Deanship of Graduates Studies

Al- Quds University

Identification of Potential Diagnostic and Prognostic Biomarkers for Triple Negative Breast Cancer (TNBC) Using Artificial Intelligence (AI)

Shahd Mustafa Yahya Qawasma

M.Sc. Thesis

Jerusalem- Palestine

1446 / 2024

Identification of Potential Diagnostic and Prognostic Biomarkers for Triple Negative Breast Cancer (TNBC) Using Artificial Intelligence (AI)

Prepared By:

Shahd Mustafa Yahya Qawasma

B.Sc.: Pharmacy Al-Quds University Palestine

Supervisor: Dr. Yousef Najajreh

Co-Supervisor: Dr. Rashid Jayousi

A thesis submitted in partial fulfilment of requirements for the degree of Master of Pharmaceutical Science from Faculty of Pharmacy/ Al-Quds University.

1446 / 2024

Al – Quds University

Deanship of Graduates Studies

Faculty of Pharmacy

Thesis Approval

Identification of Potential Diagnostic and Prognostic Biomarkers for Triple Negative Breast Cancer (TNBC) Using Artificial Intelligence (AI)

Prepared by: Shahd Mustafa Yahya Qawasma

Registration No: 22011148

Supervisor: Dr.Yousef Najajreh

Co- Supervisor: Dr. Rashid Jayousi

Master thesis submitted and accepted, Date: 18/12/2024

The names and signatures of the examining committee members are as follows:

1- **Head of Committee**: Dr. Yousef Najajreh Signature:

2- **Internal Examiner**: Dr. Nael Halawa Signature:

3- **External Examiner**: Dr. Majdi Owda Signature:

4- **Committee member**: Dr. Rashid Jayousi Signature:

Jerusalem – Palestine

1446 / 2024

Dedication

I dedicate this thesis to my dear parents, whose love, encouragement, and sacrifices have been my greatest source of strength. To my mother, for her boundless support and belief in me, and to my father, for his wisdom and constant inspiration. This achievement is a reflection of your guidance and dedication, and I am forever grateful for everything you have done**.**

Declaration

I Certify that this thesis submitted for the degree of master, is the result of my own research, except where otherwise acknowledged, and that this study (or any part of the same) has not been submitted for a higher degree to any other university or institution.

Shahd Mustafa Yahya Qawasma

18/12/2024

Acknowledgment

I am deeply grateful for my decision to embark on this master's journey; it has truly been a remarkable chapter in my life. I want to extend my heartfelt thanks to my family for their unwavering love and support, especially my mother; her support has been invaluable. I am especially thankful to my supervisor, Dr. Yousef, who, despite his numerous commitments, always found time to guide me. I also appreciate my co-supervisor, Dr. Rashid, for his valuable contributions and cooperation. A special thanks to Rosol Sharairh, an expert in artificial intelligence, whose assistance was crucial in developing my machine learning model. I feel fortunate to be part of Al-Quds University, surrounded by a friendly atmosphere, supportive colleagues, and diverse research opportunities. I'm grateful for the wonderful friends I've made over the past two years and for the lifelong friends who have supported me along the way. Lastly, I would like to express a heartfelt thanks to my friend Noora for being an integral part of my successful journey.

Abstract

Triple-negative breast cancer (TNBC) is one of the most aggressive forms of breast cancer, linked to the highest mortality rates. TNBC is characterized by the absence of estrogen, progesterone, and human epidermal growth factor receptors. MicroRNAs (miRNAs) play a key role in gene expression by interacting with messenger RNAs. Some miRNAs may act as biomarkers for diagnosing and predicting the prognosis of TNBC. The main aim of the study is to analyze the miRNA expression levels and clinical data of TNBC patients obtained from the Gene Expression Omnibus (GEO) to detect biomarkers for triple negative breast cancer (TNBC) using machine learning (ML). The study objectives are addressing the use of AI and ML to identify potential diagnostic and prognostic Biomarkers for TNBC. In addition, the process of constructing a model begins with conducting a meta-analysis, followed by differential expression analysis (DEA), which uncovers statistically significant correlations among multi-gene signatures. This study also includes comparing the miRNA expression profiles between TNBC and normal tissues, also between TNBC and non-TNBC tissues. Ultimately, build a machine learning (ML) model using a hybrid feature selection method for the biomarker selection. The study involved two datasets that were merged, resulting in a combined dataset of 4577 miRNAs. In the ML model-building stage, a combination of feature selection methods is employed to identify biomarker profiles that distinguish TNBC from normal tissue. This includes a wrapper method using recursive feature elimination (RFE), along with embedded methods using random forest (RF) and support vector machine (SVM). Our study revealed significant variations in miRNA expression between TNBC and normal tissues. In contrast, the expression of miRNAs did not differ significantly between TNBC and non-TNBC tissues. In addition, the study shows that employing Recursive Feature Elimination, SVM, and Random Forest as a hybrid feature selection algorithm for miRNA expression profiles or similar datasets with large number of features in comparing to the number of samples can effectively eliminate redundant features, identify biomarkers with diagnostic relevance, and maintain high classification accuracy. Finally, the study identified that miR-32-5p can be used as a potential biomarker for the diagnosis of TNBC, high expression of miR-32-5p is substantially related with increasing overall survival in TNBC patients.

Table of Contents

List of figures

List of tables

List of abbreviations

BC: Breast cancer

TNBC: Triple negative breast cancer

ER: Estrogen receptor

PR: Progesterone receptor

HER2: Human epidermal growth factor receptor 2

mRNA: Messenger RNA

miRNA: MicroRNA

oncomiR: Oncogenic miRNA

tsmiR: Tumor suppressor miRNAs

TSGs: Tumor suppressor genes

EMT: Epithelial–mesenchymal transition

OS: Overall survival

mTNBC: Metastatic Triple-Negative Breast Cancer

AR: Androgen receptor

SOC: Standard of care

ADC: Anti body drug conjugate

mAb: Mono-clonal antibody

PARP: Poly ADP-ribose polymerase

ICB: Immune checkpoint blockade

LAR: Luminal androgen receptor

RTK: Receptor tyrosine kinase

ABC: ATP-Binding Cassette

PFS: Progression-free survival

PM: Personalized medicine

GEO: Gene expression omnibus

NCBI: The National center for biotechnology information

AI: Artificial intelligence

ML: Machine learning

DEA: Differential expression analysis

SVM: Support vector machine

Log-FC: Log-fold change

DT: Decision tree

RF: Random forest

LR: Lasso regression

RFE: Recursive feature elimination

GA: Genetic algorithm

KNN: K-nearest neighbor

PSO: Particle swarm optimization

mRMR: The minimum redundancy maximum relevance method

Chapter One (1): Introduction

Triple-negative breast cancer (TNBC) is characterized by its high malignancy, aggressive invasiveness, rapid progression, frequent recurrence, and distant metastasis. It also has a poor prognosis, high mortality, and is resistant to traditional endocrine and targeted therapies, making it a particularly difficult challenge in breast cancer treatment and a major focus of scientific research. Recent studies have shown that certain miRNAs can directly or indirectly influence the development, progression, and recurrence of TNBC. The levels of their expression have a significant impact on the diagnosis, treatment, and prognosis of TNBC. Some miRNAs may act as biomarkers for diagnosing and predicting the prognosis of TNBC.

1.1 The study question

Why we need new and more accurate biomarkers to be discovered?

To achieve the goals of personalized therapy:

1- Individual Variability: many criteria should be considered like age, gender, ethnicity, lifestyle, pre-treatment, history of diseases.

2- Targeted Therapy: to match the right treatment for the right patient at the right time.

3- Early Diagnosis: Biomarkers can help detect diseases at earlier stage.

4- Better Prognostic Information: Personalized therapy requires knowing not only what disease someone has but also how it will progress.

5- Personalized Prevention: by identifying women at risk of TNBC disease.

1.2 Aim of the study

The main aim of the study is to analyze the miRNA expression profiles obtained from the Gene Expression Omnibus (GEO) to detect biomarkers for triple negative breast cancer (TNBC) using machine learning (ML) thus allowing for more personalized therapy.

1.3 Objectives of the study

1- To address the use of AI and ML to identify potential diagnostic and prognostic Biomarkers for TNBC.

2- To build a model starts with meta-analysis, followed by differential expression analysis (DEA) and identifies statistically significant correlations among multi-gene signatures.

3- To identify expression profiles of miRNAs between TNBC and adjacent normal TNBC (healthy tissues), also between TNBC and non-TNBC (other cancerous tissues).

4- To build a ML model using a hybrid feature selection method for the biomarker selection.

1.4 Thesis layout

The background section of the study provides a summary about TNBC, discussing miRNA roles in tumor metastasis, prognosis and treatment, and suggests a new therapeutic approaches, one of them targeting miRNA expression. Additionally, it highlights the application of AI and ML in identifying potential diagnostic and prognostic biomarkers for TNBC.

The methodology section involved three steps, which include constructing two models to identify the diagnostic biomarkers.

In the differential expression analysis and meta-analysis step, statistical tests are employed to identify genes that show varying levels of expression across different groups. These genes can be considered as potential biomarker candidates. However, like other univariate methods, statistical tests have the limitation of treating each gene independently, whereas genes are involved in complex interactions. Conversely, computational methods take these interactions into consideration. Therefore, the second step was to build a ML model using a hybrid feature selection method for the biomarker selection.

Another step included in the study was comparing the expression profiles of miRNAs between TNBC and adjacent normal tissues (healthy tissues), as well as between TNBC and non-TNBC (other cancerous tissues).

The literature review provides an overview of other studies focused on identifying new biomarkers for targeting cancer, using the two models mentioned earlier separately. Unlike the majority of

other studies, this study is not limited just to DEA based on filter methods to differentiate between the two classes, but through AI based on ML methods, it also identifies a new biomarker with diagnostic relevance.

The results of the three steps are presented below:

Step 1: After the integration of two datasets, the miRNAs were filtered using statistical tests ($P<0.05$ and $|log2FC|>2$) into 20 top ranked miRNAs, the identified 20 miRNAs were significantly associated with TNBC, with 17 miRNAs upregulated and 3 miRNAs downregulated in TNBC tissues. Four miRNAs (miR-135b-5p, hsa-miR-18a-5p, hsa-miR-18b-5p and hsa-miR-32-5p) were significantly differentially expressed in the TNBC samples compared to the normal samples, all were upregulated, and may be associated with development and progression of TNBC.

Step 2: Significant differences in miRNA expression were observed between TNBC and normal tissues. In contrast, the expression of miRNAs did not differ significantly between TNBC and non-TNBC tissues.

Step 3: In the ML model building stage, a combination of feature selection methods was employed to identify biomarker profiles that differentiates TNBC from normal samples. This includes a wrapper method using recursive feature elimination (RFE), along with embedded methods using random forest (RF) and support vector machine (SVM). The study's findings suggest that miR-32-5p, along with other identified miRNAs, could serve as potential diagnostic and prognostic biomarkers for TNBC.

Chapter Two (2): Background

2.1 Breast cancer (BC)

Breast cancer is the most widely diagnosed cancer and stands as the second leading cause of cancer mortality among women globally(Bertucci et al., 2000). Breast cancer is a disease that originates in breast tissue, mainly in the milk ducts (ductal carcinoma accounts for roughly 80% of cases). When cancer develops in the ductal regions, it is called ductal carcinoma, whereas cancer that arises in the lobules is referred to as lobular carcinoma (Medina, Oza, Sharma, Arriaga, Hernández, et al., 2020).

2.1.1 Breast cancer diagnosis

Diagnosing breast cancer depends on three distinct types of assessments :(A) clinical examination; (B) radiological imaging tests, which include mammography, magnetic resonance imaging (MRI), and ultrasonography; and (C) immunohistopathological evaluations (Aebi et al., 2011).

2.1.2 Breast cancer subtypes

The classification of breast cancer cell types is considered as below:

(1) Luminal A subtype, ERα+/PR+ or −/HER-2-(47%);

(2) Luminal B subtype, ER+/PR+/HER-2+(26%);

(3) HER-2 enriched subtype ER- and or/PR−/HER-2+(10%);

(4) Basal-like subtype ER− and/or PR−, HER2−, CK5/6+, CK14+, CK17+ and EGFR+(13%);

(5) Normal breast-like type ER− and/or PR−, HER2−, CK5/6−, CK14−, CK17−, EGFR−(4%) (Dai et al., 2015).

These five subtypes exhibit varying abilities to metastasize to distant organs, distinct pathways with favored metastatic locations, and different survival responses following a relapse (Kennecke et al., 2010).

2.1.3 Breast cancer therapy challenges

Despite advances in understanding cancer biology over the past few decades, treating breast cancer continues to be challenging due to factors such as disease heterogeneity, diverse therapeutic targets, therapeutic resistance, residual disease, and the risk of recurrence even after targeted treatments (Al-Mahmood et al., 2018). Breast cancer is primarily a sporadic disease influenced by both genetic and epigenetic factors. Genomic instability in breast cancer leads to mutations, copy number variations, and genetic rearrangements, while epigenetic changes involve alterations in gene expression profiles through DNA methylation, histone modifications, and microRNAs (miRNAs) (Rahman et al., 2019).

2.1.4 mRNA profile as prognostic marker of breast cancer

The expression levels of estrogen receptors (ER) and HER2 serve as key indicators for prognostic evaluations in breast cancer. Various tumor types display distinct gene expression signatures, which can assist in categorizing cancers into different prognostic groups. mRNA profiling has indicated that mRNA expression correlates with cancer progression and has helped in the identification of effective cancer biomarkers. Aberrant expression of oncogenes, such as KRAS and MYC, along with tumor suppressor genes (TSGs) like APC, BRAF, and TP53, is frequently linked to cancer development, often as a result of chromosomal instability, the accumulation of mutations, and DNA methylation alterations. Furthermore, all aspects of mRNA biogenesis including transcription, splicing, post-transcriptional regulation, translation, and mRNA stability can be affected during the progression of cancer (M. Li et al., 2015).

2.1.5 Breast cancer-linked microRNAs (miRNAs)

MicroRNAs (miRNAs) are small non-coding RNA molecules that serve as essential post-transcriptional gene regulators in numerous biological processes. Typically, miRNAs inhibit gene expression by binding to specific messenger RNAs (mRNAs), leading to either mRNA degradation or translational repression, depending on the degree of complementarity between the miRNA and the target mRNA sequences. Various cellular pathways involved in breast cancer development, including cell proliferation, apoptotic response, metastasis, cancer recurrence, and chemoresistance, are regulated by either oncogenic miRNAs (oncomiRs) or tumor suppressor miRNAs (tsmiRs)(Loh et al., 2019). MicroRNAs are potential excellent biomarkers of breast carcinomas, and could be utilized for innovative therapies tailored to specific patients. MiRNA expression profiling may be a valuable resource for classifying breast cancer subtypes and for assessing prognosis and therapeutic outcomes (Dai et al., 2015).

2.1.6 Breast cancer metastasis and invasion

Metastasis begins when tumor cells invade nearby host tissue. To do this, they must first change how they stick to each other and to the extracellular matrix (ECM). Cancer cells have weaker interactions with the ECM, making them less adhesive than non-cancerous cells. This reduced adhesion helps them invade and spread through surrounding blood or lymphatic systems(Seyfried & Huysentruyt, 2013). The cadherin family is important for cell-to-cell adhesion and plays a key role in breast cancer metastasis. Epithelial–mesenchymal transition (EMT) is a crucial aspect of

the breast cancer metastasis process, allowing cancer cells to gain stem-like characteristics and enhancing their migratory and invasive abilities. Epithelial-mesenchymal transition (EMT) is associated with a decrease in E-cadherin expression, leading to reduced cell localization and cell-cell contact (Jeanes et al., 2008). Additionally, research suggests that lower levels of E-cadherin in cancer cells may enhance the Wnt/β-catenin signaling pathway (Tian et al., 2011).

2.1.7 Breast cancer therapy

Breast cancer treatment focuses on inducing tumor regression in the breast and preventing the cancer from spreading to distant sites through metastasis. Therapeutic strategies for breast cancer can be categorized as either systematic or local. Systemic therapies, which include endocrine therapy and chemotherapy, typically target both nonmetastatic and metastatic breast cancer, whereas local therapies like surgery and radiation are commonly employed for nonmetastatic cases. Often, a combination of systemic and local treatments is used, either concurrently or in a sequential manner (Waks & Winer, 2019).

The selection of treatment strategies mainly relies on molecular subtyping, which is usually determined by the expression levels of estrogen receptors and human epidermal growth factor receptor 2 (HER2), along with the stage of breast cancer at diagnosis (Tang et al., 2016).

2.2. Triple Negative Breast Cancer (TNBC)

Triple-negative breast cancer is defined by its negative status for estrogen and progesterone receptors (ER/PR) and human epidermal growth factor receptor 2 (HER2). It is known for its specific molecular characteristics, aggressive tendencies, particular patterns of metastasis, and the absence of targeted treatments (Anders et al., 2008).

2.2.1 Triple Negative Breast Cancer (TNBC): A challenging BC subtype

Recent advancements in cancer genomics have clarified the intrinsic subtypes of breast cancer, enabling the use of targeted therapies such as endocrine therapy and anti-HER2 therapy for patients with hormone receptor-positive (ER/PR) or HER2-positive tumors. The success of these targeted therapies has significantly improved outcomes for breast cancer patients (Ahn et al., 2016; "Effects of Chemotherapy and Hormonal Therapy for Early Breast Cancer on Recurrence and 15-Year Survival: An Overview of the Randomised Trials.," 2005; Piccart-Gebhart et al., 2005). In contrast, nonspecific chemotherapy continues to be the main treatment strategy for triple-negative breast cancer (TNBC), which does not express estrogen receptors (ER), progesterone receptors, or HER2. TNBC comprises 15% to 20% of breast cancer cases and is considered invasive and aggressive, often leading to metastasis and presenting challenges in treatment. It disproportionately affects younger women at higher rates (Ahn et al., 2016; Haffty et al., 2006; *Le Cancer Du Sein «triple Négatif»*, n.d.). Patients with TNBC are at a higher risk of experiencing distant metastasis and early recurrence within 2 to 3 years after treatment compared to those with other breast cancer subtypes, leading to lower survival rates.Thus, TNBC poses a serious challenge for treatment among breast cancer subtypes (Garrido-Castro et al., 2019).

2.2.2 Metastatic TNBC

The metastatic potential of all breast cancer subtypes is ultimately similar, but growth rates and tumor distributions differ, leading to variations in natural history and clinical progression, especially in the short term. Once metastatic TNBC develops, patients tend to have a shorter median time from relapse to death (Hudis & Gianni, 2011).

The primary reason for treatment failure in metastatic cases is multidrug resistance to standard therapies, which can be either inherited (present before drug exposure) or acquired (developed as a result of treatment) (Andre & Zielinski, 2012; O'Driscoll & Clynes, 2006). It is well established that triple-negative breast cancer (TNBC) is among the most aggressive subtypes, frequently linked to unfavorable patient outcomes due to its tendency to metastasize to secondary organs, including the lungs, brain, and bones. The complex molecular nature of the metastatic process, coupled with the absence of effective targeted therapies for TNBC metastasis, has spurred considerable research efforts in recent years aimed at identifying the molecular "drivers" responsible for this lethal progression (Neophytou et al., 2018).

