analyzing these selected gene sets, we can derive valuable insights and contribute to the development of more accurate models for predicting disease outcomes, including the prognosis of breast cancer and the classification of subtypes.

### 2.1.4 PAM50 Gene Signature

When addressing gene sets within the breast cancer subtypes problem, the PAM50 gene signature is a critical point of reference and is widely considered a state-of-the-art method. The PAM50 signature, which is part of the Prosigna Breast Cancer Prognostic Gene Signature Assay, is a molecular classification method that supports the identification of breast cancer subtypes [Wallden et al., 2015]. This signature uses the expression levels of 50 selected genes (Table 2.1) to classify breast cancer into four intrinsic subtypes: Luminal A, Luminal B, HER2-enriched, and Basal-like. Accurate classification of these subtypes helps inform treatment decisions and is associated with different prognostic outcomes [Orrantia-Borunda et al., 2022].

From an informatics perspective, the PAM50 gene signature illustrates how selected features, in this case genes, can be used in predictive models that produce practical results. The selection of these 50 genes from thousands present in the human genome addresses the curse of dimensionality (Subsection 2.1.2) in breast cancer prognosis. This highlights the importance of feature selection techniques in machine learning and data analysis.

**Table 2.1.** PAM50 Gene Signature

| Gene Set | Genes |
|---|---|
| PAM50 | ACTR3B, ANLN, BAG1, BCL2, BLVRA, CCNB1, CCNE1, CDC20, CDC6, CDH3, CXXC5, EGFR, ERBB2, ESR1, EXO1, FOXA1, FOXC1, GRB7, KIF2C, KRT14, KRT17, KRT5, MAPT, MELK, MKI67, MMP11, MYBL2, MYC, NAT1, ORC6, PGR, PHGDH, SFRP1, SLC39A6, SLPI, SNAI2, STAT1, STK15, SYK, TFF3, TP53, TUBE1, TYMS, UBE2C, UBE2T |

The benefits of using a gene set for gene selection in machine learning models are significant, particularly in addressing the challenges of high-dimensional data. This underscores the importance of the signature of the PAM50 gene, which can be highlighted through the following key components:

- **Prognostic Value and Interpretability:** The PAM50 gene signature demonstrates how gene filtering and selection can reduce data complexity while maintaining relevance for clinical applications. The ability to classify tumors using 50 genes, rather than thousands, enables the construction of models that offer both accuracy and interpretability. This is especially important in medical applications where transparency is required for clinical decision-making.

- **Efficiency in Data Processing:** Gene expression data often involve high-dimensional datasets with limited samples. PAM50 demonstrates that by focusing on biologically relevant features, it is possible to improve computational efficiency while still capturing essential data characteristics. The careful selection of these genes ensures that models remain computationally feasible and provide useful insights for clinical decision-making.

- **Role in Predictive Modeling:** In breast cancer subtype classification, the PAM50 gene set remains a reference point due to its validation in both research and clinical settings. It serves as a benchmark for newer methods aimed at improving upon or complementing this signature. Using PAM50 as a starting point allows researchers to focus on identifying additional or alternative gene sets that might enhance classification performance, especially for subtypes with a poorer prognosis.

- **Clinical Application:** The PAM50 signature, as implemented in tools like the Prosigna Assay, is used to estimate the risk of breast cancer recurrence and guide treatment planning. It supports clinical decision-making by identifying patients who may benefit from more targeted or aggressive therapies. The implementation of this signature demonstrates how bioinformatics models can be applied in real-world clinical scenarios, affecting treatment strategies.

- **Foundation for Further Gene Selection Research:** The effectiveness of the PAM50 signature offers a starting point for further research in gene selection. Our approach builds upon this, focusing on improving subtype classification by identifying equally or more informative genes, without combining them into latent variables as done with methods like PCA [Ringnér, 2008]. This approach helps ensure that models remain interpretable, which is essential for understanding which genes are driving cancer progression and therapeutic responses.

In summary, the PAM50 gene signature serves as a widely accepted method for breast cancer subtype classification. It shows the importance of feature selection in addressing high-dimensional data challenges while maintaining the interpretability necessary for clinical application. By using PAM50 as a foundation, we aim to identify other gene sets that may provide comparable or improved performance, without compromising the clarity and clinical relevance that the PAM50 signature offers.

## 2.2 Machine Learning

In this proposal, we will use different terms when referring to topics related to machine learning. Therefore, it is necessary to define what each of these terms represents. Next, we will introduce some methods and algorithms utilized in this proposal.

### 2.2.1 Standard Scaler

The Standard Scaler is used when the attributes of the input dataset have significantly different ranges or are measured in various units. It is crucial to normalize the data by removing the mean and scaling to unit variance Bisong and Bisong [2019].

Applying the Standard Scaler transforms the data into a distribution with a mean value of 0 and a standard deviation of 1. For multivariate data, this transformation is performed feature-wise, that is, independently for each feature.

Standardization is defined as follows:

$$z = \frac{x - \mu}{\sigma}, \tag{2.1}$$

where $z$ represents the standardized value, $x$ is the original value of a data point, $\mu$ is the mean of the dataset, and $\sigma$ is the standard deviation of the dataset.

The mean is calculated as follows:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} (x_i), \tag{2.2}$$

where $\mu$ is the mean of the dataset, $N$ is the number of data points, and $x_i$ represents each individual data point in the dataset.

The standard deviation is determined as:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}, \tag{2.3}$$

where $\sigma$ represents the standard deviation of the dataset, $N$ is the number of data points, $x_i$ is each individual data point in the dataset, and $\mu$ is the mean of the dataset.

By applying the Standard Scaler, different datasets can be combined and analyzed, independent of the normalizations applied. Each feature is scaled according to its subset of values.

### 2.2.2 Support Vector Machine (SVM)

Support Vector Machines (SVM) are supervised learning algorithms widely used for classification and regression tasks. In the context of classification, the aim of SVM is to find the optimal hyperplane that separates the different classes in the feature space. In the case of binary classification, the goal is to maximize the margin between two classes while minimizing the classification error. The decision function of a linear SVM can be represented as:

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b, \tag{2.4}$$

where $\mathbf{x}$ is the input feature vector, $\mathbf{w}$ is the weight vector, and $b$ is the bias term. To find the optimal hyperplane, the SVM algorithm minimizes the following objective function:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i, \tag{2.5}$$

subject to the constraints:

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i = 1, \ldots, N, \tag{2.6}$$

where $N$ is the number of samples, $y_i$ is the class label of the sample $i$, $\xi_i$ is the slack variable, and $C$ is a regularization parameter that controls the trade-off between margin maximization and error minimization.

The algorithm for binary SVM is as follows:

**Algoritmo 1** Binary Support Vector Machine

```
procedure SVM($\mathbf{X}$, $\mathbf{y}$, $kernel$, $C$)
    Select $kernel$ and set parameter $C$
    Solve the dual optimization problem to obtain $\boldsymbol{\alpha}$
    Calculate $\mathbf{w}$ and $b$ from $\boldsymbol{\alpha}$
    return $\mathbf{w}$, $b$, and support vectors
end procedure
procedure SVM_PREDICT($\mathbf{w}$, $b$, $\mathbf{x}$, $kernel$, support vectors)
    Compute decision function $f(\mathbf{x})$
    return class label based on sign of $f(\mathbf{x})$
end procedure
```

To extend SVM for multi-class classification problems, we can use the One-vs-One (OvO) or One-vs-All (OvA) approaches. The algorithm for multi-class SVM using the One-vs-All approach is presented below:

**Algoritmo 2** Multi-Class Support Vector Machine

```
procedure MultiClass_SVM($\mathbf{X}$, $\mathbf{y}$, $kernel$, $C$)
    Initialize an empty list $classifiers$
    for each class $i$ do
        Prepare binary class labels for class $i$ vs. all other classes
        Train an SVM classifier using $\mathbf{X}$, the binary class labels, $kernel$, and $C$
        Add the trained classifier to the list $classifiers$
    end for
    return $classifiers$
end procedure
procedure MultiClass_SVM_PREDICT($classifiers$, $\mathbf{x}$, $kernel$)
    Initialize an empty list $scores$
    for each classifier in $classifiers$ do
        Compute the decision function $f(\mathbf{x})$ for the classifier
        Add the score to the list $scores$
    end for
    return class label with the highest score
end procedure
```

In the One-vs-One approach, an SVM classifier is trained for every pair of classes. For a problem with $K$ classes, a total of $K(K-1)/2$ classifiers are trained [Liu et al., 2017b]. The predicted class label is determined by majority voting among all classifiers. The concise algorithm for multi-class SVM using the One-vs-One approach is shown below:

**Algoritmo 3** One-vs-One Multi-Class Support Vector Machine

```
procedure OvO_MultiClass_SVM($\mathbf{X}$, $\mathbf{y}$, $kernel$, $C$)
    Initialize an empty list $classifiers$
    for each pair of distinct classes $i$ and $j$ do
        Prepare binary class labels and data for class $i$ vs. class $j$
        Train an SVM classifier using the binary data, class labels, $kernel$, and $C$
        Add the trained classifier to the list $classifiers$
    end for
    return $classifiers$
end procedure
procedure OvO_MultiClass_SVM_predict($classifiers$, $\mathbf{x}$, $kernel$)
    Initialize an empty list $votes$
    for each classifier in $classifiers$ do
        Compute the decision function $f(\mathbf{x})$ for the classifier
        Add the predicted class label to the list $votes$
    end for
    return class label with the highest number of votes
end procedure
```

### 2.2.3 SelectKBest

The SelectKBest method is a popular feature selection technique that selects the top K features based on their univariate statistical significance. This method helps to reduce the complexity of the problem, improve the performance of the model, and enable a better interpretation of the results [Lazaros et al., 2022].

In the context of analysis of gene expression data, the ANOVA F value is often used as a score function to classify characteristics according to their discriminatory power between two classes [Venkat et al., 2023]. The F-value measures the ratio of the variability between groups to the variability within groups, thus capturing the feature's ability to distinguish between classes.

The F-value for a given feature is computed as follows:

$$\begin{aligned} \text{F-value} = &\frac{(\text{mean}(X_{class1}) - \text{mean}(X_{total}))^2}{\text{var}(X_{class1}) + \text{var}(X_{class2})} \\ &+ \frac{(\text{mean}(X_{class2}) - \text{mean}(X_{total}))^2}{\text{var}(X_{class1}) + \text{var}(X_{class2})}, \end{aligned} \tag{2.7}$$

where $X_{class1}$ and $X_{class2}$ represent the characteristic values of the first and second classes, respectively, and $X_{total}$ represent the characteristic values of all samples.

The high-level algorithm for the SelectKBest method using the ANOVA F-value as the score function can be represented as follows:

**Algoritmo 4** SelectKBest with ANOVA F-value

```
procedure SELECTKBEST($\mathbf{X}$, $\mathbf{y}$, $K$)
    Initialize an empty list $F\_values$
    for each feature $f$ in $\mathbf{X}$ do
        Compute the F-value for feature $f$ using Equation (10)
        Add the computed F-value to the list $F\_values$
    end for
    Sort the features based on their F-values in descending order
    Select the top $K$ features with the highest F-values
    return the indices of the selected features
end procedure
```

This algorithm includes one main procedure, SelectKBest, which takes the input data $\mathbf{X}$, the class labels $\mathbf{y}$, and the number of top features to select $K$. The procedure initializes an empty list to store the F-values, computes the F-values for each feature using Equation 2.7, sorts the features based on their F-values, and selects the top K features with the highest F-values. The algorithm returns the indices of the selected features.

### 2.2.4 t-SNE

t-SNE is a popular dimensionality reduction technique for visualizing high-dimensional data in lower-dimensional spaces, such as two or three dimensions. The main objective of t-SNE is to preserve the data structure by maintaining the pairwise relationships between data points when projecting them onto a lower-dimensional space [Maaten and Hinton, 2008], example in Figure 2.2.

The t-SNE algorithm involves three main steps:

1. Compute pairwise affinities $p_{ij}$ in the high-dimensional space using a Gaussian distribution:

$$p_{ij} = \frac{\exp\left(-||x_i - x_j||^2/2\sigma^2\right)}{\sum_{k\neq l}\exp\left(-||x_k - x_l||^2/2\sigma^2\right)} \tag{2.8}$$

2. Compute pairwise affinities $q_{ij}$ in the low-dimensional space using a Student's t-distribution with one degree of freedom:

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k\neq l}(1 + ||y_k - y_l||^2)^{-1}} \tag{2.9}$$

3. Minimize the Kullback-Leibler divergence (KL divergence) Joyce [2011] between the two distributions concerning the positions of the points in the map using the

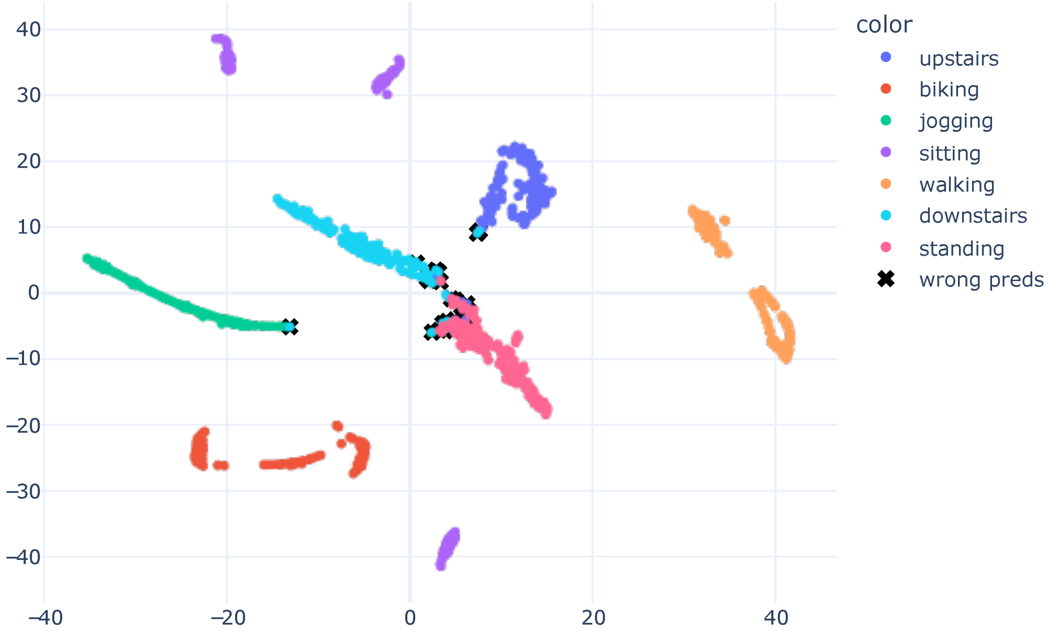


**Figure 2.2.** t-SNE example, figure adapted from Aquino et al. [2023].

following cost function:

$$C = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{2.10}$$

In these formulas, $x_i$ and $x_j$ are data points in the high-dimensional space, while $y_i$ and $y_j$ are data points in the low-dimensional space. The variable $\sigma$ represents the variance of the Gaussian distribution in the high-dimensional space, and it is usually chosen using a binary search for each data point. The pairwise probabilities $p_{ij}$ and $q_{ij}$ represent the likelihood of data points $i$ and $j$ being similar in the high-dimensional and low-dimensional spaces, respectively.

By minimizing the KL divergence (cost function $C$), t-SNE ensures that the relationships between data points are preserved when projecting the data onto a lower-dimensional space. This makes t-SNE an effective method for visualizing complex datasets and uncovering hidden structures within the data.

There are two primary steps in the t-SNE algorithm. First, t-SNE builds a probability distribution across pairs of high-dimensional objects, assigning a greater probability to comparable features and a lower likelihood of dissimilar points. Second, t-SNE minimizes the Kullback-Leibler divergence (KL divergence) Joyce [2011] between the two distributions concerning the positions of the points in the map by defining a comparable probability distribution over the points in the low-dimensional map.

### 2.2.5 Multiclass Problem

Multiclass classification problems can be approached in two ways. The first option involves using a classification method that can handle all classes simultaneously. The second option decomposes the original problem into binary subproblems, which offers two main benefits: (a) it enables the use of inherently binary classifiers, such as SVM, for multiclass problems, and (b) simplifies decision-making by breaking down classification functions. In certain cases, decomposition can improve accuracy [Fürnkranz, 2001].

Two types of binary decomposition exist for multiclass problems: One-vs-One (OvO) and One-vs-All (OvA). Decomposition is also referred to as Round Robin [Fürnkranz, 2001]. As described in Section 2.2.2, the SVM method can be adapted for multiclass problems using the OvO and OvA approaches.

The One-vs-All (OvA) approach begins by dividing all samples into two sets: one containing the target class (+1") and another with samples from all other classes (−1"). A model, denoted as $f(\cdot)$, is then trained and applied to estimate the test group labels. If the classifier favors class 1, the corresponding sample gains a vote. Otherwise, all samples in class 1 receive a vote.

In the next round, the procedure is repeated, but the samples from the previous round's class 1 are now incorporated into class 1, and a distinct class is selected to compose class 1. The model $f(\cdot)$ is trained and evaluated again, with votes assigned to all classes following the previous rule. This process continues until all classes have been designated as class 1.

