NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

JOSÉ DIOGO TEOTÓNIO PINTO ROMANO

BSc in Computer Science and Engineering

LEARNING TO MAP MICRORNA BIOMARKERS SIGNATURES TO BREAST CANCER SUBTYPES

MASTER IN COMPUTER SCIENCE AND ENGINEERING

NOVA University Lisbon

Draft: *December 30, 2025*

NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

DEPARTMENT OF
COMPUTER SCIENCE

LEARNING TO MAP MICRORNA BIOMARKERS SIGNATURES TO BREAST CANCER SUBTYPES

JOSÉ DIOGO TEOTÓNIO PINTO ROMANO

BSc in Computer Science and Engineering

Adviser: David Semedo
Assistant Professor, NOVA School of Science and Technology

Co-adviser: Bárbara Mendes
Post-Doctoral Researcher, NOVA Medical School

MASTER IN COMPUTER SCIENCE AND ENGINEERING

NOVA University Lisbon

Draft: *December 30, 2025*

Abstract

Accurate classification of breast cancer subtypes is essential for enabling personalized and effective treatment strategies. However, the manual stratification of patients based solely on clinical and pathological criteria presents significant challenges, due to the molecular complexity and heterogeneity of breast cancer. In this context, the integration of molecular biomarkers such as microRNAs (miRNAs), small non-coding RNAs with key roles in post-transcriptional regulation, offers a promising avenue for improving diagnostic precision and the possibility for personalized treatments.

This dissertation investigates how Machine Learning and Deep Learning techniques can be used to map miRNA expression profiles to the intrinsic molecular subtypes of breast cancer: Luminal A, Luminal B, HER2-enriched, and Basal-like (Triple-Negative). The main research direction is centred on evaluating the effectiveness of combining miRNA data with clinical variables to develop robust and interpretable subtype classifiers. Particular attention is given to comparing discriminative approaches (e.g., Logistic Regression, XGBoost) with correlation-based or latent space methods (e.g., DIABLO), in order to identify the most appropriate modelling strategies for multi-omics integration and evaluate the impact of incorporating data form different natures.

Preliminary results suggest that ML models can successfully capture miRNA expression signatures associated with certain subtypes, although Luminal B and HER2-enriched remain challenging to distinguish. These findings represent an initial step toward the broader goal of this work: to contribute to the development of reliable, subtype-specific decision support tools that incorporate molecular and clinical data for more precise breast cancer stratification and personalised treatment planning.

Keywords: microRNAs, breast cancer, biomarkers, machine learning applied to Medicine, multi-omics data

Resumo

A classificação precisa dos subtipos de cancro da mama é essencial para permitir estratégias de tratamento personalizadas e eficazes. No entanto, a estratificação manual de pacientes com base apenas em critérios clínicos e patológicos apresenta desafios significativos, devido à complexidade molecular e heterogeneidade do cancro da mama. Neste contexto, a integração de biomarcadores moleculares, tais como microRNAs (miRNAs), pequenos RNAs não codificantes com papéis fundamentais na regulação pós-transcricional, oferece uma via promissora para melhorar a precisão do diagnóstico e a possibilidade de tratamentos personalizados.

Esta dissertação investiga como as técnicas de Machine Learning e Deep Learning podem ser usadas para mapear perfis de expressão de miRNA para os subtipos moleculares intrínsecos do cancro da mama: Luminal A, Luminal B, HER2-enriquecido e Basal-like (Triplo-Negativo). A principal direção da pesquisa está centrada na avaliação da eficácia da combinação de dados de miRNA com variáveis clínicas para desenvolver classificadores de subtipos robustos e interpretáveis. É dada especial atenção à comparação de abordagens discriminatórias (por exemplo, regressão logística, XGBoost) com métodos baseados em correlação ou espaço latente (por exemplo, DIABLO), a fim de identificar as estratégias de modelação mais adequadas para a integração multi-ómica e avaliar o impacto da incorporação de dados de diferentes naturezas.

Os resultados preliminares sugerem que os modelos de ML podem capturar com sucesso as assinaturas de expressão de miRNA associadas a determinados subtipos, embora o Luminal B e o HER2-enriched continuem a ser difíceis de distinguir. Estas descobertas representam um primeiro passo para o objetivo mais amplo deste trabalho: contribuir para o desenvolvimento de ferramentas de apoio à decisão fiáveis e específicas para cada subtipo, que incorporem dados moleculares e clínicos para uma estratificação mais precisa do cancro da mama e um planeamento personalizado do tratamento.

Palavras-chave: microRNAs, cancro da mama, biomarcadores, aprendizagem automática aplicado à Medicina, dados multi-ômics

Contents

List of Figures v

Acronyms vii

1 Introduction 1
- 1.1 Motivation & Problem Statement . . . 1
  - 1.1.1 Breast Cancer - A Global Health Challenge . . . 1
  - 1.1.2 Can we improve the classification of Breast Cancer subtypes? . . . 2
  - 1.1.3 How to identify the most relevant miRNAs? . . . 3
- 1.2 Challenges and research hypothesis . . . 3
- 1.3 Expected Contributions . . . 5
- 1.4 Document Organization . . . 5

2 Background and Related Work 6
- 2.1 Biological Background . . . 6
  - 2.1.1 DNA & RNA - The Genetic Code . . . 7
  - 2.1.2 MicroRNAs - The Regulators of Gene Expression . . . 8
  - 2.1.3 Cancer - A Complex Disease . . . 9
- 2.2 Biological Context . . . 10
  - 2.2.1 Breast Cancer & its Subtypes . . . 10
  - 2.2.2 Treatment limitations . . . 12
  - 2.2.3 Nucleic acids as gene therapies . . . 14
- 2.3 Related work . . . 15
  - 2.3.1 Leveraging AI models for Cancer Classification . . . 16
  - 2.3.2 Machine Learning (ML) unravelling microRNAs (miRNAs) as biomarkers . . . 24
  - 2.3.3 Comparison with thesis approach . . . 32

3 Methodology and Pilot Study 34
- 3.1 Data Acquisition and Characterization . . . 34

3.2 Data Pre-processing Pipeline 34
3.3 The Supervised Learning Framework 34
3.4 Unsupervised Exploration: Finding Hidden Patterns 34

4 Scalling the Analysis: the Expanded Cohort **35**
4.1 Transition to the Larger Dataset 35
4.2 Applying the Refined Pipeline 35
4.2.1 Model Evaluation 35
4.3 Model Interpretability and Biomarker Discovery 36

5 Results and Discussion **37**
5.1 Comparative Performance 37
5.2 miRNA Biomarker Evaluation 37
5.3 Practical Implications for Clinicians 37
5.4 Limitations of the Study 37

6 Conclusions and Future Work **38**
6.1 Concluding Remarks 38
6.2 Future Work 38

Bibliography **39**

Appendices

A Appendix 1 Covers Showcase **44**

Annexes

I Unveiling microRNA Biomarkers for Breast Cancer Sub-typing **45**

List of Figures

1.1 Age-standardized incidence rate (ASR, per 100,000 inhabitants) of breast cancer in both sexes in 2022. The data represent global estimates based on International Agency for Research on Cancer [23]. 2

1.2 Visual comparison of the estimated number of new cases and deaths caused by Breast Cancer in 2020 (in blue) and 2040 (in red). [23] 2

2.1 Structure of the Deoxyribonucleic Acid (DNA) double helix 7

2.2 Illustration of the transcription mechanism: (a) initiation, (b) elongation, and (c) termination 7

2.3 The figure shows the process of gene expression: DNA is transcribed into mRNA, which is then translated into protein by the ribosome. miRNAs are shown as regulators acting on the mRNA before translation. 8

2.4 In (A) we have a normal breast tissue, while in (B) we can see the presence of a malignant tumor [7]. 11

2.5 Structure of an deep neural network (DNNs). It shows the input, hidden, and output layers, with connections between neurons responsible for processing information [6]. 17

2.6 Schematic of the CNN used by Esteva et al. [15] with Inception v3 architecture, adapted to classify skin lesions based on clinical images. The network generates a probability distribution over clinical classes, based on a structured medical taxonomy. 17

2.7 A subset of the hierarchical taxonomy developed in the study by Esteva et al. [15], with diseases organized by clinical and visual similarity into three major groups: benign, malignant, and non-neoplastic. 18

2.8 Performance evaluation of the convolutional neural network (CNN) in skin lesion classification. (a) Confusion matrices of CNN and two dermatologists. The concentration on the diagonal indicates correct classifications; CNN shows less dispersion and better overall performance [15]. (b) Reliability of the CNN demonstrated by *AUC* curves on a larger, independent dataset [15]. 19

2.9 t-SNE projection of the internal representations of the last hidden layer of the CNN [15]. The different classes of lesions are grouped into distinct clouds, revealing the model's ability to extract relevant discriminative features. . . . 19
2.10 Examples of underfitting, proper fitting, and overfitting. From left to right: the model underfits the data, fits it appropriately, and overfits by capturing noise instead of the underlying pattern [2]. 20
2.11 Illustration of the k-Nearest Neighbors (k-NN) classification process. The top-left panel shows the initial labeled data (Class A in yellow, Class B in purple) and a new unlabeled sample (?). The top-right panel demonstrates the calculation of distances from the new sample to all existing points. The bottom panel shows the selection of the k=3 nearest neighbors and class assignment based on majority voting, resulting in the classification of the new sample. 21
2.12 Example of a Decision Tree for Classification Based on Attributes: Income, Age, Student Status, and Credit Rating (CR). The tree predicts a binary decision outcome (Yes/No) using hierarchical decision rules [25]. 22
2.13 Visualization of a Support Vector Machine (SVM) classifier. The red line represents the optimal hyperplane that separates two classes (blue and green points), while the dashed lines indicate the margins. 22
2.14 Fluxogram of the methodology used in this study, including all the results obtained on each step. [9] 26
2.15 Differential regulation levels of the 29 miRNAs selected by the SVM model in the study by Azari et al. [9]. Positive logFC (log fold-change) values indicate overexpressed miRNAs (magenta) and negative values indicate underexpressed miRNAs (green) in gastric cancer samples compared to healthy tissue. 27
2.16 (a) ROC curve comparing the diagnostic performance of individual microRNAs (hsa-miR-29c and hsa-miR-93) versus their combination. The combined model (blue) shows superior sensitivity and specificity, indicating improved discriminative power for gastric cancer classification.
(b) Venn diagram illustrating the overlap of predicted gene targets among five microRNAs (miR-21, miR-133, miR-204, miR-146, and miR-29c) associated with gastric cancer. The central overlap of 426 genes indicates shared regulatory targets across all five miRNAs, suggesting involvement in common biological pathways. Unique and partially overlapping regions highlight miRNA-specific and combinatorial regulatory potential, supporting their relevance as diagnostic and prognostic biomarkers. 28
2.17 Workflow chart used in the following study. 30
2.18 (a) Entropy formula. (b) Information Gain formula. 31

Acronyms

AI Artificial Intelligence *(pp. 3, 16, 19, 20)*

BC Breast Cancer *(pp. 1–6, 11, 12, 16, 19, 20, 25, 29)*

DL Deep Learning *(pp. 3, 5, 16, 24)*

DNA Deoxyribonucleic Acid *(pp. v, 6, 7)*

miRNAs microRNAs *(pp. iii, v, vi, 2–6, 8, 9, 15, 16, 24, 25, 27–33)*

ML Machine Learning *(pp. iii, 3, 5, 16, 20, 24, 25, 29, 32, 33)*

RNA Ribonucleic Acid *(pp. 6–8, 14)*

1

Introduction

In this chapter, we will present the motivation and problem statement that led to the development of this dissertation, highlighting the global health challenge posed by breast cancer and the potential of microRNAs as biomarkers for the development of personalized treatments. We will also outline the challenges and research hypothesis that guide this work, as well as the expected contributions and organization of the document.

1.1 Motivation & Problem Statement

1.1.1 Breast Cancer - A Global Health Challenge

Breast Cancer (BC) is currently one of the biggest public health challenges worldwide. In 2022, an article from Bray et al. [11] showed that more than 2.3 million new cases of BC were diagnosed, resulting in around 665,000 global deaths . Other studies estimate that BC will continue to not only be the most commonly diagnosed cancer but also to increase in incidence, with projections indicating that by 2040, the number of deaths will almost double and the number of new cases will be around 3.2 million [8]. Figure 1.1 and 1.2 underline the high incidence and mortality associated with the disease, highlighting the geographical variations in disease burden and the ongoing need to develop more effective strategies for its diagnosis and treatment.

BC is characterized by marked biological heterogeneity, manifested in multiple molecular subtypes that exhibit distinct clinical behaviors. Each subtype exhibits substantial differences in terms of tumor aggressiveness, metastatic potential, and behavior to specific therapies as demonstrated by Prat et al. [38] and Perou et al. [36]. However, in the work of Testa, Castelli, and Pelosi [45], we get a comprehensive explanation of why accurate classification of these subtypes is essential to enable personalized therapeutic approaches, with a direct impact on treatment efficacy and disease prognosis.