2.2.3 Subtypes of TNBC

TNBC is not just one type of breast cancer; it is a heterogeneous collection of various subtypes. Recent classifications divide TNBC into four main subtypes based on genomic profiling:(I) luminal androgen receptor (LAR), (II) basal-like, (III) immune-enriched, and (IV) mesenchymal (Bharaj et al., 2021; Perou, 2011).

2.2.4 Therapeutic approaches for TNBC

TNBC is generally treated through a combination of surgery, radiation, and chemotherapy. Because hormones do not support tumor growth in TNBC, it is unlikely to respond to molecularly targeted therapies, such as endocrine agents like anastrozole (Arimidex), exemestane (Aromasin), letrozole (Femara), and fulvestrant (Faslodex). Furthermore, TNBC does not typically respond to HER2-targeted treatments, including trastuzumab (Herceptin) and lapatinib (Tykerb). Consequently, chemotherapy and radiation are the main therapeutic options available (Brown, n.d.).

2.2.4.1 surgery

Numerous studies have examined the prognostic differences between mastectomy and lumpectomy. For TNBC, breast preservation is typically the preferred surgical approach, as the type of surgical intervention does not significantly alter prognosis or the risk of local tumor recurrence, making breast-conserving surgery a suitable option for many patients (Frasci et al., 2009; Freedman et al., 2009). National Comprehensive Cancer Network guidelines suggest that a lumpectomy followed by radiation therapy can be effective. Nonetheless, neoadjuvant therapy is regarded as the gold standard for TNBC and is generally advised before surgical intervention (Medina, Oza, Sharma, Arriaga, Hernández Hernández, et al., 2020).

2.2.4.2 radiotherapy

TNBC is recognized as a distinct subtype that responds well to radiotherapy. However, there are no established treatment protocols specifically for the use of radiotherapy in TNBC (Dragun et al., 2011; Medina, Oza, Sharma, Arriaga, Hernández Hernández, et al., 2020; Panoff et al., 2011).

2.2.4.3 Chemotherapy

Chemotherapy-based approaches may improve treatment effectiveness following surgery and radiation therapy to eradicate tumors(Kumar et al., 2023). TNBC is highly responsive to chemotherapy, which is established as the standard of care (SOC) for managing this disease. Commonly used chemotherapeutic agents comprise anthracyclines (such as doxorubicin, a DNA intercalating agent and topoisomerase II blocker), alkylating agents (like cyclophosphamide), taxanes (anti-microtubule agents), and anti-metabolites such as fluorouracil (5-FU). The current SOC for newly diagnosed early-stage TNBC involves neoadjuvant chemotherapy, followed by surgery. However, there is no standard chemotherapy regimen for patients with relapsed or refractory TNBC. Responses to treatment are often short in duration, leading to rapid relapse as well as visceral and brain metastases are common. Available therapies for advanced TNBC include anti-metabolites capecitabine and gemcitabine, non-taxane microtubule inhibitor eribulin, and DNA cross-linker platinum-based agents, Doxorubicin (DOX), Paclitaxel (PTX), Docetaxel, PARP inhibitors, Checkpoints inhibitors (e.g. atezolizumab, pembrolizumab). Recently introduced combinations, such as Bevacizumab with taxanes, are also being explored. The median progression-free survival (PFS) with chemotherapy ranges from 1.7 to 3.7 months; the median OS from the onset of metastasis is 10 to 13 months. In clinical trials, advanced TNBC patients treated with single-agent taxane or platinum-based chemotherapy exhibited a median PFS of 4 to 6 months and a median OS of 11 to 17 months (Plevritis et al., 2018; Won & Spruck, 2020).

2.2.5 Novel strategies for the management of TNBC

The aggressiveness and heterogeneity of TNBC present considerable challenges for developing effective treatment regimens. Apart from radiotherapy, surgery, and chemotherapy, new treatment strategies are being explored for TNBC. This section focuses on novel strategies for treating TNBC that target cell signaling pathways, receptor-mediated therapy, epigenetic modifications, mitochondrial functions, nucleic acids, peptide-based therapies, and immunotherapy (Kumar et al., 2023). New treatment alternatives for patients with advanced TNBC have recently become available, especially in situations where surgery is not viable(Won & Spruck, 2020). A variety of promising new therapies for TNBC are presently being researched, demonstrating different degrees of success. These include: (I) immunotherapy, (II) antibody-drug conjugates, (III) PARP inhibitors, and (IV) targeted therapies that inhibit cell signaling through the androgen receptor (AR) or the PIK3/AKT/mTOR pathway (Bharaj et al., 2021). Additionally, research is investigating combination therapies that combine these various classes of agents (Bharaj et al., 2021).

2.2.5.1 Immunotherapy for TNBC

TNBC is amenable to immunotherapy strategies largely due to several factors, including tumor immune infiltration, the presence of neoantigens resulting from mutational burden and higher genomic instability, as well as high levels of immune markers such as PD-L1 and programmed

cell death protein-1 (PD-1), these markers are strongly associated with tumor response, relapse rates, and overall patient outcomes. Immunotherapy has been effective in treating numerous cancers, rendering immunotherapeutic approaches for TNBC very promising. Among the different immunological strategies, molecular and cellular immunotherapies have shown considerable potential, supported by findings from early studies (Jia et al., 2017). On March 8, 2019, the FDA approved atezolizumab for use in patients with unresectable locally advanced or metastatic TNBC that is positive for programmed death-ligand 1 (PD-L1), marking it as the first immune checkpoint inhibitor (ICB) monoclonal antibody approved for this type of cancer. Subsequently, on November 13, 2020, pembrolizumab, in combination with chemotherapy, received FDA approval for locally recurrent unresectable or metastatic TNBC patients with a positive PD-L1 expression (CPS ≥ 10) (Yun Li et al., 2022).

2.2.5.2 Antibody-drug conjugates for TNBC

Targeted therapy using antibody-drug conjugates (ADCs) has emerged as a promising strategy for treating TNBC (C. Zhang et al., 2022). ADCs consist of three distinct components: (1) antibody that targets a tumor antigen,(2) a highly potent cytotoxic payload, and (3) a linker that connects the first two components (Nagayama et al., 2020). New ADC drugs, including sacituzumab govitecan (IMMU-132) and trastuzumab deruxtecan (DS-8201a), are employed in the later stages of clinical development for patients with metastatic breast cancer, including TNBC (Nagayama et al., 2020).

2.2.5.3 PARP inhibitors for TNBC

Poly ADP-ribose polymerase (PARP) enzymes play a crucial role in various biological functions, including DNA repair, genome maintenance, and apoptosis. As a result, PARPs have emerged as promising targets for cancer therapies, leading to the development of several PARP inhibitors (PARPi), which are regarded as one of the innovative treatment options for TNBC. Poly ADP-ribose polymerase inhibitors (PARPi) mainly target PARP1, which is responsible for the recognition of single-strand DNA breaks (Rose et al., 2020). Several PARP inhibitors, including olaparib, rucaparib, iparib, veliparib, talazoparib, niraparib, and geparixto, have been evaluated for their clinical effectiveness both as standalone treatments and in combination with other chemotherapeutic drugs (Yun Li et al., 2022).

2.2.5.4 PI3K/AKT/mTOR pathway inhibition for TNBC

The PI3K/AKT/mTOR signaling cascade plays a vital role in the survival, metabolism and proliferation of breast tumors(Kumar et al., 2023). mTOR is a type of serine/threonine protein kinase. Two complexes, mTORC1 and mTORC2, mediate the effects of the serine/threonine protein kinase mTOR. mTORC1 is responsible for regulating protein translation, whereas mTORC2 activates AKT(Ortega et al., 2020). In TNBC, mutations in the INPP4B and PTEN genes are closely linked to the regulation of the PI3K/AKT/mTOR signaling pathway and its response to chemotherapy. Inhibitors targeting PI3K/AKT/mTOR are emerging as promising therapeutic options in breast cancer treatment, aiming to suppress the PI3K/AKT/mTOR signaling pathway, which is a primary contributor to resistance against ER- and HER2-targeted therapies (Kumar et al., 2023; Ying Li et al., 2021).

2.2.5.5 Androgen receptor targeted therapy for TNBC

The androgen receptor (AR) status in triple-negative breast cancer (TNBC) patients serves as a biomarker for preselecting individuals for trials focused on antiandrogen therapies. Generally, these antiandrogen therapies are well accepted and have proven safety records (Mina et al., 2017). Multiple antiandrogen therapies are currently undergoing evaluation in clinical trials, both as standalone treatments and in combination with other pathway inhibitors and cytotoxic therapies. The integration of antiandrogen-targeted therapy with immune checkpoint blockade (ICB) for AR-positive triple-negative breast cancer (TNBC) is especially noteworthy. Furthermore, beyond AR-positive TNBC subtypes, partial luminal androgen receptor (LAR) subtype cell lines display a significant frequency of PIK3CA mutations linked to AR dependency, similar to estrogen receptor-positive breast cancers. Additionally, preclinical data have shown a synergistic effect when combining bicalutamide with a PI3K inhibitor. In summary, AR-targeted therapy holds significant potential for treating AR-positive TNBC subtypes (Yun Li et al., 2022).

2.2.5.6 Receptor tyrosine kinase (RTK) pathways for TNBC

The receptor tyrosine kinase (RTK) pathways are crucial for regulating various cellular processes and comprise a diverse array of proteins. These proteins and their associated receptors play a vital role in cell growth through tyrosine phosphorylation. Overexpression of RTK proteins is linked to enhanced cell proliferation, angiogenesis, and tumor cell migration. Key protein receptors in this RTK family include the epidermal growth factor receptor, fibroblast growth factor receptor, and vascular endothelial growth factor receptors (Dev et al., 2021; Kumar et al., 2023). RTKs are often overexpressed and/or deregulated in TNBC, contributing to tumor progression and decreased survival rates in patients. As a result, targeting RTKs presents a promising therapeutic approach for treating TNBC (Jaradat et al., 2024).

2.2.5.7 CDK inhibitors for TNBC

The therapeutic efficacy of CDK 4/6 inhibitors in TNBC is closely linked to their substrate, the Rb protein, which is in turn associated with AR positivity (> 10%) (Matutino et al., 2018). Previous analyses indicate that the activation of AR signaling is a key characteristic of the luminal androgen receptor (LAR) subtype of TNBC. Consequently, combining CDK inhibitors with AR inhibitors represents a promising strategy, such as the use of palbociclib alongside AR inhibitors like bicalutamide in AR-positive metastatic TNBC (NCT02605486). Additionally, ribociclib combined with bicalutamide is being explored for advanced AR-positive TNBC (NCT03090165), and abemaciclib is being tested in Rb protein-positive metastatic TNBC (NCT03130439). Other novel CDK inhibitors, including dinaciclib, PF-06873600, and trilaciclib, have also been analyzed in clinical trials to determine their antitumor activity in TNBC (*Cinical Trials.Gov*, n.d.; Yun Li et al., 2022).

2.3 MicroRNA biogenesis and functions

MicroRNA (miRNA) biogenesis involves the production and maturation of miRNAs from precursor molecules. This process begins with the transcription of DNA into long primary transcripts known as pri-miRNAs in the nucleus. The enzyme Drosha, along with DGCR8,

processes these pri-miRNAs into shorter hairpin structures called pre-miRNAs, which are then exported to the cytoplasm by Exportin-5. In the cytoplasm, the enzyme Dicer further processes the pre-miRNAs, resulting in double-stranded RNA molecules that are unwound to produce mature miRNAs **(Figure 2.1)**. These mature miRNAs play a critical role in regulating gene expression by binding to complementary sequences in messenger RNAs (mRNAs), leading to mRNA degradation or inhibition of translation, thereby reducing protein production. MiRNAs are involved in various biological processes, including cell proliferation, differentiation, apoptosis, and development, helping maintain cellular homeostasis and respond to environmental stimuli. Dysregulation of miRNA expression is linked to various diseases, particularly cancer, where some miRNAs can act as oncogenes or tumor suppressors. Additionally, due to their stability in biological fluids and specific expression patterns, miRNAs have potential as diagnostic and prognostic biomarkers for various diseases. Overall, understanding miRNA biogenesis and functions is crucial for exploring their roles in gene regulation and therapeutic applications (Zeng et al., 2021).

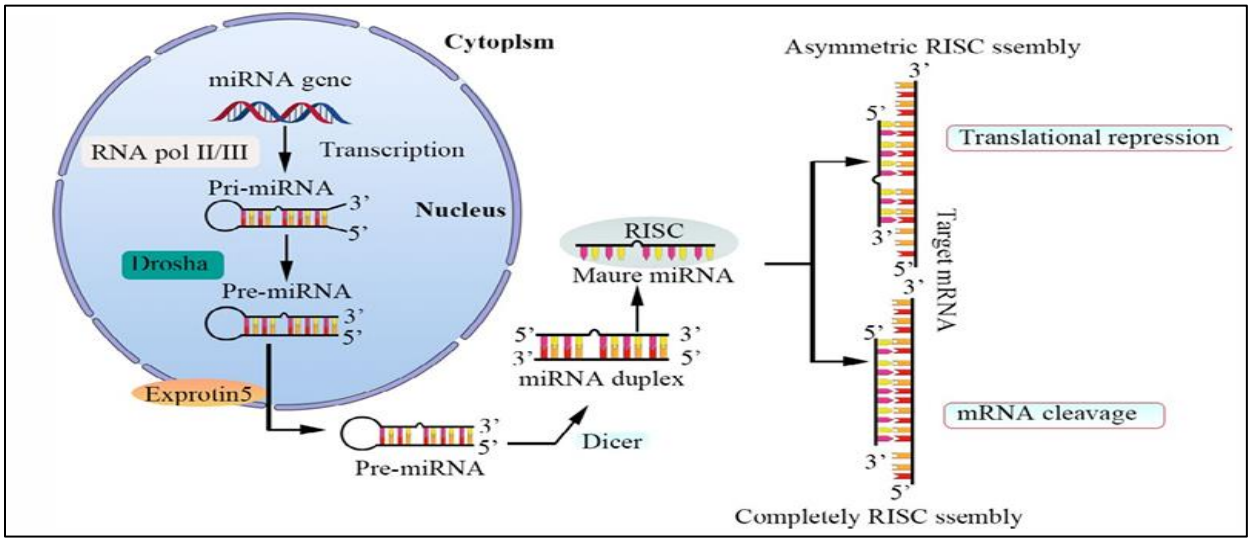

Figure 2.1 1Schematic of miRNA biogenesis

MicroRNAs (miRNAs) regulate gene expression mainly through a mechanism known as RNA interference. They become part of a protein complex called the RNA-induced silencing complex (RISC), where the miRNA directs the complex to its target messenger RNA (mRNA). When the miRNA has a strong match with the mRNA, RISC can cut the mRNA, resulting in its degradation and stopping protein synthesis. If the match is less exact, RISC can block the translation of the mRNA without destroying it **(Figure 2.1)**. Additionally, miRNAs can influence gene expression by interacting with transcription factors (Catalanotto et al., 2016).

2.3.1 miRNAs in the carcinogenesis of TNBC

Recently, the significance of microRNA in cancer therapy has attracted a lot of interest due to their possible use as diagnostic biomarkers(Rastogi et al., 2008). A meta-analysis has identified several miRNAs associated with TNBC(Iorio et al., 2005; Lü et al., 2017; Rhodes et al., 2015). The role of miRNAs as regulators of gene expression is depicted in **Figure 2.2** MiRNAs are important in how genes are turned on and off, especially in triple-negative breast cancer (TNBC). They can either help tumors grow (called oncogenic miRNAs or oncomiRs) or prevent tumor growth (tumor-suppressive miRNAs). OncomiRs stop the action of tumor suppressor genes, while tumor-suppressive miRNAs target genes that promote cancer. Changes in miRNA levels affect how cells survive and spread, influencing processes like metastasis, proliferation, and apoptosis, as well as other biological pathways in TNBC (Loh et al., n.d.).

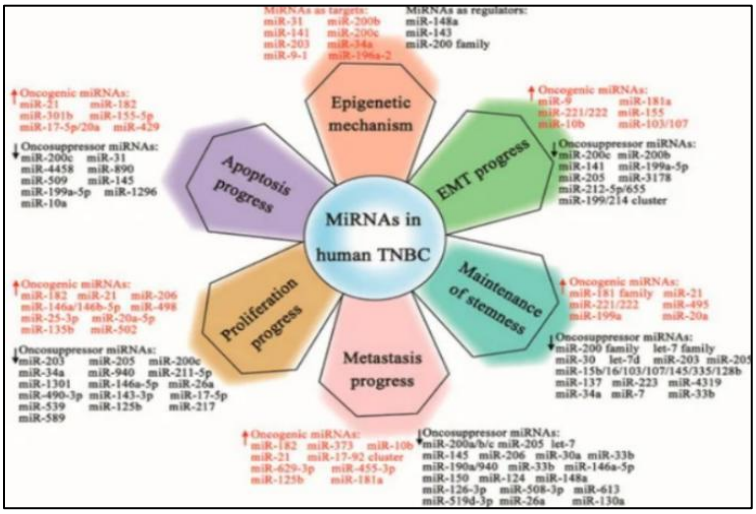

Figure 2.2 1The multiple roles of miRNAs in TNBC

2.3.2 Regulation of multidrug resistance by microRNAs in TNBC therapy

Multidrug resistance (MDR) is a major obstacle in effective cancer treatment and is commonly linked to the increased efflux of anticancer drugs mediated by proteins from the ATP-Binding Cassette (ABC) transporter family, such as P-glycoprotein (P-gp) (Gisel et al., 2014). Over the past few decades, multiple mechanisms have been discovered that play a role in both intrinsic and acquired multidrug resistance (MDR). The main mechanisms include: (1) Overexpression of MDR transporters; (2) Impairments in the apoptotic process; (3) Induction of autophagy; (4) Changes in drug metabolism; (5) Modifications in drug targets and DNA repair; and (6) Disruption of redox balance (An et al., 2017). MicroRNAs (miRNAs) are essential regulators of multidrug resistance (MDR), influencing many of the biological processes previously discussed. As a result, miRNAs could be valuable as potential biomarkers and/or targets for overcoming MDR in cancer treatment (An et al., 2017).

2.4 Artificial intelligence (AI) applications in healthcare

AI models in healthcare are applied in various ways, including risk assessment, personalized medicine, diagnosis, predicting therapy responses, and prognosis. They enhance decision-making, improve patient care, and optimize treatment plans by analyzing vast amounts of data efficiently (Rodríguez et al., 2022). AI innovations are being applied in healthcare in several impactful ways. For instance, they can predict mortality rates after cardiac surgery (Nilsson et al., 2006), facilitate the development of intelligent prostheses (Ortiz-Catalan et al., 2013), and diagnose skin cancer with accuracy comparable to or surpassing that of dermatologists (Esteva et al., 2017).

