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: *July 3, 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: *July 3, 2025*

Abstract

Regardless of the language in which the dissertation is written, usually there are at least two abstracts: one abstract in the same language as the main text, and another abstract in some other language.

Keywords: One keyword, Another keyword, Yet another keyword, One keyword more, The last keyword

Resumo

Independentemente da língua em que a dissertação está escrita, geralmente esta contém pelo menos dois resumos: um resumo na mesma língua do texto principal e outro resumo numa outra língua.

Palavras-chave: Primeira palavra-chave, Outra palavra-chave, Mais uma palavra-chave, A última palavra-chave

Contents

List of Figures

Acronyms

AI Artificial Intelligence *(pp. iii, 3, 12, 13)*

BC Breast Cancer *(pp. iii, 1–3, 5, 12, 13)*

DL Deep Learning *(pp. 3, 5, 12)*

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

miRNAs microRNAs *(pp. 2–5, 12)*

ML Machine Learning *(pp. iii, 3, 5, 12, 15)*

RNA Ribonucleic Acid *(pp. 6–8)*

1 Introduction

1.1 Motivation & Problem Statement

1.1.1 Breast Cancer - A Global Health Challenge

BC is currently one of the biggest public health challenges worldwide. In 2022, more than 2.3 million new cases of BC were diagnosed, resulting in around 665,000 global deaths (Bray et al. [10]). Other studies estimate that BC will continue to not only bethe 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 (Arnold et al. [6]). Figure 1.2 and 1.1 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.

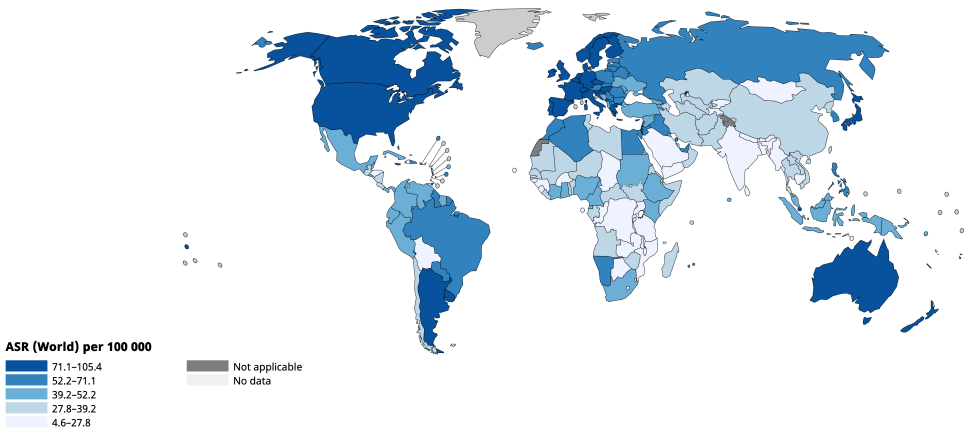
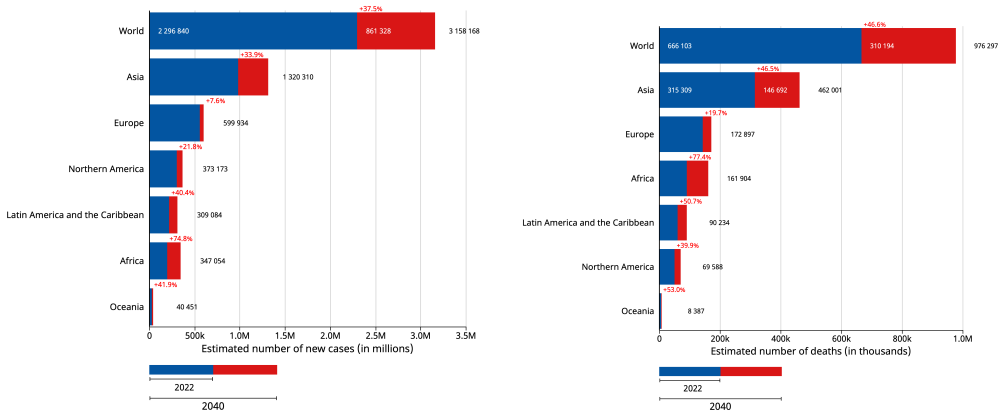


Figure 1.1: Age-standardized incidence rate (ASR, per 100,000 inhabitants) of breast cancer in both sexes in 2022. The data represent global estimates based on International Agency for Research on Cancer [20].