The final decision is determined by applying a majority vote. While the number of iterations for this decomposition method increases linearly with the number of classes, each evaluated decision function is generally simpler than in the multiclass case.

In our Multi-layer approach, we applied a modified OvA model. In the first round, samples are classified between the target and remaining classes. In the second round, the procedure is repeated, but samples that were previously classified as class 1 are discarded, and a different class is chosen to compose class 1. This process repeats until all classes have been evaluated in one class. The remaining samples not-classified are redistributed between the layers based on the max probability achieved.

This adaptation was developed to address the computational complexity of the OvA model. While the primary disadvantage of the OvO method is its exponential time complexity $\mathcal{O}(n^2)$, our method exhibits a linear time complexity $\mathcal{O}(n)$, where the total number of rounds corresponds to the number of classes.

## 2.3 Evaluation Metrics

To assess the performance of the classification model, various evaluation metrics are employed, including accuracy, precision, recall, F1 score, and AUC. These metrics provide insights into the model's ability to correctly classify instances and balance between precision and recall.

### 2.3.1 ROC-AUC Curve

The Receiver Operating Characteristic (ROC) curve is a graphical tool used to evaluate the performance of binary classification models [Hanley et al., 1989]. It is plotted by comparing the True Positive Rate (TPR) (also called sensitivity) against the False Positive Rate (FPR) at various threshold settings. In this context:

- True Positive Rate (TPR): The proportion of actual positive instances that are correctly identified by the model (also known as sensitivity or recall).
- False Positive Rate (FPR): The proportion of negative instances that are incorrectly classified as positive.

Figure 2.3 illustrates a typical ROC curve with different classifiers. In the plot, the x-axis represents the False Positive Rate (FPR), ranging from 0 to 1, and the y-axis represents the True Positive Rate (TPR), ranging from 0 to 1. And here a brief explanation of each curve:

- **Perfect Classifier (blue curve)**: The point at the top-left corner (1, 0) represents a perfect classifier, which achieves a TPR of 1 and an FPR of 0. This indicates that the model perfectly distinguishes between positive and negative instances. In the plot, the blue curve reaches this point, demonstrating ideal performance.
- **Better Classifiers (green and blue curves)**: the green curve represents a well-performing classifier, where the True Positive Rate increases rapidly with a low False Positive Rate. The closer the curve is to the top-left corner, the better the model's performance. Hence, the blue and green curves represent classifiers that achieve good results.
- **Random Classifier (red dashed line)**: The red dashed line indicates the performance of a random classifier. This classifier performs no better than chance and has an AUC (Area Under the Curve) of 0.5. The diagonal line (from (0, 0) to (1, 1)) represents a model that randomly guesses the class labels.

- **Worse Classifiers (orange line)**: Any curve falling below the red dashed line, like the orange curve, represents a classifier performing worse than random. This might occur if the classifier is systematically misclassifying instances (e.g., incorrectly predicting the opposite class).

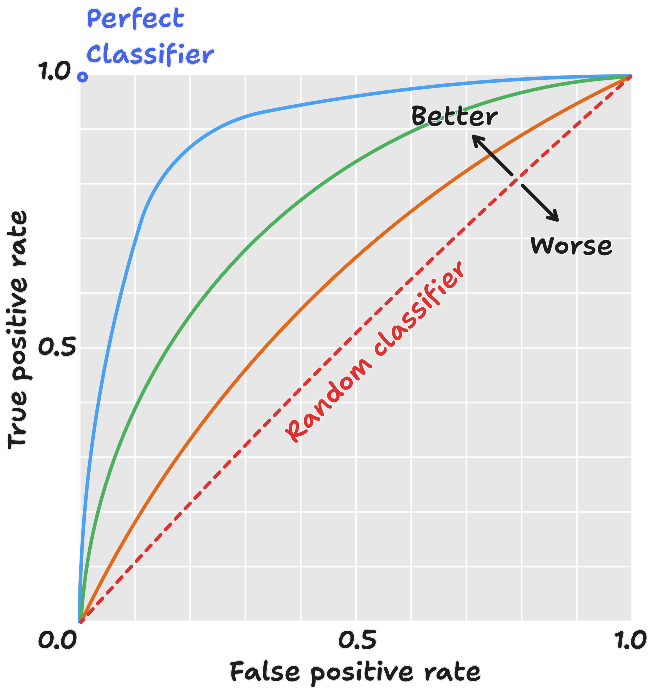


**Figure 2.3.** ROC curve explanation, figure adapted from Fawcett [2006].

The AUC (Area Under the Curve) is a single scalar value that summarizes the overall performance of the classifier. An AUC of 1 indicates a perfect classifier. An AUC of 0.5 indicates a classifier that performs no better than random guessing. An AUC below 0.5 suggests that a classifier performs worse than random.

The ROC-AUC curve is particularly useful when comparing multiple classifiers, as it summarizes performance across all possible thresholds, providing a comprehensive evaluation of a model's discriminative ability [Fawcett, 2006]. In domains such as medical diagnostics, where it is critical to minimize both false positives and false negatives, this metric helps in selecting the most appropriate model for deployment.

## 2.3.2 Classification Metrics

Evaluating the performance of classification models is essential in determining their effectiveness, especially when dealing with high-dimensional datasets like gene expression data. Several common metrics are used to assess the performance of classification models, each offering a different perspective on the model's success in correctly classifying instances. Below is a detailed description of the key metrics used in both binary and multiclass classification problems.

- **Accuracy** (Binary and Multiclass): Accuracy measures the proportion of correctly classified instances out of the total number of instances [Naidu et al., 2023]. It is a general metric that provides a broad view of model performance but may not be suitable when the classes are imbalanced. For example, if 90% of the data belongs to one class, a model that predicts the majority class most of the time will achieve a high accuracy but may fail to recognize the minority class.

  $$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.11}$$

  In multiclass settings, accuracy measures the percentage of correctly classified samples across all classes.

- **Precision** (Binary and Multiclass): Precision, also known as the positive predictive value, measures the proportion of true positives out of all instances predicted as positive [Naidu et al., 2023]. This metric is especially useful when the cost of false positives is high. For instance, in medical applications like cancer detection, incorrectly predicting a disease-free patient as having the disease (false positive) can lead to unnecessary anxiety and treatments.

  $$Precision = \frac{TP}{TP + FP} \tag{2.12}$$

  In multiclass problems, precision can be calculated for each class (one-vs-all approach), or an average precision (macro or weighted) can be taken across all classes [Sokolova and Lapalme, 2009].

- **Recall** (Sensitivity, Binary and Multiclass): Recall, or sensitivity, measures the proportion of true positive instances that were correctly identified [Naidu et al., 2023]. It is particularly important in situations where missing actual positives (false negatives) is costly, such as in diagnosing a patient who has cancer. A low recall would mean that many positive cases were not detected.

  $$Recall = \frac{TP}{TP + FN} \tag{2.13}$$

  Like precision, recall in multiclass settings can be calculated for each class individually, or averaged across classes using macro, micro, or weighted averages.

- **F1 Score** (Binary and Multiclass): The F1 score is the harmonic mean of precision and recall, providing a single metric that balances the trade-off between the two. It is particularly useful in scenarios where precision and recall are both critical, and one cannot be sacrificed for the other. The F1 score is especially helpful for imbalanced datasets, where accuracy alone can be misleading.

  $$\textit{F1 Score} = 2 \times \frac{\textit{Precision} \times \textit{Recall}}{\textit{Precision} + \textit{Recall}} \tag{2.14}$$

  In multiclass classification, the F1 score can be computed for each class and then averaged (macro or weighted) across all classes [Opitz and Burst, 2019].

- **AUC (Area Under the Curve)** (Binary, Multiclass via extension): The AUC (Area Under the Curve) measures the overall ability of a binary classifier to distinguish between the positive and negative classes. It is derived from the ROC (Receiver Operating Characteristic) curve, which plots the true positive rate (recall) against the false positive rate at various threshold settings. An AUC of 1 indicates perfect classification, while 0.5 indicates random guessing. A higher AUC reflects better performance.

  For multiclass classification, AUC can be calculated using strategies such as one-vs-one or one-vs-rest for each class, and then averaging the results [Hand and Till, 2001].

In these formulas, $TP$, $TN$, $FP$, and $FN$ represent true positives, true negatives, false positives, and false negatives, respectively. These metrics offer a comprehensive evaluation of the model's performance and allow for selecting the best models based on the overall metrics achieved.

Accuracy is applicable to both binary and multiclass problems. However, it can be misleading in imbalanced datasets, making other metrics like F1 score or AUC more relevant. Precision and recall can be extended to multiclass classification by calculating them for each class independently or using averaging methods (e.g., macro-precision, micro-precision). F1 score is ideal for problems with imbalanced classes, providing a balance between precision and recall. AUC is primarily used for binary classification but can be adapted to multiclass problems using strategies like OvA.

## 2.4 Final Remarks

In this chapter, we presented the fundamental concepts used in developing this work, including concepts related to biology and computing; we presented concepts of gene expression, gene sets, and machine learning fundamentals. We introduced the concept of multiclass problem decomposition and metrics to measure classification quality. In the next chapter, related works will be presented using 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) PAM50 Gene Signature, (ii) Gene Selection in Expression Data and (iii) Breast Cancer Subtype Classification methods.

## 3.1 Our Baseline: PAM50 Gene Signature

The PAM50 gene signature has become the cornerstone for classifying breast cancer into its intrinsic subtypes (Luminal A, Luminal B, HER2-enriched, and Basal-like). Its widespread adoption in clinical and research settings is a testament to its value in prognosticating outcomes and informing treatment decisions [Liu et al., 2016, Ochoa et al., 2020]. However, despite its utility, several limitations have emerged that suggest the need for alternative or supplementary gene sets, particularly from a classification and machine learning perspective.

One key limitation is the use of unsupervised statistical methods in deriving the PAM50 signature. While this approach has successfully identified genes that correlate with breast cancer subtypes, it does not necessarily prioritize the genes that are most effective for machine learning classification models [Qian et al., 2021]. In other words, the gene set may not be fully optimized for maximizing the accuracy or robustness of subtype predictions, particularly in the context of high-dimensional data, such as gene expression datasets, where the number of features (genes) far exceeds the number of samples. This imbalance can lead to suboptimal classifier performance or even overfitting, where noise is mistaken for meaningful patterns [Huang et al., 2018].

Additionally, the fixed gene set size of PAM50 may not be ideal for all types of classification problems. While PAM50 uses 50 genes, there is no guarantee that this is the optimal number of features for every model or dataset. In many cases, classification performance could be improved by using more flexible gene sets—either larger or smaller, depending on the specific characteristics of the dataset and model used [Ochoa et al., 2020]. Methods such as feature selection and recursive feature elimination allow us to dynamically select the most informative genes based on the data at hand, potentially improving classification accuracy [Liu et al., 2016]. This approach provides the flexibility to tailor the gene set to each specific machine learning task, which is a significant advantage over static gene signatures like PAM50.

Another limitation is related to the evolving understanding of multiclass classification. While PAM50 has shown effectiveness in binary and certain multiclass classification tasks (e.g., distinguishing Luminal A from Luminal B), its performance may degrade when applied to more nuanced classifications or when subtypes with less distinct gene expression patterns are involved. Given the complexity of breast cancer subtypes, alternative gene sets that are specifically tailored to maximize classification performance across different subtypes could be more effective than relying solely on PAM50 [Qian et al., 2021]. Expanding beyond PAM50 provides an opportunity to explore gene sets that are more adaptable and focused on improving classification metrics such as accuracy, precision, recall, and F1 score.

In summary, while the PAM50 gene signature remains an essential tool for breast cancer subtype classification, there is a clear need for more flexible and adaptable gene sets. By focusing on improving classification performance through dynamic feature selection and model-specific optimization, we can overcome the limitations inherent in static gene sets like PAM50. This approach will ensure more accurate and robust classification, enhancing the ability to guide treatment strategies and predict patient outcomes [Liu et al., 2016, Ochoa et al., 2020, Qian et al., 2021].

## 3.2 Gene Selection in Expression Data

Feature selection plays a crucial role in reducing the dimensionality of gene expression datasets, improving classification accuracy, and minimizing computational cost. Gene expression data typically involve thousands of genes but only a limited number of samples, making feature selection essential to avoid overfitting and to enhance model performance. Various methods have been proposed to address the challenges of feature selection in high-dimensional gene expression data.

Liu et al. [2018] introduced a gene selection method that combines double radial basis function (RBF) kernels with weighted analysis to extract relevant genes from gene expression data. This method reduces redundancy and irrelevance in the gene sets, which is critical for improving classification accuracy in both two-class and multiclass phenotypes. The authors applied their method to four benchmark datasets and demonstrated its superiority in terms of accuracy, true positive rate, false positive rate, and computational efficiency. The combination of RBF kernels allows for greater flexibility in learning from heterogeneous data, making it highly applicable to complex gene expression profiles [Liu et al., 2018].

Another widely used approach is the hybrid method proposed by Yang et al. [2008], which integrates correlation-based feature selection (CFS) with binary particle swarm optimization (BPSO). This method identifies a subset of genes that are strongly correlated with the target class but minimally correlated with each other. When tested on six cancer-related gene expression datasets, the approach improved classification accuracy while reducing the number of selected genes and computational resources. The authors used K-nearest neighbors (KNN) as the classifier and demonstrated that the hybrid method outperformed other conventional feature selection techniques in terms of classification accuracy and computational cost [Yang et al., 2008].

Additionally, hybrid approaches that combine filter and wrapper methods have gained attention for their ability to leverage the strengths of both strategies. For example, Chandrakar et al. [2021] proposed a hybrid feature selection method that uses a modified genetic algorithm to identify the most informative gene subsets in the lung cancer data. This approach uses ensemble learning techniques to ensure robust classification performance, demonstrating that combining filter-based gene selection with wrapper-based optimization can significantly enhance the classification accuracy in gene expression data.

Finally, in the work titled “A Gene Selection Method Based on Outliers for Breast Cancer Subtype Classification”. Mendonca-Neto et al. [2021] proposed an innovative gene selection method using outlier detection to improve the classification of breast cancer subtypes. This method focuses on identifying outlier genes that play a crucial role in subtype differentiation. By applying this approach, the authors improved the classification accuracy and robustness of breast cancer subtype models, which made it particularly effective in handling heterogeneous data, such as gene expression profiles. The method was validated using several datasets, demonstrating improved performance compared to traditional methods.

In summary, these feature selection methods highlight the importance of balancing gene redundancy reduction, computational efficiency, and classification accuracy. As gene expression datasets continue to grow in size and complexity, the use of advanced methods such as kernel-based approaches, hybrid methods, and weighted feature selection techniques is critical for extracting the most informative genes and improving model performance.

## 3.3 Breast Cancer Subtype Classification

Classifying breast cancer into its molecular subtypes—Basal-like, HER2-enriched, Luminal A, and Luminal B—is essential for personalized treatment and prognosis. Several methods and models have been proposed to address this problem, ranging from traditional machine learning approaches to advanced deep learning frameworks, each with varying degrees of success depending on the dataset and classification approach used.

Lee et al. [2020] employed a deep learning model that combines an attention mechanism and network propagation to classify cancer using a pathway-based feature selection approach. The study utilized five TCGA cancer datasets, including breast invasive carcinoma (BRCA), and achieved an average classification accuracy of 66.91% for BRCA. The authors selected 5,515 genes for their classification task. Although their pathway-based approach captures key biological features, the relatively low accuracy highlights the challenge of using large gene sets to distinguish between breast cancer subtypes, which often show subtle gene expression differences.

Similarly, Graudenzi et al. [2017] proposed a Support Vector Machine (SVM)-based framework for breast cancer subtype classification. They developed a feature selection method based on pathway activity, identifying 400 genes enriched in four breast cancer subtypes. Their classifier achieved an overall accuracy of 85.00%, demonstrating that SVM combined with biologically informed feature selection can yield effective results in classifying subtypes. This work highlights the potential of pathway-based methods for improving the interpretability of gene expression data while maintaining high classification performance.

Mostavi et al. [2020] introduced a 1D-Convolutional Neural Network (1D-CNN) model for predicting breast cancer subtypes. Their feature selection process relied on basic statistical techniques, including mean and standard deviation, to select 7,091 genes. The model achieved an average accuracy of 88.42% across the five subtypes. Despite using relatively simple feature selection methods, the 1D-CNN was effective at capturing local dependencies in gene expression data, underscoring the strength of CNNs in processing structured biological data.

Recent studies have focused on integrating multi-omics data for breast cancer subtype classification. Choi and Chae [2023] developed the moBRCA-net, a deep learning framework that incorporates multiple omics layers, such as gene expression, CpG methylation, and microRNA data, to enhance subtype classification. Their framework employs a self-attention mechanism to identify the most relevant features from these diverse data sources. This multi-omics approach improved accuracy in identifying Basal-like, HER2-enriched, and Luminal subtypes, demonstrating the power of integrating multi-layered biological data to address the complexity of breast cancer classification.

Similarly, Tafavvoghi et al. [2024] explored a deep learning model applied to H&E-stained whole-slide images for molecular subtype prediction. Their approach, which combines One-vs-Rest (OvR) classification with XGBoost, achieved a macro F1 score of 0.73 in distinguishing between subtypes, including the challenging Basal-like and Luminal B classes. This demonstrates the potential for integrating histopathological images with machine learning to complement traditional gene expression-based classification.