2.4.1 Artificial intelligence (AI) for TNBC

Recent progress in machine learning within AI has resulted in highly effective methods for modeling various types of molecular data (Eraslan et al., 2019; C. Xu & Jackson, 2019). The strength of these methods comes from their capacity to uncover relationships that traditional statistical methods might miss. This makes them essential for integrating the diverse molecular profiles associated with current triple negative breast cancer (TNBC) subtype systems (Ensenyat-Mendez et al., 2021; Jaber et al., 2020). These advancements are likely to drive a transformative shift in precision diagnosis and treatment for TNBC, addressing the complexities of this particularly aggressive form of breast cancer. By utilizing comprehensive data integration, AI models could enable more personalized and effective therapeutic strategies, ultimately leading to improved patient outcomes and survival rates. This progress represents a critical step forward in the fight against one of the most challenging types of breast cancer (Ensenyat-Mendez et al., 2021).

2.4.2 Challenges of using AI in clinical pharmacology

Currently, the primary challenges in drug discovery involve issues related to chemical structures, such as toxicity, side effects, and intellectual property, as well as selecting the appropriate drug target and dosage for specific patient populations. Despite significant advancements in clinical pharmacology, the disconnect between late preclinical and clinical data continues to impede the effective use of AI and machine learning technologies. Pharmaceutical companies typically do not share pharmacokinetic and pharmacodynamic data for most candidate drugs or their combinations unless they have received approval for human use. As a result, only a limited amount of drug discovery data is available for training AI and machine learning models, particularly regarding true negative data. This scarcity affects the ability to accurately differentiate between effective and ineffective drugs and hinder the development of robust AI and machine learning models in this field. This issue affects not only the differentiation between “drugs” and “nondrugs” but also impacts AI and machine learning models aimed at identifying novel target-disease associations, understanding the reasons behind discontinued clinical trials, the withdrawal of drugs, and accessing comprehensive datasets from successful trials. Nevertheless, it is clear that clinical development is undergoing significant transformation, and the usage of AI and machine learning is rapidly increasing (Zhavoronkov et al., 2020).

2.4.3 Artificial intelligence & tissue biomarkers

In the framework of Personalized Medicine (PM), a biomarker is a measurable biological characteristic that provides critical information about an individual's health status, disease risk, or

therapeutic response. These biomarkers can include genetic variations, proteins, metabolites, and other biological factors that can be quantified and analyzed. In oncology, biomarkers are especially valuable, as they can characterize tumors and inform decisions regarding targeted therapies. For instance, specific mutations in cancer cells may determine whether a patient is likely to benefit from particular treatments. Moreover, the use of biomarkers can streamline clinical trials by identifying suitable patient populations, ultimately accelerating the development of new therapies. AI can integrate data from various sources, including genomic, proteomic, and imaging data, to provide a comprehensive view of a patient's condition. This holistic approach enhances the understanding of how tissue biomarkers correlate with treatment responses and disease progression. The growing number of tissue biomarkers and the complexity associated with their assessment strongly support the use of AI-based tools in the evaluation process. As a result, AI enhances the role of pathology in Personalized Medicine (PM), enabling more precise diagnoses and tailored treatment strategies that align with individual patient profiles (Zhavoronkov et al., 2020).

2.5 Machine learning

Machine learning (ML) is a key component of artificial intelligence. ML involves algorithms that learn from data to make predictions by building models based on sample inputs. These algorithms identify patterns and relationships within the data, it is particularly useful for tasks where developing explicit algorithms is challenging. Common traditional ML methods include k-nearest neighbors (kNNs), logistic regression (LR), support vector machines (SVMs), gradient boosting machines (GBMs), and random forests (RF). The performance of these methods can differ based on the task type (regression or classification) as well as types and amount of data to handle (Zhavoronkov et al., 2020).

2.5.1 Machine learning (ML) based analysis of microRNA–target interactions

The identification of miRNA target sites on mRNAs is a crucial step in understanding the role of miRNAs in cellular processes. Recently, various high-throughput experimental techniques have emerged to identify these miRNA targets (J. Li & Zhang, 2019; Martinez-Sanchez & Murphy, 2013), The most prevalent and straightforward method involves assessing alterations in mRNA levels after the overexpression or inhibition of miRNAs in cultured cells (P. Xu et al., 2020). However, this method has significant limitations. First, the data can include indirect signals from downstream genes influenced by direct miRNA targets. Second, the precise binding site sequences are often unknown and need to be predicted within the relevant mRNA. Third, the experimental conditions may not accurately represent the physiological context of miRNA function, failing to capture endogenous targeting patterns. Lastly, this approach might miss signals related to translation efficiency inhibitions that affect gene expression but do not manifest as changes in mRNA levels (Fabian et al., 2010). The limited number of experimentally identified miRNA–target interactions has led to the increased use of computational predictions to broaden the range of identified miRNA–target relationships (Or & Veksler-Lublinsky, 2021). Technical challenges in applying high-throughput experimental methods have resulted in confirmed direct miRNA target datasets being restricted to a few model organisms. Nevertheless, machine learning (ML)-based target prediction models have been effectively trained and tested on some of these dataset (Ben Or & Veksler-Lublinsky, 2021).

2.5.2 Main types of tasks within ML: supervised and unsupervised

In modeling, machine learning allows machines to learn from mistakes, analyze data, identify patterns, and make informed decisions with little human involvement. The field of machine learning encompasses several categories, including supervised, unsupervised, semi-supervised, and reinforcement learning (Lewis, 2000). The two most commonly used categories are supervised and unsupervised learning. Supervised learning involves training a model on a labeled dataset, where the algorithm learns to map inputs to known outputs. This category is divided into classification and regression tasks, with classification yielding categorical outcomes and regression producing continuous values. Various supervised learning algorithms, including DT, LR, ANN, SVM and RF have proven effective in predicting prognostic and predictive biomarkers and in classifying the molecular subtypes of various cancers (Hastie et al., 2009). In contrast, unsupervised learning involves analyzing uncategorized and unlabeled data to identify patterns and structures (Al-Tashi et al., 2023).

2.5.3 Machine learning techniques used

Machine learning techniques, such as feature selection methods, are increasingly being used in biomarker discovery. Feature selection typically requires fewer assumptions than traditional statistical tests and can account for interactions between genes and their combined effects. This allows for the identification of genes that may be weak biomarkers on their own but exhibit strong joint power when considered together (X. Zhang et al., 2021). Another useful machine learning technique in biomarker discovery is classification. While classification isn't directly used to identify biomarkers, it assesses potential biomarkers selected by feature selection methods or statistical tests. True biomarkers should effectively differentiate samples from treated and control groups, making them informative for classification. The ability to enhance a classifier's prediction accuracy is commonly used as an evaluation metric for candidate methods. Moreover, the choice of classification algorithm can significantly impact the evaluation outcomes of biomarker discovery methods (X. Zhang et al., 2021). Finally, feature selection algorithms are important machine learning techniques that enhance classification performance. By identifying and retaining the most relevant genes while eliminating irrelevant or redundant ones, these algorithms improve the accuracy and efficiency of models. This process not only reduces dimensionality but also facilitates better interpretation of results, ultimately leading to more effective and reliable predictions in various applications, such as disease classification and biomarker discovery (Statnikov et al., 2008).

2.5.3.1 Feature selection methods

Feature selection methods can be categorized into filter, wrapper, and embedded methods. Generally, the filter method operates independently of classification models, relying solely on the inherent characteristics of the data to assess feature importance scores. Compared to other feature selection methods, it has lower time complexity, enabling more flexible integration with other algorithms for data preprocessing, including noise removal and dimensionality reduction (Hsu et al., 2011; Kavitha et al., 2020; Y. Zhang et al., 2008). Wrapper methods typically involve the use of a classifier to assess the value of a subset of features. These methods treat classifiers as a core component of the algorithm, evaluating the importance of selected features based on the classifier's performance, which often leads to improved model accuracy. Common wrapper methods include

stability selection, recursive feature elimination (RFE), genetic algorithms (GA), K-nearest neighbors (KNN), and particle swarm optimization (PSO) (Haury et al., 2012; Wang et al., 2017; Yan & Zhang, 2015). Embedded methods share a similar concept with wrapper methods, as they both involve classifiers. However, in embedded methods, the feature subset selection is integrated directly within the classifiers themselves, meaning that feature selection occurs simultaneously with classifier training. Common embedded methods include support vector machine (SVM), decision trees (DT), random forest algorithms (RF), and Lasso regression (LR) (Stein et al., 2005; Yu et al., 2021). Recent studies have indicated that hybrid feature selection methods can leverage the efficiency of filter methods alongside the accuracy of wrapper methods, resulting in enhanced performance. Additionally, some research has addressed the data imbalance issues often found in microarray datasets (Almugren & Alshamlan, 2019).

Alshamlan et al. introduced a new feature selection algorithm called the minimum redundancy maximum relevance (mRMR) method, which they combined with an Artificial Bee Colony (ABC) algorithm to filter biomarkers from gene microarray data. The mRMR method serves as a filtering technique to reduce the number of features and enhance the efficiency of the ABC algorithm. Using a support vector machine (SVM) as the classifier, the study compared mRMR-ABC with mRMR combined with a genetic algorithm (mRMR-GA) and mRMR with a particle swarm optimization algorithm (mRMR-PSO) across five datasets. The results indicated that the mRMR-ABC method achieved higher classification accuracy while utilizing fewer features (Alshamlan et al., 2015).

2.5.3.2 Machine learning classification algorithms: SVM and RF

Machine learning utilizing the maximization of the separating margin, known as support vector machine (SVM) is a powerful classification tool. The SVM classifier identifies a decision boundary, or hyperplane, that maximally separates data points into classes within the feature space (Huang et al., 2018). SVM offer a range of advantages that contribute to their popularity in classification tasks. They are particularly effective in high-dimensional spaces, succeeding even when the number of features exceeds the number of samples. SVMs reduce the risk of overfitting by maximizing the margin between classes, which helps to create a more generalized model. They can utilize different kernel functions to address non-linear relationships, providing versatility in handling complex datasets. Additionally, SVMs give a clear geometric representation by identifying the optimal hyperplane for class separation and are efficient in memory usage, as they only depend on a subset of the training data, known as support vectors (Huang et al., 2018). However, using SVM alone to evaluate feature selection methods can be biased; therefore, a random forest classifier is also employed.

The Random Forest algorithm relies on bagging, which creates random subsets of the dataset to build basic decision trees. This process is repeated to form a forest of trees that collectively provide predictions. At each tree node, the algorithm decides how to split the data based on one or more feature values. For classification tasks, the output is determined by the majority vote of the trees, often resulting in better performance than individual decision trees. Additionally, random forests are robust against overfitting, as adding more trees does not significantly enhance test performance beyond a certain point (Huang et al., 2018; Lavanya et al., 2023). The random forest algorithm is a powerful choice for classifying microarray data, thanks to several key advantages. It effectively manages situations where the number of predictors outnumbers the observations and automatically

selects relevant genes, making it robust against irrelevant data. Moreover, it accounts for interactions among predictors and utilizes ensemble learning, enabling it to learn both simple and complex classification patterns effectively. Random forest is suitable for both binary and multicategory classification tasks and typically requires minimal parameter adjustment, with its default settings often delivering excellent performance. These strengths make it an important tool for analyzing microarray data in clinical setting (Statnikov et al., 2008).

2.5.3.3 Feature selection algorithm: Recursive feature elimination (RFE)

Recursive feature elimination (RFE) is a sequential feature selection process where features are removed one at a time, or in small groups, through multiple iterations. The goal of RFE is to identify the most relevant features by progressively narrowing down the feature set. Initially, an estimator is trained using all available features to assess the importance of each variable, which can be derived from the coefficients of a linear regression model (coef_) or the feature importance from decision trees (feature_importance_). Subsequently, the least important feature or group of features is removed, and a new machine learning model is trained with the remaining features (Barzani et al., 2024).

Chapter Three (3): Literature Review

Zhong and his colleagues reported a study that provides an integrated analysis of gene expression profiles in triple-negative breast cancer (TNBC) compared to non-TNBC, identifying several key genes involved in the pathogenesis of TNBC. The research suggests that these differentially expressed genes (DEGs) may influence the initiation and progression of TNBC in various ways. Some of these key genes are novel, and their specific roles in TNBC are not yet fully understood. The findings from this study could offer valuable insights into the pathogenesis of TNBC (Ganggayah et al., 2019).

Ganggayah and his colleagues reported a study where machine learning models were developed using breast cancer data from the University Malaya Medical Centre to identify key prognostic factors for breast cancer survival. All algorithms, including decision trees, random forests, neural networks, extreme boosting, logistic regression, and support vector machines, produced similar accuracies, with random forest slightly outperforming the others. The goal of applying ML models within the same institution where the patients in the training set were treated was to reduce hidden variables, thereby minimizing bias from differing surgical techniques or dosage regimens (Boeri et al., 2020).

Rhodes and his colleagues reported a study showing that the tumor-suppressive miRNA family, miR-200, is not expressed in triple-negative breast cancer (TNBC) cell lines. They found that overexpression of miR-200b-3p represses epithelial-mesenchymal transition (EMT), evidenced by reduced migration and increased CDH1 expression. However, despite the loss of migratory capacity following miR-200b-3p re-expression, no subsequent decrease in the conventional miR-200 family targets and EMT markers ZEB1/2 was observed (Rhodes et al., 2015).

Iorio et al. conducted a study in which a microRNA expression signature was identified to differentiate between normal and cancerous breast tissues. They utilized a miRNA microarray to analyze expression profiles from 10 normal and 76 neoplastic breast tissue samples. Each tumor sample came from a single specimen, while 6 of the 10 normal samples were pooled from five different normal breast tissue RNAs, leading to a total of 34 normal breast samples being examined. To identify miRNAs with significantly different expression between normal and tumor samples, they employed ANOVA and class prediction statistical tools (Iorio et al., 2005).

Xu et al. and his colleagues found that miR-21 is overexpressed in NSCLC cell lines compared to the lung epithelial cell line BEAS-2B. Additionally, high levels of miR-21 are linked to the expression of PTEN, RECK, and Bcl-2 (L. Xu et al., 2014).

Rehman et al. applied four distinct feature-selection techniques, namely Information Gain (IG), Chi-Squared (CHI2), and Least Absolute Shrinkage and Selection Operator (LASSO), to identify the most relevant and effective miRNAs for distinguishing between normal and cancerous tissues. After selecting the features, they utilized Random Forest (RF) and Support Vector Machine (SVM) algorithms to classify cancerous cells. Their findings indicated that miRNAs with higher ranks, based on their feature-selection analysis, led to better classifier performance. Conversely, as the rank of the miRNAs decreased, the performance of the classifiers diminished, suggesting that the miRNAs with higher rankings held greater importance as biomarkers with varying degrees of discriminative power (Rehman et al., 2019).

Indra Waspada and his colleagues conducted a comparative analysis of various supervised machine learning techniques to identify the most suitable method for classifying cancer cells based on microRNA gene expression. The effectiveness of each method was evaluated primarily by measuring the accuracy of their results. Among the methods tested, Decision Trees (DT), Naive Bayes (NB), neural networks, and deep learning (DL). The researchers aimed to determine which approach was best suited for gene analysis. Their findings revealed that the deep learning (DL) method, which was based on a multilayer feed-forward artificial neural network (ANN) and trained using stochastic gradient descent with backpropagation, outperformed the other conventional machine learning techniques. This approach showed superior accuracy and efficiency in classifying cancer cells, highlighting its potential for more reliable gene expression analysis compared to traditional methods (Waspada et al., 2017).

Jayanta K. Pal and his colleagues highlight several significant advancements in blood-based breast cancer diagnosis. First, they demonstrate that moderately high miRNA expression levels are essential for their inclusion in machine learning models. This is further validated by density plots of miRNA expressions, which show that the selected miRNAs predominantly occupy the high expression range in the Gaussian plot. Second, the study reveals that the dysregulation patterns of these miRNAs show both upregulation and downregulation in cancerous conditions, suggesting their potential roles as either suppressors or enhancers of the disease phenotype. Evaluation results show strong performance using this miRNA set. In conclusion, if miRNA expression levels are sufficiently high, they can be effective biomarkers for blood-based breast cancer diagnosis (Pal & Rami, 2024).

Wengong Si and his colleagues focus on the miRNAs involved in regulating drug resistance across various cancers, investigating the mechanisms behind the altered expression of these miRNAs. They also conduct an in-depth exploration of the molecular targets of miRNAs and the signaling pathways involved. Gaining a comprehensive understanding of the role of miRNAs in drug resistance will aid in developing more effective strategies to regulate them, ultimately facilitating the clinical translation of miRNAs and advancing their potential as a promising approach in cancer therapy (Si et al., 2019).

Nina Petrovic and her colleagues proposed that the expression levels of specific miRNAs could enhance treatment efficacy by identifying the stage of chemotherapy or radiotherapy sensitivity. The use of miRNAs, either alone or in combination with conventional therapeutic strategies, has

the potential to significantly improve the success of cancer treatments in the future (Petrovic & Ergun, 2018).

Many researches focused on identifying new biomarkers for targeting cancer. Unlike the majority of other studies, this study is not limited just to DEA based on filter methods to differentiate between the two classes (healthy and disease), but through AI based on ML methods, it also identifies a new biomarker with diagnostic relevance.

Chapter Four (4): Methodology

4.1 Model workflow

The present study demonstrates an unconventional and important workflow that deduced the candidate diagnostic and prognostic biomarkers associated with TNBC. The workflow involved several stages: data acquisition, preprocessing, integration, filtration and differential expression analysis (DEA) using multiple statistical tests. A schematic representation of the model workflow is included to illustrate the step-by-step process (**Figure 4.3**).

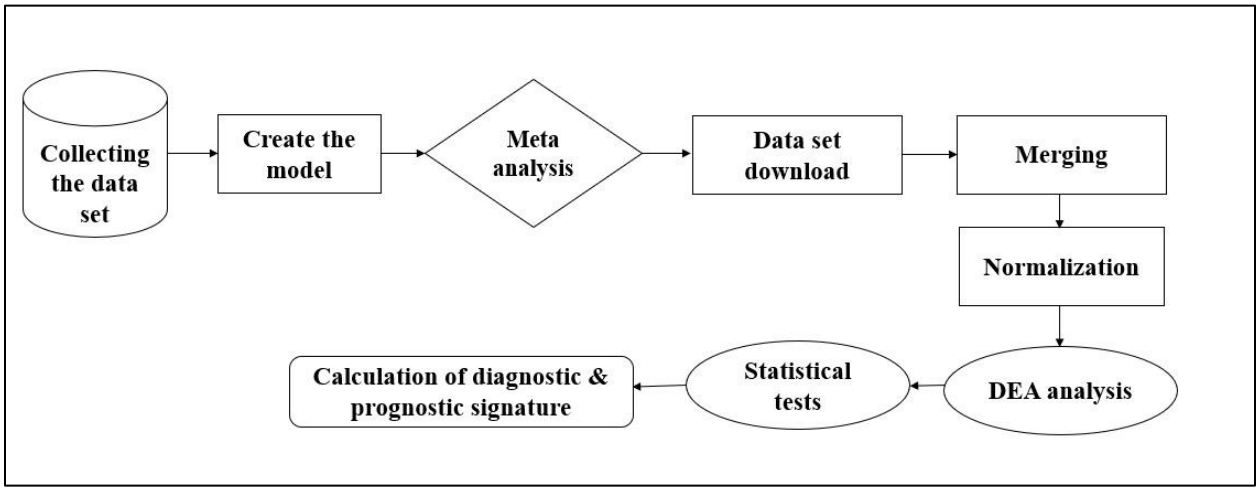

Figure 4.3 1Schematic representation of the model creation workflow

4.2 Searching GEO Database

The Gene Expression Omnibus (GEO) is a database managed by the National Center for Biotechnology Information (NCBI). Currently, all records in GenBank are generated through direct submissions from the original authors, who voluntarily provide their data for public access or as part of the publication process. The NCBI GEO is a pivotal resource for the scientific community, designed to manage a wide array of expression data types, including gene expression profiles, microarray datasets, and RNA sequencing information (Https://www.ncbi.nlm.nih.gov/gds/, n.d.).