(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). (International Agency for Research on Cancer [20])

BC is characterized by marked biological heterogeneity, manifested in multiple molecular subtypes that exhibit distinct clinical behaviors (Perou et al. [29]). Each subtype exhibits substantial differences in terms of tumor aggressiveness, metastatic potential, and behavior to specific therapies (Prat et al. [31]). Thus, accurate classification of these subtypes is essential to enable personalized therapeutic approaches, with a direct impact on treatment efficacy and disease prognosis (Testa, Castelli, and Pelosi [36]).

1.1.2 Can we improve the classification of Breast Cancer subtypes?

Among the emerging candidates for robust biomarkers for the classification of BC subtypes are microRNAs (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 (Hammond [16]) and are implicated in the regulation of multiple physiological and pathological processes, including various human diseases (Ho, Clark, and Le [17]).

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 Mendes et al. [25] Ho, Clark, and Le [17] and Muñoz et al. [26]. In addition to these aspects, a seminal study by Blenkiron et al. [9] 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 AI, 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 (Luo et al. [23]). 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-based biomarkers — 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. [7] 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 (Ho, Clark, and Le [17] and Muñoz et al. [26]), 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 (Singh et al. [34]), a widely used model. 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, I will be able to explore the applicability of *Conformal Prediction* (Angelopoulos and Bates [4]), which provides statistically based confidence intervals for each prediction (which is widely used in areas where risk must be justified and well-founded, such as finance and healthcare).

The latter is an approach that is still gaining ground in healthcare and, considering our problem, it makes sense to be able to give a prediction based on a confidence interval, giving our model greater transparency and reliability, something particularly relevant in clinical contexts where error must be minimized and uncertainty well characterized.

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

1. Biological heterogeneity:

Biological heterogeneity is characterized by the diversity of living organisms, including species, genotypes, and populations, which exhibit a variety of biological characteristics, such as morphology, physiology, genetics, and biogeography. The human body is a highly complex system in which the behavior of each component depends on its interaction with countless other parts. Exposure to the same treatment by two bodies can result in completely different reactions.

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. [7] where the implementation of a control set was fundamental to the identification of discriminative markers.

The data for this study were previously selected by Dr. Bárbara Mendes and her team based on purely biological selection criteria. These criteria included the scientific interest of the group and molecular characteristics that were considered relevant in the context of the research. Consequently, the dataset I was given consists of 256 biopsies and 888 miRNAs features, as well as 38 clinical data features. This presents a scenario of high dimensionality with a limited number of examples on which to train.

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.

Furthermore, this problem cannot be addressed with data augmentation techniques used in the field of ML, such as *SMOTE* (Synthetic Minority Over-sampling Technique - Blagus and Lusa [8]). As discussed earlier in the topic 1.2, the nature of this data makes reliable synthetic generation unfeasible, which can lead to artificially inconsistent samples ("Extra-terrestial beings", in other words).

Having a control set would be an important step toward increasing the generalizability of our model. This data set 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 (Azari et al. [7]) 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 will integrate and compare different ML and DL approaches (still underexplored for this decease) applied to a problem of high biological and statistical complexity.

In addition to the classification process, this framework will include a statistical analysis component aimed at validating the predictions made in order to give robustness to the decision made by the model. This robustness is particularly relevant in a clinical setting, where transparency and reliability of predictions are essential for possible future translation into medical practice.

Throughout the work, a critical and comparative analysis of the approaches explored 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

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 microRNA reaserch and its applications in breast cancer 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 microRNAs in gene expression regulation, and the characteristics of breast cancer and its subtypes. The second section will review the related work in the field of microRNA research, focusing on the use of microRNAs as biomarkers and subtype classifiers in breast cancer, as well as the challenges and limitations of current approaches.

2.1 Biological Context

The modern understanding of how genetic information is stored, interpreted, and regulated in cells is based on a fundamentalprinciple known as the Central Dogma of Molecular Biology. This concept, first formulated in Pray [32] and Watson and Crick [37], 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) [35]. 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.

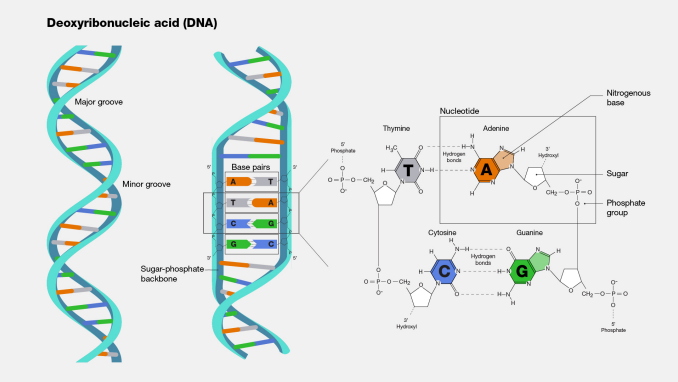


Figure 2.1: Structure of the DNA double helix

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. [2]. 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 [3].