The methods discussed demonstrate a range of approaches to breast cancer subtype classification, from classical SVM models with feature selection to modern deep learning techniques using multi-omics data and imaging. Traditional machine learning methods, such as SVM with pathway-based selection, continue to perform well, particularly for subtypes like Luminal A and Luminal B. However, deep learning frameworks face challenges due to the curse of dimensionality, where the high number of features (genes) and limited samples can lead to overfitting and reduced model generalization. As more datasets are created and become available, researchers will be better positioned to explore deep learning approaches, such as CNNs and attention mechanisms, while effectively addressing the dimensionality issues through robust feature selection and integration of additional data types, particularly for harder-to-classify subtypes like Basal-like and HER2-enriched.

## 3.4 Discussion

In this section, we have reviewed several studies that contribute to the understanding of breast cancer subtypes classification using various techniques, such as gene selection, pathway activity-based feature selection, and hierarchical classification. These studies have demonstrated the potential of these methods to improve the accuracy and performance of breast cancer classification.

The PAM50 gene signature, which has been widely adopted in breast cancer prognosis and treatment, has been shown to have certain limitations. Studies have pointed out that using unsupervised statistical methods in deriving the PAM50 signature may result in the inclusion of genes that lack biological relevance, leading to the exclusion of important genes that could potentially improve the performance and clinical utility of the PAM50 gene signature [Huang et al., 2018, Ochoa et al., 2020, Qian et al., 2021].

In conclusion, the reviewed studies showcase a range of promising techniques and methodologies that could further improve the accuracy and performance of breast cancer subtype classification. Future research should focus on addressing the limitations of existing methods for gene selection, such as the PAM50 gene signature, and exploring the potential of combining other gene selections and classification techniques to achieve better results.

## 3.5 Final Remarks

In this chapter, we presented a brief review of related works that address the problem of cancer classification. We highlight the problems of the current strategy using PAM50 Gene Signature and would like to investigate this opportunity to develop methods with better performance than the state-of-the-art.

| **Reference** | **Problem** | **Data Source** | **Methods** | **Results** |
|---|---|---|---|---|
| Liu et al. [2016], Ochoa et al. [2020] | Limitations of PAM50 Gene Signature for breast cancer subtype classification | Gene Expression | Unsupervised statistical methods | Limitations identified, alternative gene sets suggested |
| Liu et al. [2018] | Gene selection in expression data using RBF kernels | Gene Expression | Double RBF kernels with weighted analysis | Improved accuracy, true positive rate, false positive rate, and computational efficiency |
| Yang et al. [2008] | Hybrid feature selection using CFS and BPSO | Gene Expression | Correlation-based feature selection (CFS) and binary particle swarm optimization (BPSO) | Increased classification accuracy with fewer genes |
| Mendonca-Neto et al. [2021] | Gene selection method based on outliers for breast cancer subtype classification | Gene Expression | Outlier detection | Improved classification accuracy and robustness for breast cancer subtypes |
| Lee et al. [2020] | Breast cancer subtype classification using deep learning with pathway-based selection | Gene Expression | Deep learning model with attention mechanism and network propagation | 66.91% accuracy for breast cancer subtype classification |
| Graudenzi et al. [2017] | SVM-based breast cancer subtype classification with pathway activity feature selection | Gene Expression | Support Vector Machine (SVM) with pathway-based feature selection | 85.00% accuracy for breast cancer subtype classification |
| Mostavi et al. [2020] | 1D-CNN for breast cancer subtype classification | Gene Expression | 1D-Convolutional Neural Network (CNN) | 88.42% accuracy for breast cancer subtype classification |
| Choi and Chae [2023] | Multi-omics integration for breast cancer subtype classification using deep learning | Gene Expression, CpG Methylation, microRNA | Deep learning framework with self-attention mechanism | Improved accuracy in identifying Basal-like, HER2-enriched, and Luminal subtypes |
| Tafavvoghi et al. [2024] | Breast cancer subtype prediction from histopathology images using OvR and XGBoost | H&E Stained Whole-Slide Images | One-vs-Rest classification with XGBoost | Macro F1 score of 0.73 for breast cancer subtype classification |

**Table 3.1.** Summary of reviewed works, problems, data sources, methods, and results in breast cancer subtype classification.

# 4

# Few-Shot Genes Selection

This chapter introduces a proposed approach, termed “Fewer-Shot Gene Selection”, aimed at identifying more effective gene combinations for breast cancer subtype classification.

In recent years, the PAM50 gene signature, a set of 50 genes used to classify breast cancer into four intrinsic molecular subtypes, has become a valuable prognostic tool in breast cancer research. It provides critical insights into tumor subtypes and informs therapeutic decision-making Bastien et al. [2012]. However, as mentioned in Chapter 3, the PAM50 approach has certain limitations, including its reliance on a relatively large number of genes, which can increase both costs and complexity in research and clinical settings Huang et al. [2018]. This underscores the need for a more efficient method to classify breast cancer subtypes using a smaller, yet effective set of genes.

This chapter builds on the work presented in the publication *Few-shot genes selection: Subset of PAM50 genes for the classification of breast cancer subtypes*, published in BMC Bioinformatics, which investigates the potential of achieving a precise classification of breast cancer subtypes using a reduced set of genes derived from the PAM50 signature [Okimoto et al., 2024].

The primary objective of this research is to explore the potential for precise breast cancer subtype classification using a reduced gene set derived from the PAM50 gene signature. Using a method known as “Few-Shot Gene Selection”, we randomly select smaller subsets from the PAM50 set and evaluate their performance using the F1 Score and a linear model, specifically the Support Vector Machine (SVM) classifier. This approach aims to determine whether a more compact gene set can maintain classification accuracy while reducing complexity.

The main contributions of this chapter are:

1. An experimental evaluation of the effect of using fewer genes during the gene selection phase;
2. A direct comparison between the “Few-Shot Gene Selection” method and the PAM50 gene selection approach;
3. A 2D visualization using t-distributed stochastic neighbor embedding (t-SNE), providing insights into how the model interprets the data, particularly in cases of misclassification at subtype boundaries.

Through this chapter, we aim to demonstrate that selecting fewer but more representative genes can yield a more accurate classification of breast cancer subtypes, addressing the limitations inherent in the PAM50 approach.

## 4.1 Method Description

In this study, we introduce a novel approach termed “Fewer-Shot Genes Selection,” aimed at refining breast cancer subtype classification. This method strategically generates multiple gene subsets derived from the well-established PAM50 signature, subsequently evaluating their effectiveness in classification. A key motivation for this approach is the significant practical implications of reducing the number of genes required for accurate classification. In clinical settings, a condensed gene set simplifies the diagnostic process, reducing the burden of identifying a wide array of genes, thereby improving the efficiency and potentially the accuracy of classification. By rigorously assessing the performance of these gene subsets, our goal is to identify combinations that either match or exceed the classification accuracy of the PAM50 signature, ultimately enhancing the breast cancer subtype classification process. The comprehensive pipeline of our proposed method is illustrated in Figure 4.1.

The method is composed of the following key steps:

1. The dataset is normalized using the Standard Scaler.
2. The dataset is filtered based on the PAM50 gene signature.
3. The filtered dataset is divided into training, validation, and test sets to prevent overfitting during the gene selection process.
4. The “Few-Shot Genes Selection” method is applied to the training, validation, and test sets, with the goal of identifying optimal gene subsets based on the size of the gene signature.

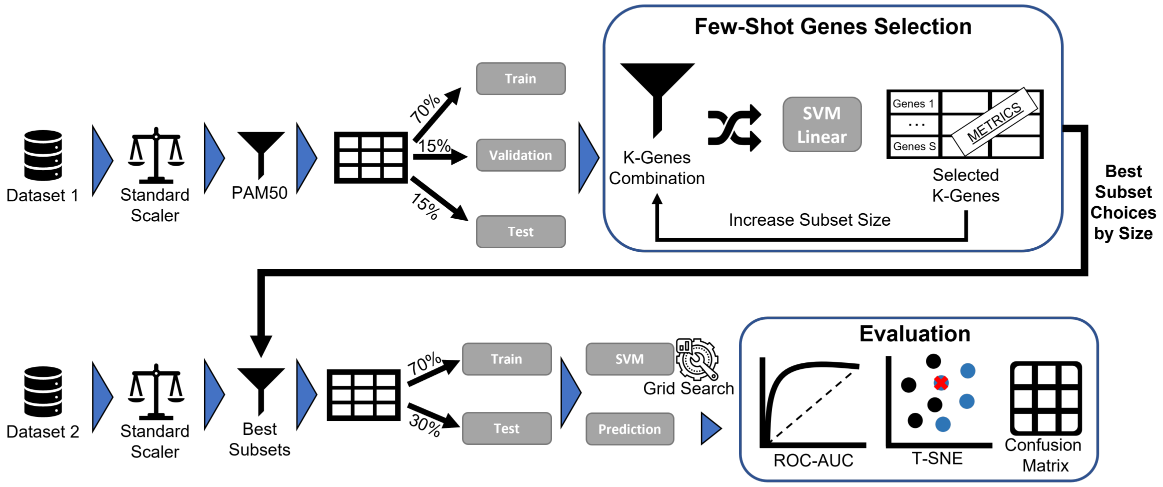


**Figure 4.1.** Overview of the proposed approach.

5. The optimal gene subsets are then evaluated in a second pipeline, where a second dataset is normalized using the Standard Scaler.

6. Each optimal gene subset is used to filter the second dataset, producing filtered subsets of varying sizes.

7. Each filtered subset is evaluated using a Support Vector Machine (SVM) model with a Grid Search for hyperparameter optimization. The second dataset is split only into training and test sets, as no validation portion is required, ensuring no data leakage from the gene selection phase in the first dataset.

8. Test predictions are evaluated using t-Distributed Stochastic Neighbor Embedding (t-SNE), a confusion matrix, and the Receiver Operating Characteristic (ROC) curve, with the Area Under the Curve (AUC) as the performance metric.

This methodology, which splits Dataset 1 into *train*, *validation*, and *test* sets to ensure unbiased gene subset selection, and Dataset 2 into only *train* and *test* sets for evaluation, allows for the identification of optimal gene subsets and their evaluation without introducing bias from the datasets.

### 4.1.1 Gene Selection

In this study, we employed a “Few-shot gene selection” approach, where gene subsets ranging from 10% to 80% of the total signature length were evaluated. This strategy was chosen to optimize the trade-off between model accuracy and interpretability, while also reducing computational complexity. Specifically, for the PAM50 gene signature, this approach involved evaluating subsets of sizes ranging from 5 to 40 genes. A Support

Vector Machine (SVM) with a linear kernel was used for this analysis, with the model trained on 70% of the data, validated on 15%, and tested on the remaining 15%.

To rigorously evaluate gene subsets within the “Few-shot gene selection” framework, we conducted an extensive set of experiments, performing 1 million trials for each combination of subset size and experimental setup. Given the vast number of potential combinations, it was infeasible to assess every possibility exhaustively. Consequently, we organized our evaluation into three separate experiments, each encompassing 1 million distinct combinations, resulting in a total of 3 million combinations per subset size. For each subset size (ranging from 5 to 40 genes), the Support Vector Machine (SVM) model generated an $F_1$ score based on the selected subset during the Few-shot gene selection phase. To ensure that the selection of optimal subsets was not due to random chance, we calculated the p-value by comparing the distribution of $F_1$ scores across the three experiments. This statistical comparison was carried out for each subset size individually, from size 5 to 40, yielding a p-value for every subset size. In all cases where the p-values are greater than 0.05, we confirm that we can use the best combination for that subset size.

This extensive experimentation enabled the exploration of a large search space, allowing us to identify the optimal gene subsets that maximized the combined $F_1$ scores from both the validation and test data for the ACES and TCGA datasets independently. Following this validation process, the results from the three experiments for each subset size were aggregated, and the gene sets corresponding to the highest combined $F_1$ scores on the validation and test datasets were selected. These optimal subsets were then cross-evaluated, whereby the best gene sets for one dataset were evaluated on the other, ensuring that each subset size achieved robust results without overfitting to any single dataset.

## 4.2 Data and Methods

In this section, we describe the datasets used in our experiments, along with the preprocessing steps and the machine learning algorithms employed. We also outline the evaluation metrics applied in this study.

### 4.2.1 Experiment Datasets

To accurately classify breast cancer subtypes without overfitting, large datasets are essential [Ein-Dor et al., 2006]. In this study, we utilized samples from 12 ACES studies (n = 1606) [Staiger et al., 2013] and the TCGA breast invasive cancer dataset (n

= 532) [Chin et al., 2012], resulting in a total of 2138 samples from 13 independent investigations. The METABRIC dataset [Curtis et al., 2012], which requires ethical approval, was excluded. Consequently, this study represents one of the most comprehensive compilations of publicly available gene expression data for breast cancer subtypes.

Our dataset captures a substantial proportion of the biological heterogeneity observed among breast cancer patients, as well as technical biases resulting from variations in platforms and study-specific sample preparation methods [Allott et al., 2016]. This diversity aids in training models capable of generalization, which is critical for ensuring the applicability of the final classification model in real-world scenarios [Kourou et al., 2015, Ransohoff, 2005, Stretch et al., 2013]. To focus on the relevant molecular subtypes, inclusion criteria were defined to concentrate on the following subtypes: **Basal**, **HER2**, **Luminal B**, and **Luminal A**. Samples classified as Normal-like or those without defined subtypes were excluded. This refinement reduced the dataset to 2027 samples, with **1512 samples from ACES** and **515 from TCGA**.

The primary objective of this experiment research is to support post-surgical treatment decisions for patients diagnosed with breast cancer by classifying tumors into specific subtypes using small gene signatures. This classification is crucial for determining the most suitable treatment modalities. Normal samples were excluded, as our study focuses on treatment strategies for tumor subtypes, rather than early detection or the transformation of normal tissue into cancer. Additional details on the clinical variables are provided in the Supplementary Files.

### 4.2.2 Experiment Settings and Computation Time

Our experiments were conducted on a custom-built server equipped with a 12th Gen Intel(R) Core(TM) i9-12900KF processor, featuring 16 cores. The system also included 64 GB of RAM and a 1 TB NVMe SSD for storage. All computations were performed in an environment running Ubuntu 20.04.6 LTS.

Regarding computation time, the experiment utilizing the TCGA dataset as a filter required approximately 48 hours to complete, while the experiment using the ACES dataset took around 52 hours to generate the list of optimal gene subsets.

We used the Python programming language and the scikit-learn[1] package for the machine learning algorithms employed in all experiments presented in this manuscript.

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

### 4.2.3 Data Preprocessing

#### PAM50 Gene Selection

For gene filtering, we applied the PAM50 Gene Selection (Subsection 2.1.4), which serves as the foundation for our analysis. This study explores the potential for accurate breast cancer subtype classification using a reduced gene set derived from the PAM50 signature.

#### Standard Scaler

To ensure consistency across the 13 breast cancer studies in our dataset, we normalized the data using the Standard Scaler (Subsection 2.2.1). This method standardizes each feature by removing the mean and scaling to unit variance, ensuring that the data is adjusted for machine learning applications, with each feature treated independently across all samples.

### 4.2.4 SVM

We decided to utilize traditional machine learning instead of employing deep learning. While deep learning methods have shown significant advancements in the cancer domain [Alanni et al., 2019, Gao et al., 2019, Lyu and Haque, 2018], the characteristics of our dataset—limited sample size and high-dimensional gene expression features—make deep learning prone to overfitting and high-variance gradient updates [Liu et al., 2017a].

The choice of using SVM - Suport Vector Machines (Subsection 2.2.2) was informed by previous studies, where the SVM method achieved the best results for representative genes [Mendonca-Neto et al., 2022].

### 4.2.5 Evaluation

We evaluated the performance of the "Few-Shot Genes Selection" approach by comparing it against the well-established PAM50 gene signature. To comprehensively assess the performance of our model, several key metrics were computed: accuracy, precision, recall, F1 score, and AUC. These metrics provide a holistic view of the model's effectiveness in classifying the gene expression data.

- **Accuracy**: The proportion of correctly classified instances out of the total number of instances.

- **Precision**: The proportion of true positive instances among the instances predicted as positive.
- **Recall** (Sensitivity): The proportion of true positive instances among the actual positive instances.
- **F1 Score**: The harmonic mean of precision and recall.
- **AUC**: The Area Under the Curve (AUC) of the ROC curve, a scalar value that measures the classifier's overall performance.

In addition to the quantitative metrics, the following visualization and comparative techniques were employed:

1. **ROC-AUC Curve**: The Receiver Operating Characteristic (ROC) curve was used to calculate the AUC for PAM50 and our top 3 subset sizes. Despite the multiclass nature of our model, we applied the One-vs-Rest approach for each class to compute the ROC-AUC scores.
2. **Confusion Matrix**: Confusion matrices were generated to evaluate the classification accuracy of both the PAM50 signature and the selected gene subsets, helping identify discrepancies between the two approaches.
3. **T-SNE**: t-Distributed Stochastic Neighbor Embedding (t-SNE) was utilized to visualize the high-dimensional gene expression data in a lower-dimensional space, allowing for a comparative analysis of the clustering patterns produced by the PAM50 signature and our method.