4.3 Create a model

The created model was aimed to identify a statistically significant miRNA signature by merging two separate miRNA expression datasets. This approach strengthens the robustness of our analysis and facilitates the identification of relevant miRNA patterns. This model integrates a (A) meta-analysis (multiple dataset integration), (B) differential gene expression analysis, and (C) identifies statistically significant correlations among multi-gene signatures.

4.3.1 Meta analysis

Meta-analysis refers to the statistical integration of findings from two or more independent studies. Its potential benefits include enhanced precision, the capacity to address questions that individual studies may not cover, and the potential to clarify disputes from conflicting results. In our study meta-analysis (dataset download, normalization and merging) is used to identify a gene expression signature by merging two microarray datasets.

4.3.1.1 Dataset download

The workflow of the datasets that obtained from microarray profiling available in the GEO database is presented in **Figure 4.4**. The datasets of (GSE100453) are 24 tissue samples obtained from 6 TNBC patients and 18 other breast cancer patients, and (GSE154255) are 20 tissue samples obtained from 3 TNBC patients and 14 other breast cancer patients with 3 normal tissues. Microarray expression datasets were obtained using these keywords: microRNA, Breast cancer, Triple Negative Breast Cancer, Gene expression, Biomarker, Machine learning as shown in **Figure 4.4.** Some factors for selecting the studies for the research include the power of the study (determined by the number of samples), disease prevalence, the analytical validity of the biomarker test, and the preplanned analysis strategy. It's worth noting that using GEO for analysis of microRNA gene expression data requires a significant amount of expertise and resources. It's important to work with a team of experts, including biologists, bioinformaticians, and data scientists, to ensure that the research is following best practices and producing valid, reliable results.

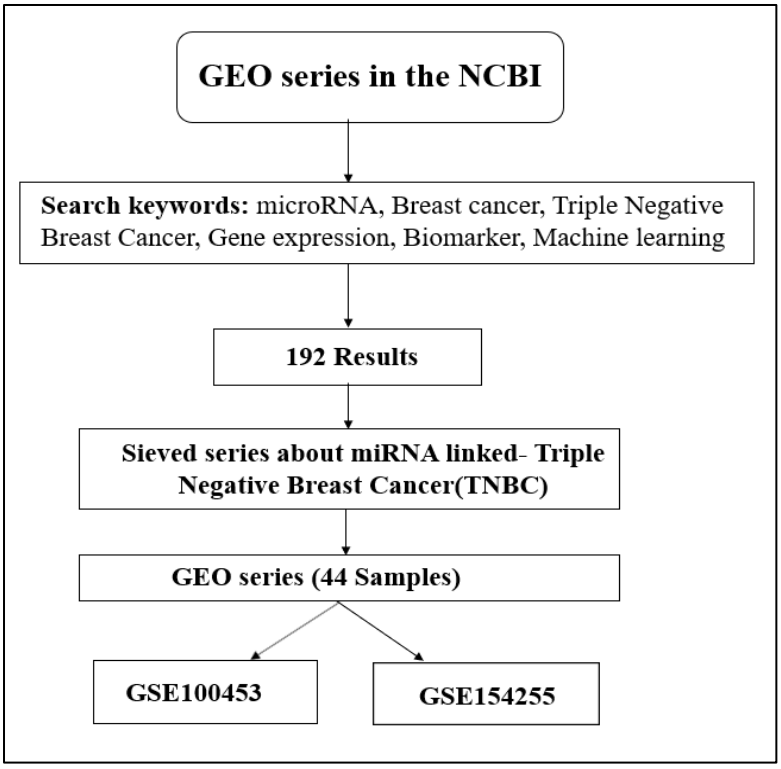

Figure 4.4 1Schematic represents the process for choosing the datasets of our study

4.3.1.2 Import the data

The datasets were manually reviewed to ensure they met the following criteria: (A) they should consist of miRNA gene expression profiles from tissues (excluding cell lines), (B) only studies that involved human samples were selected; and (C) studies should not include any drug treatment. Comprehensive details, such as accession numbers and sample sizes, are listed in **Table 4.1**.

Table 4.1 1Summary for the selected datasets of our study

Study accession	Sample#	TNBC#	TNBC adjacent normal#	Non-TNBC#	miRNAs #
GSE100453	24	6	0	18	2571
GSE154255	20	3	3	14	2007

4.3.1.3 Data Processing and merging

To achieve the best representation of the data, it's important to find a lower-dimensional representation, as high-dimensional data can degrade the performance of the applied methods. A large number of features increases computational complexity. Moreover, slowing down the classification process. Additionally, accurate diagnosis relies heavily on the selection of appropriate features. Data preprocessing techniques, such as merging datasets, resulted in a combined dataset of 4577 miRNAs.

4.3.1.4 The inclusion criteria

The clinical information was utilized for survival analysis based on the following criteria: (A) complete follow-up data were available for periods ranging from 1 to 60 months (30 to 1825 days); (B) all clinical data were thorough (patients with uncertain or missing information were excluded); and (C) the integrity of the miRNA-seq data was confirmed (patients without individual miRNA values were excluded).

4.3.1.5 Normalization

Normalization procedures were applied to address variations in expression levels across different studies. A maximum filter approach was utilized to remove low-expressing genes, thereby reducing noise and enhancing the detection of differentially expressed miRNAs. The final dataset included 189 miRNAs that met the inclusion criteria and statistical thresholds ($P<0.05$, $|\log2FC|>2$).

4.3.2 Differential Expression Analysis

Following the quantification of genes, differential expression analysis (DEA) is performed to identify genes that show varying expression levels between two sample classes (e.g. healthy, disease).

4.3.3 Statistical tests

The expression levels of a specific gene across all samples were assessed. This study aims to determine whether this gene shows different expression patterns between the two groups.

The Fold Change (FC) measures the difference in gene expression between the TNBC and control groups. Log Fold Change is employed so that a positive value indicates the gene is upregulated, while a negative value shows downregulation.

The significance of differentially expressed genes is usually assessed through the P-value derived from statistical tests, which are univariate and treat each gene independently, ignoring the interactions between genes. Genes with the lowest P-values are considered the most significant.

4.4 Visualization

The results of DEA can be visualized in several ways with the heatmap being one of the most widely used. A heatmap is effective for visualizing gene expression across samples from various conditions. Specifically, a correlation heatmap displays a two-dimensional correlation matrix between two distinct dimensions, utilizing colored cells to represent data, often on a monochromatic scale. The values of the first dimension are represented in the rows, while those of the second dimension appear in the columns. This type of visualization is commonly employed to highlight the top differentially expressed genes from the DEA results.

4.5 Make predictions

These models could serve as an additional resource for assessing the prognosis of breast cancer patients in our routine clinical practice (clinical association analysis) (Boeri et al., 2020). This study hypothesizes that meta-analysis of publicly available gene expression datasets can identifies novel diagnostic and prognostic biomarker in triple negative breast cancer patients.

4.6 Evaluate and improve

Gene expression analysis often faces challenges such as selection bias, inadequate sample quality, and poor sample size estimation, all of which can impact the statistical power and validity of further analyses. Meta-analysis of diverse gene expression datasets has been shown to enhance statistical power and reduce selection biases, facilitating the identification of diagnostic and prognostic biomarkers. However, the selection of differentially expressed genes through meta-analysis largely depends on univariate p-value statistics, which complicates the task of identifying gene sets with non-redundant information and determining the optimal number of genes that accurately reflect the data. This challenge limits the development of diagnostic and prognostic signatures that take into account various feature selections and covariates, such as patient characteristics (e.g., survival) and histology. To overcome these obstacles, this study implements a meta-analysis that merges multiple gene expression datasets into a single array and subsequently apply machine learning techniques to identify biomarker signatures.

4.7 Machine learning analysis with python

This study will implement a machine learning analysis using Python, specifically focusing on algorithms that offer attribute importance measures, such as (coef_) or (feature_importance_). In this analysis, Recursive Feature Elimination (RFE) will be applied along with Random Forest (RF) and Support Vector Machine (SVM) models, focusing on identifying important features from datasets related to triple-negative breast cancer (TNBC) and adjacent normal tissues will be employed. The Scikit-learn library is used to facilitate this analysis. In Scikit-learn, RFE involves removing features based on their importance, repeating steps 2-4 until a stopping criterion is met. The stopping criterion in Scikit-learn is typically defined as an arbitrary number of final features. The steps for RFE in Scikit-learn are as follows:1) Train a machine learning model. 2) Derive feature importance. 3) Remove the least important feature(s). 4) Re-train the machine learning model on the remaining features. 5) Repeat until the desired number of features is reached.

4.8 Feature selection and machine learning model development

The study employed a hybrid approach, combining filter, wrapper, and embedded methods to select the most relevant features. Specifically; Filter Method: Statistical tests were used to rank features based on their correlation with TNBC. Besides, wrapper Method: Recursive Feature Elimination (RFE) was utilized to iteratively remove the least significant features, guided by the performance of a predictive model. Finally, embedded Method: Random Forest (RF) and Support Vector Machine (SVM) models were employed to simultaneously select features and classify samples.

4.9 Validation of machine learning model

Model validation involves procedures and actions that evaluate the accuracy of a machine learning model after training, offering insights into its reliability. Various methods for model validation include train/test split, K-fold cross-validation, leave-one-out cross-validation, and nested cross-validation. The choice of the appropriate method depends on the size and structure of the dataset, as well as the specific objectives of the analysis. In this study, train/test split validation was used.

Chapter Five (5): Results and Discussion

5.1 Meta analysis: Data processing and Normalization

The two data sets were merged and filtrated, resulting in 189 miRNAs that met both the inclusion and cutoff criteria ($P<0.05$ and $|log2FC|>2$) for this study. This process, referred to as filter feature selection, narrows down the number of genes to hundreds based on the correlation between each sample and gene expressions according to the established correlation threshold (see **Appendix 1**).

Due to the diverse normalization techniques, statistical tests, and other specific factors employed in different studies, various differential expression analysis (DEA) tools can yield significantly different results for the same dataset.

5.2 Differential Expression Analysis

The fold change in expression levels ($|Log\text{-}FC| > 2$) was utilized to quantify the differences in expression between TNBC and adjacent normal tissues, as well as between TNBC and non-TNBC tissues. Notably, significant differences in miRNA expression were observed between TNBC and adjacent normal tissues (**Figure 5.7**). In contrast, no significant differences in miRNA expression were found between TNBC and non-TNBC tissues **(Figure 5.9)**. The study identified 20 miRNAs significantly associated with TNBC, with 17 miRNAs upregulated and 3 miRNAs downregulated in TNBC tissues (**Table 5.2**).

5.3 Clinical independence prognosis analysis for the model

To evaluate the independent prognostic ability of the model, the miRNAs meeting $p < 0.05$ were considered statistically significant. 20 miRNAs with $p < 0.05$ were considered as independent prognostic factors to predict the prognosis of TNBC, these miRNAs were considered independent risk factors associated with over survival (OS) time in the model (**Table 5.2**). Identifying independent prognostic factors is essential as it assists clinicians in determining appropriate treatment strategies.

5.4 Identification and analysis of differentially expressed miRNAs in TNBC

A total of 20 miRNAs significantly associated with OS time (with $p < 0.05$). The fold-change (FC) was used to indicate the difference in miRNA expression between TNBC and Non-TNBC (|Log-FC|>2). 3 miRNAs were down-regulated and 17 miRNAs were up-regulated (**Table 5.2; Figure 5.5)**. All of them were highly expressed in TNBC tissue.

The top-ranked miRNAs expression correlation heatmap is shown in **Figure 5.6**, strong miRNA expression correlation was investigated for the 20 miRNAs.

Table 5.2 1The selected miRNAs with their adjusted p-value, p-value, log-fold change and Gene expression.

miRNA-ID	Adj-P-Val	P-Value	Log-FC	Gene expression
hsa-miR-135b-5p	0.00392	0.00000153	5.261439	Up-regulated
hsa-miR-521	0.02096	0.00001631	4.769512	Up-regulated
hsa-miR-18a-5p	0.08956	0.00013068	2.67759	Up-regulated
hsa-miR-18b-5p	0.08956	0.00014512	2.565207	Up-regulated
hsa-miR-522-3p	0.08956	0.00020727	3.760554	Up-regulated
hsa-miR-517c-3p	0.08956	0.00020995	3.66183	Up-regulated
hsa-miR-516a-5p	0.08956	0.00025061	3.490997	Up-regulated
hsa-miR-32-5p	0.08956	0.00030142	4.011881	Up-regulated
hsa-miR-182-3p	0.08956	0.00031363	3.654194	Up-regulated
hsa-miR-214-5p	0.09725	0.0003784	-3.81504	Down-regulated
hsa-miR-4291	0.16164	0.00069183	1.696896	Up-regulated
hsa-miR-17-3p	0.22697	0.00105977	2.223859	Up-regulated
hsa-miR-557	1	0.00137	-3.9875	Down-regulated
hsa-miR-545-3p	0.30919	0.00156399	2.47794	Up-regulated
hsa-miR-1181	0.33641	0.00190402	4.533229	Up-regulated
hsa-miR-301b-3p	0.33641	0.00196346	3.35425	Up-regulated
hsa-miR-934	0.34634	0.00215619	3.512173	Up-regulated
hsa-miR-198	0.38338	0.002536	-3.46075	Down-regulated
hsa-miR-17-5p	0.39435	0.00277314	1.740012	Up-regulated
hsa-miR-1825	0.39435	0.00299908	3.681377	Up-regulated

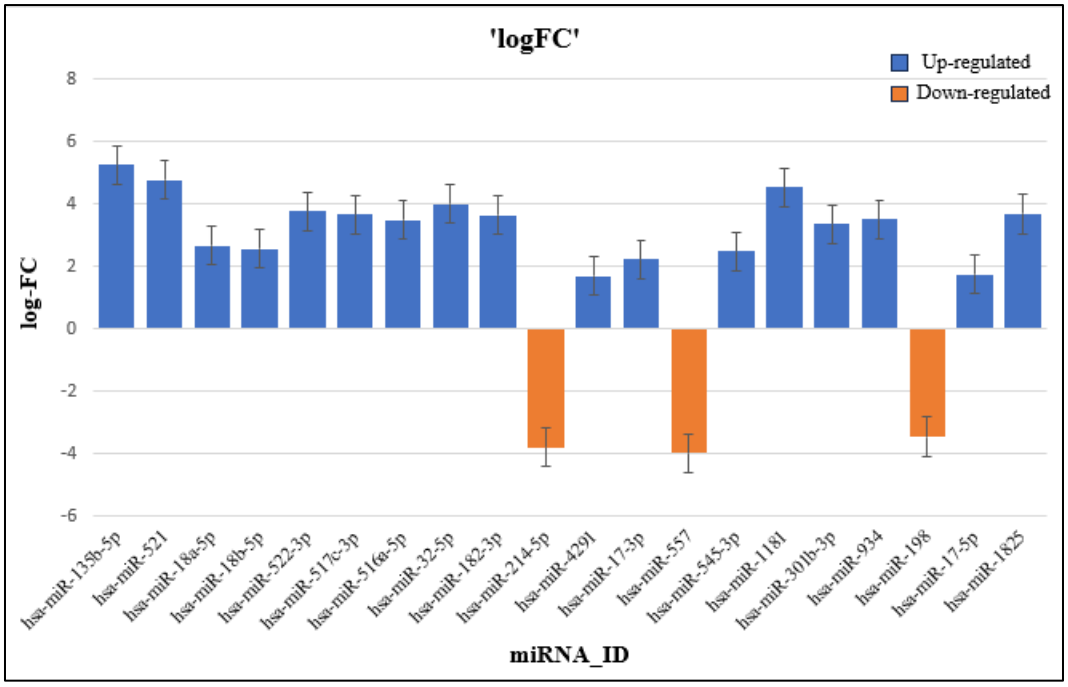

Figure 5.5 1The selected miRNAs with their adjusted p-value, p-value, log-fold change and Gene expression

A heatmap of the selected 20 differentially expressed miRNAs. Orange represents upregulation, and yellow represents downregulation (**Figure 5.6**).

	hsa-miR-135b-5p	hsa-miR-521	hsa-miR-18a-5p	hsa-miR-18b-5p	hsa-miR-522-3p	hsa-miR-517c-3p	hsa-miR-516a-5p	hsa-miR-32-5p	hsa-miR-182-3p	hsa-miR-214-5p	hsa-miR-4291	hsa-miR-17-3p	hsa-miR-557	hsa-miR-545-3p	hsa-miR-1181	hsa-miR-301b-3p	hsa-miR-934	hsa-miR-198	hsa-miR-17-5p	hsa-miR-1825
hsa-miR-135b-5	1																			
hsa-miR-521	0.970757547	1																		
hsa-miR-18a-5	0.864181583	0.93521372	1																	
hsa-miR-18b-5	0.852945982	0.92626813	0.99699527	1																
hsa-miR-522-3	0.829719987	0.93794969	0.963713983	0.960577243	1															
hsa-miR-517c-3	0.829717461	0.93727397	0.967717405	0.964899071	0.999853299	1														
hsa-miR-516a-5	0.814016671	0.92638678	0.968941082	0.966947414	0.998891544	0.999405756	1													
hsa-miR-32-5p	0.788347422	0.91281944	0.936481702	0.93321761	0.995689855	0.994444085	0.993914739	1												
hsa-miR-182-3	0.78941017	0.91217253	0.932574072	0.930436863	0.997629878	0.997404293	0.998197383	0.998221878	1											
hsa-miR-214-5	-0.748165614	-0.8848895	-0.928590571	-0.926729964	-0.989917728	-0.989020409	-0.990664669	-0.99724955	-0.996780017	1										
hsa-miR-4291	0.646808335	0.77377797	0.942691201	0.949684355	0.910455051	0.91652805	0.928788682	0.896577379	0.920274728	-0.911004563	1									
hsa-miR-17-3p	0.610701609	0.76600581	0.916107392	0.921701638	0.932072621	0.935541727	0.947080045	0.934190318	0.949046696	-0.951128758	0.98367955	1								
hsa-miR-557	-0.423997546	-0.1984188	0.04380379	0.059404842	0.153145483	0.153653922	0.180808532	0.216591574	0.220436917	-0.279214744	0.34733941	0.43971421	1							
hsa-miR-545-3	0.556894248	0.73019656	0.875307694	0.880609418	0.919318222	0.921378837	0.933217772	0.931875126	0.941932307	-0.951671184	0.9575894	0.993524895	0.5106903	1						
hsa-miR-1181	0.552549285	0.73484477	0.779629667	0.778185515	0.906871971	0.902891283	0.906769959	0.941129377	0.930116648	-0.953040074	0.81534581	0.903093648	0.4829167	0.937555704	1					
hsa-miR-301b-3	0.543069302	0.72762459	0.82785385	0.830112309	0.918119557	0.917068493	0.925570898	0.943733026	0.942999585	-0.96117481	0.89478979	0.960250122	0.5253436	0.983280899	0.984638222	1				
hsa-miR-934	0.531933407	0.71904704	0.812852715	0.814706058	0.911524282	0.909940022	0.91804287	0.939820961	0.937233357	-0.957276621	0.8794322	0.950415667	0.5330058	0.976849419	0.989141966	0.989444277	1			
hsa-miR-198	-0.939500193	-0.9898479	-0.917345198	-0.907855427	-0.951268547	-0.94902464	-0.938216306	-0.937045943	-0.9311082	0.914428047	-0.76677142	-0.779447826	0.1254801	-0.755684404	-0.7917106	-0.771272724	-0.76615634	1		
hsa-miR-17-5p	0.429261804	0.6131813	0.814241493	0.823708964	0.83965727	0.84372975	0.861154784	0.852516214	0.869568042	-0.881764022	0.94978976	0.977036912	0.6124494	0.983649634	0.879227187	0.948122181	0.9396705	-0.6340531	1	
hsa-miR-1825	0.487005333	0.68205692	0.773617004	0.77543341	0.886365607	0.88406278	0.892431009	0.920320633	0.915245339	-0.939620406	0.84977736	0.930711622	0.5698939	0.964255146	0.961068917	0.985422232	0.98787629	-0.7354787	0.928987755	1

Figure 5.6 1Correlation heatmap for the expression of the top-ranked miRNAs

5.5 GSE154255 (TNBC VS. TNBC adjacent normal)

The GSE154255 dataset was processed to filter out miRNAs, leading to the identification of 10 miRNAs that met the inclusion and cut-off criteria (P<0.05 and |log2FC|>2), as presented in **Table 5.3**. Significant differences in miRNA expression were found between triple-negative breast cancers and adjacent normal tissues (healthy control samples), as shown in **Figure 5.7**. **Appendix 2** contains the data from GSE154255 before and after the filtration process.