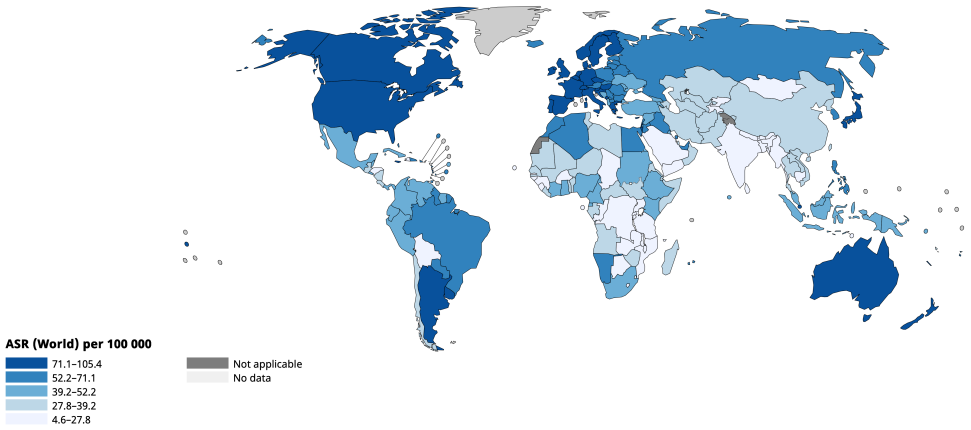
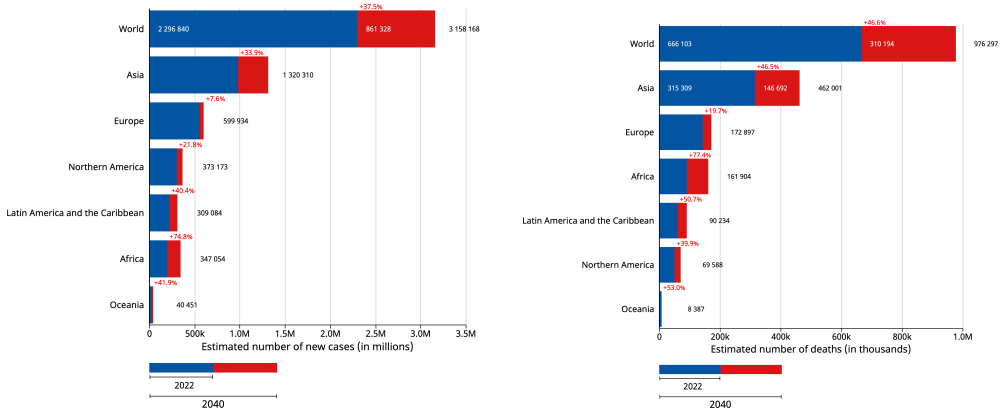


Figure 1.1: Age-standardized incidence rate (ASR, per 100,000 inhabitants) of breast cancer in both sexes in 2022. The data represent global estimates based on International Agency for Research on Cancer [23].



(a) Estimate number of new cases of Breast Cancer

(b) Estimate number of deaths caused by Breast Cancer

Figure 1.2: Visual comparison of the estimated number of new cases and deaths caused by Breast Cancer in 2020 (in blue) and 2040 (in red). [23]

1.1.2 Can we improve the classification of Breast Cancer subtypes?

Among the emerging candidates for robust biomarkers of BC subtypes are miRNAs, small non-coding RNA molecules that play a crucial regulatory role in gene expression. They are estimated to modulate the expression of about one-third of the genes in the human genome [19] and are implicated in the regulation of multiple physiological and pathological processes, including various human diseases [20].

Given their regulatory nature, several studies have demonstrated a significant association between miRNAs expression profiles and relevant clinical characteristics in the context of BC, including processes such as tumor progression and metastasis development, as seen in [20, 31, 33]. In addition to these aspects, a seminal study by Blenkiron et al.

[10] demonstrated that miRNAs expression profiles can effectively distinguish between different molecular subtypes of BC, highlighting their potential as a precise subtyping tool. This ability to discriminate between subtypes reinforces the value of miRNAs as promising clinical biomarkers.

1.1.3 How to identify the most relevant miRNAs?

The identification of the most relevant miRNAs for BC profilling represents a major analytical challenge due to the complexity of these high dimensional regulatory molecules, non-linearity interactions between and clinical phenotypes that require advanced computational approaches to be effectively modelled. Recent advances in Artificial Intelligence (AI), some of which are explored in the work of Luo et al. [29], particularly in ML and Deep Learning (DL), have demonstrated remarkable potential in extracting meaningful patterns from high-dimensional and heterogeneous (data from distinct nature) biomedical data. These approaches enable not only the accurate classification of BC subtypes but also the identification of discriminative miRNAs signatures, supporting their integration as actionable biomarkers in clinical workflows.

In this context, ML and DL models are particularly well suited for the task of robustly characterizing and explaining the profiles of miRNAs - should such biomarkers exist - with the potential to effectively discriminate between different BC subtypes, as already seen in a study done by Azari et al. [9] where ML algorithms identified potential diagnostic and prognostic miRNAs in gastric cancer, showing high accuracy in the identification of reliable biomarkers for this disease.

This reality reinforces the urgency of developing advanced computational tools that can enable more precise molecular characterization and guide personalized therapeutic decisions, ultimately improving clinical outcomes for patients with aggressive and hard-to-treat BC subtypes.

1.2 Challenges and research hypothesis

Based on the assumption that it is possible to use microRNA expression values and clinical data to map BC subtypes, as shown by Ho, Clark, and Le [20] and Muñoz et al. [33], this dissertation proposes to explore several complementary directions for this pathology where the application of AI techniques is still growing.

First, we intend to assess whether discriminative linear models perform better than latent representation models (in a context where there are two different data sources and many dimensions) - such as DIABLO [42], a widely used model in multi-omics problems. At the same time, we will investigate the impact of patient clinical information (such as age, presence or absence of metastases, hormone levels, among others) on the classification performance of the models, where we will be able to gain valuable insights into possible relationships between these features and BC subtypes.

If substantial results are obtained by any of the models, we will be able to conduct a more extensive study on our main point: whether or not there are miRNAs that are potential biomarkers for BC subtypes. In a more advanced approach, we will be able to explore the applicability of methods in the realm of unsupervised learning, such as clustering techniques, to identify potentially novel subgroups within BC that may not align with the currently established subtypes. This exploration could reveal new insights into the heterogeneity of the disease and potentially identify new therapeutic targets.

Even though the base seems promising, there are several challenges to overcome in order to achieve the desired results such as:

1. Inter-Tumor Heterogeneity:
One of the main challenges in studying breast cancer is the significant biological variation between tumours classified within the same subtype, known as *intertumoral heterogeneity*. This diversity can manifest at various levels (genomic, transcriptomic, epigenetic and phenotypic), reflecting microRNA expression profiles that vary among patients. Consequently, defining consistent molecular signatures is difficult, as apparently similar tumours may exhibit distinct biological behaviours and respond differently to therapies. This variability makes it difficult to generalise biomarkers and hinders the identification of stable patterns that can be reliably used in a clinical context.

2. Functional complexity of microRNAs:
The role of miRNAs in biological regulation and cancer progression is extremely complex and still relatively new from a scientific point of view. The action of a single microRNA is not isolated, but rather part of a network of interactions with dozens (or hundreds) of other miRNAs and contextual factors. This highly interdependent behavior raises questions about the effectiveness of overly simplistic or linear models. The application of nonlinear models allows for the discovery of complex relationships and cross-interactions between different miRNAs or between them and clinical variables. These relationships and interactions would be invisible to more traditional approaches.

3. No control set: Another relevant challenge is the absence of a control set that includes data from healthy individuals. Since this type of analysis (miRNAs expression profiling) is not routinely performed in individuals without cancer, it is difficult to define what would be a "normal level" of expression. The implementation of a control group would not only broaden the scope of the model's task (e.g., distinguishing between the presence and absence of cancer before predicting the subtype), but also optimize the robustness of biomarker identification. An illustrative example of this phenomenon is presented in the study Azari et al. [9] where the implementation of a control set was fundamental to the identification of discriminative markers.

The morphology of this dataset limits the choice of approaches to be used and requires extra caution in how we work with certain models, such as nonlinear ones, given their high adaptability in high dimensions, which, in a context of limited data, can easily lead to overfitting. This requires careful selection of algorithms and attention to the pipeline that is set up to ensure that the results obtained are robust and relevant.

Having a control would be an important step toward increasing the generalizability of our model. Using a negative sampling, it would not only allow us to expand the problem to include the distinction between healthy and sick individuals, but also improve the identification of discriminative biomarkers. The latter has already been successfully tested in other types of cancer, and a key step in the pipeline used is precisely the comparison with a control set [9] to isolate clinically relevant markers.

1.3 Expected Contributions

The main contribution of this dissertation is the development of a computational framework for the classification of BC subtypes based on miRNAs expression and patient clinical data. This framework integrates and compares different ML and DL approaches (still underexplored for this disease) applied to a problem of high biological and statistical complexity. A preliminary version of this framework, focused on discriminative models and subtype-specific biomarker identification, was presented in the article accepted at the EPIA 2025 conference under the track AI in Medicine I.

In addition to the classification process, this framework includes a statistical analysis component aimed at validating the predictions made, in order to confer robustness to the decisions generated by the model. This robustness is particularly relevant in a clinical setting, where transparency and reliability of predictions are essential for potential future translation into medical practice. The findings of this work will be tested using both *in-vivo* and *in-vitro* methods by the team of Dr. Bárbara Mendes NOVA Medical School.

Throughout the work, a critical and comparative analysis of the explored approaches will be promoted, focusing on their applicability to complex and heterogeneous biomedical data. It is thus hoped to contribute to the development of more robust, explainable computational solutions adapted to the reality of biological systems, reinforcing the potential of miRNAs as relevant molecular markers in the stratification of BC patients.

1.4 Document Organization

TO-DO

2

Background and Related Work

This chapter will provide the necessary background to understand the biological context related to this work, as well as the work that has been already made in the field of miRNAs reaserch and its applications in BC as a biomarker and a subtype classifier. The chapter is divided into two main sections: the first one will present the biological background, including the central dogma of molecular biology, the role of miRNAs in gene expression regulation, and the characteristics of BC and its subtypes. The second section will review the related work in the field of miRNAs research, focusing on the use of miRNAs as biomarkers and subtype classifiers in BC, as well as the challenges and limitations of current approaches.

2.1 Biological Background

The modern understanding of how genetic information is stored, interpreted, and regulated in cells is based on a fundamenta principle known as the Central Dogma of Molecular Biology. This concept, first formulated in Pray [39] and Watson and Crick [46], describes the unidirectional flow of genetic information in cells: from DNA to Ribonucleic Acid (RNA) and from there to protein synthesis. According to this model, genes encoded in DNA are transcribed into messenger RNA (or mRNA), which in turn is translated into proteins - the functional molecules responsible for most essential biological processes. This dogma has served as the basis for much of the research in molecular biology and biotechnology.

However, in recent decades, it has become clear that this flow of information is regulated in a much more complex way than initially thought. In particular, it has been discovered that a substantial part of the genome is transcribed into non-coding RNA, i.e., RNA that does not give rise to proteins but plays fundamental regulatory roles. It is in this context that merge, small RNA molecules with central functions in the regulation of gene expression. Their discovery has broadened the classical view of the central dogma, introducing new layers of post-transcriptional control that decisively influence normal and pathological biological phenomena.

2.1.1 DNA & RNA - The Genetic Code

At the molecular level, the genetic information of all living organisms is encoded in a molecule called deoxyribonucleic acid (DNA). DNA consists of two complementary strands arranged in a double helix structure, with each strand consisting of a sequence of nucleotides. These nucleotides are composed of a sugar-phosphate structure and one of four nitrogenous bases: adenine (A), cytosine (C), guanine (G), and thymine (T) [44]. When in the helix structure, these bases can only be linked to their corresponding base: adenine can only be linked to thymine and cytosine to guanine, and it is in the sequence of bases that the instructions necessary for the synthesis of all the proteins that govern cell structure and function are encoded.

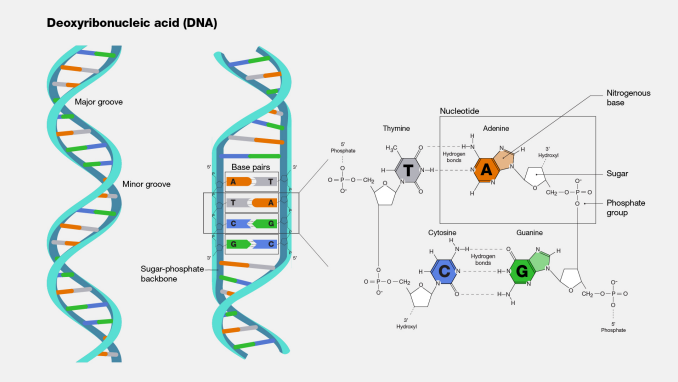


Figure 2.1: Structure of the DNA double helix

The functional units of DNA are called genes, which are discrete sequences that contain the instructions for producing proteins. However, DNA itself cannot participate directly in protein synthesis. Instead, a process called transcription is used to copy the information from a gene to a RNA molecule as seen in Alberts et al. [4]. Unlike DNA, RNA is single-stranded and uses uracil (U) instead of thymine as one of its bases.

Among the various types of RNA, the best known is messenger RNA (mRNA), which serves as an intermediary between genes and proteins. During transcription, an mRNA molecule is synthesized as a complementary copy of a gene, and this mRNA carries the genetic message from the DNA in the nucleus to the ribosomes in the cytoplasm, where protein synthesis occurs. This process, known as translation, is where the mRNA sequence is read in triplets (called codons), each of which corresponds to a specific amino acid [5].

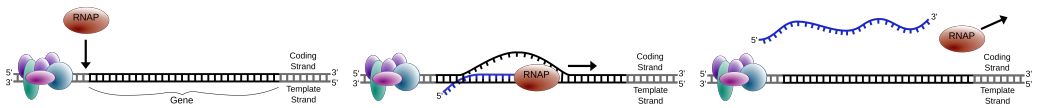


Figure 2.2: Illustration of the transcription mechanism: (a) initiation, (b) elongation, and (c) termination

The set of rules by which the nucleotide sequence in messenger RNA is translated

into a sequence of amino acids is known as the genetic code. This code is composed of triplets of nucleotides, called codons, where each codon specifies one of the twenty standard amino acids used in protein synthesis seen in Novozhilov and Koonin [35] study.

The genetic code is described as redundant but unambiguous. Redundancy means that most amino acids are encoded by more than one codon - for example, leucine is specified by six different codons - which provides a certain degree of robustness to the system. At the same time, the code is unambiguous because each codon corresponds to only one amino acid; that is, a given codon does not encode multiple amino acids [44].

