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: *June 16, 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: *June 16, 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 *(p. 2)*

BC Breast Cancer *(pp. 1, 2)*

DL Deep Learning *(p. 2)*

miRNAs microRNAs *(pp. 1, 2)*

ML Machine Learning *(p. 2)*

1 Introduction

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, more than 2.3 million new cases of BC were diagnosed, resulting in around 665,000 global deaths [7] (Bray *et al.*, 2024). 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 [4] (Arnold *et al.*, 2022). These figures underline the high incidence and mortality associated with the disease, highlighting 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 [20] (Perou *et al.*, 2000). Each subtype exhibits substantial differences in terms of tumor aggressiveness, metastatic potential, and behavior to specific therapies [22] (Prat *et al.*, 2015). Thus, accurate classification of these subtypes is essential to enable personalized therapeutic approaches, with a direct impact on treatment efficacy and patient prognosis. [25] (Testa *et al.*, 2020).

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 [11] (Hammond *et al.*, 2015) and are implicated in the regulation of multiple physiological and pathological processes, including various human diseases [12] (Ho *et al.*, 2022).

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 [12] [17] (Muñoz *et al.*, 2023) [6] (Blenkiron *et al.*, 2007). In addition to these aspects, a seminal study by Blenkiron *et al.* (2007) 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 subtyping represents a major analytical challenge due to the complexity of these high dimensional regulatory molecules, non-linearity interactions between and clinical phenotypes require advanced computational approaches to be effectively modelled. Recent advances in Artificial Intelligence (AI), particularly in Machine Learning (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-based biomarkers — should such biomarkers exist — with the potential to effectively discriminate between different BC subtypes, as already seen in a study done by [5] Azari *et al.* 2023 where ML algorithms identified potential diagnostic and prognostic n 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 Background

1.2.1 Biology

The modern understanding of how genetic information is stored, interpreted, and regulated in cells is based on a fundamental principle known as the Central Dogma of Molecular Biology. This concept, first formulated by Francis Crick in 1958, describes the unidirectional flow of genetic information in cells: from deoxyribonucleic acid (DNA) to ribonucleic acid (RNA), and from there to protein synthesis. According to this model, genes encoded in DNA are transcribed into messenger RNA (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 microRNAs (miRNAs) emerge, 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.

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) [24]. 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.

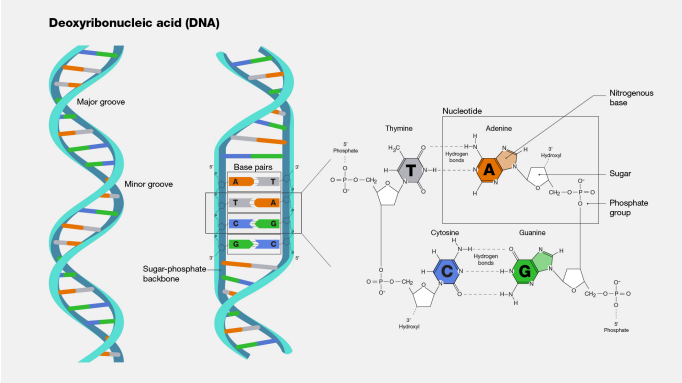

Figure 1.1: Structure of the DNA double helix

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 ribonucleic acid (RNA) molecule [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].

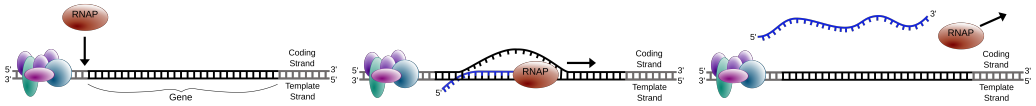

Figure 1.2: Illustration of the transcription mechanism: (a) initiation, (b) elongation, and (c) termination

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 [19].

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 [24].

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 [15].

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.

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 [10, 16, 26]. Instead of encoding proteins, miRNAs act by controlling the production of proteins from genes.

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 [9].

The action of miRNAs occurs mainly in the untranslated 3' region (3'UTR) of mRNA and is mediated by protein complexes such as RISC (RNA-induced silencing complex), which facilitates this interaction [10]. This regulation is highly efficient: 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 [9].

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 [10].

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.

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 that of the NCI, 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 [8, 18].

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 [8, 18].

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 [8].

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.

Progression to true cancer 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.

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 [23, 13]. 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 [23].

Most breast tumors originate 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 [21].

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 [23, 1].

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 [21, 14].

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 [13, 1].

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 [23].

In addition, it has been 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 [21].

1.2.2 Artificial Intelligence

2

Background

3

State of the Art

4

Premilinary Results

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 (cit. on p. 6).

[2] B. 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/ (cit. on p. 3).

[3] B. 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/ (cit. on p. 4).

[4] 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 (cit. on p. 1).

[5] 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 (cit. on p. 2).

[6] 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 (cit. on p. 2).

[7] F. 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 (cit. on p. 1).

[8] J. 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 (cit. on p. 5).

[9] G. 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 (cit. on p. 5).

[10] 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 (cit. on pp. 4, 5).

[11] S. Hammond. "An overview of microRNAs." In: *Advanced drug delivery reviews* 87 (2015), pp. 3–14. DOI: 10.1016/j.addr.2015.05.001 (cit. on p. 1).

[12] 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 (cit. on pp. 1, 2).

[13] R. Hong and B.-h. Xu. "Breast cancer: an up-to-date review and future perspectives". In: *Cancer Communications* 42 (2022), pp. 913–936. DOI: 10.1002/cac2.12358 (cit. on p. 6).

[14] 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 (cit. on p. 6).

[15] 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 (cit. on p. 4).

[16] R. C. Lee, R. L. Feinbaum, and V. 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 (cit. on p. 4).

[17] J. 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 (cit. on p. 2).

[18] National Cancer Institute. *What Is Cancer?* Accessed: 2025-06-15. 2021. URL: https://www.cancer.gov/about-cancer/understanding/what-is-cancer (cit. on p. 5).

[19] 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 (cit. on p. 4).

[20] C. 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 (cit. on p. 1).

[21] K. Polyak. "Breast cancer: origins and evolution." In: *The Journal of clinical investigation* 117 11 (2007), pp. 3155–63. DOI: 10.1172/JCI33295 (cit. on pp. 6, 7).

[22] 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 (cit. on p. 1).

[23] H. Romanowicz, B. Smolarz, and A. Z. Nowak. "Breast Cancer—Epidemiology, Classification, Pathogenesis and Treatment (Review of Literature)". In: *Cancers* 14 (2022). DOI: 10.3390/cancers14102569 (cit. on p. 6).

[24] Sunrayce Biology Authors. *Chapter 9.1: The Structure of DNA*. BC Open Textbooks. Accessed: 2025-06-16. 2015. URL: https://opentextbc.ca/biology/chapter/9-1-the-structure-of-dna/ (cit. on pp. 3, 4).

[25] 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 (cit. on p. 1).

[26] B. Wightman, I. Ha, and G. 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 (cit. on p. 4).

A

Appendix 1 Covers Showcase

B

Appendix 2 Lorem Ipsum

I

Annex 1 Lorem Ipsum