Table 5.3 1The differentially expressed miRNAs in TNBC and TNBC adjacent normal

ID_REF	TNBC adjacent normal 1	TNBC adjacent normal 2	TNBC adjacent normal 3	TNBC cancer 1	TNBC cancer 2	TNBC cancer 3
hsa-miR-1181	11.9577	16.8671	14.1585	12.7311	36.2556	19.9412
hsa-miR-135b-5p	0.1	14.8205	0.1	221.314	0.1	0.1
hsa-miR-17-3p	0.1	5.36628	0.1	17.3129	24.3551	10.4097
hsa-miR-182-3p	0.1	0.1	0.1	0.1	0.1	0.1
hsa-miR-1825	10.9786	10.3093	11.2726	9.7202	7.97509	11.862
hsa-miR-18a-5p	0.1	0.1	0.1	24.4116	30.387	16.1978
hsa-miR-18b-5p	0.1	0.1	0.1	6.54062	10.4035	7.81851
hsa-miR-198	3.6499	4.47031	10.2987	0.1	0.1	0.1
hsa-miR-214-5p	0.1	19.5547	13.3654	0.1	0.1	41.0854
hsa-miR-32-5p	0.1	0.1	0.1	2.00933	25.6167	17.5805

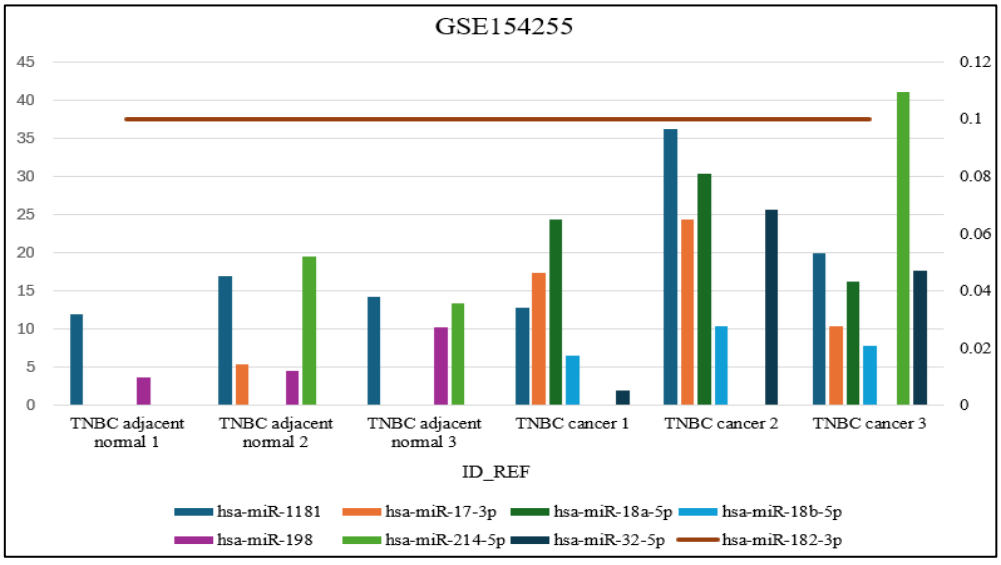

Figure 5.7 1The differentially expressed miRNAs in TNBC and TNBC adjacent normal (1)

When TNBC tissue compared with adjacent normal TNBC tissues (**Figure 5.7**), the expression of miR-214-5p and miR-198 was significantly reduced (p= 0.0003784 and p= 0.002536 respectively) whereas expression of miR-1181 and miR-135b-5p was significantly increased (p= 0.00190402 and p= 0.00000153 respectively).

Interestingly, there is one miRNA (miR-135b-5p) that was highly up-regulated for one TNBC sample in this study (**Figure 5.8**). To our knowledge, miR-135b-5p has previously been described in relation to TNBC (Naorem et al., 2019; Paszek et al., 2017; Piña-Sánchez et al., 2020).

The study also identified that hsa-miR-18a-5p, hsa-miR-18b-5p, hsa-miR-32-5p were significantly differentially expressed in the three TNBC samples, all were upregulated (**Figure 5.8**).

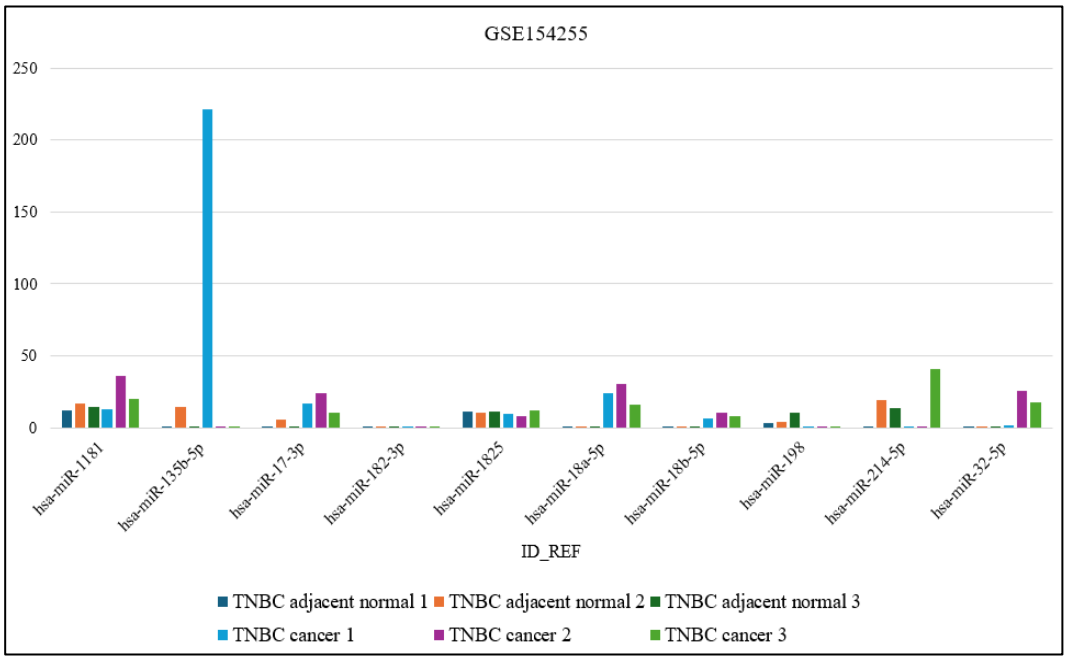

Figure 5.8 1The differentially expressed miRNAs in TNBC and TNBC adjacent normal (2)

Our study confirms that miRNA expression profile is dysregulated in TNBC patients compared to healthy controls. Four miRNAs (miR-135b-5p, hsa-miR-18a-5p, hsa-miR-18b-5p and hsa-miR-32-5p) may be associated with development and progression of TNBC.

5.6 GSE100453 (TNBC VS. Non-TNBC)

GSE100453 dataset was filtrated, a total of 20 miRNAs who met the inclusion criteria and the cut-off criteria (P<0.05 and |log2FC|>2) were selected (**Table 5.4**), The data of GSE154255 before and after filtration is shown in **Appendix 3**.

The differentially expressed miRNAs in TNBC tissue and non-TNBC tissue (luminal B and lymph node non- metastasis_1, luminal B and lymph node non-metastasis_2, luminal B and lymph node non-metastasis_3, luminal B and lymph node metastasis_1, luminal B and lymph node metastasis_2, luminal B and lymph node metastasis_3, Her2 amplification and lymph node non-metastasis_1, Her2 amplification and lymph node non-metastasis_2, Her2 amplification and lymph node non-metastasis_3, Her2 amplification and lymph node metastasis_1, Her2 amplification and lymph node metastasis_2, Her2 amplification and lymph node metastasis_3, triple negative and lymph node non-metastasis_1, triple negative and lymph node non-metastasis_2, triple negative and lymph node non-metastasis_3, triple negative and lymph node metastasis_1, triple negative and lymph node metastasis_2, triple negative and lymph node metastasis_3) is shown in **Figure 5.9 and Table 5.4.**

There was no significant difference in survival time between the low- and high-expression miRNAs. All were highly expressed in both; TNBC and non-TNBC tissue.

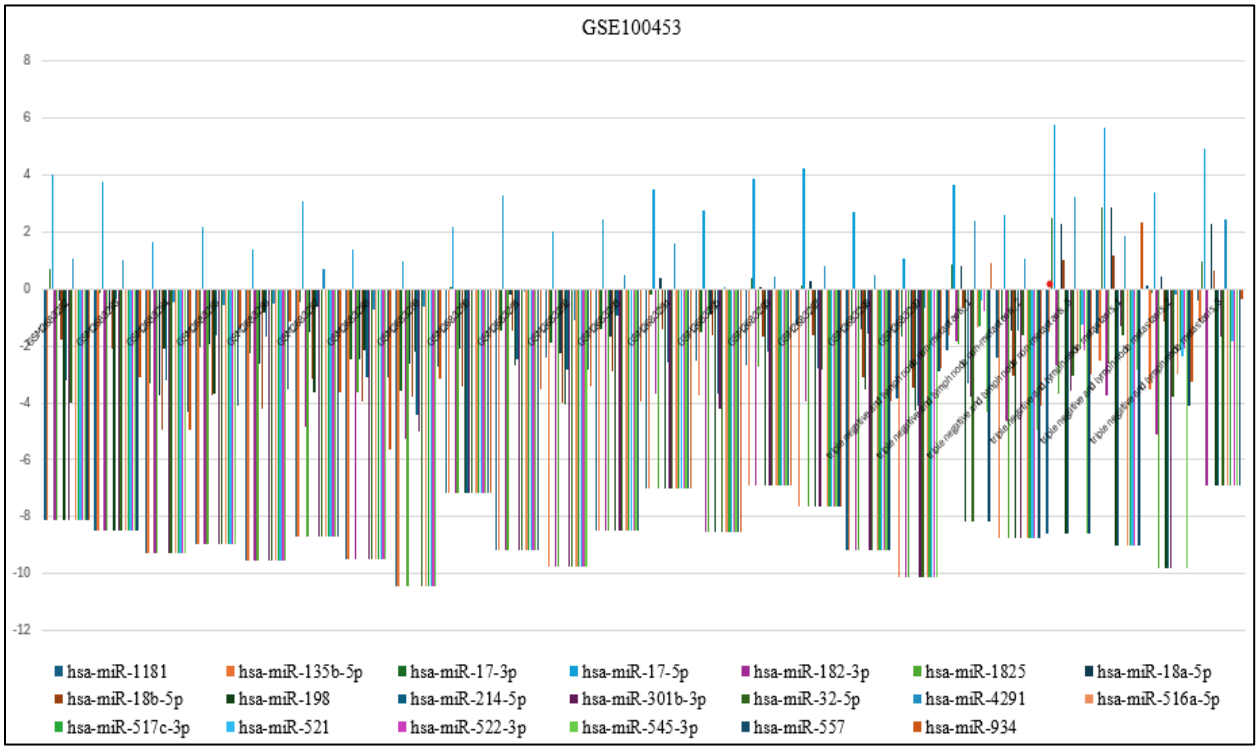

Figure 5.9 1The differentially expressed miRNAs in TNBC and non-TNBC

Table 5.4 1The differentially expressed miRNAs in TNBC and non- TNBC

	luminal B and lymph node metastasis 3	Her2 amplification and lymph node non-metastasis 1	Her2 amplification and lymph node non-metastasis 2	Her2 amplification and lymph node non-metastasis 3	Her2 amplification and lymph node metastasis 1	Her2 amplification and lymph node metastasis 2	Her2 amplification and lymph node metastasis 3	triple negative and lymph node non-metastasis 1	triple negative and lymph node non-metastasis 2	triple negative and lymph node non-metastasis 3	triple negative and lymph node metastasis 1	triple negative and lymph node metastasis 2	triple negative and lymph node metastasis 3
	8.466671	-7.0221233	-2.5434604	-2.6670396	-1.2615921	-9.184891	-3.8202271	-2.1323178	-2.4397817	-8.585818	-1.5667415	0.11698627	-0.4150531
	8.466671	-7.0221233	-3.7376356	-6.892729	-7.6175776	-9.184891	-10.087188	-1.627481	-8.751094	-1.5407138	-2.5438297	-3.5018873	-1.1366317
	1.4191	-0.2385907	-1.5356703	0.39178085	0.10270214	-1.096045	-1.6812253	-0.847733	-1.0872965	2.4765954	2.854123	-0.1585917	0.9278188
	2.4452248	5.5006602	2.7149944	3.868964	4.188595	2.6718917	1.0626884	3.645028	2.5616055	5.764859	5.6278477	3.3700662	4.877158
	8.466671	-3.700195	-8.507984	-6.892729	-3.9371295	-9.184891	-10.087188	-1.8608987	-4.6488304	-2.6984954	-3.7613955	-5.1287837	-6.9055834
	8.466671	-7.0221233	-8.507984	-2.711801	-7.6175776	-9.184691	-10.087188	-1.95225	-8.751094	-3.684691	-2.6503932	-9.777025	-0.0071032
	1.6645918	0.3866818	-0.8929262	0.04846072	0.28475666	-1.4007077	-2.8120897	-0.8182735	-1.483177	2.2403789	2.8299427	0.4465947	2.2574897
	2.9058368	-1.4196336	-1.6104565	-1.7034653	-1.6413927	-3.1224427	-3.486786	-0.6724505	-3.0551224	1.0285606	1.1391315	-1.1460981	0.6199317
	8.466671	-7.0221233	-8.507984	-6.892729	-7.6175776	-3.5268302	-4.2452173	-8.176063	-8.751094	-8.585818	-9.016602	-9.777025	-6.9055834
	0.95085764	2.380779	-3.6779008	-2.2394295	-2.772509	-1.5818901	-4.114686	-3.3223064	-1.4817653	-8.585818	-9.016602	-9.777025	-6.9055834
	8.466671	-7.0221233	-4.1965313	-6.892729	-7.6175776	-9.184891	-10.087188	-3.7779174	-8.751094	-3.576489	-1.3134379	-9.777025	-1.6875346
	8.466671	-7.0221233	-8.507984	-6.892729	-2.8375983	-9.184891	-10.087188	-8.176063	-1.6364269	-3.0738761	-1.6590195	-3.8004115	-6.9055834
	0.49777365	1.5636919	0.03682232	0.44944215	0.8109665	0.4726729	-0.6700255	2.3766737	1.0579443	3.2231216	1.8309584	-0.2162313	2.4415593
	8.466671	-7.0221233	-8.507984	-6.892729	-7.6175776	-9.184891	-10.087188	-1.381712	-8.751094	-2.2855134	-9.016602	-2.973901	-6.9055834
	8.466671	-7.0221233	-8.507984	-6.892729	-7.6175776	-9.184891	-10.087188	-1.3384347	-8.751094	-2.1031215	-9.016602	-2.174571	-6.9055834
	8.466671	-7.0221233	-8.507984	-6.892729	-7.6175776	-9.184891	-10.087188	-0.4046034	-8.751094	-1.2589602	-9.016602	-2.3751388	-1.836917
	8.466671	-7.0221233	-8.507984	-6.892729	-7.6175776	-9.184891	-10.087188	-0.7731428	-8.751094	-2.153517	-9.016602	-2.097126	-6.9055834
	8.466671	-7.0221233	-8.507984	-6.892729	-7.6175776	-9.184891	-10.087188	-4.3008833	-4.9598427	-8.585818	-2.8636475	-9.777025	-6.9055834
	8.466671	-7.0221233	-8.507984	-6.892729	-7.6175776	-9.184891	-2.879212	-8.176063	-8.751094	-8.585818	-9.016602	-4.098671	-6.9055834
	3.935452	-7.0221233	-8.507984	-6.892729	-7.6175776	-9.184891	-2.8059967	0.8762708	-4.1293	-3.0149346	2.3357096	-3.3696712	-0.3716199