Another fundamental characteristic of the genetic code is its universality. With very few exceptions, the same codons specify the same amino acids in virtually all living organisms, from bacteria to humans. This evolutionary conservation has been fundamental in enabling the development of many molecular biology tools and biotechnological applications prooved by Koonin and Novozhilov [26].

Although the focus of molecular biology for decades has been on the coding sequence of the genome (that is, the genes that give rise to proteins) it is now known that a large part of the human genome is transcribed into RNA that does not code for proteins. These non-coding RNA (ncRNA) molecules play crucial regulatory roles in controlling gene expression. One of the most studied groups within this class are miRNAs, which appear to be central elements in the fine-tuning of the genetic regulation process.

2.1.2 MicroRNAs - The Regulators of Gene Expression

miRNAs are small non-coding RNA molecules, approximately 20 to 25 nucleotides in length, that play a key role in regulating gene expression at the post-transcriptional level [18, 27, 48]. Instead of encoding proteins, they control the production of proteins from genes.

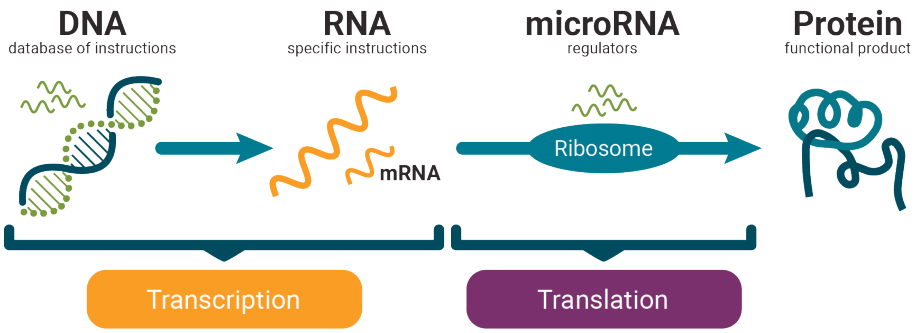


Figure 2.3: The figure shows the process of gene expression: DNA is transcribed into mRNA, which is then translated into protein by the ribosome. miRNAs are shown as regulators acting on the mRNA before translation.

In simple terms, **miRNAs function as molecular switches that bind to messenger RNA (mRNA) molecules**, blocking their translation into protein or promoting their degradation. This mechanism depends on the degree of complementarity between the miRNAs sequence and that of the target mRNA:

- When there is high complementarity, the mRNA tends to be degraded;
- When complementarity is partial, the miRNAs generally acts by inhibiting translation without destroying the mRNA.

A study made by Calaf et al. [13] demonstrates the high effienciency of this mechanism of regulation: a **single miRNAs can control dozens to hundreds of different genes**, and it is estimated that more than 60% of human coding genes are targeted for regulation by miRNAs.

Given this broad regulatory capacity, miRNAs play a central role in multiple cellular processes such as proliferation, differentiation, apoptosis, and stress response. Consequently, changes in miRNAs expression profiles are associated with several diseases, including cancer, neurodegenerative and cardiovascular diseases. In an oncological context, miRNAs can act as oncogenes (promoting tumor growth) or as tumor suppressors, depending on the biological context and cell type as shown by Gulyaeva and Kushlinskiy [18].

Due to their specificity, stability, and direct involvement in relevant molecular mechanisms, miRNAs have been extensively investigated as promising biomarkers for diagnosis, prognosis, and classification in various diseases - including cancer - because if we understand how these regulators are present in our body, we can ease the clinical decision-making process and improve the treatment of patients.

2.1.3 Cancer - A Complex Disease

Cancer is a disease characterized by the uncontrolled proliferation of transformed cells, which can invade neighboring tissues and spread to other parts of the body through processes such as metastasis. This definition, based on Brown et al. [12] and National Cancer Institute [34], has recently been expanded to recognize the role of natural selection in the evolution of cancer: it is a cellular system that continuously evolves, adapting to internal and external pressures to ensure its survival.

Under normal conditions, the body's cells divide only when necessary, die when damaged or obsolete, and are replaced by new ones. However, in cancer, this biological balance is disrupted: abnormal cells gain the ability to multiply independently of the body's signals and to resist programmed cell death (apoptosis). These transformed cells become autonomous units that not only ignore normal growth controls but also interact with the tumor microenvironment to promote their own survival, using angiogenesis, immune evasion, and other adaptive mechanisms [12, 34].

The result is a heterogeneous cell population, subject to natural selection within the human body. Cells that acquire adaptive advantages (e.g., higher proliferation rate, drug resistance, or migration ability) tend to prevail, making cancer a constantly evolving disease [12].

Although cancer can arise in virtually any tissue, not all cellular changes are malignant. There are precancerous conditions, such as *hyperplasia or dysplasia*, which represent an increase in the number of cells or changes in their morphology, but which do not yet invade surrounding tissues.

The advance of cancer stages is a complex process that involves the **acquisition of invasive and metastatic capacity - properties that distinguish malignant tumors from benign ones**. This process can be silent for years, until more severe symptoms arise, often related to the invasion of vital organs.

2.2 Biological Context

Breast cancer represents both one of the greatest challenges in modern medicine and one of the most dynamic areas of biomedical research. Its relevance stems not only from its high incidence and mortality rates, but above all from its biological complexity, which distinguishes it from many other neoplasms. This type of tumour is not uniform, resulting from the accumulation of multiple genetic and epigenetic alterations in breast epithelial cells that deregulate fundamental processes such as proliferation, differentiation, and apoptosis. This molecular instability translates into significant clinical and biological heterogeneity, in which different tumours can have profoundly different gene expression profiles, progression patterns, and metastatic behaviours. This diversity, both intertumoral and intratumoral, constitutes a significant obstacle to the accurate stratification of the disease and the definition of universal therapeutic strategies, given that even patients diagnosed with the same clinical subtype may respond differently to the same treatments.

Although therapeutic advances in recent decades have transformed the prognosis for many patients, with the introduction of more effective surgical therapies, combined drug regimens, hormone blockade, and therapies targeting specific molecular targets, substantial limitations remain, particularly related to phenomena such as acquired resistance, clonal adaptation, and tumour recurrence. In this context, there is a clear need to explore new therapeutic avenues capable of overcoming these limitations. Among the emerging strategies, nucleic acid-based therapies stand out, which aim to directly modulate the gene expression involved in tumour progression, constituting a paradigm shift in the treatment of breast cancer.

2.2.1 Breast Cancer & its Subtypes

Breast cancer is the most commonly diagnosed cancer in women worldwide and is one of the leading causes of cancer death in developed and developing countries as seen in Hong and Xu [21] and Romanowicz, Smolarz, and Nowak [41]. It is estimated that **one in eight women** will be diagnosed with this disease during their lifetime, although it can also affect men - albeit with a much lower incidence.

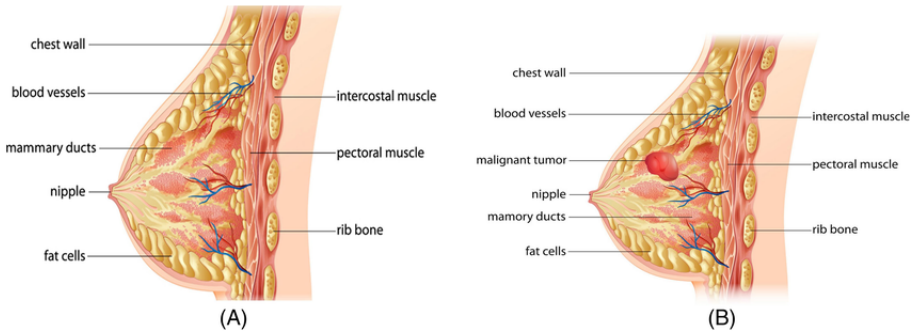


Figure 2.4: In (A) we have a normal breast tissue, while in (B) we can see the presence of a malignant tumor [7].

Most breast tumors are originated in the epithelial cells of the ducts or lobules of the breast, which acquire malignant properties after the accumulation of genetic and epigenetic changes. These events alter the normal control of cell proliferation, differentiation, and apoptosis, allowing for unregulated tumor growth [37].

The development of the disease is associated with a set of well-established risk factors, which include:

- Age and family history of the disease;
- Hereditary genetic mutations, especially in the *BRCA1* and *BRCA2* genes;
- Prolonged exposure to endogenous or exogenous hormones (e.g., early menarche, late menopause, hormone therapy);
- Environmental and behavioral factors, such as obesity, physical inactivity, alcohol consumption, and a diet rich in saturated fats [1, 41].

From a molecular and clinical point of view, **BC is highly heterogeneous**. Each tumor may have unique combinations of genetic alterations, signaling pathways, and gene expression profiles, which are reflected in different clinical behaviors, degrees of aggressiveness, and response to treatment [22, 37]. This diversity explains why early detection remains crucial for prognosis: when diagnosed in its early stages, BC has survival rates of over 90%. However, in more advanced stages, especially when metastases appear, controlling the disease becomes substantially more difficult and the therapeutic goal shifts from curative to palliative [1, 21]. The therapeutic approach is therefore typically multimodal, combining surgery, radiotherapy, chemotherapy, hormone therapy, and targeted or biological therapies, depending on the characteristics of the tumor and the patient's general condition. The most significant advance in the last decade has been the transition from a uniform model to a personalized treatment approach, tailored to the molecular subtype and individual risk as studied by Romanowicz, Smolarz, and Nowak [41].

Table 2.1: Summary of intrinsic BC subtypes and typical characteristics of each one. *References:* [1], [22], [38], [21], [30].

Subtype	**Receptors / HER2**	**Prolif.**	**Prognosis**	**Treatment**
Luminal A	ER+/PR+, HER2−	Low	Favorable	Endocrine only
Luminal B	ER+, PR↓, HER2±	High	Intermediate	Hormone ± Chemo ± Anti-HER2
HER2-enriched	HER2+, ER−, PR−	High	Improved	Anti-HER2 + Chemo
Basal-like / TNBC	ER−, PR−, HER2−	High	Poor	Chemo; ± PARP/IO (selected)

Acronyms: Chemo = Chemotherapy , IO = Immunotherapy .

It is also important to note that breast tumors are not static entities. As recognized by Polyak [37], phenomena such as intra-tumor heterogeneity and clonal evolution enable tumors to adapt to the selective pressure of treatments, frequently leading to therapeutic resistance and disease progression. This dynamic nature of the disease further complicates clinical management and reinforces the need for accurate molecular characterization. Indeed, given the molecular complexity and clinical diversity of breast tumors, several studies have established that BC is not a single disease but rather a collection of biologically distinct entities that arise from a common anatomical site [1, 38]. This heterogeneity is reflected in major differences in tumor progression, metastatic behavior, response to therapy, and long-term prognosis.

To better capture this complexity and inform clinical decision-making, researchers have proposed molecular classification systems that subdivide breast tumors into intrinsic subtypes [36, 38]. These are typically defined as *Luminal-A*, *Luminal-B*, *Basal-like* (or Triple-negative) and *HER2-enriched*, and rely on the expression status of three key hormonal biomarkers: **estrogen receptor (ER)**, **progesterone receptor (PR)**, and **human epidermal growth factor receptor 2 (HER2)**, in combination with proliferation indices (e.g., Ki-67) and gene expression patterns. The relationship between BC subtypes and these hormone receptors is summarized in Table 2.1. This classification underpins modern precision oncology approaches and has profound implications for both therapy and prognosis.

Recent evidence by Polyak [37] and Yeo and Guan [51] suggests that multiple subtypes can coexist within the same tumor (a phenomenon called intra-tumor heterogeneity). This complexity contributes to therapeutic resistance and disease progression.

2.2.2 Treatment limitations

Despite remarkable progress in recent decades, breast cancer treatment continues to face significant limitations that compromise both clinical efficacy and patients' quality of life [28]. The development of new therapies and molecular stratification of the disease have

substantially improved prognosis in several contexts, but significant obstacles remain. Among the most critical are the lack of effective therapeutic options for particularly aggressive subtypes, such as triple-negative breast cancer, acquired resistance to different treatment modalities, the toxicity associated with existing regimens, and difficulties in personalizing therapy for specific patient groups. These factors, analyzed below, reveal the complexity of the disease and the continuing need for more effective and safer solutions [36, 38].

One of the most striking limitations when it comes to breast cancer treatment is undoubtedly the triple-negative subtype, recognized for its aggressiveness in the patient's body and poor prognosis, as seen in the study by Howlader et al. [22] and Loibl et al. [28]. The absence of estrogen, progesterone, and HER2 receptors (hence the term triple negative) immediately eliminates the possibility of using hormonal or targeted therapies, leaving chemotherapy as the main treatment option. Although many patients respond positively in the initial period, relapses are frequent and, in most cases, there are no effective options to counteract the gradual weakening of the body given the aggressiveness of the treatments [14]. Trials with anti-angiogenic agents, such as bevacizumab, have shown only marginal benefits and no consistent impact on overall patient survival. Epidemiological studies find that TNBC has the lowest specific survival rate among the subtypes, reflecting the lack of a therapy with clear and targeted targets.

Therapeutic resistance is another cross-cutting challenge. In the study by Sun et al. [43], hormone-sensitive tumors, acquired resistance to endocrine therapies (therapies that aim to block the production of hormones that lead to tumor growth or prevent those that are produced from acting on tumor receptors) such as aromatase inhibitors and selective estrogen receptor modulators, often mediated by alternative signaling pathways such as PI3K/AKT/mTOR. In the HER2-positive subtype, despite the success of trastuzumab, resistance often emerges in a metastatic context, associated with mutations in HER2 or activation of parallel pathways such as MEK/ERK [14, 28]. Trials such as HERA have demonstrated clear gains in disease-free survival with the addition of trastuzumab, but have also confirmed that the benefit is not uniform across all patients. In cases of TNBC, resistance to chemotherapy is even more problematic: patients who do not achieve a complete pathological response have very high relapse rates, and lifestyle factors such as obesity or smoking have been associated with lower therapeutic efficacy.