These metrics and visualization techniques facilitated a detailed comparison of the "Few-Shot Genes Selection" approach with the PAM50 gene signature, highlighting the strengths and weaknesses of the selected gene subsets.

## 4.3 Results

In this section, we examine the findings of our cross-validation approach. This technique involves determining the optimal subset for one dataset and subsequently applying it to the model on the second dataset, then reversing the process to select from the second dataset and applying it to the first dataset. Our aim is to capture less complex attributes in the Selection Phase using a linear model while employing a Grid Search to train non-linear SVM models in the Evaluation Phase, the parameters applied in the SVM models are shown in Table 4.1.

In the Data Preprocessing step we had to do a gene intersection between datasets ACES and TCGA, in order to select one dataset and ensure that we had the genes in the other dataset. When we applied the PAM50 gene in the intersection, we found that the ACES dataset did not have 6 genes presented in the PAM50 Gene Signature: *ANLN, CXXC5, GPR160, NUF2, TMEM45B, UBE2T*. This does not discourage us from using the PAM50 gene signature because the list still had 44 important genes to use as features in our models.

**Table 4.1.** SVM parameters used in our experiments. Parameters missing were set as default values in scikit-learn version 1.4.0.

| Parameter | Value |
|---|---|
| C | 1 |
| kernel | linear |
| gamma | scale |
| decision function shape | one-vs-rest |

### 4.3.1 TCGA to ACES

Initially, we utilized the TCGA dataset with 515 samples as the filtering dataset. After passing this dataset to the first pipeline with the "Few-Shot Genes Selection" process, we retained the best subsets for each subset length then we applied the ACES dataset for the evaluation process.

#### ROC-AUC

In the ROC curve, we highlighted the top 3 subsets which achieved AUC (Area Under Curve) equal to or even higher than the baseline PAM50, highlighted in red in Figure 4.2, Subset size 36 (S-36), Subset size 35 (S-35) and Subset size 34 (S-34).

As we can see, there are selections that can compete with the PAM50 Gene Signature baseline, the gray lines in the plot also show us other selections which contain subsets with fewer genes.

#### Metrics

The results in Table 4.2 present the performance of four highlighted filtering methods (PAM50, S-36, S-35, and S-34) across the four subtypes (Basal, Her2, LumA, and

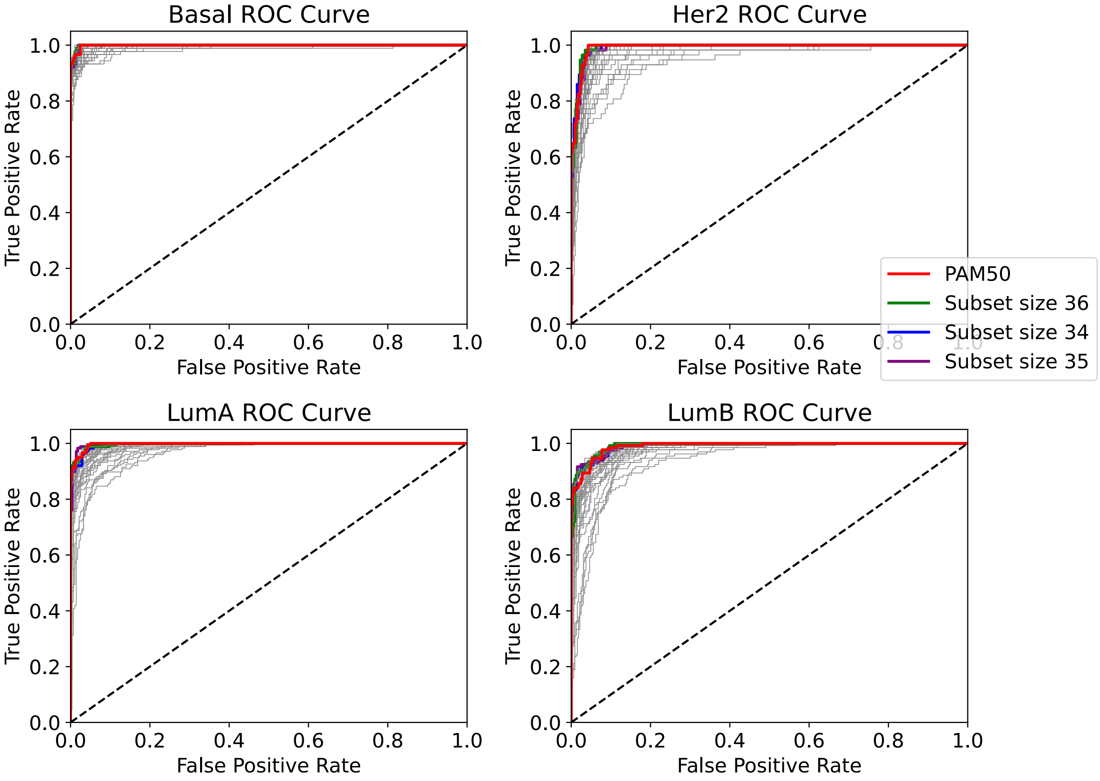


**Figure 4.2.** ROC curve for the evaluation of the prediction samples from ACES in a model with features filtered by TCGA.

| | | | Basal | | | | | Her2 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | Prec | Rec | F1 | AUC | Acc | Prec | Rec | F1 | AUC |
| PAM50 | 0.930 | **0.966** | 0.944 | 0.955 | **0.999** | 0.930 | 0.845 | **0.860** | **0.852** | **0.992** |
| Subset 36 | 0.927 | 0.946 | **0.978** | **0.961** | **0.999** | 0.927 | 0.855 | 0.825 | 0.839 | **0.992** |
| Subset 35 | **0.943** | 0.935 | **0.978** | 0.956 | **0.999** | **0.943** | 0.852 | 0.807 | 0.829 | 0.991 |
| Subset 34 | 0.927 | 0.955 | 0.955 | 0.955 | **0.999** | 0.927 | **0.882** | 0.789 | 0.833 | **0.992** |

| | | | LumA | | | | | LumB | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | Prec | Rec | F1 | AUC | Acc | Prec | Rec | F1 | AUC |
| PAM50 | 0.930 | 0.955 | 0.966 | 0.960 | 0.997 | 0.930 | 0.908 | 0.902 | 0.905 | 0.990 |
| Subset 36 | 0.927 | 0.954 | 0.949 | 0.952 | 0.997 | 0.927 | 0.909 | 0.909 | 0.909 | **0.992** |
| Subset 35 | **0.943** | **0.967** | **0.989** | **0.978** | **0.998** | **0.943** | **0.953** | 0.917 | **0.934** | 0.991 |
| Subset 34 | 0.927 | 0.944 | 0.960 | 0.952 | 0.996 | 0.927 | 0.904 | **0.924** | 0.914 | 0.991 |

**Table 4.2.** Classification results for the four subtypes TCGA to ACES.

LumB), reporting Accuracy (Acc), Precision (Prec), Recall (Rec), F1 Score (F1), and AUC for each case.

In the **Basal** subtype, S-35 attained the highest accuracy (0.943), while S-36 reached the best F1 Score (0.961). PAM50, however, secured the highest precision (0.966) and AUC (0.999), with both S-36 and S-35 achieving the best recall (0.978). For **Her2**, S-35 displayed the highest accuracy (0.943) but lower precision (0.852) and F1 Score (0.829) compared to PAM50 and S-36. Conversely, PAM50 and S-34 tied for the highest AUC (0.992), and PAM50 demonstrated the best recall (0.860), with S-34 leading in precision (0.882).

In **LumA**, S-35 outperformed other methods with the highest accuracy (0.943), precision (0.967), recall (0.989), F1 Score (0.978), and AUC (0.998). For **LumB**, S-35 achieved the highest accuracy (0.943) and F1 Score (0.934), and S-36 obtained the highest AUC (0.992). Additionally, S-35 had the highest precision (0.953), and S-34 presented the best recall (0.924).

Overall, S-35 consistently performed well across all subtypes, obtaining the highest accuracy and F1 Score in most situations. Nevertheless, PAM50 exhibited better precision for Basal and the highest metrics for Her2. Comparing the results, S-35 can be considered the best-performing method in most cases, with S-36 and PAM50 remaining competitive in some aspects.

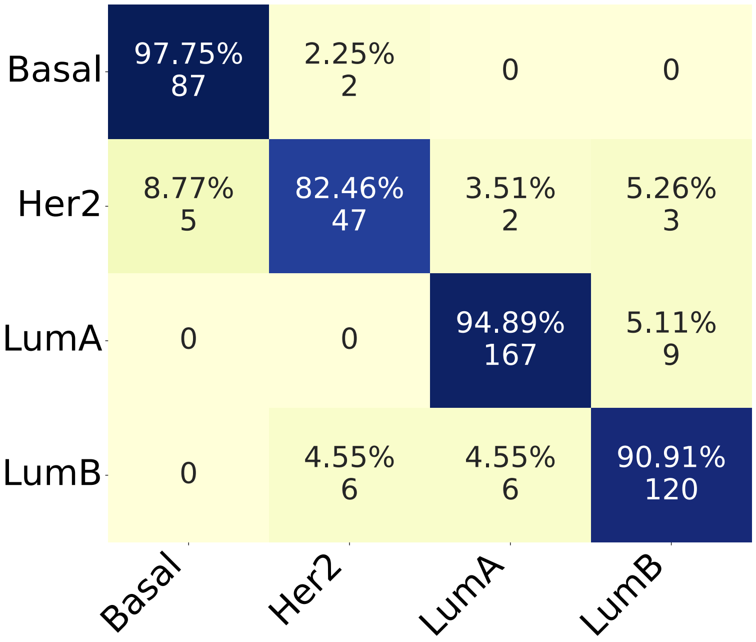
### Confusion Matrix and t-SNE

In this section, we delve deeper into the comparison between the subset size of 36, which achieved the highest mean AUC for the four subtypes, and the established PAM50 gene signature. By analyzing the confusion matrices and t-SNE plots, we aim to better understand the strengths and limitations of our gene subset selection method.

**Confusion Matrix Analysis** The Confusion Matrix is a crucial tool for evaluating the classification performance of our model. It shows how well the model predicted the correct subtypes and where it made errors. In Figure 4.3, we present the Confusion Matrix for the subset size of 36 genes, derived from the TCGA dataset and applied to the ACES cohort. This subset achieved a strong overall performance, particularly excelling in the Her2 subtype.

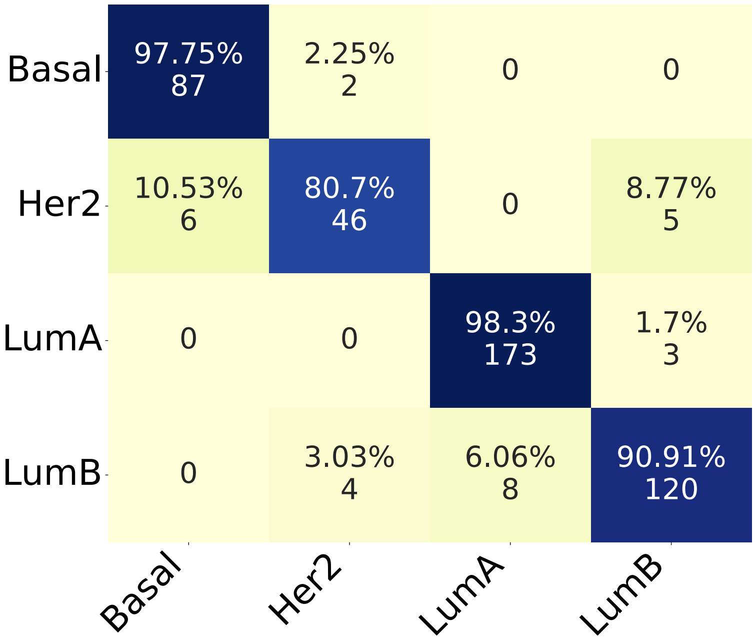
Comparing both Confusion Matrices (Figure 4.3 and Figure 4.4), it becomes evident that while both models performed similarly in predicting the Basal and LumB subtypes, there are key differences in the results for the LumA and Her2 subtypes. Specifically, the PAM50 signature (Figure 4.4) showed better accuracy in identifying LumA subtypes, whereas our method performed better in the Her2 subtype classification. These distinctions highlight the potential of our subset selection approach in certain subtypes, even when compared to the widely used PAM50 gene signature.

These results reflect the importance of examining the confusion matrix to understand which subtypes are better predicted by each model. For example, while PAM50 is highly effective for LumA, our method may offer a more nuanced advantage in predicting Her2, where traditional methods may fall short.



| | Basal | Her2 | LumA | LumB |
|---|---|---|---|---|
| Basal | 97.75%<br>87 | 2.25%<br>2 | 0 | 0 |
| Her2 | 8.77%<br>5 | 82.46%<br>47 | 3.51%<br>2 | 5.26%<br>3 |
| LumA | 0 | 0 | 94.89%<br>167 | 5.11%<br>9 |
| LumB | 0 | 4.55%<br>6 | 4.55%<br>6 | 90.91%<br>120 |

**Figure 4.3.** Confusion Matrix for subset size 36, filtering from TCGA and applying on ACES.



| | Basal | Her2 | LumA | LumB |
|---|---|---|---|---|
| Basal | 97.75%<br>87 | 2.25%<br>2 | 0 | 0 |
| Her2 | 10.53%<br>6 | 80.7%<br>46 | 0 | 8.77%<br>5 |
| LumA | 0 | 0 | 98.3%<br>173 | 1.7%<br>3 |
| LumB | 0 | 3.03%<br>4 | 6.06%<br>8 | 90.91%<br>120 |

**Figure 4.4.** Confusion Matrix for PAM50 applying on ACES.

**t-SNE Visualization Analysis** t-SNE plots offer a different perspective on the data by reducing its dimensionality and providing a visual representation of how well the samples are grouped by subtype. Figures 4.5 and 4.6 present the t-SNE visualizations for both our subset of 36 genes and the PAM50 signature, respectively.

When examining the t-SNE plot for the subset size 36 (Figure 4.5), we can observe distinct separations between most subtypes, which aligns with the performance observed in the confusion matrix. This visualization allows us to explore how the samples are grouped, offering insight into the areas where the model misclassifies or struggles to draw clear subtype boundaries.

The PAM50 t-SNE plot (Figure 4.6) similarly shows a clear separation between the subtypes. However, it's essential to note that t-SNE is a dimensionality reduction technique, and while it provides an intuitive understanding of the subtype clusters, some overlap or boundary errors are expected due to the complexity of projecting high-dimensional gene expression data into a two-dimensional space.

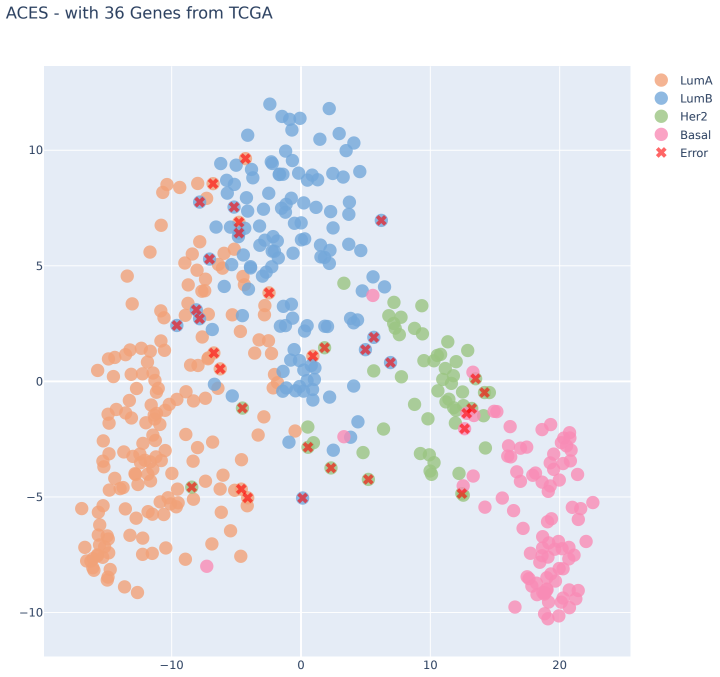


**Figure 4.5.** t-SNE Visualization for subset size 36, filtering from TCGA and applying on ACES.

While neither t-SNE plot presents a perfect separation between subtypes, we can still identify areas where the models struggle, especially around subtype boundaries. The errors visible in the confusion matrices are also reflected here, where some samples appear close to the borders of different subtype clusters, indicating that these samples may have been misclassified.