ID_REF	luminal A and lymph node non-metastasis 1	luminal A and lymph node non-metastasis 2	luminal A and lymph node non-metastasis 3	luminal A and lymph node metastasis 1	luminal A and lymph node metastasis 2	luminal A and lymph node metastasis 3	luminal B and lymph node non-metastasis 1	luminal B and lymph node non-metastasis 2	luminal B and lymph node non-metastasis 3	luminal B and lymph node metastasis 1	luminal B and lymph node metastasis 2
hsa-miR-1181	-8.1022215	-8.468683	-9.269625	-8.953807	-9.551994	-8.691883	-9.485067	-10.417484	-7.1771107	-9.156488	-2.425933
hsa-miR-135b-5p	-8.1022215	-8.468683	-9.269625	-8.953807	-9.551994	-8.691883	-9.485067	-10.417484	-7.1771107	-9.156488	-9.727636
hsa-miR-17-3p	0.69978285	-0.1495361	-3.3117428	-2.0746284	-2.278935	-0.4483361	-2.4668224	-3.6038568	0.0020577	-1.477142	-1.8730044
hsa-miR-17-5p	4.021016	3.763351	1.64957	2.1793928	1.3933334	3.053259	1.3745489	0.9481944	2.1529338	3.2527642	1.9781537
hsa-miR-182-3p	-8.1022215	-8.468683	-9.269625	-8.953807	-9.551994	-4.830414	-9.485067	-5.2876196	-7.1771107	-9.156488	-9.727636
hsa-miR-1825	-8.1022215	-8.468683	-9.269625	-8.953807	-9.551994	-8.691883	-3.6172867	-10.417484	-7.1771107	-9.156488	-9.727636
hsa-miR-18a-5p	-0.4242849	-0.8129411	-3.757477	-1.958327	-2.6480951	-1.5500855	-2.456437	-2.6254745	-2.0936337	-0.2237148	-2.2672276
hsa-miR-18b-5p	-1.8060963	-2.122616	-4.9213	-3.7418408	-4.218742	-3.1380005	-3.940016	-3.7905788	-3.402482	-1.4591208	-4.0010605
hsa-miR-198	-8.1022215	-8.468683	-2.1209245	-3.6662312	-0.8388066	-3.6058636	-2.1458583	-2.2075624	-7.1771107	-2.6895854	-4.0411243
hsa-miR-214-5p	-3.188389	-0.3874697	-3.2039318	-1.64172	-1.6623311	-0.6157021	-3.0894506	-4.4407377	-7.1771107	-2.4483478	-2.8332872
hsa-miR-301b-3p	-8.1022215	-8.468683	-9.269625	-8.953807	-9.551994	-8.691883	-9.485067	-4.9747696	-7.1771107	-9.156488	-9.727636
hsa-miR-32-5p	-4.009033	-8.468683	-9.269625	-8.953807	-9.551994	-8.691883	-9.485067	-10.417484	-7.1771107	-9.156488	-9.727636
hsa-miR-4291	1.0796461	1.0075102	-0.4763074	-0.6015119	-0.5291576	0.6668453	-0.7574563	-0.6463413	-0.0759463	-0.1141691	-1.1165113
hsa-miR-516a-5p	-8.1022215	-8.468683	-9.269625	-8.953807	-9.551994	-8.691883	-9.485067	-10.417484	-7.1771107	-9.156488	-9.727636
hsa-miR-517c-3p	-8.1022215	-8.468683	-9.269625	-8.953807	-9.551994	-8.691883	-9.485067	-10.417484	-7.1771107	-9.156488	-9.727636
hsa-miR-521	-8.1022215	-8.468683	-9.269625	-8.953807	-9.551994	-8.691883	-9.485067	-10.417484	-7.1771107	-9.156488	-9.727636
hsa-miR-522-3p	-8.1022215	-8.468683	-9.269625	-8.953807	-9.551994	-8.691883	-9.485067	-10.417484	-7.1771107	-9.156488	-9.727636
hsa-miR-545-3p	-8.1022215	-8.468683	-9.269625	-8.953807	-9.551994	-8.691883	-9.485067	-10.417484	-7.1771107	-9.156488	-9.727636
hsa-miR-557	-8.1022215	-8.468683	-4.2999997	-4.087893	-3.5109308	-8.691883	-3.125581	-2.7488403	-7.1771107	-9.156488	-2.8540184
hsa-miR-934	-8.1022215	-3.1220167	-4.925513	-1.3855734	-1.1313314	-3.6535347	-5.6400426	-3.136967	-7.1771107	-3.5353044	-3.4157744

5.7 Building of the ML model

In the ML model-building stage, the aim of feature-selection methods is finding the best subset of features. After feature selection the data can be easily represented in a form that can be used by a ML algorithm.

5.7.1 Choosing of the feature selection methods

A combination of feature selection methods was employed to identify biomarker profiles that differentiates TNBC from normal samples (**Appendix 4**).

To identify the genes most strongly associated with TNBC, recursive feature elimination (RFE) as a wrapper method were used, along with support vector machine (SVM) and random forest (RF) methods as embedded methods (**Appendix 5**).

Throughout the feature selection process, a filtering method was employed to reduce the number of features, thereby enhancing the efficiency of the machine learning algorithm. Following this, a hybrid feature selection strategy combining both wrapper and embedded methods was implemented (**Figure 5.10**).

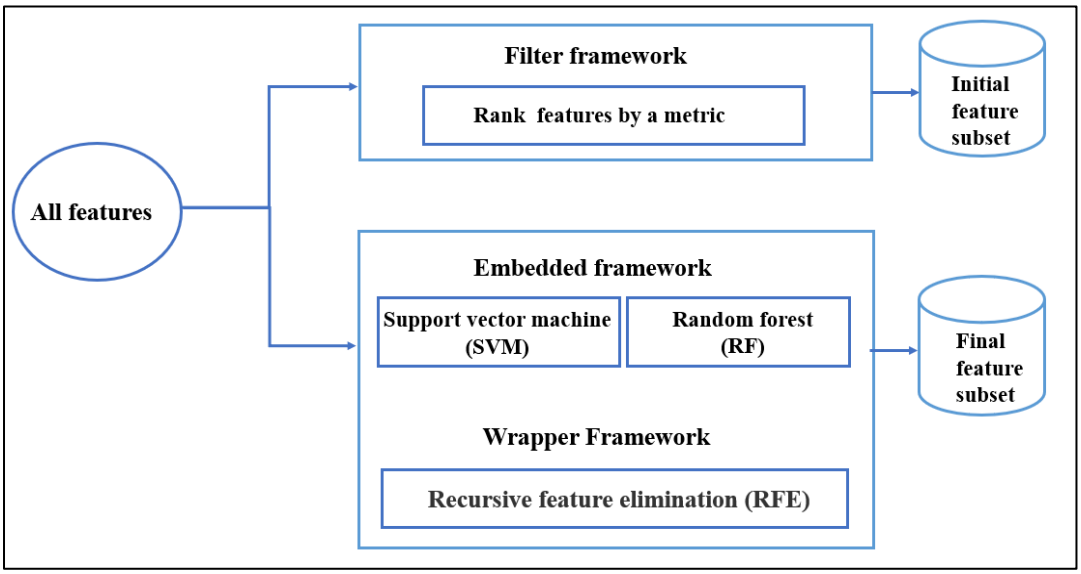

Figure 5.10 1Feature selection methods

Our framework takes advantage of hybrid feature selection methods. The filter method concentrates on the statistical attributes of the input data, selecting features based on their correlations, independent of any classification model.

In contrast, the wrapper method assesses subsets of features collectively, employing predictive models to evaluate their effectiveness. Recursive feature elimination (RFE) does not engage with

classification models; instead, it aims to minimize the feature space needed for training traditional machine learning models like RF and SVM (Sanz et al., 2018).

Embedded method, on the other hand, is integrated into the classifier construction process. RF and SVM classification models significantly reduced the selected features, these methods build the model and perform feature selection simultaneously. Embedded methods lower risk of overfitting, faster running time and are less computationally expensive as compared to wrapper methods (Sanz et al., 2018).

All the methods mentioned earlier have been used to identify a particular subset of features as potential candidate biomarkers. By using multiple feature selection techniques, the study aims to capitalize on their strengths while minimizing their weaknesses. The study hypothesizes that the features consistently identified across all methods will yield the most significant biomarkers.

5.7.2 Selection of the ML algorithms

While SVM and Random Forest classifiers are generally effective at managing a large number of irrelevant genes, the recursive feature elimination (RFE) algorithm was added to improve classification performance even more. SVMs are focused on identifying the most relevant features, reducing noise, and improving overall model accuracy. However, they can struggle with very large datasets due to their high computational demands, which can lead to increased memory usage and longer processing times. SVMs may also be affected by noisy data, as outliers can significantly influence the decision boundary.

On the other hand, Random Forests are skilled at handling irrelevant features but can be prone to overfitting if not properly tuned, especially in high-dimensional contexts. Additionally, as the number of features grows, the interpretability of Random Forest models may diminish.

By applying the RFE feature selection algorithm to focus on significant genes, our study aims to optimize the performance of these classifiers while addressing their potential limitations, ultimately leading to more accurate and interpretable results.

5.8 Machine learning with python

Python is downloaded(*Python 3.12.5*, n.d.). Python is a versatile programming language widely used for machine learning and data analysis. Machine learning with Python utilizes several libraries and frameworks to build and deploy models effectively. Key libraries include NumPy for numerical calculations, Pandas for data manipulation, and Scikit-learn for implementing machine learning algorithms. These tools support tasks like data preprocessing, model training, evaluation, and deployment, simplifying the process of implementing machine learning solutions.

5.8.1 Loading data for the ML model

NumPy is a Python library that have been used for working with arrays. For Creating a NumPy array:

```
import numpy as np
```

Pandas is a Python library that have been used to analyze data. For Loading a CSV file into a pandas data frame:

```
import pandas as pd
```

```
breast_cancer = pd. read_excel ('/content/ML data csv.xlsx')
```

Sicket learn (sklearn) is an open-source python library of popular ML algorithms that will allow us to build these types of systems. For installation from sicket learn different instructions were used:

```
from sklearn.preprocessing import MinMaxScaler
```

```
from sklearn.model_selection import train_test_split
```

```
from sklearn.ensemble import GradientBoostingRegressor
```

```
from sklearn.feature_selection import RFE
```

```
from sklearn.ensemble import RandomForestClassifier
```

```
from sklearn.svm import SVC
```

5.8.2 Normalization of the ML model

Normalization of the machine learning model is a crucial step that ensures the data is scaled appropriately for optimal model performance. Min-max scaling is a common normalization technique that rescales the features to a fixed range (usually 0 to 1). For scaling the values to a specific value range without changing the shape of the original distribution; where x is a raw value, x' is the normalized value, min is the smallest value in the column, and max is the largest value:

```
X = breast_cancer.drop('ID_REF',axis=1)
```

```
y = breast_cancer['ID_REF']
```

```
scaler = MinMaxScaler()
```

```
scaled_X = scaler.fit_transform(X)
```

```
scaled_X = pd.DataFrame(scaled_X, columns=X.columns)
```

5.8.3 Validation of the ML model

Model validation is a crucial technique for evaluating model performance. The Train/Test method involves splitting the dataset randomly into two subsets: a training set and a testing set. The model is trained using the training set, which typically consists of 80% of the original data, while the

testing set, comprising the remaining 20%, is used to assess the model's accuracy. This approach helps ensure that the model's performance is reliable and generalizable to unseen data. For evaluating the model and measure if the model is good enough, a Train/Test split method is used:

```
X_train,X_test,y_train,y_test = train_test_split(scaled_X, y,
random_state=0)
```

5.8.4 Training a recursive feature elimination

RFE is a wrapper-type feature selection algorithm, meaning it utilizes a different machine learning algorithm at its core, which is wrapped by RFE to assist in feature selection. As a transformation method, RFE requires configuring the class with the chosen algorithm through the "estimator" argument and specifying the number of features to select using the "n_features_to_select" argument.

```
from feature_engine.selection import RecursiveFeatureElimination
```

Using of (feature_engine.selection) function helps drop subsets of features with low predictive value, streamlining the feature set. Importing of (GradientBoostingRegressor) optimizes the training process by utilizing a sequential flow, enhancing model performance through boosting techniques.

5.8.5 Training a random forest

Random Forest (RF) is used to train multiple decision tree classifiers on microarray data which consists of 130 features and 6 samples. Averaging is employed to enhance the accuracy and reduce overfitting. The number of trees in the forest is set to (n_estimators=10). RFE feature selection algorithm is used to select features, specifying the number of features to retain as two over two rounds. Random forest model will be fitted using the selected features, employing the fit function to train the model on the training data. This study reports an accuracy of 100%.

```
rfe_method = RFE (RandomForestClassifier(n_estimators=10,
random_state=10), n_features_to_select=2, step=2, )
rfe_method.fit(X_train, y_train)
RFE(estimator=RandomForestClassifier(n_estimators=10,
random_state=10), n_features_to_select=2, step=2)

X_train.columns[(rfe_method.get_support())]

Index (['hsa-miR-32-5p', 'hsa-miR-6089'], dtype='object')

rfe_method.score(X_train,y_train)
```

5.8.6 Training a support vector machine

Support Vector Machine (SVM) is used for classification, requiring two input arrays: X of shape (n_samples,n_features) for training samples, and y of shape (n_samples) for class labels (strings or integers). Linear kernel function will be employed since the data is nearly separable. RFE

feature selection algorithm is used to select features, choosing to eliminate four features in one round. The fit function will be used to train the model on the training data.

```
estimator = SVC (kernel="linear")

selector = RFE (estimator, n_features_to_select=4, step=1)

selector = selector.fit(X_train, y_train)

selected_col= selector.support

cols = np.where(selected_col == True)

X_train.columns[cols]

Index (['hsa-miR-18a-5p', 'hsa-miR-32-5p', 'hsa-miR-495-3p', 'hsa-miR
542-5p'], dtype='object')
```

5.8.7 Visualization of the ML model

Matplotlib is a low-level graph plotting library in Python that functions as a visualization utility. Seaborn, built on top of Matplotlib, offers a high-level interface for creating attractive and informative statistical graphics, making it easier to visualize complex datasets. For visualizing statistical relationships, finding correlations heatmap and show the relationships between the columns:

```
import matplotlib. pyplot as plt

import seaborn as sns
```

There are several ways to draw a scatter plot in seaborn, to customize Seaborn Correlation Heatmap:

```
sns.heatmap(scaled_X.corr())
```

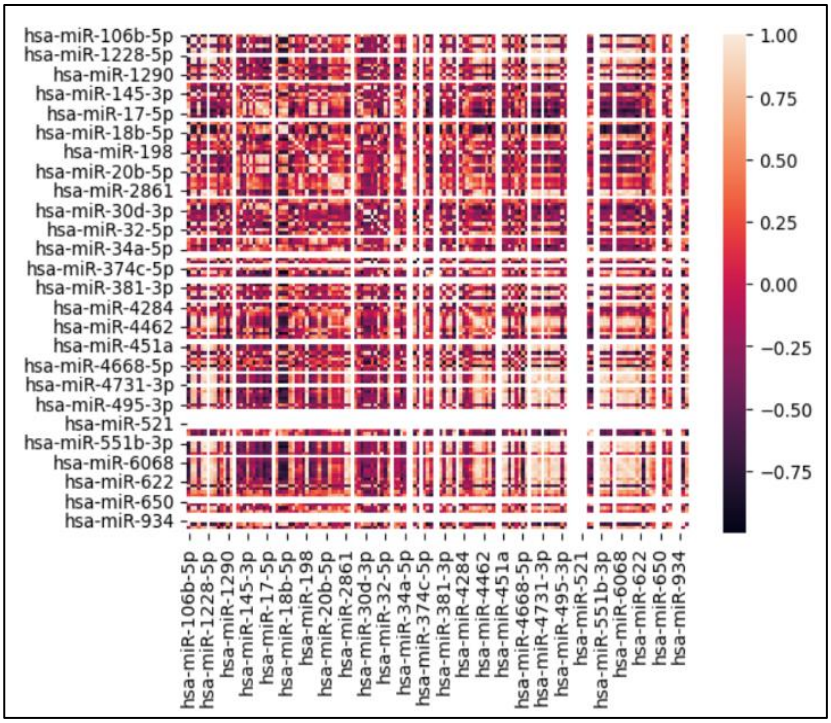

Figure 5.11 1A heatmap representing the correlations for the expression levels of the candidate genes

A heatmap was created with a dendrogram using the Seaborn library. The miRNA expression profiles obtained from the normal and TNBC samples were clustered generated a heatmap dendrogram. This visualization reveals distinctive patterns and clear separation. According to the scale, red to white indicates strong positively correlations and dark red to black indicates strong negatively correlations (**Figure 5.11**).

5.9 Potential biomarkers for TNBC identified by our study

After conducting differential gene expression analysis, four genes were identified as significantly differentially expressed in TNBC samples compared to normal samples: miR-135b-5p, hsa-miR-18a-5p, hsa-miR-18b-5p, and hsa-miR-32-5p. Furthermore, the Random Forest classifier selected the miRNAs: hsa-miR-32-5p and hsa-miR-6089, while the SVM classifier identified: hsa-miR-18a-5p, hsa-miR-32-5p, hsa-miR-495-3p, and hsa-miR-542-5p as key miRNAs.

Interestingly, noticeable variations were observed in the candidate genes identified by each technique. However, the common candidate across all methods was miR-32-5p, which is strongly correlated with TNBC.

5.10 Clinical relevance and prognostic potential

The study's findings suggest that miR-32-5p, along with other identified miRNAs, could serve as potential diagnostic and prognostic biomarkers for TNBC. The strong correlation between these miRNAs and TNBC underscores their potential utility in clinical settings, where they could be used to guide treatment decisions and monitor disease progression.

5.11 miRNA signature and model performance

The hybrid feature selection approach successfully identified miRNAs that were strongly correlated with TNBC. The final model, incorporating RFE, RF, and SVM, achieved high classification accuracy, with miR-32-5p being the most robust biomarker. The heatmap dendrogram **(Figure 5.11)** clearly distinguished between TNBC and normal samples, highlighting the distinct expression patterns of the selected miRNAs.

5.12 Comparative analysis of ML algorithms

As mentioned earlier in this study, TNBC is a highly heterogeneous disease, posing a significant challenge in developing accurate and computationally efficient algorithms for patient classification to aid therapeutic decision-making. Although this study did not specifically investigate the use of machine learning , several studies have highlighted its application in breast cancer, demonstrating its superiority. For instance, Hiba Asri and her colleagues conducted a performance comparison of various ML algorithms (SVM, Decision Tree, Naive Bayes, and k-Nearest Neighbors) on breast cancer datasets, finding that SVM achieved the highest accuracy (97.13%) with the lowest error rate(Asri et al., 2016). Mehmet Fatih Akay and his colleagues confirmed that SVM has a superior diagnostic accuracy. The performance of their method was assessed using metrics such as classification accuracy, sensitivity, specificity, positive and negative predictive values, receiver operating characteristic (ROC) curves, and confusion matrix. Their results indicated that the SVM model achieved the highest classification accuracy at 99.51%(Akay, 2009). The Support Vector Machine algorithm demonstrated superior accuracy in classifying breast cancer into triple-negative and non-triple-negative categories, showing fewer misclassification errors compared to the other three algorithms evaluated in J. Pers. research(J. Wu & Hicks, 2021). Experimental results from a study by Dazhong Wu and his colleagues demonstrated that Random Forests can produce more accurate predictions compared to feed-forward back propagation (FFBP) artificial neural networks (ANNs) and support vector regression (SVR)(D. Wu et al., 2017).The four models (LR, SVM, NB, KNN) experienced an underfitting problem, while underfitting was not an issue for the Random Forest model(Bagui et al., 2017). The RFE technique was employed to select promising features from the available data, resulting in a significant improvement in classifier performance (Sachdeva et al., 2022). C. Aroef and his colleagues found that Random Forest and SVM achieved accuracies of 90% and 95%, respectively. Their results indicated that SVM outperformed Random Forest in terms of accuracy (Aroef et al., 2020).