The toxicity associated with treatments remains a central limitation in breast cancer. Despite advances in therapeutic efficacy, many regimens have adverse effects that significantly affect patients' quality of life and, in some cases, can even be severe [28]. Chemotherapy, for example, is associated with severe toxic reactions in some patients. Hormone therapy, evaluated in the study by Sun et al. [43], used in hormone receptor-positive tumors, may increase the risk of cardiovascular complications and other systemic effects, while hormone inhibitors are linked to bone health problems. Even targeted therapies, considered more selective, are not risk-free and can cause long-term complications. This delicate balance between clinical efficacy and toxic impact represents a constant

dilemma, underscoring the need to develop safer and more personalized approaches.

Finally, important limitations remain in terms of treatment individualization. The ideal approach for certain patient groups, such as those with low hormone receptor expression, remains unclear, as these cases do not easily fit into traditional clinical categories or available trials. Metastatic disease remains one of the greatest challenges, remaining essentially incurable and associated with reduced median survival, regardless of subtype [22]. Furthermore, there is still no consensus on the best sequence of therapies for some patient profiles, and clinical practice often faces the dilemma between undertreatment and overtreatment: in many contexts, more intensive regimens are chosen to reduce the risk of failure, but this results in adverse effects that significantly impact quality of life [38].

In summary, the current limitations of breast cancer treatment reflect the interaction of multiple factors: subtypes with poor prognosis (such as TNBC), resistance phenomena that reduce therapeutic efficacy, toxicities that compromise quality of life, and the absence of clear consensus on the individualization of care. These challenges explain why, despite advances, metastatic disease remains incurable and why many patients are still subjected to therapeutic regimens that do not offer the ideal balance between efficacy and safety. Overcoming these barriers will depend on the ability to integrate new biomarkers, innovative treatment strategies, and truly personalized approaches—points that also form the basis for research into computational methods and artificial intelligence applied to oncology.

2.2.3 Nucleic acids as gene therapies

The use of nucleic acids as therapeutic agents has emerged as one of the most promising approaches in contemporary biomedicine, based on their unique ability to directly modulate gene expression. Molecules such as plasmid DNA (pDNA), messenger RNA (mRNA), small interfering RNAs (siRNAs), antisense oligonucleotides (ASOs), and microRNAs (miRNAs) allow precise intervention in the molecular pathways responsible for the origin and progression of multiple diseases [31]. Unlike conventional drugs, which act mainly at the protein or metabolic level, these molecules act upstream, on the regulation of gene expression. This feature makes it possible to correct mutations, restore the function of tumor suppressor genes, silence unwanted transcripts, or reintroduce lost cellular functions. Thus, nucleic acid-based therapeutics offer the potential to treat diseases previously considered untreatable, including monogenic genetic disorders, persistent viral infections, and various types of cancer.

Therapeutic strategies based on nucleic acids can be organized into three broad categories: **gene editing**, which uses systems such as CRISPR-Cas to correct or inactivate mutated genes; **gene addition**, which involves introducing functional copies of missing genes; and **gene silencing**, which uses small RNA molecules to block the translation of pathogenic transcripts. Despite their enormous potential, the clinical application of

these technologies faces significant obstacles related to biological stability and the effective delivery of molecules to target cells. Nucleic acids are structurally fragile, as seen in Mendes et al. [31] and Mendes et al. [32], degrade rapidly by the action of nucleases, and, due to their negative charge and high molecular weight, have low diffusion through the plasma membrane. To overcome these limitations, nanotransport systems - such as lipid, polymeric, and inorganic nanoparticles - have been developed that not only protect these molecules from degradation but also facilitate their entry into cells and release at the site of action.

The development of these delivery systems has been instrumental in the clinical translation of gene therapies. A prime example is the central role of lipid nanoparticles (LNPs) in the success of mRNA vaccines against SARS-CoV-2, which has consolidated the viability and global acceptance of this technology. Although they face challenges related to safety, toxicity, and low accumulation in tumor tissues, other platforms, such as polymeric and inorganic nanoparticles, offer additional advantages, namely greater structural versatility and unique physicochemical properties. Thanks to their ability to selectively and sustainably modulate biological processes, nucleic acids now play a central role in the development of precision cancer therapies. This is particularly relevant in breast cancer, one of the most prevalent neoplasms worldwide and marked by profound molecular heterogeneity. Subtypes such as luminal, HER2-positive, or triple-negative tumors have distinct genetic and epigenetic profiles, which determine not only the progression of the disease but also the response to conventional therapies Mendes et al. [32]. In this scenario, nucleic acid-based therapies offer a unique opportunity: instead of resorting to nonspecific chemotherapeutic agents, highly selective interventions can be designed to correct specific alterations in each tumor subtype - from restoring inactivated suppressor genes to blocking active oncogenes or regulating key elements of the tumor microenvironment.

Breast cancer is thus a privileged model for preclinical research on nucleic acid delivery. Recent trials highlight the use of lipid and inorganic nanoparticles in breast tumor models, precisely because of their ability to transport mRNA, siRNA, or miRNA to tumor cells and, in some cases, to the surrounding stroma. Although tumor accumulation efficiency remains low - often less than 5% of the administered dose - results in animal models show significant reductions in tumor growth, reinforcing the translational potential of this approach. In this context, breast cancer emerges not only as one of the most urgent targets due to its high incidence, but also as a paradigmatic case for the therapeutic exploration of nucleic acids, where the need for personalized solutions is coupled with the challenge of resistance and clinical variability associated with currently available therapies.

2.3 Related work

This section will present a critical review of computational approaches developed to date to explore the potential of miRNAs as biomarkers in the context of oncology, covering both

BC and other malignant neoplasms. The contributions of ML and DL models applied to the task of classifying different cancer subtypes will also be analyzed, with a special focus on methodologies that integrate molecular data with data from other nature, like clinical characteristics for example.

ML, a branch of AI, involves developing computational models that learn from data to make predictions or decisions. These models are typically trained using either supervised learning, where the target outcomes are known and used during training, or unsupervised learning, in which no explicit labels or outcome variables are provided. In both paradigms, the goal is to uncover meaningful patterns in the data that can be used to generate predictive insights, such as detecting the presence of cancer, estimating survival probabilities, or stratifying patients into risk categories. ML techniques are particularly valuable when dealing with unstructured or complex clinical datasets, as is often the case in oncology.

In recent years, the application of ML algorithms to the field of biomedicine has led to significant advances in the analysis of complex and high-dimensional data, including the expression of miRNAs in cancer [17]. In this context, several studies have explored the use of computational models for the classification of tumor subtypes and/or the identification of discriminative biomarkers, with promising results but also with important limitations.

In this context, we will review and analyze scientific works that have leveraged ML algorithms in contexts similar to the stratification of BC subtypes based on miRNAs expression profiles, complementing them with data of other types (multi-omics data). For each study, it will be important to define the specific work in question so that we can analyze the methodology, algorithms used, and results obtained, all based on the specific context of the research in question, in order to capture and consolidate a ground on which we can work. At the end of the review, we will discuss how these contributions inform and substantiate the methodological choices made in the present work, justifying, whenever possible, the algorithmic and experimental choices based on the available scientific evidence.

2.3.1 Leveraging AI models for Cancer Classification

The classification of different types of cancer using computational models has been one of the most explored areas within the application of AI to medicine. Let's take a look at the work of Esteva et al. [15], a remarkable advance in this "new" relationship between computers and dermatology, where deep neural networks (Figure 2.5) have demonstrated capabilities comparable to those of human experts in the diagnosis of malignant skin lesions.

This work was made possible by the use of an architecture based on convolutional neural networks (CNNs) - a type of DNN that is particularly effective in image processing (Figure 2.6). CNNs work by applying convolutional filters that extract visual patterns at different levels of complexity, allowing the model to identify relevant features directly

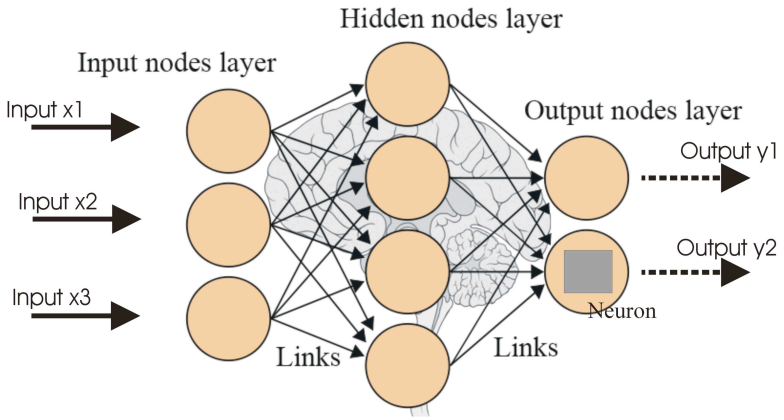


Figure 2.5: Structure of an deep neural network (DNNs). It shows the input, hidden, and output layers, with connections between neurons responsible for processing information [6].

from the image pixels, without the need for specialized preprocessing [3].

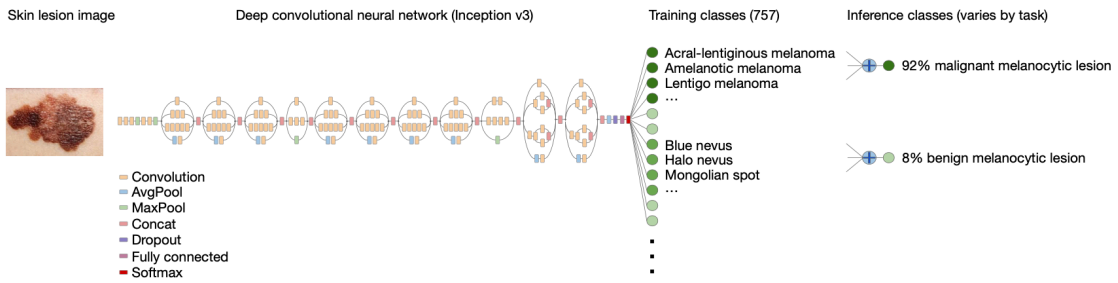


Figure 2.6: Schematic of the CNN used by Esteva et al. [15] with Inception v3 architecture, adapted to classify skin lesions based on clinical images. The network generates a probability distribution over clinical classes, based on a structured medical taxonomy.

The biggest problem with this methodology is that these architectures require a large number of cases (positive and negative) in order to learn the necessary patterns. In this case, 129,450 clinical images covering more than 2,000 different diseases were used. Some factors that determined the good results of this model were:

1. The photographic variability of the samples on which it was trained, since they covered not only images taken with mobile phones, but also dermoscopy images;

2. The manipulation of images during training, enlarging and inverting them to increase the adaptability and robustness of the model;

3. The use of a structured medical taxonomy (Figure 2.7), built on clinical and visual criteria, which allowed for the organization of more than 2,000 diseases into a hierarchy of 757 fine-grained training classes, such as *acrolentiginous melanoma* and

amelanotic melanoma.

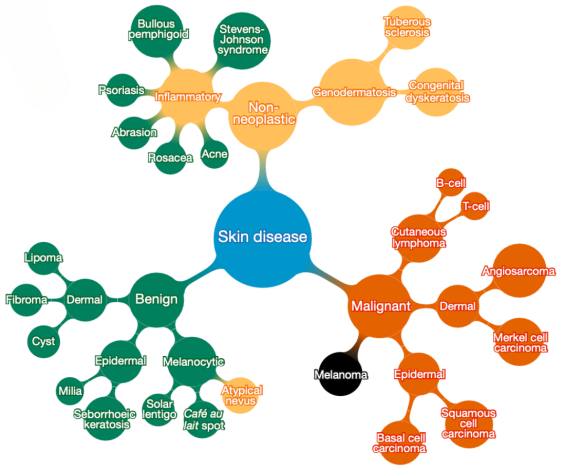
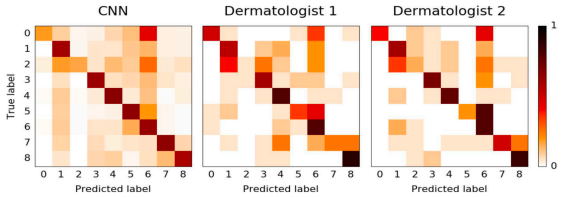


Figure 2.7: A subset of the hierarchical taxonomy developed in the study by Esteva et al. [15], with diseases organized by clinical and visual similarity into three major groups: benign, malignant, and non-neoplastic.

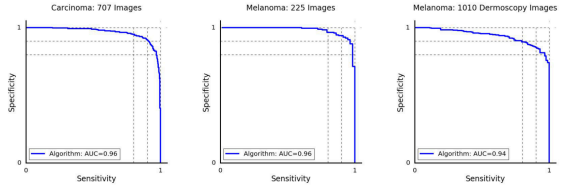
The result? A computational model that not only achieved performance comparable to that of certified dermatologists, but in several scenarios even demonstrated superiority over average human performance, verifiably by this confusion matrix (Figure 2.8a). The trained convolutional neural network was able to classify two critical clinical cases with high accuracy: keratinocytic carcinomas versus benign seborrheic keratoses, and malignant melanomas versus benign nevi. In these binary scenarios, it obtained areas under the curve (*AUC*) of 0.96 and 0.94, respectively (Figure 2.8b) - values higher than those obtained by dermatologists in the same tasks. *AUC* is a metric that quantifies a model's ability to distinguish between classes, with values close to 1 indicating excellent performance.

Furthermore, in more complex scenarios with multiple classes (three and nine disease categories), the model maintained its**remarkable levels of accuracy** (72.1% and 55.4%), surpassing or equaling human experts (the levels itself are not extraordinary, but considering that surpases the group of experts, it becomes a great achievement). The robustness of the methodology, that is, its ability to maintain performance under different conditions or test data, was also confirmed in larger test sets, where the network's performance remained stable, with **minimal variations in evaluation metrics**. From a technical standpoint, it was an efficient, scalable system with **potential for application in mobile devices**, which gives it relevant clinical applicability, especially in contexts with limited access to specialists.