Ultimately, these visualizations and matrices provide complementary insights: the confusion matrix quantifies the classification performance, while the t-SNE visualizations offer a more intuitive understanding of how well the model differentiates

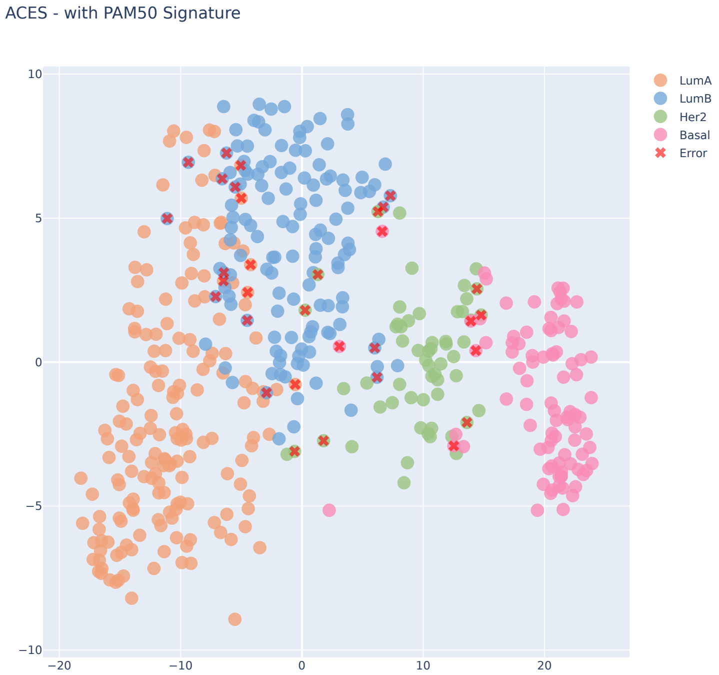


**Figure 4.6.** t-SNE Visualization in the ACES Prediction data after PAM50 filtering.

between subtypes. By analyzing both, we can better understand where our method excels and where improvements may be needed.

### 4.3.2 ACES to TCGA

After doing the TCGA filtering and ACES dataset evaluation, we cross-evaluated, using the ACES dataset with 1512 samples as the filtering dataset. After choosing the best subsets for each subset length, we applied the TCGA dataset for the evaluation process. Our goal by doing so is to have a full understanding of how fewer genes impact the model's performance when we compare it against the PAM50 Signature.

#### ROC-AUC

In the ROC curve, we highlighted the top 3 subsets which achieved AUC (Area Under Curve) equal to or even higher than the baseline PAM50, highlighted in red in Figure 4.7, Subset size 37 (S-37), Subset size 36 (S-36) and Subset size 32 (S-32).

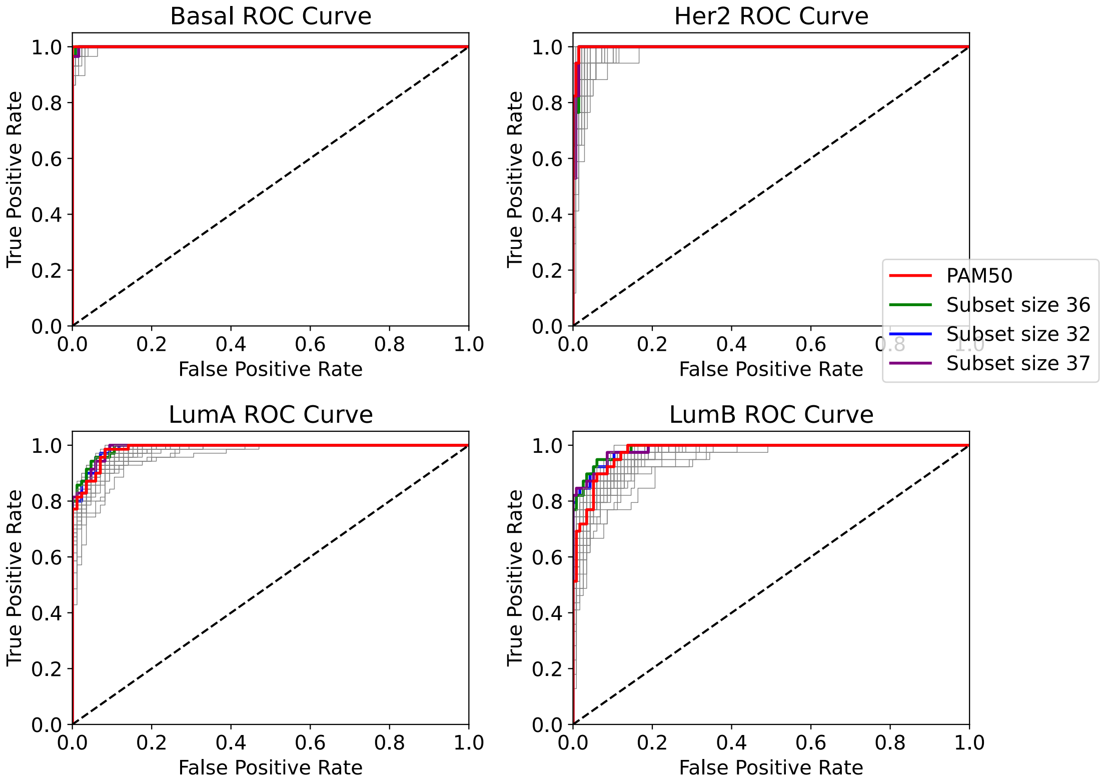


**Figure 4.7.** ROC curve for the evaluation of the prediction samples from TCGA in a model with features filtered by ACES.

As we can see, the results achieved by the subsets can compete with the PAM50 baseline. In this particular ROC-AUC visualization, where we have less data in the prediction with the TCGA dataset, we can have a better view of the subsets that achieved greater curves than the baseline.

### Metrics

Table 4.3 presents the performance of four filtering methods (PAM50, S-37, S-36, and S-32) across four subtypes (Basal, Her2, LumA, and LumB), reporting Accuracy (Acc), Precision (Prec), Recall (Rec), F1 Score (F1), and AUC for each case.

| | Basal | | | | | Her2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | Prec | Rec | F1 | AUC | Acc | Prec | Rec | F1 | AUC |
| PAM50 | 0.910 | **1.000** | **0.931** | **0.964** | **1.000** | 0.910 | 0.850 | **1.000** | 0.919 | 0.999 |
| Subset 37 | 0.910 | **1.000** | 0.897 | 0.945 | 0.999 | 0.910 | 0.842 | 0.941 | 0.889 | 0.996 |
| Subset 36 | **0.935** | **1.000** | **0.931** | **0.964** | **1.000** | **0.935** | 0.850 | **1.000** | 0.919 | 0.996 |
| Subset 32 | 0.923 | **1.000** | **0.931** | **0.964** | 0.999 | 0.923 | **0.895** | **1.000** | **0.944** | **0.997** |

| | LumA | | | | | LumB | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | Prec | Rec | F1 | AUC | Acc | Prec | Rec | F1 | AUC |
| PAM50 | 0.910 | 0.914 | 0.914 | 0.914 | 0.988 | 0.910 | 0.868 | 0.846 | 0.857 | 0.978 |
| Subset 37 | 0.910 | 0.915 | 0.929 | 0.922 | 0.990 | 0.910 | 0.872 | **0.872** | 0.872 | 0.985 |
| Subset 36 | **0.935** | **0.944** | **0.957** | **0.950** | **0.992** | **0.935** | **0.919** | **0.872** | **0.895** | **0.988** |
| Subset 32 | 0.923 | 0.929 | 0.929 | 0.929 | 0.989 | 0.923 | 0.872 | **0.872** | 0.872 | 0.986 |

**Table 4.3.** Classification results for the four subtypes ACES to TCGA.

In the Basal subtype, S-36 achieved the highest accuracy (0.935), while PAM50, S-36, and S-32 shared the best precision (1.000), recall (0.931), F1 Score (0.964), and AUC (1.000). For Her2, S-36 displayed the highest accuracy (0.935) and shared the best recall (1.000) with PAM50 and S-32. S-32 attained the highest precision (0.895) and F1 Score (0.944), as well as the highest AUC (0.997).

In LumA, S-36 outperformed other methods with the highest accuracy (0.935), precision (0.944), recall (0.957), F1 Score (0.950), and AUC (0.992). For LumB, S-36 secured the highest accuracy (0.935), precision (0.919), F1 Score (0.895), and AUC (0.988), while sharing the best recall (0.872) with S-37 and S-32.

Overall, S-36 consistently performed well across all subtypes, achieving the highest accuracy, F1 Score, and AUC in most cases, and sharing the best results in precision and recall with other methods. S-32 also demonstrated competitive performance in several aspects, particularly for the Her2 subtype.

### Confusion Matrix and t-SNE

In this analysis, we switch datasets to evaluate the performance of our model when filtering from the ACES cohort and applying the results to the TCGA dataset. Interestingly, as in the previous evaluation, the subset size of 36 once again achieved the highest mean AUC for the four subtypes, reinforcing its robustness across datasets.

**Confusion Matrix Analysis** The Confusion Matrix allows us to break down the classification performance of our model by subtype, helping us visualize where the model excels and where it encounters challenges. Figure 4.8 presents the Confusion Matrix for the subset size of 36 genes, this time filtered from ACES and applied on the TCGA dataset.

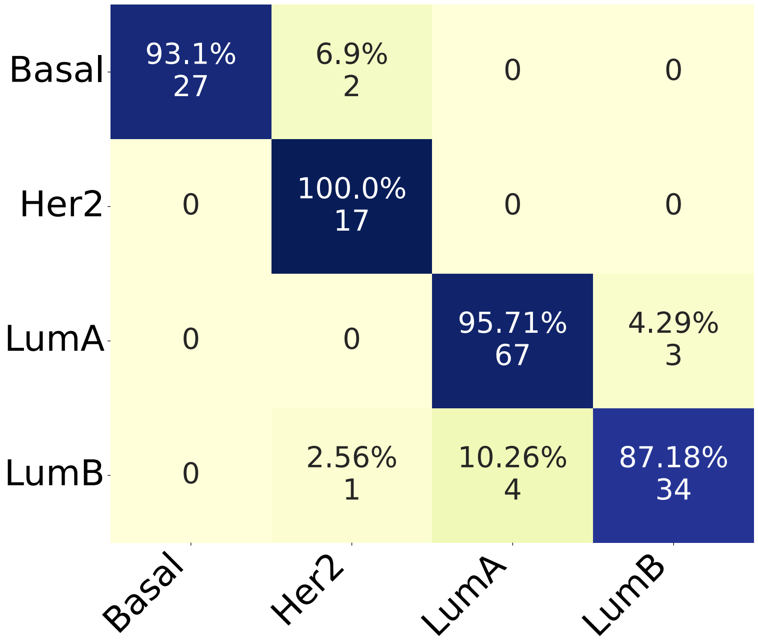
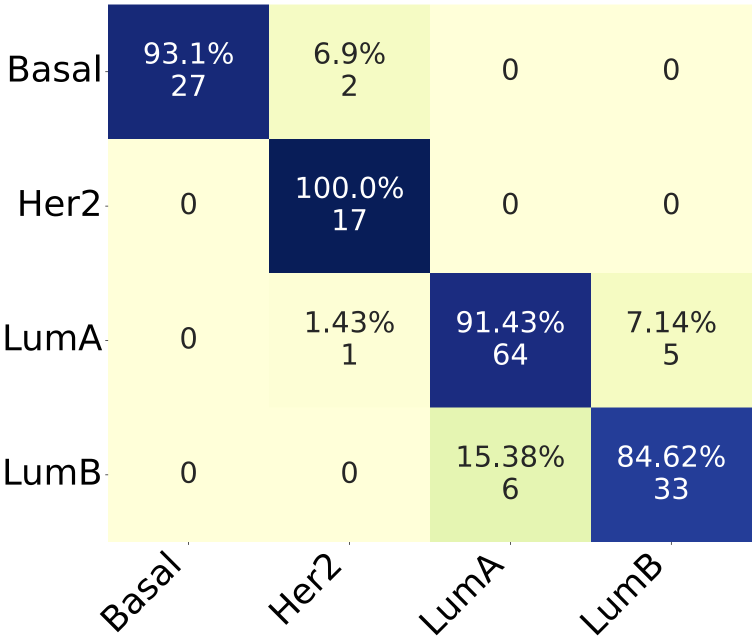


**Figure 4.8.** Confusion Matrix for subset size 36, filtering from ACES and applying on TCGA.

Comparing the results between our method (Figure 4.8) and the PAM50 gene signature (Figure 4.9), we can observe some significant differences. Both methods performed similarly well in classifying the Basal and Her2 subtypes. However, our subset size of 36 genes showed a clear advantage in classifying the LumA and LumB subtypes, outperforming PAM50 in these categories. This suggests that the gene subset selection approach not only transfers well between datasets but also offers improved classification for specific subtypes, such as LumA and LumB, where PAM50 may be less effective.

This comparison highlights the flexibility of our subset selection approach across different datasets and underscores its potential in improving the classification accuracy of LumA and LumB subtypes, which may be more challenging for traditional methods like PAM50. The detailed breakdown provided by the confusion matrix enables us to pinpoint these improvements, showing that the subset size of 36 offers significant advantages in these specific subtypes.



|  | Basal | Her2 | LumA | LumB |
|---|---|---|---|---|
| Basal | 93.1%<br>27 | 6.9%<br>2 | 0 | 0 |
| Her2 | 0 | 100.0%<br>17 | 0 | 0 |
| LumA | 0 | 1.43%<br>1 | 91.43%<br>64 | 7.14%<br>5 |
| LumB | 0 | 0 | 15.38%<br>6 | 84.62%<br>33 |

**Figure 4.9.** Confusion Matrix for PAM50 applying on TCGA.

**t-SNE Visualization Analysis** To complement the insights from the confusion matrices, we turn to the t-SNE visualizations, which provide a two-dimensional projection of the high-dimensional gene expression data. Figures 4.10 and 4.11 present the t-SNE visualizations for both the subset size 36 and the PAM50 signature, applied to the TCGA dataset.

In both visualizations, we can observe a clear separation between subtypes, indicating that both the subset size 36 and the PAM50 signature successfully differentiate between most of the breast cancer subtypes. However, a noteworthy point in this comparison is that the inversion of the y-axis between the two plots is due to the non-linear transformation inherent to the t-SNE algorithm, rather than any biological or feature-related discrepancy.

Despite the similar separations in the t-SNE plots, when cross-referenced with the confusion matrices, we can identify subtle differences in the classification errors, particularly around the boundaries between subtypes. These boundary errors are visible in the t-SNE plots, where some samples appear clustered near the edges of different subtypes, suggesting potential misclassifications. This aligns with the errors observed in the confusion matrices, offering a more intuitive view of where the model struggles in drawing clear distinctions between subtypes.

Although the t-SNE plots offer similar overall separations between subtypes, the detailed analysis provided by combining the t-SNE visualizations with the confusion matrices helps clarify how well the models perform across the datasets. The compar-

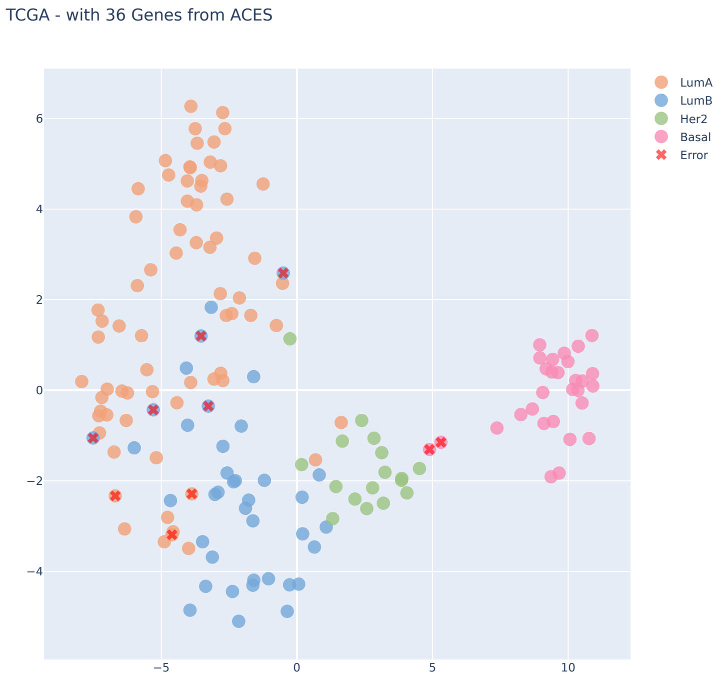


**Figure 4.10.** t-SNE Visualization for subset size 36, filtering from ACES and applying on TCGA.

ison highlights where each model succeeds and where it struggles, particularly at the subtype boundaries. The visual representation of these errors provides an intuitive understanding of the limitations of both methods and emphasizes the potential for further refinement in gene subset selection.