The primary distinction and innovative aspect of our investigation is the application of Recursive Feature Elimination (RFE) alongside Random Forest (RF) and Support Vector Machine (SVM) classifiers to differentiate between triple-negative breast cancer (TNBC) and normal samples. The clinical significance of this research lies in the potential of machine learning algorithms to improve diagnostic accuracy and identify women at high risk of developing TNBC. Ultimately, this study

highlights the effectiveness of combining RFE with SVM and RF as a hybrid feature selection approach for analyzing miRNA expression profiles and other datasets with a high number of features compared to the number of samples.

5.13 Comparative analysis of the key miRNAs

This study identified several miRNAs associated with TNBC, including miR-135b-5p, hsa-miR-18a-5p, hsa-miR-18b-5p, hsa-miR-6089, hsa-miR-495-3p, hsa-miR-542-5p, and hsa-miR-32-5p. A systematic search of the PubMed database for relevant studies were conducted for exploring these miRNAs and their connections to TNBC and other cancer subtypes, selecting several articles for in-depth analysis. Among the differentially expressed microRNAs, miR-135b-5p exhibited a specific correlation with TNBC prognosis but not with non-TNBC (Bao et al., 2019). A Naïve Bayes classifier utilizing miRNA signatures was effective in distinguishing TNBC from non-TNBC samples in the test dataset, demonstrating high sensitivity and specificity. The analysis indicated that hsa-miR-135b-5p and hsa-miR-18a-5p are critical for improving diagnostics and therapeutics for TNBC (Naorem et al., 2019). Additionally, a study comparing the characteristics of plasma exosomes in Diffuse Large B-cell Lymphoma (DLBCL) patients and healthy individuals found a significant decrease in miR-6089 expression in the plasma exosomes of DLBCL patients, suggesting it could serve as a diagnostic miRNA signature (Caner et al., 2021). Functional experiments, including CCK-8, EdU, Transwell, and apoptosis assays, were conducted to assess the effects of miR-495-3p. The results revealed that miR-495-3p inhibits colorectal cancer progression by downregulating HMGB1 expression, indicating its potential as a therapeutic target for colorectal cancer (J. L. Zhang et al., 2022). Finally, a study identified miR-542-5p as a key miRNA in osteosarcoma by analyzing three overlapping Gene Expression Omnibus datasets, showing significantly higher levels of miR-542-5p in osteosarcoma tissues compared to healthy bone (Zhu et al., 2020).

5.14 The magic and mystery of microRNA-32-5p

The study identified that miR-32-5p can be used as a potential biomarker for the diagnosis of TNBC; miR-32-5p showed a significant overexpression in TNBC samples compared to normal samples.

The aberrant expression of miR-32-5p has been linked to various diseases. Haiqiu Liao and his colleagues identified miR-32-5p as an oncomiR in prostate cancer (Liao et al., 2015). Additionally, Filip and his team confirmed that miR-32-5p could serve as a potential diagnostic biomarker to differentiate benign prostatic hyperplasia (BPH) from noncancerous regions in cancerous prostate tissue (Ambrozkiewicz et al., 2020). Jin-Xing and his colleagues identified SMAD3 as a target of miR-32-5p, suggesting that miR-32-5p acts as a tumor suppressor by inhibiting SMAD3, thus positioning it as a potential therapeutic target for lung adenocarcinoma (J.-X. Zhang et al., 2021). Zhang et al. reported significant downregulation of miR-32-5p in ovarian cancer tissues and cells (R.-R. Zhang et al., 2020). Furthermore, Gladys and her colleagues found that miR-32-5p is involved in pathways related to obesity and insulin resistance (Wojciechowska et al., 2022).

In 2017, a research group utilized CRISPR/Cas9 technology to create the first miR-32-5p knockout mouse (see **Table 5.5** for the reference sequence), identifying it as a potential key miRNA in

vascular calcification (VC) in both mice and humans. By 2019, they demonstrated that the knockout alleviates lipopolysaccharide-induced depressive-like behavior in mice by inhibiting astrocyte overactivity. So far, studies have linked miR-32-5p to VC, atherosclerosis, diabetes, depression, and inflammation. There's a growing focus on its role in tumorigenesis and tumor progression, as miR-32-5p is involved in regulating tumor cell apoptosis, proliferation, and migration. While many studies classify it as an oncomiR, some conflicting reports suggest it may also function as a tumor suppressor miRNA.(Zeng et al., 2021).

In further detail, location and sequence of miR-32 is shown in **figure 5.12** (*MiRBase: Stem-Loop Hsa-Mir-32*, n.d.).

Table 5.5 1Reference sequence for miR-32-5p knockout in mice

sgRNA name	Oligo name	sgRNA sequence	Target
miR-32-5p-sgRNA	Forward primier	caccggtactaagttgcatgttgtca	tactaagttgcatgttgtcacgg
	Reverse primer	aaactgacaacatgcaacttagtacc	

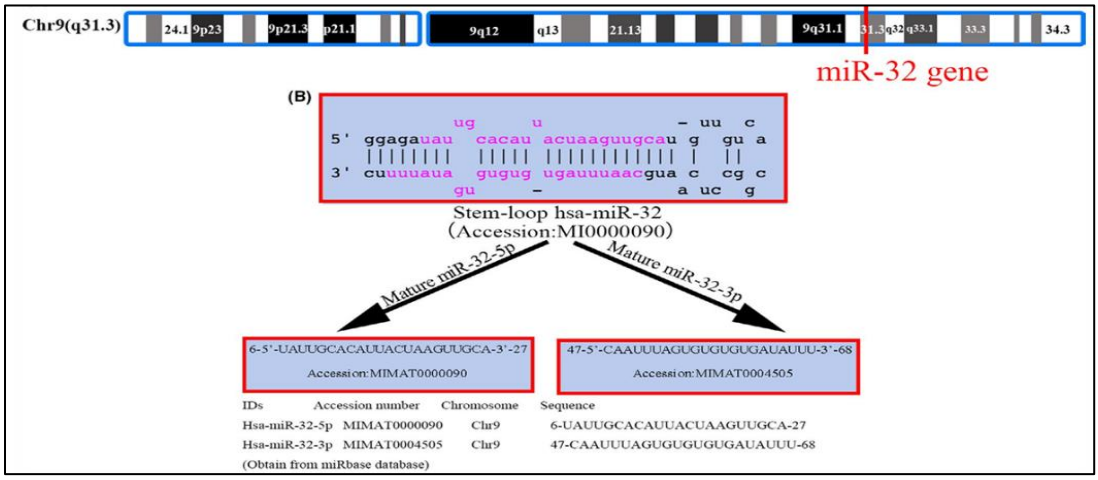

Figure 5.12 1Location and sequence of miR-32.

(A) miR-32 location. MiR-32 is located on chromosome 9q31 (in the NR_029506.1 noncoding region); (B) miR-32 sequence. The stem- loop and the maturation sequences of miR-32-5p and miR-32- 3p.

Conclusions

Identifying new biomarkers for triple-negative breast cancer (TNBC) using machine learning is a complex task that requires careful consideration of several factors. Some general steps should be followed to approach this problem:

1. Data collection and preprocessing: The first step is to collect a dataset of microRNA gene expression profiles from TNBC patients and healthy controls. This dataset should be large enough to capture the diversity of gene expression patterns in TNBC.

2. Model selection: to select an appropriate machine learning model for the study task.

3. Training the model: Once a model is selected. The model will learn to identify patterns in the microRNA gene expression profiles that are associated with TNBC.

4. Testing and evaluation: Once the model has been trained, the performance should be evaluated on a separate testing set that the model has not seen before.

5. Interpretation: Finally, the results of the ML model should be interpretable to identify potential new microRNA gene biomarkers for TNBC. This can involve analyzing the most important features identified by the model, exploring the relationship between these features, and validating the potential biomarkers.

Normalization methods are an important step in the analysis of high-throughput genomic data. They help to ensure that the data is comparable across samples, platforms, and batches, and can improve the accuracy and reliability of gene expression measurements. The choice of normalization method depends on the type and size of data, and should be carefully considered to ensure that the results are accurate and reliable.

It is also important to carefully account for batch effects in genomic studies to ensure the accuracy and reliability of the results. Failure to do so can lead to erroneous conclusions and hinder the progress of scientific research.

In summary, biomarker discovery is an evolving field with continuous innovative ideas emerging. While no single method has proven perfect, largely due to data dependency and a lack of universal

evaluation standards. Ongoing dedicated efforts are clearly advancing the field. This progress holds great promise for enhancing our understanding of disease diagnosis, prevention, and therapy.

Key factors to consider in biomarker discovery studies using archived specimens include the patient population represented in the specimen archive, the power of the study (determined by the number of samples), disease prevalence, the analytical validity of the biomarker test, and the preplanned analysis strategy. These factors are crucial for generating credible and clinically relevant outcomes.

The clinical, pathological, and molecular heterogeneity of TNBC indeed underscores the need for comprehensive analyses to establish a unified classification system. This is essential for tailoring appropriate therapies and enhancing patient outcomes. The structural and functional characterization of miRNAs provides valuable insights into their roles in cancer development and progression. By leveraging machine learning techniques, this study aimed to identify specific miRNAs that could serve as potential targets for new therapies in TNBC, offering hope for more effective treatment strategies.

Overall, miRNA target prediction algorithms are important tools for identifying potential target genes of miRNAs, which can provide valuable insights into the molecular mechanisms of disease and potential therapeutic targets. However, experimental validation is critical to confirm the accuracy and specificity of miRNA target predictions.

Based on the findings of this study, using recursive feature elimination (RFE) feature selection technique along with the random forest (RF) and SVM classification model (hybrid feature selection algorithm) can effectively filter redundant features, select biomarkers with diagnostic significance, and ensure high classification accuracy.

Having 100% accuracy doesn’t mean that the model not doing well in real world, too small dataset makes the model easily run and easily learn the relationships leading to perfect accuracy.

Finally, this study identified miR-32-5p as a potential biomarker for TNBC diagnosis, with high expression levels significantly associated with improved overall survival in patients.

limitations

DNA microarray-based gene expression profiling shows promise for diagnosing and prognosing TNBC. However, several limitations complicate the analysis, including small sample sizes, biased case-control distributions, the presence of multiple breast cancer subtypes, variability in populations, and differences across platforms. These challenges hinder the consistent identification of gene signatures.

The use of miRNAs as potential targets for addressing proliferation and metastasis presents a significant opportunity to enhance treatment options for TNBC, potentially reducing toxicity and minimizing drug resistance. However, key challenges remain in miRNA-based therapies for TNBC, particularly in creating effective miRNA mimics and developing suitable delivery systems that avoid off-target effects.

AI and ML algorithms can effectively identify patterns in large datasets, discovering new biomarkers, and enhance the accuracy of existing ones. The rise of digital biomarkers, driven by the increasing use of digital health technologies, offers objective and measurable data for healthcare evaluations. However, several challenges exist in the development of AI and ML algorithms, including bias from incomplete data and the need for regulatory frameworks.

While hybrid feature selection methods can enhance model accuracy, they also pose a risk of overfitting, particularly due to the high dimensionality of miRNA data compared to the limited number of samples.

Traditional statistical tests are popular for identifying differentially expressed genes due to their simplicity and interpretability, but they often assume that genes operate independently. In biological contexts, genes typically collaborate within pathways and networks, leading to correlated datasets. Additionally, many of these tests rely on specific distributional assumptions, which can be problematic, particularly when dealing with limited biological replicates. This highlights the need for more sophisticated methods that account for gene interactions and variability in biological data.

Dysregulated miRNAs play a crucial role in the development of triple-negative breast cancer (TNBC), influencing gene expression and cancer-related pathways. This dysregulation positions these miRNAs as promising candidates for cancer biomarkers in TNBC diagnosis, prognosis, and therapy prediction. However, their clinical application as disease-specific markers has been hindered by the need for optimized detection strategies. As our understanding of miRNAs' functional roles continues to evolve, there is growing evidence supporting their involvement in tumorigenesis. This offers a promising avenue for developing innovative approaches in the clinical management of TNBC patients.

Recommendations

The identification and validation of biomarkers demand comprehensive planning and teamwork among clinicians, scientists, statisticians, and epidemiologists. Success in this field relies on cross-disciplinary approaches, emphasizing the importance of cohesive teams of collaborative scientists. By fostering such partnerships, we can expedite the translation of groundbreaking scientific discoveries from the lab to clinical practice, ultimately enhancing patient care and outcomes.

Future steps should focus on community-driven standardization initiatives that engage researchers, practitioners, and regulators. These initiatives are essential for developing comprehensive documentation and validation standards, establishing minimum requirements, and creating study-type-specific guidelines. Such efforts will enhance the quality of biomarker stratification and prediction projects, ultimately leading to more reliable and effective applications in clinical settings.

Pharmaceutical collaborations can greatly accelerate research by leveraging industry expertise, resources, and real-world data. Joint research projects can lead to the development of new drugs or enhancements of existing ones, while access to proprietary datasets, such as clinical trial results and electronic health records, can enrich the research process. Moreover, collaboration provides access to advanced laboratory equipment and high-throughput screening platforms, enabling more efficient experiments. Utilizing real-world clinical data helps validate findings and refine predictive models, ensuring that research outcomes are applicable and impactful in clinical settings.

Fostering interdisciplinary collaboration can significantly enhance research efforts in drug discovery by integrating diverse expertise and resources. Biologists and chemists contribute vital domain-specific knowledge on drug-target interactions, molecular biology, and chemical properties. Meanwhile, computational scientists and data engineers focus on developing and implementing machine learning algorithms, managing data integration, and optimizing computational workflows. Pharmacologists and clinical researchers provide crucial insights into drug efficacy, safety profiles, and clinical trial data.

Additionally, utilizing collaborative tools and platforms, such as shared data repositories, facilitates access to large datasets and promotes data sharing among collaborators. This integrated approach not only drives advancements in drug discovery but also accelerates the translation of research findings into clinical applications, ultimately benefiting patient care and therapeutic outcomes.

Project time plan

Table 9.6 1The project time plan

Activities	**Start date**	**End date**
Identifying research gap and topic development	September	October
Write a first draft proposal	October	November
Formulate the research questions	November	December
Introduction, literature review and methodology write up	December	January
Pre thesis defense	January	February
Data collection	February	March
Data analysis	March	May
Building of ML model	May	June
Results and discussion write up	July	September
Completion of the final Proposal	October	December
Final thesis defense	December	January

Budget

Table 10.7 1The project budget

Items needed for the project/activities	**Total cost**
1. Computer use/data storage	0₪
2. Data managers or analysts	0₪
3. Consultant Services	0₪
4. Data analysis	0₪
5. Data access	0₪
6. Memory for Data storage	0₪
7. Software and Programs- (R, Python, and others)	0₪
Total budget	0₪

References

- Aebi, S., Davidson, T., Gruber, G., & Cardoso, F. (2011). Primary breast cancer: Esmo clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, *22*(SUPPL. 6), vi12–vi24. https://doi.org/10.1093/annonc/mdr371
- Ahn, S. G., Kim, S. J., Kim, C., & Jeong, J. (2016). Molecular classification of triple-negative breast cancer. *Journal of Breast Cancer*, *19*(3), 223–230. https://doi.org/10.4048/jbc.2016.19.3.223
- Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications*, *36*(2), 3240–3247.
- Al-Mahmood, S., Sapiezynski, J., Garbuzenko, O. B., & Minko, T. (2018). Metastatic and triple-negative breast cancer: challenges and treatment options. *Drug Delivery and Translational Research*, *8*(5), 1483–1507. https://doi.org/10.1007/s13346-018-0551-3
- Al-Tashi, Q., Saad, M. B., Muneer, A., Qureshi, R., Mirjalili, S., Sheshadri, A., Le, X., Vokes, N. I., Zhang, J., & Wu, J. (2023). Machine learning models for the identification

of prognostic and predictive cancer biomarkers: a systematic review. *International Journal of Molecular Sciences*, *24*(9), 7781.

- Almugren, N., & Alshamlan, H. (2019). A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. *IEEE Access*, *7*, 78533–78548.
- Alshamlan, H., Badr, G., & Alohali, Y. (2015). mRMR-ABC: a hybrid gene selection algorithm for cancer classification using microarray gene expression profiling. *Biomed Research International*, *2015*.
- Ambrozkiewicz, F., Karczmarski, J., Kulecka, M., Paziewska, A., Cybulska, M., Szymanski, M., Dobruch, J., Antoniewicz, A., Mikula, M., & Ostrowski, J. (2020). Challenges in cancer biomarker discovery exemplified by the identification of diagnostic MicroRNAs in prostate tissues. *BioMed Research International*, *2020*(1), 9086829.
- An, X., Sarmiento, C., Tan, T., & Zhu, H. (2017). Regulation of multidrug resistance by microRNAs in anti-cancer therapy. *Acta Pharmaceutica Sinica B*, *7*(1), 38–51.
- Anders, C., Carey, L. A., & Carey, L. (2008). Understanding and Treating Triple-Negative Breast Cancer. *Oncology*, *22*(11), 1233–1243.
- Andre, F., & Zielinski, C. C. (2012). Optimal strategies for the treatment of metastatic triple-negative breast cancer with currently approved agents. *Annals of Oncology*, *23*, vi46–vi51.
- Aroef, C., Rivan, Y., & Rustam, Z. (2020). Comparing random forest and support vector machines for breast cancer classification. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *18*(2), 815–821.
- Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*, *83*, 1064–1069.
- Bagui, S., Fang, X., Kalaimannan, E., Bagui, S. C., & Sheehan, J. (2017). Comparison of machine-learning algorithms for classification of VPN network traffic flow using time-related features. *Journal of Cyber Security Technology*, *1*(2), 108–126.
- Bao, C., Lu, Y., Chen, J., Chen, D., Lou, W., Ding, B., Xu, L., & Fan, W. (2019). Exploring specific prognostic biomarkers in triple-negative breast cancer. *Cell Death & Disease*, *10*(11), 807.
- Barzani, A. R., Pahlavani, P., Ghorbanzadeh, O., Gholamnia, K., & Ghamisi, P. (2024). Evaluating the Impact of Recursive Feature Elimination on Machine Learning Models for Predicting Forest Fire-Prone Zones. *Fire*, *7*(12), 440.
- Ben Or, G., & Veksler-Lublinsky, I. (2021). Comprehensive machine-learning-based analysis of microRNA–target interactions reveals variable transferability of interaction rules across species. *BMC Bioinformatics*, *22*, 1–27.
- Bertucci, F., Houlgatte, R., Benziane, A., Granjeaud, S., Adélaïde, J., Tagett, R., Loriod, B., Jacquemier, J., Viens, P., Jordan, B., Birnbaum, D., & Nguyen, C. (2000). Gene expression profiling of primary breast carcinomas using arrays of candidate genes. *Human Molecular Genetics*, *9*(20), 2981–2991. https://doi.org/10.1093/hmg/9.20.2981
- Bharaj, U. K., Lohmann, A. E., & Blanchette, P. S. (2021). *Triple negative breast cancer: emerging light on the horizon—a narrative review*.
- Boeri, C., Chiappa, C., Galli, F., De Berardinis, V., Bardelli, L., Carcano, G., Rovera, F., Berardinis, V. De, Bardelli, L., Carcano, G., & Rovera, F. (2020). Machine Learning techniques in breast cancer prognosis prediction: A primary evaluation. *Cancer Medicine*, *9*(9), 3234–3243. https://doi.org/10.1002/cam4.2811