Internal analyses further reinforced confidence in the model, showing that it learned consistent clinical representations: the network tended to **group diseases with similar visual characteristics** (Figure 2.9) and focused its attention on the damaged areas of the



(a) (Confusion matrices comparison)



(b) AUC curves analysis

Figure 2.8: Performance evaluation of the convolutional neural network (CNN) in skin lesion classification. (a) Confusion matrices of CNN and two dermatologists. The concentration on the diagonal indicates correct classifications; CNN shows less dispersion and better overall performance [15]. (b) Reliability of the CNN demonstrated by *AUC* curves on a larger, independent dataset [15].

images, ignoring irrelevant regions such as background or healthy skin - promising evidence of *automated clinical focus* with real practical utility.

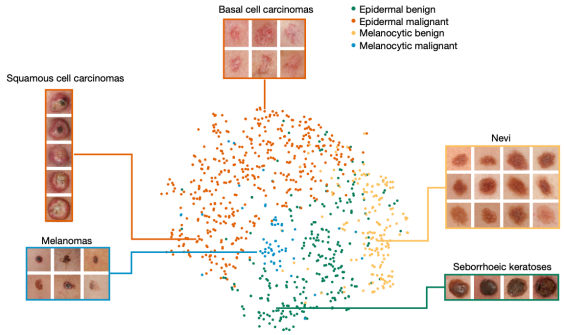


Figure 2.9: t-SNE projection of the internal representations of the last hidden layer of the CNN [15]. The different classes of lesions are grouped into distinct clouds, revealing the model's ability to extract relevant discriminative features.

The effectiveness demonstrated in the previous project shows the magnitude of the benefits that AI can bring to the world of medicine, helping doctors diagnose and stratify diseases with an accuracy that, in some cases, surpasses that of human specialists. This capability is not limited to imaging data: it also extends to the field of molecular data and dermatology, as demonstrated by the study by Wu et al. [49], which applies machine learning algorithms to the task of classifying BC subtypes (in this case, the goal was to distinguish between Triple Negative, or basal-like, and non-Triple Negative tumors, since TNBC is the most deadly cancer with the most difficult prognosis, as we saw in the table).

In the study by Wu et al. [49], when working with gene expression data from thousands of patients, additional challenges arise related to the high dimensionality of the data, requiring robust feature selection methods and predictive models capable of dealing with complex and often non-linear correlations. This type of study brings us closer not only to the context of this thesis, but also to the type of challenges we will encounter and how we can take advantage of the methodologies used in this paper to achieve good results. From all the algorithms that were tested, Support Vector Machine (*SVM*) stood out for its performance. This approach allowed the authors to achieve high levels of accuracy, sensitivity, and specificity, demonstrating the potential of AI in classifying BC subtypes based on genomic information. The type of information for this study was the RNA-Sequence (RNA-Seq) profiles made available by The Cancer Genome Atlas (TCGA), a public database containing thousands of tumor samples characterized at the genomic level. This dataset, after pre processing, had 934 tumor samples and over 57,000 genes per sample, a typical high-dimensionality scenario where the number of variables far exceeds the number of observations, and considering that not all of them are necessary or have a great impact, **differential expression analysis was applied** - a bioinformatics technique used to detect which genes are significantly more or less expressed between different conditions - resulting in the selection of 5,502 differentially expressed genes, which served as input for the predictive models (a large reduction of over 50,000 genes). This step corresponds to feature selection, which is essential in problems where there is a high risk of *overfitting* - that is, when the model memorizes the training data but fails to generalize to new examples.

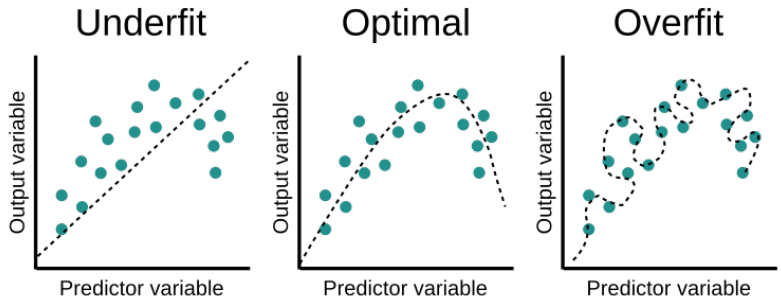


Figure 2.10: Examples of underfitting, proper fitting, and overfitting. From left to right: the model underfits the data, fits it appropriately, and overfits by capturing noise instead of the underlying pattern [2].

Now that the dataset had been reduced to a more informative and manageable subset of features, the authors moved on to the predictive modeling phase. This is a crucial moment in the ML pipeline, where the ability of different algorithms to learn discriminative patterns present in the data is tested - in this case, distinguishing between TNBC and non-TNBC tumors based on the expression levels of selected genes. Several classic supervised learning algorithms were then evaluated, representing different approaches to the classification task:

- **K-nearest Neighbors (*kNN*)**: classifies new data based on the K nearest neighbors in the feature space [53].

Figure 2.11: Illustration of the k-Nearest Neighbors (k-NN) classification process. The top-left panel shows the initial labeled data (Class A in yellow, Class B in purple) and a new unlabeled sample (?). The top-right panel demonstrates the calculation of distances from the new sample to all existing points. The bottom panel shows the selection of the k=3 nearest neighbors and class assignment based on majority voting, resulting in the classification of the new sample.

- **Naïve Bayes (*NB*)**: uses Bayes' Theorem to estimate the most likely class of a sample, assuming that the input variables are independent of each other [47].

$$P(C_k \mid \mathbf{x}) = \frac{P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k)}{P(\mathbf{x})}$$

- **Decision Tree (*DT*)**: a model that makes decisions through a hierarchical tree-shaped structure, where each internal node represents a condition on a variable, and each branch represents a possible outcome of that condition. The process continues until it reaches a leaf node, which indicates the final class or value [24].

- **Support Vector Machine (*SVM*)**: This algorithm constructs an optimal hyperplane that best separates samples from different classes. The goal of SVM is to maximize the margin between the two classes for better generalization. Support vectors are the data points that lie closest to the hyperplane and define its position.

In order to evaluate a model, performance metrics are used to assess how well it performs across different dimensions, such as accuracy, relevance, and sensitivity to different classes. These metrics provide quantitative insight into the strengths and limitations of a classification algorithm, allowing researchers and practitioners to make informed decisions when comparing models or tuning parameters. Evaluating models using multiple metrics is especially important in scenarios involving imbalanced datasets, where a single

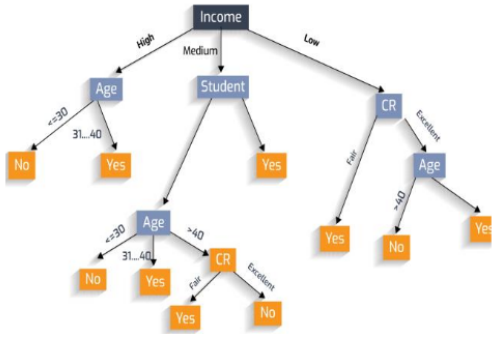


Figure 2.12: Example of a Decision Tree for Classification Based on Attributes: Income, Age, Student Status, and Credit Rating (CR). The tree predicts a binary decision outcome (Yes/No) using hierarchical decision rules [25].

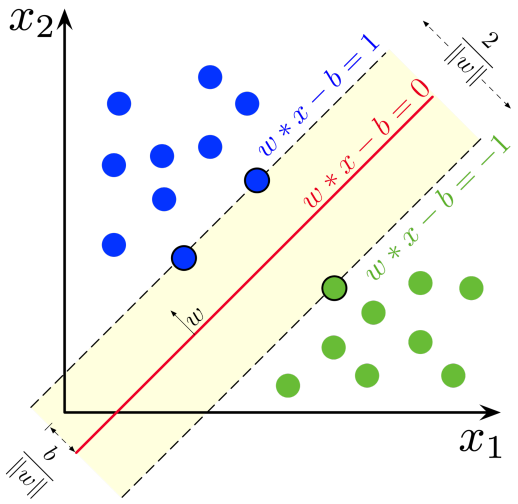


Figure 2.13: Visualization of a Support Vector Machine (SVM) classifier. The red line represents the optimal hyperplane that separates two classes (blue and green points), while the dashed lines indicate the margins.

metric (such as accuracy) may not provide a complete picture. The most commonly used evaluation metrics in classification tasks, according to Yaseen and Abdulazeez [50], are:

1. **Accuracy**
   Accuracy is the ratio of correctly predicted instances to the total number of instances. It measures the overall effectiveness of a classification model.

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

   where:

   - TP = True Positives
   - TN = True Negatives
   - FP = False Positives
   - FN = False Negatives

2. **Precision**
   Precision measures the proportion of correctly predicted positive observations to the total predicted positive observations. It reflects the model's ability to return only relevant results.
$$\text{Precision} = \frac{TP}{TP + FP}$$

3. **Recall (Sensitivity or True Positive Rate)**
   Recall is the ratio of correctly predicted positive observations to all actual positives. It indicates the model's ability to identify all relevant cases.
$$\text{Recall} = \frac{TP}{TP + FN}$$

4. **F1-Score**
   The F1-Score is the harmonic mean of Precision and Recall, providing a balance between the two. It is particularly useful when the class distribution is imbalanced.
$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

5. **Support**
   Support refers to the number of actual occurrences of each class in the dataset. While not a performance metric per se, it helps in understanding how many examples a classifier is making predictions on for each class.

Now that we are familiar with the main concepts of classification models and the metrics used to evaluate them, let us return to the study by Wu and Hicks (2021) in light of these indicators. The analysis of the results, presented in the following table, allows us to see more clearly how each algorithm performed in the task of distinguishing between TNBC and non-TNBC, highlighting the performance of SVM in virtually all scenarios evaluated.

In the complete gene set, SVM achieved 90% accuracy, 87% recall, and 90% specificity - metrics that indicate, respectively, the proportion of correct classifications, the ability to correctly identify TNBC cases, and the ability to avoid false positives. The analysis of these metrics is essential to correctly interpret the results in a clinical context, where the consequences of classification errors can be significant.

To validate the robustness of the differentially expressed gene selection approach, the authors compared it with other classic feature selection methods, such as *SVM-RFE* (an iterative technique that removes the least relevant features based on *SVM* weights), *Relief* (which weights features based on their correlation with the class), *ARCO*, and *mRMR* (which maximizes relevance and minimizes redundancy between variables). Even so, the model based on differential expression and *SVM* demonstrated better performance in most cases, confirming the soundness of the strategy adopted.

The study further deepened the analysis of feature importance by evaluating the performance of models with different sizes of gene subsets, from the initial 5,502 to only 16 genes. Interestingly, *SVM* performance remained high even with reduced sets, achieving the best results with 256 genes. This stability suggests that discriminative information is concentrated in a small subset of features, which is relevant for scenarios with a low number of samples and a large number of features, as we have in this dissertation, and quite positive because:

- it allows molecular tests to be cheaper and faster since fewer genes are needed;
- it makes models easier to validate in a clinical setting since it is simpler to obtain quality samples with few targets;
- greater interpretability for physicians.

The analysis of the two studies presented here allows us to consolidate a fundamental idea for this dissertation: ML and DL algorithms demonstrate a remarkable ability to **classify different types of cancer based on complex data**, whether imaging or molecular. From deep neural networks applied to dermatological imaging to discriminative algorithms used to analyze gene expression, these models have proven to be effective tools for supporting clinical decisions, sometimes proving superior to human experts. More than just efficient classifiers, these systems have also proven to be interpretable, robust, and applicable in real clinical scenarios. Both the image-based approach [15] and the gene expression-based approach [49] have faced and overcome challenges typical of medical practice and biomedical research: sample scarcity, high dimensionality, and the need for models with good overall performance, but also **confidence in the prediction of clinically critical cases**.

These findings provide the **conceptual support needed to explore a more specific direction**: the use of ML models to discover and validate **miRNAs as biomarkers** in oncology. Like coding genes, miRNAs carry rich and discriminative information about the biological state of cells and have shown promise in the stratification of tumor subtypes. The next section addresses precisely this line of research, focusing on how ML has been used to reveal miRNAs signatures with diagnostic and prognostic value - a foundation point for the objectives of this dissertation.

2.3.2 ML unravelling miRNAs as biomarkers

As previously discussed, artificial intelligence models have demonstrated proficiency in the task of classificating tumor type, or subtype, by either analysing clinical images, high-dimensional genomic data and other forms of data. In this thesis, our focus is into a more specific and pertinent domain: the identification of miRNAs as biomarkers in oncological contexts through ML techniques. We know for a fact that miRNAs are a class of small molecules that have been demonstrated to possess a substantial regulatory capacity over

gene expression 2.1.2, but even though this capacity distinguishes them as optimal candidates for utilization as molecular biomarkers, the number of xpressed in human tissues, in conjunction with their variability across individuals, demands the implementation of robust computational methodologies.

In this context, ML models have gained prominence as powerful tools for revealing latent patterns in miRNAs expression data, allowing the identification of subsets with diagnostic, prognostic, or tumor subtype stratification value. This research trajectory is particularly auspicious, as it proffers more readily implementable, non-invasive methodologies that can be substantiated within a clinical environment. That's what we are going to explore in this section, where we we will present three papers that illustrate different stages of this scientific effort: the first being a general reference of the type of work that we will be analyzing giving us a glampse of how ML can work in this context; the second one being a more detailed pipeline applied to the identification of miRNAs as biomarkers for gastric cancer, and lastly, the intersection of this kind of approach with which is the central carcinoma focus of this thesis.