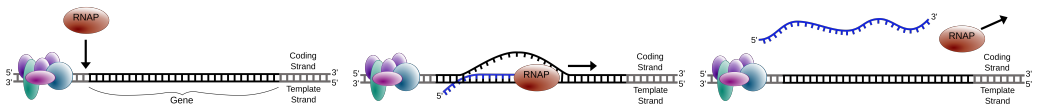


Figure 2.2: Illustration of the transcription mechanism: (a) initiation, (b) elongation, and (c) termination

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 [28] 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 [35].

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 [21].

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 microRNAs (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

MicroRNAs (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 [15, 22, 38]. Instead of encoding proteins, miRNAs act by controlling the production of proteins from genes.

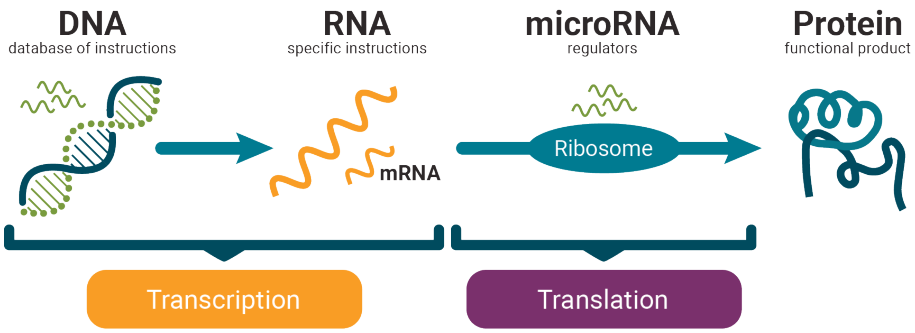


Figure 2.3: The figure shows the process of gene expression: DNA is transcribed into mRNA, which is then translated into protein by the ribosome. microRNAs are shown as regulators acting on the mRNA before translation.

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 miRNA sequence and that of the target mRNA:

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

A study made by Calaf et al. [12] demonstrates the high effienciency of this mechanism of regulation: a **single miRNA 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.

Due to this broad regulatory capacity, miRNAs play a central role in multiple cellular processes such as proliferation, differentiation, apoptosis, and stress response. Consequently, changes in miRNA 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 [15].

Due to their specificity, stability, and direct involvement in relevant molecular mechanisms, miRNAs have been extensively investigated as promising biomarkers for diagnosis, prognosis, and subtype stratification in various diseases—including cancer.

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. [11] and National Cancer Institute [27], 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 [11, 27].

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 [11].

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

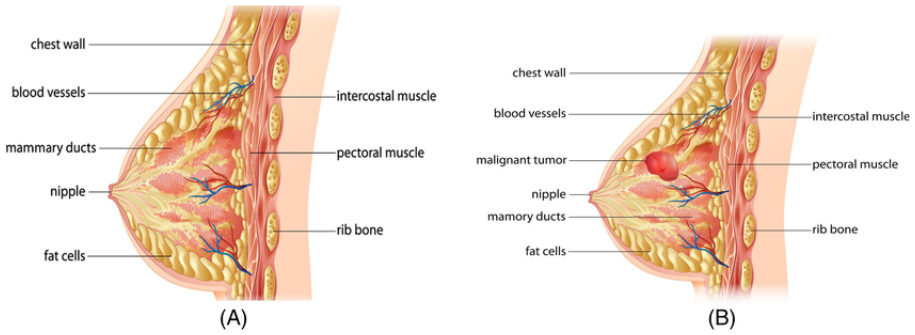


Figure 2.4: In (A) we have a normal breast tissue, while in (B) we can see the presence of a malignant tumor [5].

an increase in the number of cells or changes in their morphology, but which do not yet invade surrounding tissues.

Progression of cancer 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.

2.1.4 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 [18] and Romanowicz, Smolarz, and Nowak [33]. 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.

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 [30].

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, 33].

From a molecular and clinical point of view, **breast cancer 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 [19, 30].

Early detection is crucial for prognosis. When diagnosed in its early stages, breast cancer 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, 18].

The therapeutic approach is 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 [33].

In addition, a study made by Polyak [30] recognized that breast tumors are not static entities. Due to phenomena of intra-tumor heterogeneity and clonal evolution, tumors adapt to the selective pressure of treatments, often leading to the development of therapeutic resistance and disease progression.