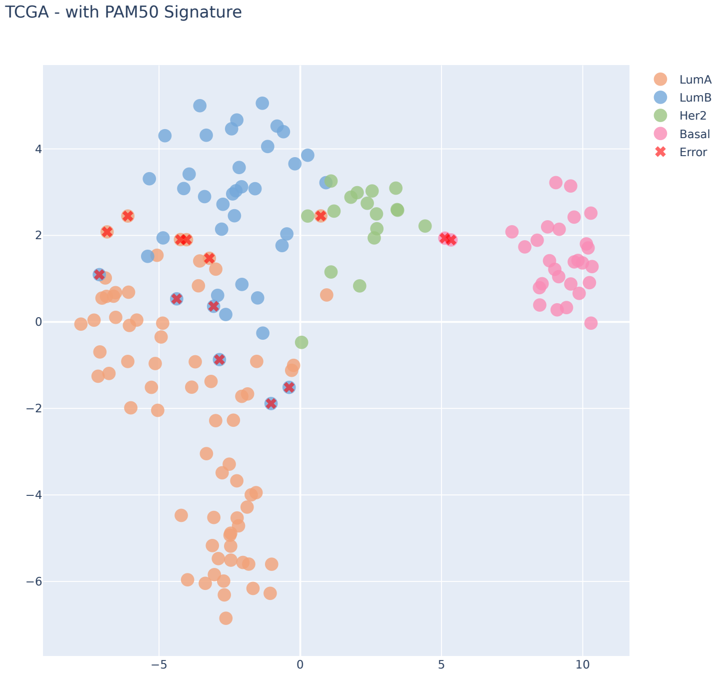


**Figure 4.11.** t-SNE Visualization in the ACES Prediction data after PAM50 filtering.

### 4.3.3 Discussion

In our study, we observed that the list of genes from the S-36 filtering method varied between the two datasets, with 30 genes being shared across both experiments, as shown in Table 4.4. While it may not be surprising that these 30 genes overlap, given that the S-36 gene set is derived from the PAM50 Signature, it is noteworthy that our method achieves comparable or even superior results. This demonstrates that the S-36 filtering method offers other viable options for gene sets, which can be valuable for clinicians when researching breast cancer subtypes. The shared genes likely play a significant role in cancer subtype classification, underscoring the robustness of the S-36 method, which retains its effectiveness across different datasets despite some differences in the gene lists.

**Table 4.4.** S-36 Signatures for ACES and TCGA datasets experiments.

| Gene Set | Genes |
|---|---|
| S-36 ACES Dataset | MAPT, BLVRA, CENPF, GRB7, BCL2, FGFR4,<br>CDC6, CCNB1, FOXC1, KRT17, SFRP1, PTTG1,<br>FOXA1, NDC80, KIF2C, MELK, MMP11, CEP55,<br>SLC39A6, KRT14, MIA, NAT1, EGFR, EXO1,<br>BIRC5, ACTR3B, CDH3, ESR1, MLPH, MDM2<br>**CDC20, MYBL2, PGR, MYC, PHGDH, RRM2** |
| S-36 TCGA Dataset | MAPT, BLVRA, CENPF, GRB7, BCL2, FGFR4,<br>CDC6, CCNB1, FOXC1, KRT17, SFRP1, PTTG1,<br>FOXA1, NDC80, KIF2C, MELK, MMP11, CEP55,<br>SLC39A6, KRT14, MIA, NAT1, EGFR, EXO1,<br>BIRC5, ACTR3B, CDH3, ESR1, MLPH, MDM2<br>**UBE2C, CCNE1, MKI67, TYMS, KRT5, BAG1** |

The fact that S-36 achieves comparable or improved performance compared to the PAM50 Signature across several evaluation metrics suggests that some genes in the PAM50 set may not be as critical for subtype classification as previously thought. This opens up the possibility of refining or even replacing certain components of the PAM50 Signature with other gene sets, such as S-36, that may offer better accuracy and flexibility in clinical research. Providing clinicians with multiple gene set options allows for more tailored approaches to breast cancer subtype classification, depending on the specific characteristics of a given patient population or dataset.

Further analysis of the molecular functions of the genes, using tools such as the Panther Classification System [Mi et al., 2019], revealed that while the S-36 Signature from TCGA shares a similar distribution of molecular functions with PAM50, the S-36 Signature from ACES lacked the "Structural Molecular Activity" provided by the

KRT5 gene, which is unique to the PAM50 set. This indicates that certain biological functions may be underrepresented in specific gene sets, and careful consideration is required when selecting genes for clinical use. Details of the molecular functions for these genes can be found in the Supplementary Files.

These findings suggest potential areas for enhancing the PAM50 Signature through improved gene selection methods. By incorporating genes that demonstrate strong classification performance across multiple datasets, such as those identified in the S-36 Signature, we can create more precise and efficient classifiers for breast cancer subtypes. The common genes between the S-36 sets derived from both ACES and TCGA datasets should be of particular interest for future research, as they may represent key drivers of cancer subtype classification and offer pathways for developing more concise, effective gene signatures.

In summary, our study highlights the potential of the S-36 filtering method as a promising alternative to the PAM50 Signature. While PAM50 remains a widely used tool, the ability of S-36 to match or exceed its performance suggests that alternative gene sets may provide additional benefits in breast cancer classification. By offering clinicians new gene set options and refining the gene selection process, we can move towards more accurate and individualized classification systems for cancer subtypes, ultimately contributing to better patient care and treatment outcomes.

## 4.4 Final Remarks

In this chapter, we developed an approach called “Fewer-Shot Genes Selection”, which utilized a large number of subsets derived from the PAM50 Signature to compare the classification performance against this established baseline. The primary goal of our approach was not to highlight the individual importance of features during classification, but rather to offer alternative gene subsets that could achieve comparable or better results.

Our approach yielded the S-36 gene subset, which surpassed the performance of the baseline PAM50 Signature. This suggests that there is significant potential for improvement in feature selection techniques for breast cancer subtype classification. By demonstrating that even subsets derived from PAM50 can outperform the original signature, we provide additional options for clinicians and researchers to explore when classifying breast cancer subtypes. The availability of such alternative gene sets opens the door for more tailored diagnostic tools that may be more suitable for specific datasets or patient populations.

The results presented in this chapter encourage us to extend our approach further. Instead of limiting ourselves to subsets derived solely from the PAM50 Signature, future work could expand into the entire feature space, exploring new ways to enhance classification accuracy. Given the limitations of the PAM50 Signature discussed in previous studies, our findings prove that even with derived subsets, we can achieve results that rival or surpass the original PAM50, which is an encouraging direction for further research.

Additionally, we aim to incorporate explainable artificial intelligence (XAI) into the evaluation process, particularly focusing on misclassified samples at the boundaries between subtypes. Understanding how certain features, or groups of features, contribute to classification errors could provide valuable insights for refining future models. XAI techniques would allow us to explain these errors by analyzing the expression of specific features, offering a deeper understanding of how certain genes influence classification performance. This approach could prove crucial in identifying patterns that lead to misclassification, especially in cases where samples fall within ambiguous subtype boundaries.

In conclusion, our focus was not merely on the importance of individual features but on providing alternative gene subsets that can offer high classification accuracy. By extending this work, we can explore the entire feature space and integrate explainable AI techniques, ultimately contributing to more precise and clinically relevant gene signatures for breast cancer subtype classification.

5

# Using Multi-layer Classification to Improve Worst Prognosis Breast Cancer Subtype Outcomes

This chapter introduces a novel approach titled "Using Multi-layer Classification to Improve Worst Prognosis Breast Cancer Subtype Outcomes." This approach evaluates class orders in a multi-layer classification framework, which forms part of a submitted paper from our ongoing research.

As illustrated in Figure 1.1 (Chapter 1), accurately classifying breast cancer subtypes is crucial to avoid misclassifying subtypes with poorer prognoses as those with better prognoses, and vice versa. Such errors could lead to suboptimal treatment strategies for patients [Dai et al., 2015]. Ensuring precise classification is essential to tailoring appropriate treatment plans and improving patient outcomes.

In recent years, the application of machine learning techniques to classify breast cancer subtypes has gained considerable momentum [Graudenzi et al., 2017, Mendonca-Neto et al., 2021, Mostavi et al., 2020, Murtaza et al., 2020]. Among these methods, multi-layer classification approaches offer several advantages over traditional models, including improved evaluation of class relationships, more accurate identification of misclassified data, and more effective isolation of individual classes [Silla and Freitas, 2011].

Building on this foundation, this Chapter introduces a multi-layer support vector machine (SVM) classification method, with a particular emphasis on optimizing feature selection through Recursive Feature Elimination (RFE). This optimization aims to enhance classification accuracy, especially for breast cancer subtypes associated with the worst prognoses.

Our contributions include:

- Developing a pipeline for multi-layer classification models that reduces gene sets for subtype-specific classification.
- Providing an interpretation of t-distributed stochastic neighbor embedding (t-SNE) for visualizing predicted samples, which allows a 2D representation that highlights misclassified samples near subtype boundaries.
- Comparing our results with the widely-used PAM50 Gene Signature, demonstrating our method's ability to offer insights into the misclassification of certain samples.

## 5.1 Proposed Approach

In Figure 5.1, we present our proposed multi-layer classification approach. One advantage of this approach is that we can prioritize specific subtypes by selecting the order in which they are classified. For instance, if we aim to maximize the results in the order: **Basal**, **Her2**, **Luminal B**, and **Luminal A**—representing the subtypes from worst to best prognosis—we can sort the generated models using this subtype order, ensuring that the classification metrics are optimized for subtypes with the worst prognoses.

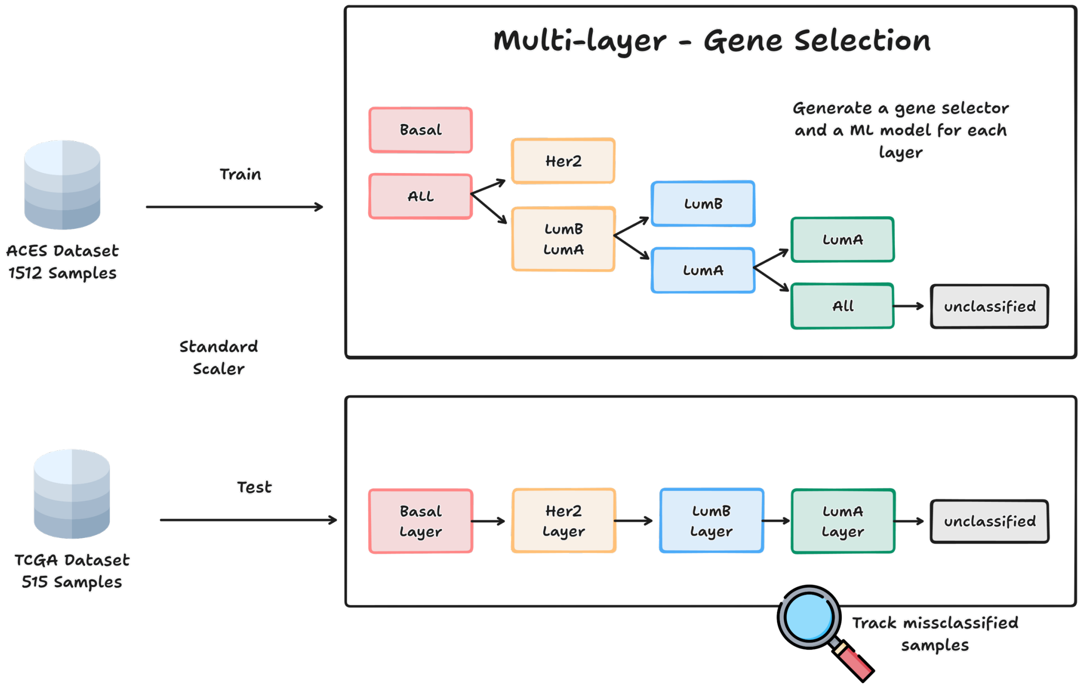


**Figure 5.1.** Approach for the Multi-Layer Gene Selection and Model Optimization.

We combine data preprocessing, feature selection, and machine learning techniques to achieve optimal results. Below is a detailed explanation of each step in our approach:

1. We begin with the ACES and TCGA datasets, which are normalized and split into training and testing sets. The larger dataset is used for training, and the smaller dataset is reserved for testing.
2. The training dataset is then input into the multi-layer process, where the subtypes are sorted in the desired order. Gene selection and model fitting are performed at each layer for the corresponding subtype.
3. For each layer, we generate a list of selected genes relevant to the specific subtype and train a model for its classification.
4. The testing dataset is then passed through the trained multi-layer process for prediction.
5. Finally, the multi-layer classification is evaluated using the testing dataset.

During the prediction phase, as the model iterates through each subtype, the predictions are progressively updated for samples classified into the current subtype, while tracking any misclassified samples. After each layer, the data is updated to retain only the samples that have not yet been classified into any subtype.

As the final output of our approach, we obtain the set of selected genes for each subtype, enabling further studies of subtype-specific gene relationships. Additionally, we identify and plot the misclassified samples and record the IDs of samples that remain unclassified.

This multi-layer process provides valuable insights into how the data is perceived during classification for each subtype, generating separate gene and classification lists for each class. Importantly, our approach maintains a high level of flexibility, allowing the selection of different gene selectors and models for each layer, ensuring adaptability to various datasets and objectives.

## 5.2 Material and Methods

In this section, we describe the datasets used in our experiments as well as the preprocessing, feature selection, machine learning methods, and evaluation metrics employed in our study.

### 5.2.1 Experiment Datasets

To accurately classify breast cancer subtypes without overfitting, we utilized samples from 12 ACES studies [Staiger et al., 2013] and the TCGA breast invasive cancer dataset [Chin et al., 2012], totaling 2138 samples. We excluded the METABRIC dataset [Curtis et al., 2012], which requires specific ethical approval. As a result, our study represents the largest publicly available gene expression compilation for breast cancer subtypes.

The datasets exhibit significant biological heterogeneity and technical biases due to varying platforms and sample preparations [Allott et al., 2016], contributing to robust model generalization for real-world applications [Kourou et al., 2015, Ransohoff, 2005, Stretch et al., 2013]. We focused on subtypes Basal, HER2, Luminal B, and Luminal A, excluding Normal-like and undefined subtypes, which resulted in 2027 samples (ACES: 1512, TCGA: 515), as shown in Table 5.1.

The study emphasizes the post-surgical treatment classification of breast cancer subtypes based on small gene signatures, aiming to guide treatment decisions. Normal samples were excluded since the focus is on treatment response rather than early detection or tumor transformation.

**Table 5.1.** ACES and TCGA Dataset.

| Subtype | ACES Dataset | TCGA Dataset |
|---|---|---|
| Luminal A | 584 | 232 |
| Luminal B | 440 | 129 |
| Basal | 297 | 96 |
| Her2 | 191 | 58 |
| Total | 1512 | 515 |

### 5.2.2 Standard Scaler

We applied the Standard Scaler to normalize the data, ensuring that each feature had a mean of 0 and a standard deviation of 1. This standardization is particularly useful when working with datasets compiled from multiple studies, as it mitigates the effect of different scales and units, allowing the machine learning algorithms to perform more effectively.

The Standard Scaler ensures unbiased treatment of features and prevents issues caused by differences in measurement scales across datasets, making it a more suitable approach than techniques such as quantile normalization or RMA, which can introduce biases unsuitable for machine learning models sensitive to input data distributions.

### 5.2.3 Feature Selection

We first applied the PAM50 gene signature as a baseline method for classifying breast cancer subtypes. To further enhance model performance, we implemented Recursive Feature Elimination (RFE) with cross-validation, which iteratively removes less significant features and selects the most relevant genes for each subtype. RFE is a widely used technique for optimizing feature selection in gene expression studies [Guyon et al., 2002, Mendonca-Neto et al., 2021].

### 5.2.4 Machine Learning Model

We employed support vector machines (SVM) with radial basis function (RBF) kernels as our primary classification algorithm. SVMs, particularly with RBF kernels, are well-suited for handling high-dimensional data such as gene expression profiles, as they efficiently manage both linear and non-linear separations in the data [Mendonca-Neto et al., 2022].

### 5.2.5 Metrics and Optimization

Our models were evaluated using standard performance metrics: classification accuracy, precision, recall, and F1 score. These metrics provided a comprehensive evaluation of model performance, ensuring that we selected the best models based on an overall balance of these criteria. In addition, confusion matrices were generated for each model to analyze the misclassification patterns in greater detail.

### 5.2.6 Visualization: T-SNE

To better understand the relationships between different breast cancer subtypes and to interpret the results of our classification models, we employed t-distributed stochastic neighbor embedding (t-SNE) for visualization. This technique reduces the high-dimensional gene expression data to a lower-dimensional space while preserving the local structure of the data [Maaten and Hinton, 2008]. The t-SNE visualizations provide insights into how well the model differentiates between subtypes and are commonly used in cancer research for understanding gene expression patterns [Abdelmoula et al., 2016, Allahyar et al., 2019, Jia et al., 2018].

## 5.3 Results

In this section, we present the results achieved by our proposed approach. We highlight the selected genes for each subtype, the classification outcomes, and a visualization of the results using t-SNE.

### 5.3.1 Selected Genes for Each Subtype

Our study aimed to identify a reduced set of genes for accurately classifying breast cancer subtypes. We applied a feature selection method for each subtype in our multi-layer approach, reducing the gene set while maintaining classification performance. Table 5.2 shows the selected genes for each subtype.

**Table 5.2.** Recursive Feature Elimination Genes for Each Subtype.

| Subtype | Genes | Length |
|---|---|---|
| Basal | FOXA1, MLPH, UBE2C, NAT1, RRM2, PHGDH, FOXC1, SFRP1, ACTR3B, MIA, KRT17, KRT14, GRB7, FGFR4, ORC6 | 15 |
| Her2 | FOXA1, ESR1, NDC80, KIF2C, SLC39A6, MAPT, NAT1, RRM2, BCL2, CENPF, FOXC1, CDH3, SFRP1, KRT5, KRT17, MYC, KRT14, ERBB2, MDM2, FGFR4 | 20 |
| LumB | FOXA1, ESR1, MYBL2, CEP55, MELK, NDC80, UBE2C, CCNB1, SLC39A6, MAPT, BIRC5, RRM2, BCL2, TYMS, PHGDH, FOXC1, CDH3, SFRP1, ACTR3B, MIA, CDC6, KRT5, KRT17, BAG1, ERBB2 | 25 |
| LumA | FOXA1, MLPH, ESR1, CDC20, CEP55, MELK, KIF2C, UBE2C, PTTG1, EXO1, MAPT, MKI67, BIRC5, NAT1, RRM2, CENPF, TYMS, CDH3, SFRP1, ACTR3B, MIA, CDC6, KRT5, MYC, BAG1, KRT14, ERBB2, MDM2, GRB7, FGFR4 | 30 |

Our results demonstrate the potential to classify breast cancer subtypes using fewer genes than those in the PAM50 gene set. The most informative genes were identified for each subtype, which can improve classification accuracy, particularly for subtypes with the worst prognosis.