The first example of work that we will address in this subsection focused on a major challenge in modern oncology: "Is there a reliable and clinically relevant molecular biomarker for the diagnosis and prognosis of gastric cancer?" Despite growing evidence of the regulatory role of miRNAs in tumor progression and aggressiveness, their biological complexity (as already mentioned in section 1.2) makes it difficult to select those with true clinical value from among the thousands that exist, and despite these barriers, considering that gastric cancer has one of the lowest survival rates in a 5 year period (between 20% and 30% survival rate) with a big share due to late detection, usually when the tumor is already in a advanced stage and metastasized. Given this scenario, the authors Azari et al. [9] propose a methodical and well-structured ML-based approach, as we can see in the workflow chart, to automate the process of discovering these miRNAs that may be biomarkers, where the result also has added robustness so that it can be replicated in a real clinical environment.

The research was based on data from 576 samples from gastric cancer patients, collected from the TCGA repository (previously used on other studies), including miRNAs expression profiles and associated clinical information - this clinical information includes patient characteristics such as age, gender, among others - adding 1,882 features whose expression in a real context is not uniform. Therefore, similar to the study on BC subtyiping with ML, a differential expression analysis was performed, resulting in a reduction to only 220 of these molecular regulators (a much more acceptable and interpretable value than the 1882 previously). Following this extensive data preprocessing, the authors set up a classification pipeline consisting of five classic ML algorithms: Support Vector Machine, Random Forest, Decision Tree, Logistic Regression, and K-Nearest Neighbors 2.3.1 (all of which have been discussed previously except for Random Forest, which consists of a set of several DTs), all of which were evaluated using metrics such as F1-score, *AUC*,

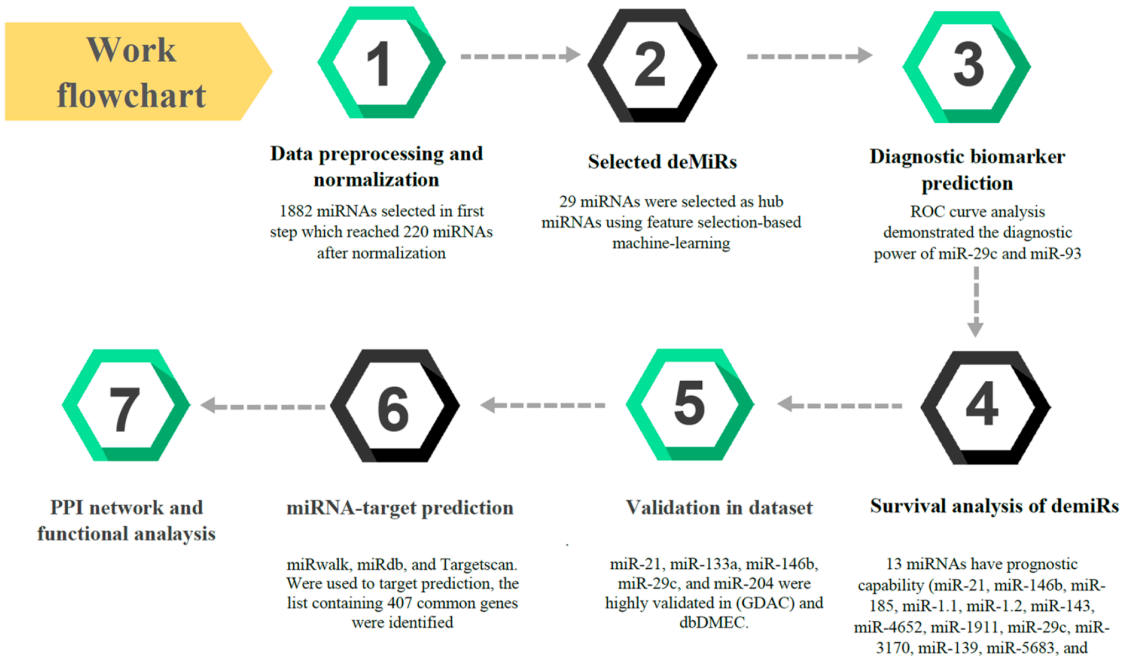


Figure 2.14: Fluxogram of the methodology used in this study, including all the results obtained on each step. [9]

and confusion matrices.

Table 2.2: Performance of the evaluated classification algorithms.

Algorithm	**Accuracy (%)**	***AUC* (%)**
DTS	88	47.0
Random Forest	**93**	39.5
SVM	**93**	**88.5**
KNN	**93**	41.7
Logistic	**93**	88.0

With regard to the performance of the predictive models (as presented in table 2.2), the results show substantial differences between the classifiers. Although four of the five models - all except Decision Trees - achieved an accuracy of 93%, the analysis of the area under the curve (*AUC*) revealed a more complex and informative picture. SVM stood out as the most balanced model, achieving an *AUC* of 88.5%, which indicates a strong discriminatory capacity (i.e. the model not only gets it right often, but also assigns probabilities reliably). In contrast, models such as KNN and RF, despite their high accuracy, obtained very low *AUCs* (41.7% and 39.5%, respectively), suggesting poor calibration of probabilistic predictions and possible overfitting to the training set. The Decision Tree (DT), with slightly lower accuracy (88%) and an *AUC* of only 47%, showed a more modest performance compared to the others, probably due to its vulnerability to variance in data sets with more noise. Finally, logistic regression also showed robust performance, with

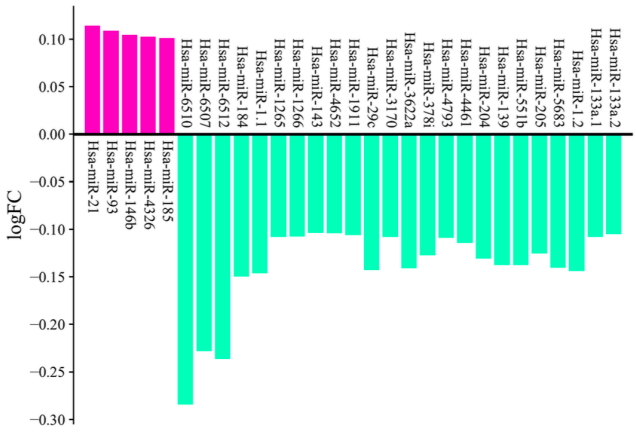


Figure 2.15: Differential regulation levels of the 29 miRNAs selected by the SVM model in the study by Azari et al. [9]. Positive logFC (log fold-change) values indicate overexpressed miRNAs (magenta) and negative values indicate underexpressed miRNAs (green) in gastric cancer samples compared to healthy tissue.

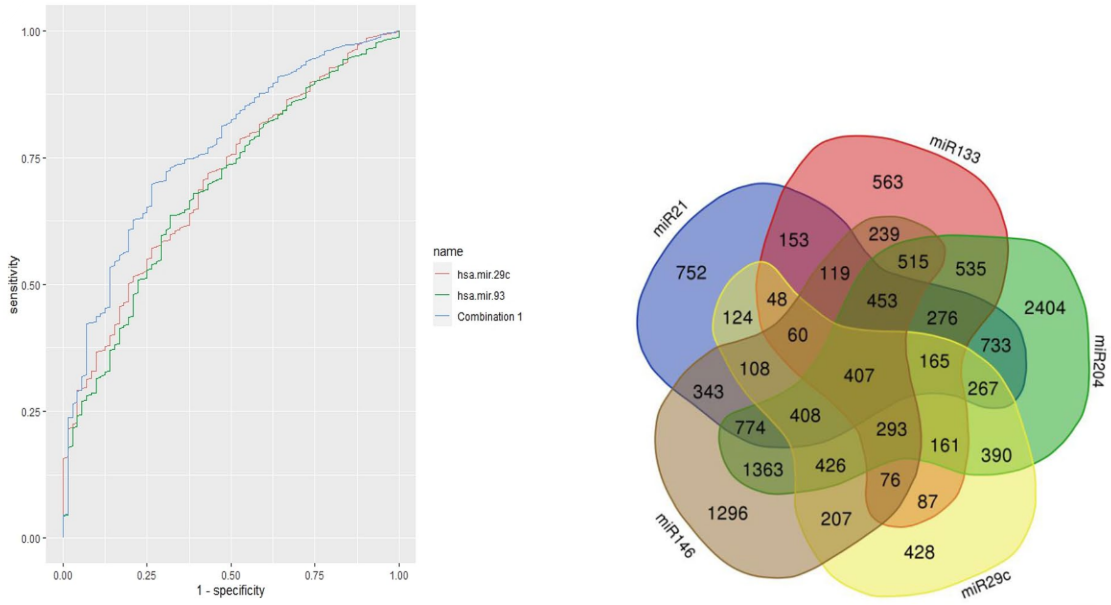
an *AUC* of 88%, very close to that obtained by SVM. However, SVM offered better generalization or sensitivity, justifying its selection as the final model (as already discussed in the work by Wu et al. [49]). This analysis highlights the importance of considering multiple metrics in model evaluation, especially in clinical contexts, where the reliability of the assigned probabilities - and not just the hit rate - can be decisive for a safe medical decision.

Considering the chosen model, a processing step was performed where the most important features were selected using a heatmap analysis, which helps to identify patterns and the relevance of miRNAs in this disease. This step resulted in the reduction of 220 candidates to only 29 (5 of which are significantly up-regulated and 24 considerably down-regulated - in Figure 2.15), which is a very important result as it makes the interpretation of the relationships between these potential biomarkers much more human-friendly.

After identifying 29 candidate miRNAs from the reduced set of 220 miRNAs, the authors proceeded to a validation and refinement stage. This phase aimed to ensure that the selected miRNAs not only stood out statistically in the training set but also maintained biological relevance and predictive robustness in independent scenarios. To this end, cross-validation was performed in external databases, such as the Global Data Assembly Centres (GDAC) and dbDMEC, which aggregate information from public repositories such as GEO, SRA, ArrayExpress, and TCGA. This process allowed us to identify five miRNAs with consistent differential expression in multiple cancer contexts: hsa-miR-21, hsa-miR-133a, hsa-miR-146b, hsa-miR-29c, and hsa-miR-204, all classified as "highly validated". These markers stood out for their high ability to discriminate between healthy and pathological states and, at the same time, predict patient prognosis, positioning themselves as a potential tool of clinical value for the diagnosis and monitoring of gastric cancer. Complementarily, ROC curve analyses were conducted to estimate the diagnostic potential of

each miRNAs, as well as survival analyses to assess their prognostic value, demonstrating a a better indicator performance when combining two miRNAs together (Figure 2.16a).

However, the study's major revelation is not limited to predictive capacity. The final set of four/five selected miRNAs (hsa-miR-21, hsa-miR-133a, hsa-miR-146b, and hsa-miR-29c + hsa-miR-204) underwent additional functional validation, which allowed the identification of the genes targeted by the selected microRNA. This identification was achieved by constructing a protein interaction network (PPI) (Figure 2.16b). This phase confirmed that the genes regulated by these miRNAs are involved in processes essential to gastric carcinogenesis, such as Wnt signaling or epigenetic regulation. hsa-miR-29c stood out above all for its simultaneous predictive value for diagnosis and prognosis, reinforcing its clinical potential as a dual biomarker.



(a) (b)

Figure 2.16: (a) ROC curve comparing the diagnostic performance of individual microRNAs (hsa-miR-29c and hsa-miR-93) versus their combination. The combined model (blue) shows superior sensitivity and specificity, indicating improved discriminative power for gastric cancer classification.

(b) Venn diagram illustrating the overlap of predicted gene targets among five microRNAs (miR-21, miR-133, miR-204, miR-146, and miR-29c) associated with gastric cancer. The central overlap of 426 genes indicates shared regulatory targets across all five miRNAs, suggesting involvement in common biological pathways. Unique and partially overlapping regions highlight miRNA-specific and combinatorial regulatory potential, supporting their relevance as diagnostic and prognostic biomarkers.

It's now a fact that identifying miRNAs that are promising biomarkers requires a study involving several different stages, as we saw in the study by Azari et al. [9], but the impact of the results obtained is of such magnitude that it fully justifies this complexity. From statistical pre-selection to functional and biological validation, each phase contributed to

ensuring that the identified miRNAs were not only statistically relevant but also biologically plausible and clinically actionable. The final panel of four miRNAs, validated at multiple levels, represents a concrete contribution to the construction of more accessible, interpretable, and potentially applicable diagnostic and prognostic tools in clinical practice. This work not only illustrates the power of ML in biomarker discovery, but also establishes a solid methodological foundation that can be replicated and adapted to other diseases - including BC, the focus of this dissertation.

If the previous study demonstrated how it is possible, through a robust machine learning pipeline, to reduce thousands of candidate miRNAs to a small panel with proven clinical potential for gastric cancer, it is natural to ask: is it possible to apply the same principles of statistical selection, biological validation, and predictive modeling in a context closer to the focus of this dissertation, namely This is precisely the proposal of the work of Rehman et al. [40], which starts from a clinically identified set of miRNAs associated with BC and seeks, with the aid of ML algorithms, not only to validate their relevance, but also to build classifiers capable of distinguishing between healthy and cancerous tissue with high accuracy. This study introduces a perspective that complements that of the previous work: while Azari et al. [9] take an exploratory approach to biomarker discovery, this new work focuses on the next step: the verification and validation of already known miRNAs 2.3, assessing the extent to which they are, in fact, discriminative and clinically actionable. It is this transition, from discovery to validation and based on the workflow chart used in this study 2.17, that we will analyze in detail in the following paragraphs.