Given the molecular complexity and clinical diversity of breast tumors, it was established in the papers Adamo et al. [1] and Prat et al. [31] that breast cancer is not a single disease but rather a collection of biologically distinct entities that arise from a common anatomical site. 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 from various studies, such as Perou et al. [29] and Prat et al. [31], have developed a molecular classification system that subdivides breast tumors into intrinsic subtypes. These subtypes are defined based on the expression status of three key biomarkers: **estrogen receptor (ER)**, **progesterone receptor (PR)**, and **human epidermal growth factor receptor 2 (HER2)** as well as other proliferation indices (e.g., Ki-67) and gene expression patterns. This classification underpins modern precision oncology approaches and has profound implications for therapy and prognosis.

As shown in the table 2.1, breast cancer can be classified into four main intrinsic subtypes based on the expression of these biomarkers and other molecular characteristics:

Recent evidence by Polyak [30] and Yeo and Guan [39] 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.

2.1.5 Nucleic acids as gene therapies

Table 2.1: Summary of intrinsic breast cancer subtypes and typical characteristics of each one.
References: [1], [19], [31], [18], [24].

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 .

2.2 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 [14]. 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.

2.2.1 Leveraging AI models for BC subtype 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. [13], 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.

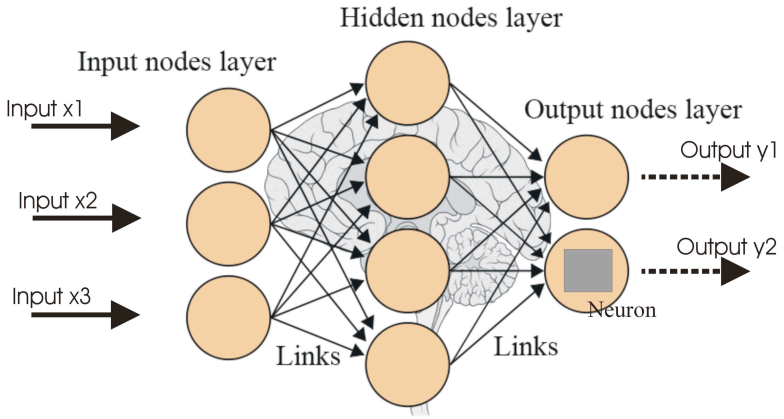


Figure 2.5: Structure of an deep neural network (DNNs). It shows the input, hidden, and output layers, with connections between neurons responsible for processing information.

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 from the image pixels, without the need for specialized preprocessing.

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;

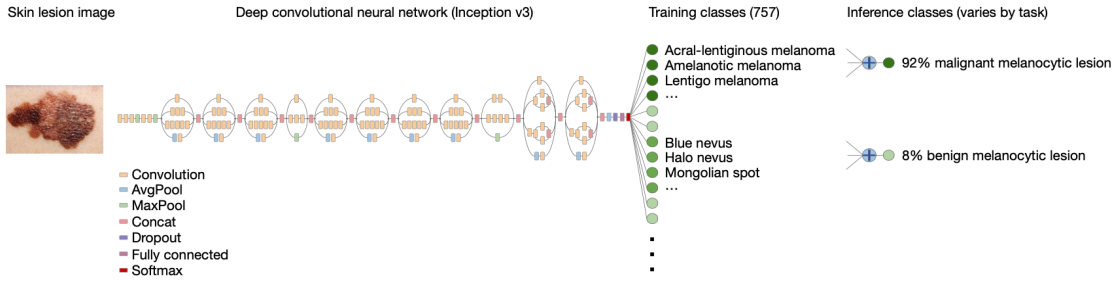


Figure 2.6: Schematic of the CNN used by Esteva et al. [13] with Inception v3 architecture, adapted to classify skin lesions based on clinical images. The network generates a probability distribution over clinical classes, based on a structured medical taxonomy.

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. [13], 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 **??**). 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 **??**) - 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.

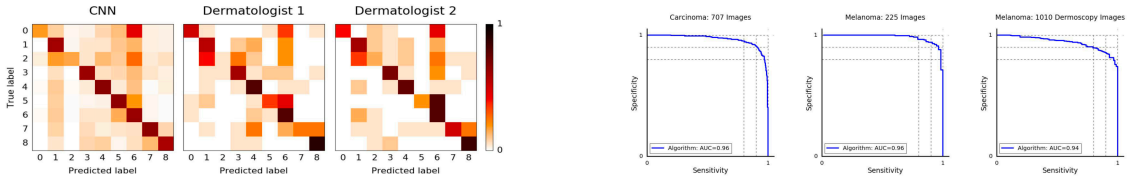


Figure 2.8: Performance evaluation of the convolutional neural network (CNN) in skin lesion classification. (a) Confusion matrices of CNN and two dermatologists. The concentration on the diagonal indicates correct classifications; CNN shows less dispersion and better overall performance [13]. (b) Reliability of the CNN demonstrated by *AUC* curves on a larger, independent dataset [13].