### 5.3.2 Classification Results

Once we had selected genes and trained the models, we evaluated the classification performance using the test set. The comparison between the proposed multi-layer method and a standard multiclass classifier is presented in Table 5.3.

The classification results can be summarized as follows:

**Table 5.3.** Comparison of Multiclass and Proposed Multi-layer Classifiers.

| Class | Metric | Multiclass | Proposed Multi-layer | Better Classifier |
|---|---|---|---|---|
| Basal | Precision | 0.950 | **0.989** | Proposed Multi-layer |
| | Recall | 1.000 | 1.000 | Tie |
| | F1-Score | 0.974 | **0.994** | Proposed Multi-layer |
| Her2 | Precision | 0.836 | **0.893** | Proposed Multi-layer |
| | Recall | **0.879** | 0.862 | Multiclass |
| | F1-Score | 0.857 | **0.877** | Proposed Multi-layer |
| LumA | Precision | **0.976** | 0.973 | Multiclass |
| | Recall | **0.875** | 0.806 | Multiclass |
| | F1-Score | **0.923** | 0.882 | Multiclass |
| LumB | Precision | **0.814** | 0.780 | Multiclass |
| | Recall | **0.914** | 0.906 | Multiclass |
| | F1-Score | **0.861** | 0.838 | Multiclass |

1. For the Basal subtype, the proposed multi-layer method demonstrated a clear improvement over the multiclass classifier. It achieved a precision of 0.989, recall of 1.00, and F1-score of 0.994, outperforming the multiclass classifier, which had a precision of 0.950 and F1-score of 0.974. The higher precision in the proposed method indicates a reduction in false positives, a crucial factor for this aggressive subtype.

2. For the Her2 subtype, the proposed multi-layer method also exhibited better performance in terms of precision and F1-score. The proposed classifier achieved a precision of 0.893 and F1-score of 0.877, compared to 0.836 and 0.857 for the multiclass classifier. While the multiclass method had a slightly higher recall (0.879 vs. 0.862), the overall balance between precision and recall favored the proposed method.

3. For the LumB subtype, both classifiers performed similarly, with the proposed method achieving a slightly lower precision (0.780) but maintaining strong recall (0.906) and a comparable F1-score (0.838) to the multiclass classifier.

4. For the LumA subtype, the multiclass classifier outperformed the proposed method, achieving a precision of 0.976 and an F1-score of 0.923, compared to the proposed method's precision of 0.973 and F1-score of 0.882.

Overall, the proposed multi-layer method showed significant improvements for the Basal and Her2 subtypes, which are known for their poorer prognoses. The increased precision for both subtypes reduces the risk of false positives, ensuring more accurate treatment recommendations. For LumA, the multiclass classifier performed better, while for LumB, both methods offered similar results. These findings highlight the potential of the proposed method to enhance classification for the most aggressive breast cancer subtypes, especially where misclassification can critically affect treatment outcomes.

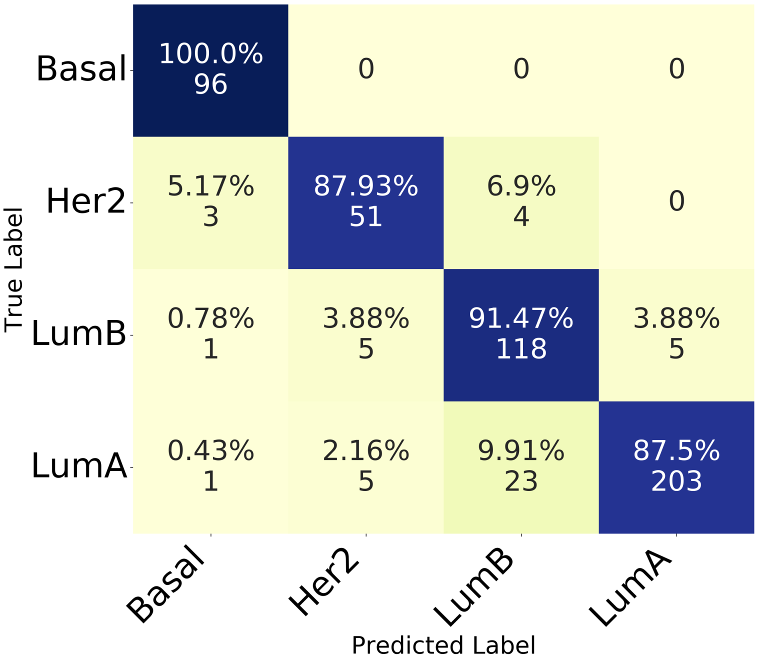


**Figure 5.2.** Confusion matrix for the multiclass classifier (training with ACES and testing with TCGA).

The confusion matrices in Figures 5.2 and 5.3 show that the proposed method was particularly effective for the Basal and Her2 subtypes, which are crucial for clinical decision-making due to their poor prognosis. The higher precision of the proposed method for these subtypes indicates fewer false positives, which is essential for reducing the risk of unnecessary aggressive treatments.

### 5.3.3 t-SNE Visualization

To assess the discriminative power of the selected gene sets, we used t-SNE to visualize the separation between the four breast cancer subtypes: Basal, Her2, LumB, and LumA.

The t-SNE plot for the Basal subtype (Figure 5.4) revealed clear separation from Non-Basal samples, indicating the effectiveness of the selected genes for this subtype.

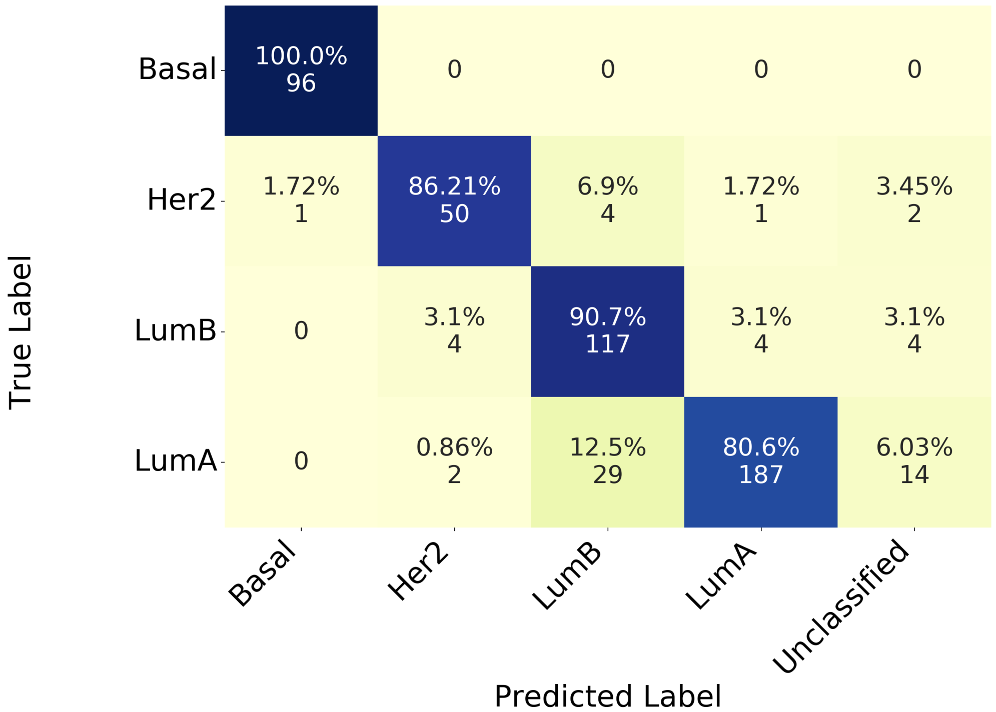


**Figure 5.3.** Confusion matrix for our proposed method (training with ACES and testing with TCGA).

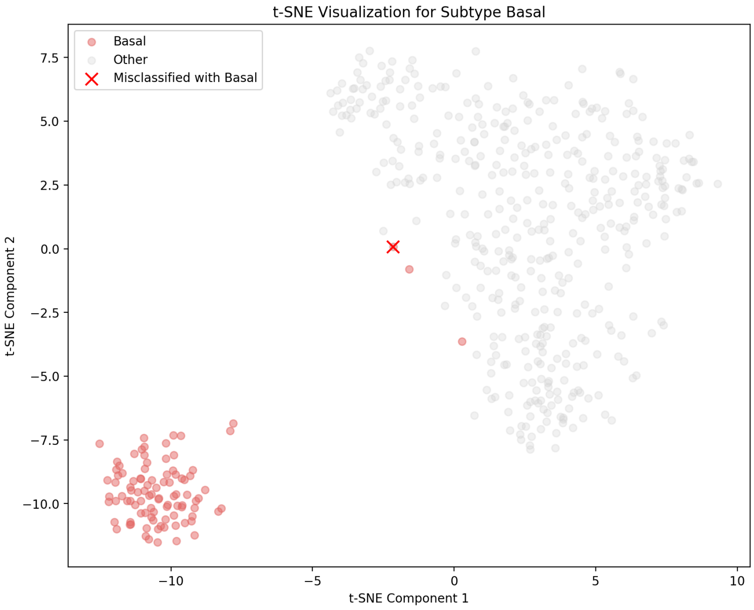


**Figure 5.4.** t-SNE for Basal subtype using 7 selected genes.

The misclassified Basal sample was positioned near the Non-Basal cluster, aligning with the confusion matrix results that showed only one misclassified sample.

For the Her2 subtype (Figure 5.5), a degree of separation was observed between Her2 and Non-Her2 samples, though less pronounced than for Basal. The misclassified

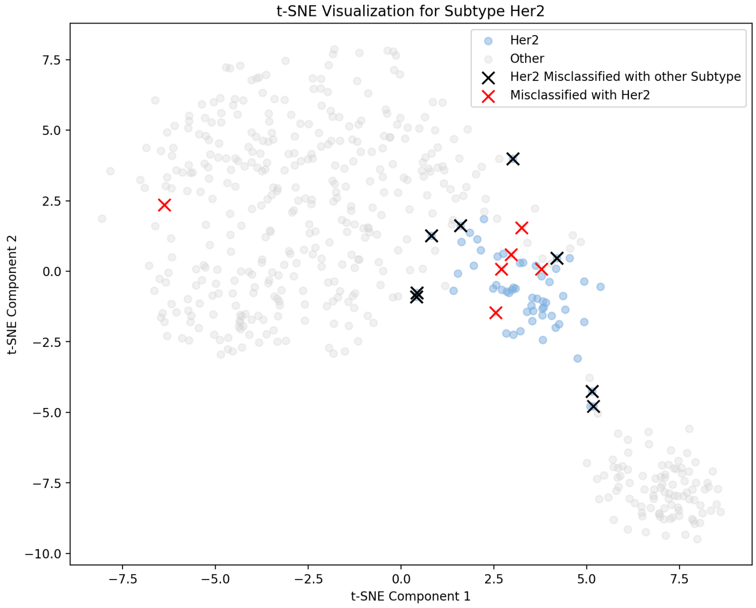


**Figure 5.5.** t-SNE for Her2 subtype using 22 selected genes.

samples tended to cluster near the boundaries of the Her2 group, indicating that further refinement of the model or features may be needed.

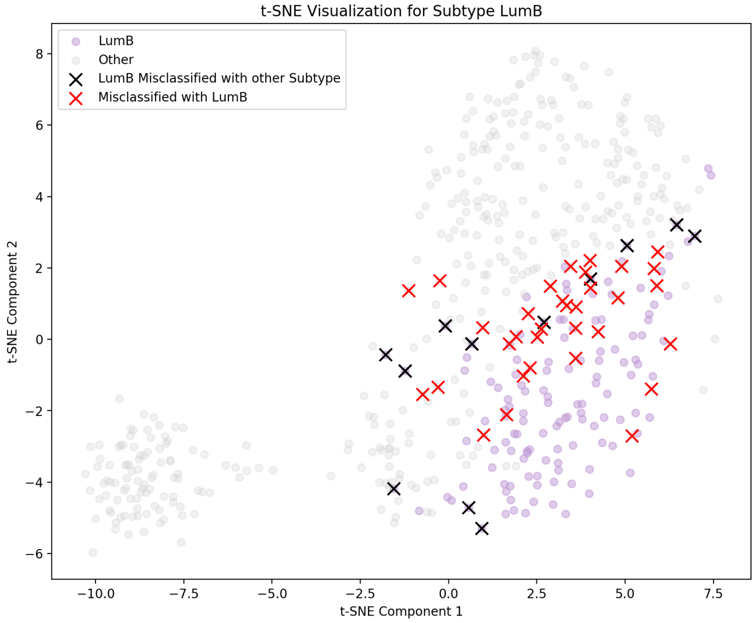


**Figure 5.6.** t-SNE for LumB subtype using 49 selected genes.

The t-SNE visualizations for Luminal A and Luminal B (Figures 5.6 and 5.7) revealed some overlap between subtype and non-subtype samples, suggesting that the gene sets for these subtypes may not be as discriminative as desired. Further feature

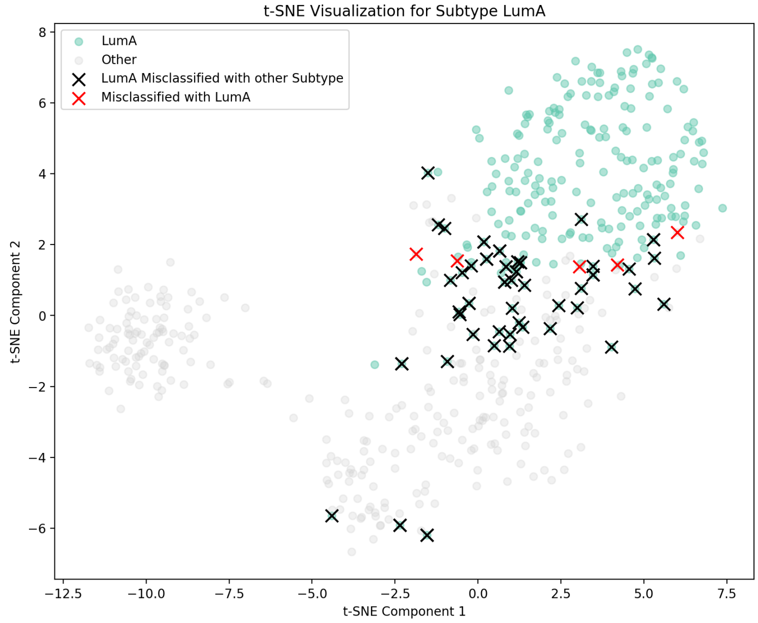


**Figure 5.7.** t-SNE for LumA subtype using 21 selected genes.

selection or alternative dimensionality reduction techniques may improve the classification of these subtypes.

### 5.3.4 Conclusion

In conclusion, our results demonstrate that the proposed multi-layer classification method outperforms a multiclass classifier for Basal and Her2 subtypes, with improved precision and fewer false positives. t-SNE visualizations confirm the discriminative power of the selected gene sets, particularly for the Basal subtype. However, there is room for improvement in classifying Luminal A and Luminal B subtypes, suggesting further refinement in feature selection or model optimization may enhance overall performance.

## 5.4 Final Remarks

In this study, we introduced a multi-layer classification process for identifying breast cancer subtypes based on gene expression data. Our approach focused on selecting a reduced subset of informative genes, optimizing classification performance while reducing the complexity of the models. Notably, the best classification results were achieved for the two worst prognosis subtypes, Basal and Her2.

The results demonstrated the effectiveness of the selected gene sets, particularly for the Basal subtype, which exhibited clear separation from Non-Basal samples. The Her2 subtype also showed a reasonable degree of separation, though there remains room for further improvement.

Looking ahead, several promising directions can be explored to enhance both the performance and interpretability of our approach. One potential avenue is the integration of Explainable Artificial Intelligence (XAI) techniques to gain deeper insights into misclassified samples. By investigating the underlying factors contributing to misclassification, we can identify areas for refinement in feature selection and classification processes, thereby improving the model's accuracy and generalizability.

In conclusion, our study has demonstrated the potential of a multi-layer classification process with optimized gene selection for identifying breast cancer subtypes. Future work should prioritize refining the classification models, incorporating additional sources of biological information, and applying XAI techniques to improve both the interpretability and performance of the models, ultimately advancing our understanding of the underlying biological mechanisms.

6

# Final Remarks

In this thesis, we presented gene selection strategies for breast cancer subtype classification, aiming to improve the performance of classification models. We have described in detail the proposed solution, which has been the primary focus of this research. In Section 6.1, we summarize the key contributions of this study. Section 6.3 outlines the next steps for future research, and in Section 6.4, we provide the publications resulting from this work.