Table 2.3: List of miRNAs clinically verified.

miRNA [16]			
hsa-mir-10b	hsa-let-7d	hsa-mir-206	hsa-mir-34a
hsa-mir-125b-1	hsa-let-7f-1	hsa-mir-17	hsa-mir-27b
hsa-mir-145	hsa-let-7f-2	hsa-mir-335	hsa-mir-126
hsa-mir-21	hsa-mir-206	hsa-mir-373	hsa-mir-101-1
hsa-mir-125a	hsa-mir-30a	hsa-mir-520c	hsa-mir-101-2
hsa-mir-17	hsa-mir-30b	hsa-mir-27a	hsa-mir-146a
hsa-mir-125b-2	hsa-mir-203a	hsa-mir-221	hsa-mir-146b
hsa-let-7a-2	hsa-mir-203b	hsa-mir-222	hsa-mir-205
hsa-let-7a-3	hsa-mir-213	hsa-mir-200c	
hsa-let-7c	hsa-mir-155	hsa-mir-31	

The study by Rehman et al. [40] focuses on validating a set of miRNAs previously identified as clinically relevant for BC, using a ML-based approach. Unlike exploratory

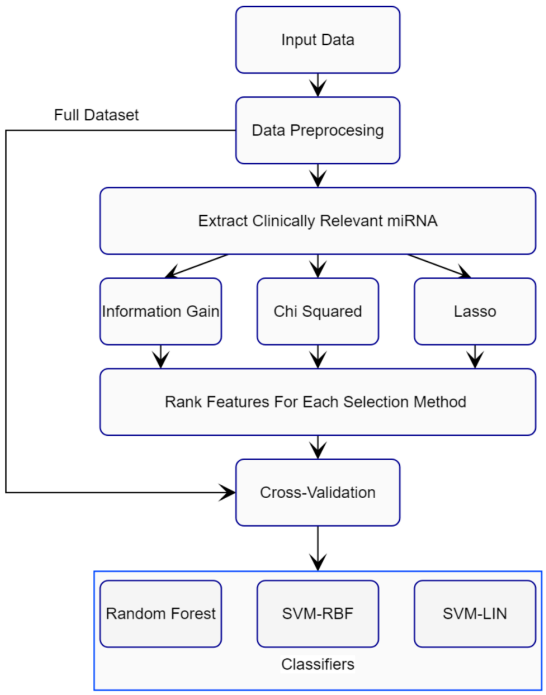


Figure 2.17: Workflow chart used in the following study.

methods, this work starts from an established list of candidates and evaluates their discriminatory power in distinguishing healthy from tumor tissue. The dataset comprises 1207 breast cancer samples with 1881 miRNAs expression profiles from the TCGA-BRCA repository, already normalized and labeled. Notably, the class distribution is balanced, reducing bias in classification and strengthening the reliability of evaluation metrics. Given the high dimensionality, an initial feature cleaning step (including removal of null values) was performed, reducing the set to 1626 features. Based on prior studies, the authors manually selected the 36 most promising miRNAs for classification. Three complementary feature selection methods were then applied:

(a) *Information Gain* - a concept from information theory, consists of reducing uncertainty (entropy) by dividing the data based on a given attribute. The greater the information gain, the better that attribute is for classifying the data.

(b) *Chi-Squared* - the objective of this statistical method is to assess whether there is a relationship between a feature and the target variable. If this is verified, then it means that this feature is potentially useful for the model.

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- O: observed frequency.
- E: expected frequency

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

(a)

$$\text{Information Gain} = \text{Entropy(before)} - \sum_{j=1}^{K} \text{Entropy}(j, \text{after})$$

(b)

Figure 2.18: (a) Entropy formula. (b) Information Gain formula.

(c) *LASSO* - this technique starts with a simple linear regression and during training automatically chooses the most important variables, eliminating irrelevant ones by forcing their coefficients to zero. All of this works with a mechanism that penalizes irrelevant features during training [52].

Now, with these 36 miRNAs, the authors moved on to the predictive modeling phase, testing the effectiveness of different classifiers with progressively smaller subsets of miRNAs. Three main classification algorithms were used (Support Vector Machine with linear kernel, SVM with RBF kernel, and Random Forest) evaluated according to several relevant metrics: accuracy, F1-score, sensitivity, specificity, and AUC. Validation was conducted using 10-fold cross-validation, ensuring the reliability of the results.

One of the most relevant results of this study was the finding that the use of only three highly informative miRNAs could support high-quality classifications, with performance comparable (or superior) to that obtained with the complete set of 1881 miRNAs. This has significant practical implications, as it makes the models simpler, more interpretable, and potentially easier to translate into a clinical context. Although the overall accuracy is very high in all scenarios (often above 99%), the authors point out that this metric can be misleading in an unbalanced dataset. For this reason, they placed particular emphasis on the analysis of sensitivity (recall) and specificity,essential metrics in the healthcare context, where the correct identification of patients (avoiding false negatives) and non-patients (avoiding false positives) is critical.

The analysis of the results shows that sensitivity remained consistently high in all models (in many cases above 0.99), reflecting the strong ability of the classifiers to correctly identify tumor samples. Specificity, in turn, showed clear improvements after feature selection, rising from more modest values (such as 0.875 in SVM-RBF without selection) to levels above 0.98 in scenarios with only three selected miRNAs (e.g., IG-3 and CHI2-3 with SVM-RBF). This improvement reveals that reducing dimensionality helped eliminate noise and avoid misclassifying healthy samples as tumorous. Additionally, the AUC metric remained consistently high, with values close to or even equal to 1.00, reinforcing the robustness of the classifiers even in scenarios with a reduced number of

variables. The results also show that the three feature selection methods (as seen in (a) , (b) and (c)) led to very similar performances, with no clear advantage of one over the others.

The results obtained in the two studies analyzed in this subsection clearly demonstrate the central role that miRNAs can play in the future of cancer diagnosis. From the discovery of new candidates through exploratory pipelines based on ML, as illustrated in the work of Azari et al. [9], to the rigorous validation of miRNAs previously identified as clinically relevant, as presented in Rehman et al. [40], these small molecular regulators consistently prove to be strong indicators of tumor presence and progression. The ability to reduce thousands of candidates to a small panel of reliable biomarkers with clinical applicability, without compromising predictive performance, represents a significant step toward more accessible, non-invasive, and interpretable diagnostic tools. These findings reinforce the fundamental premise of this dissertation: ML models, when properly designed and validated, can effectively exploit the complexity of miRNAs expression profiles, not only to detect cancer, but also to support more specific tasks such as tumor subtype stratification, with real potential for clinical translation.

2.3.3 Comparison with thesis approach

The analysis of the studies presented throughout this chapter allows us not only to understand the state of the art in the use of ML algorithms for biomarker identification, but also how these computational resources can become key elements in the classification of tumor masses, directly outlining some possible paths for this thesis.

One of the first aspects to highlight is the **importance of differential expression analysis as an initial step in feature selection**. This approach was common to several of the studies analyzed, namely Wu et al. [49] and Azari et al. [9], and proved effective in reducing dimensionality in scenarios with high feature density, such as miRNA expression data. This strategy allows the analysis to focus on more informative subsets, reducing the risk of overfitting and making the models more robust. In addition, the studies mentioned demonstrate that this initial selection can (and should) be complemented with additional computational methods. Techniques such as LASSO regression, Information Gain, Chi-Squared, and Recursive Feature Elimination (RFE) have been explored in different papers and show that combining statistical approaches with model-based methods increases the probability of finding truly discriminative features. This idea will be incorporated into the pipeline developed in this thesis, with the aim of building more reliable and interpretable models.

Another valuable methodological element drawn from the analyzed works is the **rigorous validation of the identified models and biomarkers**, either through internal cross-validation or through the use of external data. The study by Azari et al. [9], for example, stands out for having validated miRNA candidates in several public repositories such as

GEO, dbDMEC, and GDAC. This type of cross-validation with other sources, even if partial, **gives greater solidity and generalization to the results obtained**, a key aspect that will be taken into account in the structure of this dissertation, with the consideration of robust validation strategies. Regarding the **origin of the data**, the reviewed studies reinforce the relevance of using **The Cancer Genome Atlas (TCGA)** repository as a reliable, comprehensive, and widely used data source in similar studies. This will also be the base repository for this study, ensuring comparability with previous studies and data quality.

It is also important to mention that both sutides showed that **the performance of the models does not necessarily depend on all available features**. On the contrary, several models have shown excellent performance with only a fraction of the input data (such as subsets of 3 to 10 miRNAs) highlighting the feasibility of building simpler, more interpretable, and economically applicable models in a clinical context. The possibility of finding a small group of miRNAs that would have the same performance as using all of them is the real breakthrough of a thesis like this, since it would open doors to more affordable testing, leading to more prevention tests and the possibility of earlier detection of these malignous cells.

Finally, the workflow diagrams proposed in the reference studies - with special emphasis on that of Azari et al. [9] - offer a clear, sequential, and methodologically rigorous visual representation of all the steps involved in the discovery and validation of molecular biomarkers using ML. These flowcharts not only facilitate understanding of the process, but also help to ensure the reproducibility of studies and effective communication of methods between researchers. In the particular case of this dissertation, the structure presented in these works serves as direct inspiration for the design of the pipeline that will be implemented here, covering all critical phases: from data acquisition and pre-processing, through careful feature selection and predictive modeling, to internal statistical validation and model performance analysis. This systematic approach ensures alignment with best practices in the literature and maximizes the translational potential of the results obtained.

In short, the literature review allows us to consolidate a solid methodological basis, anchored in studies that have demonstrated the effectiveness of machine learning in the discovery and validation of molecular biomarkers. From this evidence, it is clear that well-structured approaches such as combining differential expression analysis, careful feature selection, and rigorous validation offer a promising path to building robust and clinically relevant predictive models. This accumulated knowledge will now be adapted and applied to the specific context of this dissertation, which seeks to identify miRNA signatures with discriminatory power for the classification of breast cancer subtypes. The next chapter describes in detail how this approach will be implemented.

3

Methodology and Pilot Study

Small introduction to the Chapter.

3.1 Data Acquisition and Characterization

3.1.1 The Primary Cohort (N=256): Source and Clinical Features.

3.1.2 miRNA Expression Profiles: Data distribution and normalization.

3.2 Data Pre-processing Pipeline

3.2.1 Handling missing values and outliers.

3.2.2 Dimensionality reduction and feature scaling.

3.3 The Supervised Learning Framework

3.3.1 Algorithm Selection: Why XGBoost, Random Forests, and SVM?

3.3.2 Optimization: Hyperparameter tuning and Cross-Validation strategy.

3.4 Unsupervised Exploration: Finding Hidden Patterns

3.4.1 Clustering Analysis (K-Means/Hierarchical): Do clinical subtypes match genetic clusters?

3.4.2 Silhouette Analysis and Cluster Stability.

4

Scalling the Analysis: the Expanded Cohort

4.1 Transition to the Larger Dataset

4.1.1 Identification and Alignment: Merging the 256-subset with the 700-entry master set.
- Why do we say that our original dataset is actually a subset of this master set? (optional)
- What are the key differences between the subset and this master set?

4.1.2 Comparative Analysis: Does the larger dataset change the data distribution?

4.2 Applying the Refined Pipeline

4.2.1 Re-training the Supervised Models.

4.2.1 Model Evaluation

mencionar no texto antes da tabela que estamos a usar a mesma arquitetura para o dataset grande

Rever esta tabela com o professor para saber se é legível

Table 4.1: Homogenous models. Comparison with previous results (miRNAs only).

	Overall				**F1-Score per subtype**			
Model	P	R	F1	Acc	Basal	HER-2	Lum A	Lum B
DIABLO	69.06 75.27 (−6.21)	70.19 67.26 (+2.93)	68.51 67.69 (+0.82)	70.19 73.47 (−3.28)	77.42 **100.00** (−22.58)	46.15 40.00 (+6.15)	78.74 80.77 (−2.03)	43.24 50.00 (−6.76)
Decision Tree	74.90 57.35 (+17.55)	75.00 56.40 (+18.6)	74.92 56.79 (+18.13)	75.00 57.14 (+17.86)	91.43 87.50 (+3.93)	58.82 36.36 (+22.46)	**81.42** 61.90 (+19.52)	51.16 41.38 (+9.78)
Logistic Reg.	72.63 80.26 (−7.63)	70.19 **75.60** (−5.41)	70.99 73.50 (−2.51)	70.19 **77.55** (−7.36)	78.79 94.12 (−15.33)	44.44 61.54 (−17.1)	76.64 **85.71** (−9.07)	**60.00** **52.63** (+7.37)
XGBoost	**77.05** **86.96** (−9.91)	**77.88** 74.40 (+3.48)	**77.21** **76.73** (+0.48)	**77.88** 75.51 (+2.37)	**100.00** **100.00** (0.00)	**70.59** **80.00** (−9.41)	81.36 76.92 (+4.44)	51.28 50.00 (+1.28)

*Note: Each cell displays the result of this experiment in the main dataset (top), the previous result (meaning, on the subset of the dataset) (middle), and the calculated delta (bottom). **Bold** indicates best-in-class for each dataset comparison.*

Table 4.2: Heterogeneous models (miRNAs + Clinical Data). Comparison with previous results.

	Overall				F1-Score per subtype			
Model	P	R	F1	Acc	Basal	HER-2	Lum A	Lum B
DIABLO	69.06 78.47 (−9.41)	70.19 72.62 (−2.43)	68.51 73.82 (−5.31)	70.19 79.59 (−9.4)	77.42 100.00 (−22.58)	46.15 40.00 (+6.15)	78.74 85.71 (−6.97)	43.24 69.57 (−26.33)
Decision Tree	71.88 62.01 (+9.87)	72.12 63.69 (+8.43)	71.97 62.45 (+9.52)	72.12 61.22 (+10.9)	84.85 88.89 (−4.04)	50.00 54.55 (−4.55)	81.03 61.90 (+19.13)	46.51 44.44 (+2.07)
Logistic Reg.	72.81 77.22 (−4.41)	71.15 75.60 (−4.45)	71.70 75.57 (−3.87)	71.15 79.59 (−8.44)	76.47 94.12 (−17.65)	47.06 54.55 (−7.49)	77.78 **86.96** (−9.18)	**61.22** 66.67 (−5.45)
XGBoost	**77.81** **90.06** (−12.25)	**78.85** **79.17** (−0.32)	**77.96** **82.17** (−4.21)	**78.85** **83.67** (−4.82)	**97.14** **100.00** (−2.86)	**70.59** **66.67** (+3.92)	**83.05** 85.11 (−2.06)	52.63 **76.92** (−24.29)

Note: Each cell displays the current result (top), the previous result from chapter **??** *(middle), and the calculated delta (bottom).* ***Bold*** *indicates best-in-class for each dataset.*

No texto a seguir à tabela, adicionar uma análise da diferença de valores entre implementações

4.2.2 Performance Metrics: Accuracy, F1-Score, and AUC-ROC for each Breast Cancer Subtype.

4.3 Model Interpretability and Biomarker Discovery

4.3.1 Global Interpretability: Feature Importance (SHAP/Gain) across the whole cohort.

4.3.2 Local Interpretability: Identifying subtype-specific miRNA signatures.

4.3.3 Clinical Relevance: Mapping the top miRNAs to known biological pathways.

5

Results and Discussion

5.1 Comparative Performance

Discussion on how model performance evolved from the small dataset to the large dataset (did it generalize better?).