Furthermore, in more complex scenarios with multiple classes (three and nine disease categories), the model maintained **remarkable levels of accuracy** (72.1% and 55.4%), surpassing or equaling human experts. 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 images, ignoring irrelevant regions such as background or healthy skin - promising evidence of *automated clinical focus* with real practical utility.

2.2.2 ML unravelling microRNAs as biomarkers

2.2.3 Comparison with thesis approach

2.2.4 Other relevant work

2.2.5 Conclusion

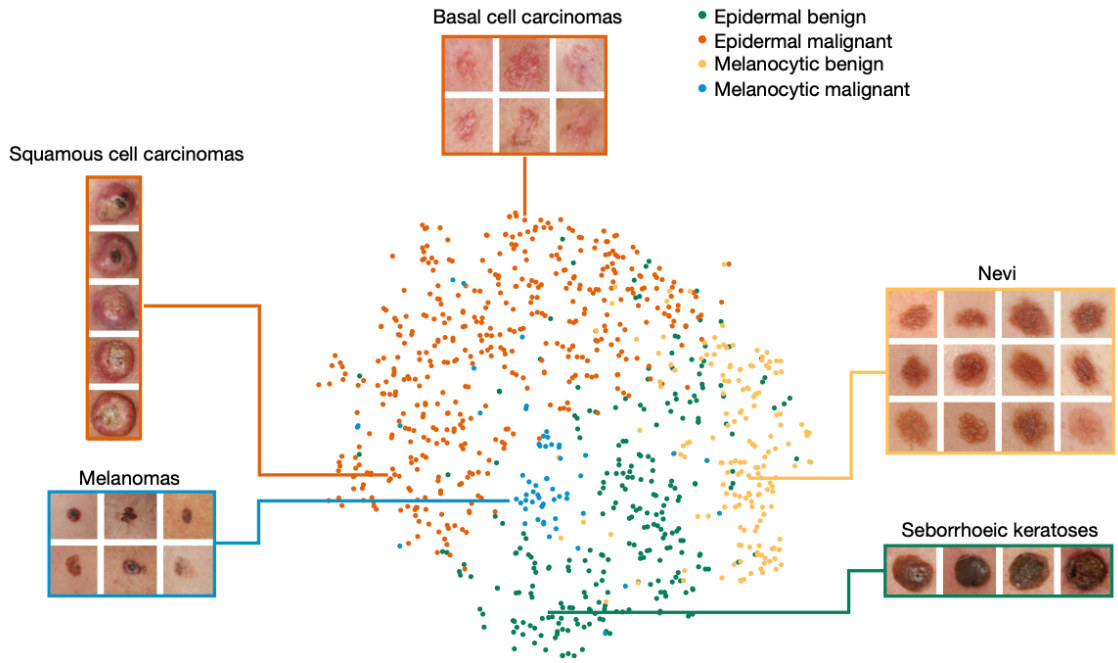


Figure 2.9: t-SNE projection of the internal representations of the last hidden layer of the CNN [13]. The different classes of lesions are grouped into distinct clouds, revealing the model's ability to extract relevant discriminative features.

3

Preliminary Work and Work Plan

4

Work Plan

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] 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/.

[3] 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/.

[4] Anastasios N. Angelopoulos and Stephen Bates. "Conformal Prediction: A Gentle Introduction". In: *Found. Trends Mach. Learn.* 16.4 (2023-03), pp. 494–591. ISSN: 1935-8237. DOI: 10.1561/2200000101. URL: https://doi.org/10.1561/2200000101.

[5] 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.

[6] 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.

[7] 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.

[8] Rok Blagus and Lara Lusa. "SMOTE for high-dimensional class-imbalanced data". In: *BMC Bioinformatics* 14.1 (2013), p. 106. DOI: 10.1186/1471-2105-14-106. URL: https://doi.org/10.1186/1471-2105-14-106.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.

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

[28] 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.

[29] 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.

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

[31] 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.

[32] 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/.

[33] 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.

[34] 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.

[35] 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/.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

A

Appendix 1 Covers Showcase

B

Appendix 2 Lorem Ipsum

I

Annex 1 Lorem Ipsum