## 6.1 Conclusion

This research has made two significant contributions: (i) the development of classification models for breast cancer subtypes based on reduced gene sets derived from the PAM50 signature; and (ii) the introduction of a multi-layer classification approach to create smaller, focused gene sets tailored for subtype classification.

The first contribution centers on the concept of "Fewer-Shot Gene Selection", where subsets derived from the PAM50 gene signature were analyzed to assess their performance against the established baseline. This approach resulted in the identification of subsets, such as the S-36 gene set, that outperformed the original PAM50 signature in classification tasks. These results emphasize the potential for refining the PAM50 signature and improving feature selection techniques to develop classifiers that are both accurate and efficient. Furthermore, by demonstrating the viability of smaller gene subsets, this work provides a foundation for more tailored diagnostic tools that may be better suited to specific datasets or patient populations. These findings open avenues for future research to explore the broader feature space beyond the PAM50 signature and incorporate advanced feature selection methodologies to further enhance classification accuracy.

The second contribution lies in the development and validation of a multi-layer classification process. By stratifying the classification task into layers and focusing on distinct gene sets for each subtype, this approach optimized classification performance while reducing the complexity of the models. The results highlighted the strength of this approach, particularly for the Basal and Her2 subtypes, which are associated with the worst prognosis. The Basal subtype demonstrated a high degree of separability from Non-Basal samples, underscoring the effectiveness of the selected gene sets. Although the Her2 subtype exhibited slightly lower performance, it still achieved reasonable separation, indicating room for further optimization.

Beyond achieving these results, this research has underscored the importance of addressing misclassified samples. Misclassification analysis, particularly for subtypes with overlapping gene expression profiles, revealed opportunities to refine both feature selection and classification processes. The integration of Explainable Artificial Intelligence (XAI) techniques is a promising future direction, as it can provide insights into the causes of misclassification and enhance the interpretability of the models. By leveraging XAI, researchers and clinicians can gain a deeper understanding of how specific features influence classification outcomes, especially for samples near subtype boundaries.

In terms of broader implications, this study contributes to the field of computational oncology by demonstrating the feasibility of improving classification performance with reduced gene sets and advanced classification strategies. The potential to derive alternative gene signatures that are both robust and clinically relevant positions this work as a stepping stone for more personalized and effective approaches to breast cancer diagnosis and treatment.

Looking forward, future work should aim to refine the models by exploring the full feature space, integrating additional sources of biological data (such as multi-omics approaches), and addressing the inherent heterogeneity of breast cancer subtypes. Further investigation into the creation of subgroups within established subtypes could provide valuable insights into the molecular mechanisms underlying breast cancer heterogeneity. This, in turn, could pave the way for more targeted therapies and improved patient outcomes.

In conclusion, this research has demonstrated the potential for leveraging advanced classification methodologies and optimized gene selection to enhance breast cancer subtype classification. By building on these contributions and addressing the challenges identified, future studies can advance the understanding of breast cancer biology and support the development of precise, interpretable, and clinically actionable diagnostic tools.

## 6.2 Limitations

While this thesis presents significant advances in breast cancer subtype classification, several limitations must be recognized.

The study relied primarily on publicly available datasets, such as TCGA and ACES, which, while extensive, may not fully capture the diversity of breast cancer populations. Additionally, ethical constraints prevented the inclusion of datasets like METABRIC, which could have further validated the findings.The classification performance for Luminal A and Luminal B subtypes, which exhibit overlapping gene expression profiles, was lower compared to Basal and Her2. This highlights a limitation in the resolution of finer distinctions between subtypes with high biological similarity.

Although the study successfully reduced gene sets using PAM50-derived subsets, it did not explore the entire feature space for potentially novel gene sets beyond the PAM50 signature, which may have restricted the discovery of alternative biomarkers. The analysis was restricted to gene expression data, without integrating multi-omics data (e.g., proteomics, epigenomics). This limits the model's ability to capture the broader molecular mechanisms underlying breast cancer subtypes.

While visualization techniques and analysis were used to assess misclassified samples, there remains room for improvement in understanding and addressing the underlying causes of these errors, particularly for samples near subtype boundaries. The models were validated on specific datasets, and their generalizability to other cohorts or clinical settings has not been fully established, which may limit their direct applicability in diverse populations.

Finally, although preliminary steps toward explainability were outlined, the lack of integration of Explainable Artificial Intelligence (XAI) limits the interpretability of model predictions, which is critical for clinical acceptance.

## 6.3 Future Directions

Building on the results obtained so far, several promising avenues for future research have emerged. The primary focus will be on improving classification performance by conducting an in-depth analysis of misclassified samples. Understanding and explaining these errors will not only refine the models but also provide insights into the limitations of current feature selection and classification methods. Specifically, we plan to employ advanced visualization techniques and Explainable Artificial Intelligence (XAI) tools to examine how the models perceive and classify these samples. By identifying the underlying patterns or inconsistencies in the data, we aim to uncover factors contributing to misclassification and develop strategies to address them.

Another key direction involves investigating the potential for creating subgroups within the established breast cancer subtypes. These subgroups may represent clusters of samples with unique molecular profiles that deviate from traditional subtype definitions. By characterizing these subgroups, particularly those near the boundaries of misclassification, we hope to uncover novel biological insights into the heterogeneity of breast cancer. This knowledge could be instrumental in enhancing the current understanding of subtype-specific characteristics and in identifying new therapeutic targets or diagnostic biomarkers.

In addition, it's important to plan to expand the scope of this research by integrating multi-omics data, such as proteomics, epigenomics, and transcriptomics, to enrich the classification framework. Multi-omics integration has the potential to capture complementary biological information, offering a more comprehensive understanding of the molecular mechanisms underlying breast cancer subtypes. This approach could improve classification accuracy, particularly for subtypes with overlapping gene expression patterns, and help to further validate the robustness of the models across diverse datasets.

Another promising direction involves extending the application of our methodology to other cancer types. By adapting the multi-layer classification framework and feature selection processes to other cancers, we can explore whether similar strategies can improve subtype classification and lead to broader clinical applications.

Lastly, incorporating uncertainty quantification methods into the classification process could enhance the reliability of predictions. By assessing the confidence levels of model predictions, particularly for samples near decision boundaries, we can provide more nuanced insights for clinical decision-making. This could be particularly valuable in guiding follow-up analyses or determining cases that warrant further investigation.

In conclusion, the future directions outlined above aim to address current challenges, expand the scope of this research, and ultimately contribute to a more nuanced and actionable understanding of breast cancer subtypes. By integrating advanced analytics, multi-omics data, and exploratory subgroup analysis, we hope to drive further improvements in classification performance, model interpretability, and clinical relevance.

## 6.4 Publications

This section presents the publications produced during the development of this thesis.

### 6.4.1 Main Publications

The following manuscript was published as a result of this work:

- *Few-shot genes selection: Subset of PAM50 genes for breast cancer subtypes classification*, published in BMC Bioinformatics. This paper investigates the possibility of achieving accurate breast cancer subtype classification using a reduced gene set derived from the PAM50 signature.

### 6.4.2 Collaboration Publications

Our collaborations with other researchers in the field resulted in the following publication:

- *Classification of breast cancer subtypes: A study based on representative genes*, published in the Journal of the Brazilian Computer Society (JBCS). This article explores various machine learning techniques for analyzing the PAM50 gene set in breast cancer subtype classification.

### 6.4.3 Submitted Papers

The following paper, based on this thesis, has been submitted and is currently under review:

- *Using Multi-layer classification to improve worst prognosis Breast Cancer Subtypes outcomes*, submitted to BMC Bioinformatics. This paper discusses a hierarchical classification approach designed to improve outcomes for the worst prognosis breast cancer subtypes.

# Bibliography

Abdelmoula, W. M., Balluff, B., Englert, S., Dijkstra, J., Reinders, M. J., Walch, A., McDonnell, L. A., and Lelieveldt, B. P. (2016). Data-driven identification of prognostic tumor subpopulations using spatially mapped t-sne of mass spectrometry imaging data. *Proceedings of the National Academy of Sciences*, 113(43):12244--12249.

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.

Allahyar, A., Ubels, J., and de Ridder, J. (2019). A data-driven interactome of synergistic genes improves network-based cancer outcome prediction. *PLoS computational biology*, 15(2):e1006657.

Allott, E. H., Geradts, J., Sun, X., Cohen, S. M., Zirpoli, G. R., Khoury, T., Bshara, W., Chen, M., Sherman, M. E., Palmer, J. R., et al. (2016). Intratumoral heterogeneity as a source of discordance in breast cancer biomarker classification. *Breast Cancer Research*, 18(1):1--11.

Aquino, G., Costa, M. G. F., and Filho, C. F. F. C. (2023). Explaining and visualizing embeddings of one-dimensional convolutional models in human activity recognition tasks. *Sensors*, 23(9):4409.

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.

Bastien, R. R., Rodríguez-Lescure, Á., Ebbert, M. T., Prat, A., Munárriz, B., Rowe, L., Miller, P., Ruiz-Borrego, M., Anderson, D., Lyons, B., et al. (2012). Pam50 breast cancer subtyping by rt-qpcr and concordance with standard clinical molecular markers. *BMC medical genomics*, 5(1):1--12.

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

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

Bhandari, N., Walambe, R., Kotecha, K., and Khare, S. P. (2022). A comprehensive survey on computational learning methods for analysis of gene expression data. *Frontiers in Molecular Biosciences*, 9:907150.

Bisong, E. and Bisong, E. (2019). Introduction to scikit-learn. *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, pages 215--229.

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.

Chandrakar, P. K., Shrivas, A. K., and Sahu, N. (2021). Design of a novel ensemble model of classification technique for gene-expression data of lung cancer with modified genetic algorithm. *EAI Endorsed Transactions on Pervasive Health and Technology*, 7(25):e2--e2.

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.

Chin, L., Park, P. J., Kucherlapati, R., Creighton, C. J., Donehower, L. A., Reynolds, S., et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61--70.

Choi, J. M. and Chae, H. (2023). mobrca-net: a breast cancer subtype classification framework based on multi-omics attention neural networks. *BMC bioinformatics*, 24(1):169.

Consortium, E. P. et al. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346--352.

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.

Dey, T. K., Mandal, S., and Mukherjee, S. (2021). Gene expression data classification using topology and machine learning models. *BMC bioinformatics*, 22(Suppl 10):627.

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

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.

Ein-Dor, L., Zuk, O., and Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences*, 103(15):5923--5928.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861--874.

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

Gao, F., Wang, W., Tan, M., Zhu, L., Zhang, Y., Fessler, E., Vermeulen, L., and Wang, X. (2019). Deepcc: a novel deep learning-based framework for cancer molecular subtype classification. *Oncogenesis*, 8(9):1--12.

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.

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.

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.

Hand, D. J. and Till, R. J. (2001). A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine learning*, 45:171--186.

Hanley, J. A. et al. (1989). Receiver operating characteristic (roc) methodology: the state of the art. *Crit Rev Diagn Imaging*, 29(3):307--335.

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

Huang, S., Murphy, L., and Xu, W. (2018). Genes and functions from breast cancer signatures. *BMC cancer*, 18(1):1--15.

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.

Jia, X., Han, Q., and Lu, Z. (2018). Analyzing the similarity of samples and genes by mg-pcc algorithm, t-sne-ss and t-sne-sg maps. *BMC bioinformatics*, 19(1):1--13.

Joyce, J. M. (2011). Kullback-leibler divergence. In *International encyclopedia of statistical science*, pages 720--722. Springer.

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.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8--17.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330.

Lazaros, K., Tasoulis, S., Vrahatis, A., and Plagianakos, V. (2022). Feature selection for high dimensional data using supervised machine learning techniques. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 3891--3894. Ieee.

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.

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, B., Wei, Y., Zhang, Y., and Yang, Q. (2017a). Deep neural networks for high dimension, low sample size data. In *Ijcai*, pages 2287--2293.

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, M. C., Pitcher, B. N., Mardis, E. R., Davies, S. R., Friedman, P. N., Snider, J. E., Vickery, T. L., Reed, J. P., DeSchryver, K., Singh, B., et al. (2016). Pam50 gene signatures and breast cancer prognosis with adjuvant anthracycline-and taxane-based chemotherapy: correlative analysis of c9741 (alliance). *NPJ breast cancer*, 2(1):1--8.

Liu, S., Xu, C., Zhang, Y., Liu, J., Yu, B., Liu, X., and Dehmer, M. (2018). Feature selection of gene expression data for cancer classification using double rbf-kernels. *BMC bioinformatics*, 19(1):1--14.

Liu, Y., Bi, J.-W., and Fan, Z.-P. (2017b). A method for multi-class sentiment classification based on an improved one-vs-one (ovo) strategy and the support vector machine (svm) algorithm. *Information Sciences*, 394:38--52.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The genotype-tissue expression (gtex) project. *Nature Genetics*, 45(6):580.

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. Acm.

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

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579--2605.

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.

Mendonca-Neto, R., Reis, J., Okimoto, L., Fenyö, D., Silva, C., Nakamura, F., and Nakamura, E. (2022). Classification of breast cancer subtypes: A study based on representative genes. *Journal of the Brazilian Computer Society*, 28(1):59--68.

Mi, H., Muruganujan, A., Huang, X., Ebert, D., Mills, C., Guo, X., and Thomas, P. D. (2019). Protocol update for large-scale genome and gene function analysis with the panther classification system (v. 14.0). *Nature Protocols*, 14(1):703--721.

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.

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.

Murtaza, G., Shuib, L., Mujtaba, G., and Raza, G. (2020). Breast cancer multi-classification through deep neural network and hierarchical classification approach. *Multimedia Tools and Applications*, 79(21):15481--15511.

Naidu, G., Zuva, T., and Sibanda, E. M. (2023). A review of evaluation metrics in machine learning algorithms. In *Computer Science On-line Conference*, pages 15--25. Springer.

Ochoa, S., de Anda-Jáuregui, G., and Hernández-Lemus, E. (2020). Multi-omic regulation of the pam50 gene signature in breast cancer molecular subtypes. *Frontiers in Oncology*, 10:845.

Okimoto, L. Y., Mendonca-Neto, R., Nakamura, F. G., Nakamura, E. F., Fenyö, D., and Silva, C. T. (2024). Few-shot genes selection: subset of pam50 genes for breast cancer subtypes classification. *BMC bioinformatics*, 25(1):92.

Opitz, J. and Burst, S. (2019). Macro f1 and macro f1. *arXiv preprint arXiv:1911.03347*.

Orrantia-Borunda, E., Anchondo-Nuñez, P., Acuña-Aguilar, L. E., Gómez-Valles, F. O., and Ramírez-Valdespino, C. A. (2022). Subtypes of breast cancer. *Breast Cancer [Internet]*.

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.

Qian, Y., Daza, J., Itzel, T., Betge, J., Zhan, T., Marmé, F., and Teufel, A. (2021). Prognostic cancer gene expression signatures: current status and challenges. *Cells*, 10(3):648.

Raghu, V. K., Ge, X., Balajee, A., Shirer, D. J., Das, I., Benos, P. V., and Chrysanthis, P. K. (2020). A pipeline for integrated theory and data-driven modeling of genomic and clinical data. *arXiv preprint arXiv:2005.02521*.

Ransohoff, D. F. (2005). Bias as a threat to the validity of cancer molecular-marker research. *Nature Reviews Cancer*, 5(2):142--149.

Ringnér, M. (2008). What is principal component analysis? *Nature biotechnology*, 26(3):303--304.

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.

Silla, C. N. and Freitas, A. A. (2011). A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1):31--72.

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

Staiger, C., Cadot, S., Györffy, B., Wessels, L. F., and Klau, G. W. (2013). Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis. *Frontiers in genetics*, 4:289.

Stretch, C., Khan, S., Asgarian, N., Eisner, R., Vaisipour, S., Damaraju, S., Graham, K., Bathe, O. F., Steed, H., Greiner, R., et al. (2013). Effects of sample size on differential gene expression, rank order and prediction accuracy of a gene signature. *PloS one*, 8(6):e65380.

Tafavvoghi, M., Sildnes, A., Rakaee, M., Shvetsov, N., Bongo, L. A., Busund, L.-T. R., and Møllersen, K. (2024). Deep learning-based classification of breast cancer molecular subtypes from h&e whole-slide images. *arXiv preprint arXiv:2409.09053*.

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

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

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.

Venkat, V., Abdelhalim, H., DeGroat, W., Zeeshan, S., and Ahmed, Z. (2023). Investigating genes associated with heart failure, atrial fibrillation, and other cardiovascular diseases, and predicting disease using machine learning techniques for translational research and precision medicine. *Genomics*, 115(2):110584.

Wallden, B., Storhoff, J., Nielsen, T., Dowidar, N., Schaper, C., Ferree, S., Liu, S., Leung, S., Geiss, G., Snider, J., et al. (2015). Development and verification of the pam50-based prosigna breast cancer gene signature assay. *BMC medical genomics*, 8(1):1--14.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, C. G. A. R., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113--1120.

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.

Yang, C.-S., Chuang, L.-Y., Ke, C.-H., and Yang, C.-H. (2008). A hybrid approach for selecting gene subsets using gene expression data. In *2008 IEEE Conference on Soft Computing in Industrial Applications*, pages 159--164. Ieee.

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.