5.2 miRNA Biomarker Evaluation

Consistency of ”important” miRNAs across different models.

Are these miRNAs known in literature, or are they novel candidates for personalized medicine?

5.3 Practical Implications for Clinicians

How these models can aid in faster, data-driven subtyping.

5.4 Limitations of the Study

Data imbalance, biological noise in miRNA, etc.

6

Conclusions and Future Work

6.1 Concluding Remarks

Summary of the effectiveness of classical ML on miRNA data.

6.2 Future Work

Integrating Genetic data and Protein data to deepen the biological mechanics with AI algorithms (ML and DL)

Testing Deep Learning architectures as data continues to grow.

Bibliography

[1] B. Adamo et al. "Clinical implications of the intrinsic molecular subtypes of breast cancer." In: *Breast* 24 Suppl 2 (2015), S26–35. DOI: 10.1016/j.breast.2015.07.008.

[2] AI Lab MTI Vietnam. *Imagem ilustrativa de análise de dados*. https://ailab.mti-vietnam.vn/wp-content/uploads/2020/12/image-2.png. Imagem retirada do site. 2020. (Visited on 2025-07-03).

[3] Saad Albawi, T. A. Mohammed, and Saad Al-Zawi. "Understanding of a convolutional neural network". In: *2017 International Conference on Engineering and Technology (ICET)* (2017), pp. 1–6. DOI: 10.1109/ICENGTECHNOL.2017.8308186.

[4] Bruce Alberts et al. *Molecular Biology of the Cell*. Accessed via NCBI Bookshelf. New York, NY, 2002. URL: https://www.ncbi.nlm.nih.gov/books/NBK26887/.

[5] Bruce Alberts et al. *Molecular Biology of the Cell*. Accessed via NCBI Bookshelf. New York, NY, 2002. URL: https://www.ncbi.nlm.nih.gov/books/NBK26887/.

[6] Analytics Vidhya. *Processo de análise de dados*. https://www.analyticsvidhya.com/wp-content/uploads/2016/08/Artificial-Intelligence-Neural-Network-Nodes.jpg. Imagem retirada do site. s.d. (Visited on 2025-07-03).

[7] Arul Edwin Raj Anthony Muthu, Sundaram Muniasamy, and Thirassama Jaya. "Thermography based breast cancer detection using self-adaptive gray level histogram equalization color enhancement method". In: *International Journal of Imaging Systems and Technology* 31 (2020-10). DOI: 10.1002/ima.22488.

[8] M. Arnold et al. "Current and future burden of breast cancer: Global statistics for 2020 and 2040". In: *The Breast : Official Journal of the European Society of Mastology* 66 (2022), pp. 15–23. DOI: 10.1016/j.breast.2022.08.010.

[9] H. Azari et al. "Machine learning algorithms reveal potential miRNAs biomarkers in gastric cancer". In: *Scientific Reports* 13 (2023). DOI: 10.1038/s41598-023-32332-x.

[10] C. Blenkiron et al. "MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype". In: *Genome Biology* 8 (2007), R214–R214. DOI: 10.1186/gb-2007-8-10-r214.

[11] Freddie Bray et al. "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: A Cancer Journal for Clinicians* 74.3 (2024), pp. 229–263. DOI: https://doi.org/10.3322/caac.21834. eprint: https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21834. URL: https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21834.

[12] Joel S Brown et al. "Updating the Definition of Cancer". In: *Molecular Cancer Research* 21 (2023), pp. 1142–1147. DOI: 10.1158/1541-7786.MCR-23-0411.

[13] Gloria M. Calaf et al. "The Role of MicroRNAs in Breast Cancer and the Challenges of Their Clinical Application". In: *Diagnostics* 13 (2023). DOI: 10.3390/diagnostics13193072.

[14] P. Eroles et al. "Molecular biology in breast cancer: intrinsic subtypes and signaling pathways." In: *Cancer treatment reviews* 38 6 (2012), pp. 698–707. DOI: 10.1016/j.ctrv.2011.11.005.

[15] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". In: *Nature* 542.7639 (2017), pp. 115–118. DOI: 10.1038/nature21056. URL: https://doi.org/10.1038/nature21056.

[16] S. W. Fu, L. Chen, and Y. G. Man. "miRNA Biomarkers in Breast Cancer Detection and Management". In: *Journal of Cancer* 2 (2011), pp. 116–122. DOI: 10.7150/jca.2.116. URL: https://doi.org/10.7150/jca.2.116.

[17] Maryellen L. Giger. "The Role of Artificial Intelligence in Early Cancer Diagnosis". In: *Radiologic Clinics of North America* 59.6 (2021), pp. 933–946. DOI: 10.1016/j.rcl.2021.06.006. URL: https://doi.org/10.1016/j.rcl.2021.06.006.

[18] L. Gulyaeva and N. E. Kushlinskiy. "Regulatory mechanisms of microRNA expression". In: *Journal of Translational Medicine* 14 (2016). DOI: 10.1186/s12967-016-0893-x.

[19] S. Hammond. "An overview of microRNAs." In: *Advanced drug delivery reviews* 87 (2015), pp. 3–14. DOI: 10.1016/j.addr.2015.05.001.

[20] P. T. Ho, I. Clark, and L. Le. "MicroRNA-Based Diagnosis and Therapy". In: *International Journal of Molecular Sciences* 23 (2022). DOI: 10.3390/ijms23137167.

[21] R. Hong and Bing-he Xu. "Breast cancer: an up-to-date review and future perspectives". In: *Cancer Communications* 42 (2022), pp. 913–936. DOI: 10.1002/cac2.12358.

[22] N. Howlader et al. "Differences in Breast Cancer Survival by Molecular Subtypes in the United States". In: *Cancer Epidemiology, Biomarkers and Prevention* 27 (2018), pp. 619–626. DOI: 10.1158/1055-9965.EPI-17-0627.

[23] International Agency for Research on Cancer. *Global Cancer Observatory - Cancer Today: Breast cancer incidence heatmap (2022)*. https://gco.iarc.fr/today/en/dataviz/maps-heatmap?mode=population&types=0&cancers=20. Accessed: 2025-06-22. 2022.

[24] Bahzad Taha Jijo and Adnan Mohsin Abdulazeez. "Classification Based on Decision Tree Algorithm for Machine Learning". In: *Journal of Applied Science and Technology Trends* 2.1 (2021), pp. 20–28. ISSN: 2708-0757. DOI: 10.38094/jastt20165. URL: http://www.jastt.org/index.php/index.

[25] Bahzad Taha Jijo and Adnan Mohsin Abdulazeez. "Classification Based on Decision Tree Algorithm for Machine Learning". In: *Journal of Applied Science and Technology Trends* 02.01 (2021), pp. 20–28. ISSN: 2708-0757. DOI: 10.38094/jastt20165. URL: http://www.jastt.org/index.php/index.

[26] E. Koonin and A. Novozhilov. "Origin and Evolution of the Universal Genetic Code." In: *Annual review of genetics* 51 (2017), pp. 45–62. DOI: 10.1146/annurev-genet-120116-024713.

[27] Rosalind C. Lee, Rhonda L. Feinbaum, and Victor Ambros. "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14". In: *Cell* 75 (1993), pp. 843–854. DOI: 10.1016/0092-8674(93)90317-3.

[28] S. Loibl et al. "Breast cancer". In: *The Lancet* 397 (2021), pp. 1750–1769. DOI: 10.1016/S0140-6736(20)32381-3.

[29] Yuxun Luo et al. "Machine learning in the development of targeting microRNAs in human disease". In: *Frontiers in Genetics* 13 (2023), p. 1088189. DOI: 10.3389/fgene.2022.1088189. URL: https://doi.org/10.3389/fgene.2022.1088189.

[30] D. Mahalingam, E. Vagia, and M. Cristofanilli. "The Landscape of Targeted Therapies in TNBC". In: *Cancers* 12 (2020). DOI: 10.3390/cancers12040916.

[31] Bárbara B. Mendes et al. "Nanodelivery of nucleic acids". In: *Nature Reviews Methods Primers* 2.24 (2022). Article citation ID: (2022) 2:24. DOI: 10.1038/s43586-022-00104-y. URL: https://doi.org/10.1038/s43586-022-00104-y.

[32] Bárbara B. Mendes et al. "A large-scale machine learning analysis of inorganic nanoparticles in preclinical cancer research". In: *Nature Nanotechnology* (2024-06). DOI: 10.1038/s41565-024-01673-7.

[33] Juan P Muñoz et al. "The Role of MicroRNAs in Breast Cancer and the Challenges of Their Clinical Application". In: *Diagnostics* 13 (2023). DOI: 10.3390/diagnostics13193072.

[34] National Cancer Institute. *What Is Cancer?* Accessed: 2025-06-15. 2021. URL: https://www.cancer.gov/about-cancer/understanding/what-is-cancer.

[35] A. Novozhilov and E. Koonin. "Origin and evolution of the genetic code: The universal enigma". In: *IUBMB Life* 61 (2008). DOI: 10.1002/iub.146.

[36] Charles M. Perou et al. "Molecular portraits of human breast tumours". In: *Nature* 406.6797 (2000), pp. 747–752. DOI: 10.1038/35021093. URL: https://doi.org/10.1038/35021093.

[37] K. Polyak. "Breast cancer: origins and evolution." In: *The Journal of clinical investigation* 117 11 (2007), pp. 3155–63. DOI: 10.1172/JCI33295.

[38] A. Prat et al. "Clinical implications of the intrinsic molecular subtypes of breast cancer." In: *Breast* 24 Suppl 2 (2015), S26–35. DOI: 10.1016/j.breast.2015.07.008.

[39] Leslie A. Pray. "Discovery of DNA structure and function: Watson and Crick". In: *Nature Education* 1.1 (2008), p. 100. URL: https://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397/.

[40] Oneeb Rehman et al. "Validation of miRNAs as Breast Cancer Biomarkers with a Machine Learning Approach." In: *Cancers* (2019). DOI: 10.3390/cancers11030431.

[41] H. Romanowicz, B. Smolarz, and Anna Zadrożna Nowak. "Breast Cancer—Epidemiology, Classification, Pathogenesis and Treatment (Review of Literature)". In: *Cancers* 14 (2022). DOI: 10.3390/cancers14102569.

[42] Amrit Singh et al. "DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays". In: *Bioinformatics* 35.17 (2019), pp. 3055–3062. DOI: 10.1093/bioinformatics/bty1054. URL: https://doi.org/10.1093/bioinformatics/bty1054.

[43] Yi-Sheng Sun et al. "Risk Factors and Preventions of Breast Cancer". In: *International Journal of Biological Sciences* 13 (2017), pp. 1387–1397. DOI: 10.7150/ijbs.21635.

[44] Sunrayce Biology Authors. *Chapter 91: The Structure of DNA*. BC Open Textbooks. Accessed: 2025-06-16. 2015. URL: https://opentextbc.ca/biology/chapter/9-1-the-structure-of-dna/.

[45] U. Testa, G. Castelli, and E. Pelosi. "Breast Cancer: A Molecularly Heterogenous Disease Needing Subtype-Specific Treatments". In: *Medical Sciences* 8 (2020). DOI: 10.3390/medsci8010018.

[46] J. Watson and F. Crick. "The structure of DNA." In: *Cold Spring Harbor symposia on quantitative biology* 18 (1953), pp. 123–31. DOI: 10.1101/SQB.1953.018.01.020.

[47] Thomas J. Watson. "An empirical study of the naive Bayes classifier". In: 2001. URL: https://api.semanticscholar.org/CorpusID:14891965.

[48] Bruce Wightman, Iva Ha, and Gary Ruvkun. "Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans". In: *Cell* 75 (1993), pp. 855–862. DOI: 10.1016/0092-8674(93)90318-4.

[49] Jiande Wu et al. "Breast Cancer Type Classification Using Machine Learning." In: *Journal of Personalized Medicine* (2021). DOI: 10.3390/jpm11020061.

[50] Mohammed Yaseen and Adnan Mohsin Abdulazeez. "Performance Evaluation Metrics in Machine Learning Models: A Comparative Study". In: *Journal of Soft Computing and Data Mining* 2.2 (2021), pp. 21–31. ISSN: 2710-139X. DOI: 10.30880/jscdm.2021.02.02.003.

[51] S. Yeo and J. Guan. "Breast Cancer: Multiple Subtypes within a Tumor?" In: *Trends in cancer* 3 11 (2017), pp. 753–760. DOI: 10.1016/j.trecan.2017.09.001.

[52] Huaqing Zhang et al. "Feature Selection for Neural Networks Using Group Lasso Regularization". In: *IEEE Transactions on Knowledge and Data Engineering* 32.4 (2020), pp. 659–673. DOI: 10.1109/TKDE.2019.2893266.

[53] Lili Zhu and P. Spachos. "Support Vector Machine and YOLO for a Mobile Food Grading System". In: *Internet of Things* 13 (2021-01), p. 100359. DOI: 10.1016/j.iot.2021.100359.

A

Appendix 1 Covers Showcase

I

Unveiling microRNA Biomarkers for Breast Cancer Sub-typing

This annex includes the full version of the paper titled **"Unveiling microRNA Biomarkers for Breast Cancer Sub-typing using Discriminative Models"**, submitted and accepted to **EPIA 2025** under the category **AI for Medicine**.