- Brown, C. H. (n.d.). *Triple-NegativeBreastCancer*. https://www.uspharmacist.com/article/triple-negative-breast-cancer
- Caner, V., Cetin, G. O., Hacioglu, S., Baris, I. C., Tepeli, E., Turk, N. Sen, Bagci, G., Yararbas, K., & Cagliyan, G. (2021). The miRNA content of circulating exosomes in DLBCL patients and in vitro influence of DLBCL-derived exosomes on miRNA expression of healthy B-cells from peripheral blood. *Cancer Biomarkers*, *32*(4), 519–529.
- Catalanotto, C., Cogoni, C., & Zardo, G. (2016). MicroRNA in Control of Gene Expression: An Overview of Nuclear Functions. In *International Journal of Molecular Sciences* (Vol. 17, Issue 10). https://doi.org/10.3390/ijms17101712
- *Cinical Trials.gov*. (n.d.). https://clinicaltrials.gov/study/NCT02659631?tab=results
- Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., & Shi, B. (2015). Breast cancer intrinsic subtype classification, clinical use and future trends. *American Journal of Cancer Research*, *5*(10), 2929–2943.
- Dev, S. S., Abidin, S. A. Z., Farghadani, R., Othman, I., & Kinases, R. N. R. T. (2021). Their Signaling Pathways as Therapeutic Targets of Curcumin in Cancer., 2021, 12. *DOI: Https://Doi. Org/10.3389/Fphar*, 772510.
- Dragun, A. E., Pan, J., Rai, S. N., Kruse, B., & Jain, D. (2011). Locoregional recurrence in patients with triple-negative breast cancer: preliminary results of a single institution study. *American Journal of Clinical Oncology*, *34*(3), 231–237. https://doi.org/10.1097/COC.0b013e3181dea993
- Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. (2005). *Lancet (London, England)*, *365*(9472), 1687–1717. https://doi.org/10.1016/S0140-6736(05)66544-0
- Ensenyat-Mendez, M., Llinàs-Arias, P., Orozco, J. I. J., Íñiguez-Muñoz, S., Salomon, M. P., Sesé, B., DiNome, M. L., & Marzese, D. M. (2021). Current Triple-Negative Breast Cancer Subtypes: Dissecting the Most Aggressive Form of Breast Cancer . In *Frontiers in Oncology* (Vol. 11, p. 2311). https://www.frontiersin.org/article/10.3389/fonc.2021.681476
- Eraslan, G., Avsec, Ž., Gagneur, J., & Theis, F. J. (2019). Deep learning: new computational modelling techniques for genomics. *Nature Reviews. Genetics*, *20*(7), 389–403. https://doi.org/10.1038/s41576-019-0122-6
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118.
- Fabian, M. R., Sonenberg, N., & Filipowicz, W. (2010). Regulation of mRNA translation and stability by microRNAs. *Annual Review of Biochemistry*, *79*, 351–379.
- Frasci, G., Comella, P., Rinaldo, M., Iodice, G., Di Bonito, M., D'Aiuto, M., Petrillo, A., Lastoria, S., Siani, C., Comella, G., & D'Aiuto, G. (2009). Preoperative weekly cisplatin-epirubicin-paclitaxel with G-CSF support in triple-negative large operable breast cancer. *Annals of Oncology : Official Journal of the European Society for Medical Oncology*, *20*(7), 1185–1192. https://doi.org/10.1093/annonc/mdn748
- Freedman, G. M., Anderson, P. R., Li, T., & Nicolaou, N. (2009). Locoregional recurrence of triple-negative breast cancer after breast-conserving surgery and radiation. *Cancer*, *115*(5), 946–951. https://doi.org/10.1002/cncr.24094
- Ganggayah, M. D., Taib, N. A., Har, Y. C., Lio, P., & Dhillon, S. K. (2019). Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC*

Medical Informatics and Decision Making, *19*(1), 48. https://doi.org/10.1186/s12911-019-0801-4

- Garrido-Castro, A. C., Lin, N. U., & Polyak, K. (2019). Insights into Molecular Classifications of Triple-Negative Breast Cancer: Improving Patient Selection for Treatment. *Cancer Discovery*, *9*(2), 176–198. https://doi.org/10.1158/2159-8290.CD-18-1177
- Gisel, A., Valvano, M., El Idrissi, I. G., Nardulli, P., Azzariti, A., Carrieri, A., Contino, M., & Colabufo, N. A. (2014). miRNAs for the detection of multidrug resistance: overview and perspectives. *Molecules*, *19*(5), 5611–5623.
- Haffty, B. G., Yang, Q., Reiss, M., Kearney, T., Higgins, S. A., Weidhaas, J., Harris, L., Hait, W., & Toppmeyer, D. (2006). Locoregional relapse and distant metastasis in conservatively managed triple negative early-stage breast cancer. *Journal of Clinical Oncology*, *24*(36), 5652–5657. https://doi.org/10.1200/JCO.2006.06.5664
- Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Tibshirani, R., & Friedman, J. (2009). Overview of supervised learning. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 9–41.
- Haury, A.-C., Mordelet, F., Vera-Licona, P., & Vert, J.-P. (2012). TIGRESS: trustful inference of gene regulation using stability selection. *BMC Systems Biology*, *6*, 1–17.
- Hsu, H.-H., Hsieh, C.-W., & Lu, M.-D. (2011). Hybrid feature selection by combining filters and wrappers. *Expert Systems with Applications*, *38*(7), 8144–8150.
- Https://www.ncbi.nlm.nih.gov/gds/. (n.d.). *NCBI*.
- Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics & Proteomics*, *15*(1), 41–51.
- Hudis, C. A., & Gianni, L. (2011). Triple-negative breast cancer: an unmet medical need. *The Oncologist*, *16*, 1–11.
- Iorio, M. V., Ferracin, M., Liu, C. G., Veronese, A., Spizzo, R., Sabbioni, S., Magri, E., Pedriali, M., Fabbri, M., Campiglio, M., Ménard, S., Palazzo, J. P., Rosenberg, A., Musiani, P., Volinia, S., Nenci, I., Calin, G. A., Querzoli, P., Negrini, M., & Croce, C. M. (2005). MicroRNA gene expression deregulation in human breast cancer. *Cancer Research*, *65*(16), 7065–7070. https://doi.org/10.1158/0008-5472.CAN-05-1783
- Jaber, M. I., Song, B., Taylor, C., Vaske, C. J., Benz, S. C., Rabizadeh, S., Soon-Shiong, P., & Szeto, C. W. (2020). A deep learning image-based intrinsic molecular subtype classifier of breast tumors reveals tumor heterogeneity that may affect survival. *Breast Cancer Research : BCR*, *22*(1), 12. https://doi.org/10.1186/s13058-020-1248-3
- Jaradat, S. K., Ayoub, N. M., Al Sharie, A. H., & Aldaod, J. M. (2024). Targeting Receptor Tyrosine Kinases as a Novel Strategy for the Treatment of Triple-Negative Breast Cancer. *Technology in Cancer Research & Treatment*, *23*, 15330338241234780. https://doi.org/10.1177/15330338241234780
- Jeanes, A., Gottardi, C. J., & Yap, A. S. (2008). Cadherins and cancer: how does cadherin dysfunction promote tumor progression? *Oncogene*, *27*(55), 6920–6929. https://doi.org/10.1038/onc.2008.343
- Jia, H., Truica, C. I., Wang, B., Wang, Y., Ren, X., Harvey, H. A., Song, J., & Yang, J.-M. (2017). Immunotherapy for triple-negative breast cancer: Existing challenges and exciting prospects. *Drug Resistance Updates*, *32*, 1–15.
- Kavitha, K. R., Prakasan, A., & Dhrishya, P. J. (2020). Score-based feature selection of

gene expression data for cancer classification. *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 261–266.
- Kennecke, H., Yerushalmi, R., Woods, R., Cheang, M. C. U., Voduc, D., Speers, C. H., Nielsen, T. O., & Gelmon, K. (2010). Metastatic behavior of breast cancer subtypes. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *28*(20), 3271–3277. https://doi.org/10.1200/JCO.2009.25.9820
- Kumar, H., Gupta, N. V., Jain, R., Madhunapantula, S. V, Babu, C. S., Kesharwani, S. S., Dey, S., & Jain, V. (2023). A review of biological targets and therapeutic approaches in the management of triple-negative breast cancer. *Journal of Advanced Research*, *54*, 271–292. https://doi.org/https://doi.org/10.1016/j.jare.2023.02.005
- Lavanya, C., Pooja, S., Kashyap, A. H., Rahaman, A., Niranjan, S., & Niranjan, V. (2023). Novel biomarker prediction for lung cancer using random forest classifiers. *Cancer Informatics*, *22*.
- *le cancer du sein «triple négatif»*. (n.d.).
- Lewis, R. J. (2000). An introduction to classification and regression tree (CART) analysis. *Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California*, *14*.
- Li, J., & Zhang, Y. (2019). Current experimental strategies for intracellular target identification of microRNA. *ExRNA*, *1*(1), 1–8.
- Li, M., Fu, S., & Xiao, H. (2015). Genome-wide analysis of microRNA and mRNA expression signatures in cancer. *Acta Pharmacologica Sinica*, *36*(10), 1200–1211.
- Li, Ying, Zhan, Z., Yin, X., Fu, S., & Deng, X. (2021). Targeted therapeutic strategies for triple-negative breast cancer. *Frontiers in Oncology*, *11*, 731535.
- Li, Yun, Zhang, H., Merkher, Y., Chen, L., Liu, N., Leonov, S., & Chen, Y. (2022). Recent advances in therapeutic strategies for triple-negative breast cancer. *Journal of Hematology & Oncology*, *15*(1), 121.
- Liao, H., Xiao, Y., Hu, Y., Xiao, Y., Yin, Z., & Liu, L. (2015). microRNA-32 induces radioresistance by targeting DAB2IP and regulating autophagy in prostate cancer cells. *Oncology Letters*, *10*(4), 2055–2062.
- Loh, H.-Y., Norman, B. P., Lai, K.-S., Mohd, N., Nik, A., Rahman, A., Banu, N., Alitheen, M., Osman, M. A., & My, N. B. M. A.). (n.d.). *Molecular Sciences The Regulatory Role of MicroRNAs in Breast Cancer*. https://doi.org/10.3390/ijms20194940
- Loh, H.-Y., Norman, B. P., Lai, K.-S., Rahman, N. M. A. N. A., Alitheen, N. B. M., & Osman, M. A. (2019). The Regulatory Role of MicroRNAs in Breast Cancer. *International Journal of Molecular Sciences*, *20*(19). https://doi.org/10.3390/ijms20194940
- Lü, L., Mao, X., Shi, P., He, B., Xu, K., Zhang, S., & Wang, J. (2017). MicroRNAs in the prognosis of triple-negative breast cancer: A systematic review and meta-analysis. *Medicine*, *96*(22), e7085. https://doi.org/10.1097/MD.0000000000007085
- Martinez-Sanchez, A., & Murphy, C. L. (2013). MicroRNA target identification—experimental approaches. *Biology*, *2*(1), 189–205.
- Matutino, A., Amaro, C., & Verma, S. (2018). CDK4/6 inhibitors in breast cancer: beyond hormone receptor-positive HER2-negative disease. *Therapeutic Advances in Medical Oncology*, *10*, 1758835918818346.
- Medina, M. A., Oza, G., Sharma, A., Arriaga, L. G., Hernández Hernández, J. M., Rotello, V. M., & Ramirez, J. T. (2020). Triple-Negative Breast Cancer: A Review of

Conventional and Advanced Therapeutic Strategies. *International Journal of Environmental Research and Public Health*, *17*(6). https://doi.org/10.3390/ijerph17062078

- Medina, M. A., Oza, G., Sharma, A., Arriaga, L. G., Hernández, J. M. H., Rotello, V. M., & Ramirez, J. T. (2020). Triple-negative breast cancer: A review of conventional and advanced therapeutic strategies. *International Journal of Environmental Research and Public Health*, *17*(6). https://doi.org/10.3390/ijerph17062078
- Mina, A., Yoder, R., & Sharma, P. (2017). Targeting the androgen receptor in triple-negative breast cancer: current perspectives. *OncoTargets and Therapy*, 4675–4685.
- *miRBase: stem-loop hsa-mir-32*. (n.d.). https://mirbase.org/hairpin/MI0000090
- Nagayama, A., Vidula, N., Ellisen, L., & Bardia, A. (2020). Novel antibody–drug conjugates for triple negative breast cancer. *Therapeutic Advances in Medical Oncology*, *12*, 1758835920915980. https://doi.org/10.1177/1758835920915980
- Naorem, L. D., Muthaiyan, M., & Venkatesan, A. (2019). Identification of dysregulated miRNAs in triple negative breast cancer: a meta-analysis approach. *Journal of Cellular Physiology*, *234*(7), 11768–11779.
- Neophytou, C., Boutsikos, P., & Papageorgis, P. (2018). Molecular mechanisms and emerging therapeutic targets of triple-negative breast cancer metastasis. *Frontiers in Oncology*, *8*, 31.
- Nilsson, J., Ohlsson, M., Thulin, L., Höglund, P., Nashef, S. A. M., & Brandt, J. (2006). Risk factor identification and mortality prediction in cardiac surgery using artificial neural networks. *The Journal of Thoracic and Cardiovascular Surgery*, *132*(1), 12–19.
- O'Driscoll, L., & Clynes, M. (2006). Biomarkers and multiple drug resistance in breast cancer. *Current Cancer Drug Targets*, *6*(5), 365–384. https://doi.org/10.2174/156800906777723958
- Or, G. Ben, & Veksler-Lublinsky, I. (2021). Comprehensive machine-learning-based analysis of microRNA–target interactions reveals variable transferability of interaction rules across species. *BMC Bioinformatics*, *22*(1), 1–27.
- Ortega, M. A., Fraile-Martínez, O., Asúnsolo, Á., Buján, J., García-Honduvilla, N., & Coca, S. (2020). Signal transduction pathways in breast cancer: the important role of PI3K/Akt/mTOR. *Journal of Oncology*, *2020*(1), 9258396.
- Ortiz-Catalan, M., Brånemark, R., & Håkansson, B. (2013). BioPatRec: A modular research platform for the control of artificial limbs based on pattern recognition algorithms. *Source Code for Biology and Medicine*, *8*, 1–18.
- Pal, J. K., & Rami, B. R. (2024). Machine learning based identification of candidate miRNA biomarkers for micro-invasive breast cancer diagnosis. *BioRxiv*, 2008–2024.
- Panoff, J. E., Hurley, J., Takita, C., Reis, I. M., Zhao, W., Sujoy, V., Gomez, C. R., Jorda, M., Koniaris, L., & Wright, J. L. (2011). Risk of locoregional recurrence by receptor status in breast cancer patients receiving modern systemic therapy and post-mastectomy radiation. *Breast Cancer Research and Treatment*, *128*(3), 899–906. https://doi.org/10.1007/s10549-011-1495-1
- Paszek, S., Gabło, N., Barnaś, E., Szybka, M., Morawiec, J., Kołacińska, A., & Zawlik, I. (2017). Dysregulation of microRNAs in triple-negative breast cancer. *Ginekologia Polska*, *88*(10), 530–536. https://doi.org/{}
- Perou, C. M. (2011). Molecular stratification of triple-negative breast cancers. *The Oncologist*, *16*(S1), 61–70.

- Petrovic, N., & Ergun, S. (2018). miRNAs as potential treatment targets and treatment options in cancer. *Molecular Diagnosis & Therapy*, *22*, 157–168.
- Piccart-Gebhart, M. J., Procter, M., Leyland-Jones, B., Goldhirsch, A., Untch, M., Smith, I., Gianni, L., Baselga, J., Bell, R., Jackisch, C., Cameron, D., Dowsett, M., Barrios, C. H., Steger, G., Huang, C.-S., Andersson, M., Inbar, M., Lichinitser, M., Láng, I., … Gelber, R. D. (2005). Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *The New England Journal of Medicine*, *353*(16), 1659–1672. https://doi.org/10.1056/NEJMoa052306
- Piña-Sánchez, P., Valdez-Salazar, H.-A., & Ruiz-Tachiquín, M.-E. (2020). Circulating microRNAs and their role in the immune response in triple-negative breast cancer (Review). *Oncol Lett*, *20*(5), 224. https://doi.org/10.3892/ol.2020.12087
- Plevritis, S. K., Munoz, D., Kurian, A. W., Stout, N. K., Alagoz, O., Near, A. M., Lee, S. J., Van Den Broek, J. J., Huang, X., & Schechter, C. B. (2018). Association of screening and treatment with breast cancer mortality by molecular subtype in US women, 2000-2012. *Jama*, *319*(2), 154–164.
- *python 3.12.5*. (n.d.). https://www.python.org/
- Rahman, M. M., Brane, A. C., & Tollefsbol, T. O. (2019). MicroRNAs and Epigenetics Strategies to Reverse Breast Cancer. *Cells*, *8*(10). https://doi.org/10.3390/cells8101214
- Rastogi, P., Anderson, S. J., Bear, H. D., Geyer, C. E., Kahlenberg, M. S., Robidoux, A., Margolese, R. G., Hoehn, J. L., Vogel, V. G., Dakhil, S. R., Tamkus, D., King, K. M., Pajon, E. R., Wright, M. J., Robert, J., Paik, S., Mamounas, E. P., & Wolmark, N. (2008). Preoperative chemotherapy: updates of National Surgical Adjuvant Breast and Bowel Project Protocols B-18 and B-27. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *26*(5), 778–785. https://doi.org/10.1200/JCO.2007.15.0235
- Rehman, O., Zhuang, H., Muhamed Ali, A., Ibrahim, A., & Li, Z. (2019). Validation of miRNAs as breast cancer biomarkers with a machine learning approach. *Cancers*, *11*(3), 431.
- Rhodes, L. V, Martin, E. C., Segar, H. C., Miller, D. F. B., Buechlein, A., Rusch, D. B., Nephew, K. P., Burow, M. E., & Collins-Burow, B. M. (2015). Dual regulation by microRNA-200b-3p and microRNA-200b-5p in the inhibition of epithelial-to-mesenchymal transition in triple-negative breast cancer. *Oncotarget*, *6*(18), 16638–16652. https://doi.org/10.18632/oncotarget.3184
- Rodríguez, F. A. R., Flores, L. G., & Vitón-Castillo, A. A. (2022). Artificial intelligence and machine learning: present and future applications in health sciences. *Seminars in Medical Writing and Education*, *1*, 9.
- Rose, M., Burgess, J. T., O'Byrne, K., Richard, D. J., & Bolderson, E. (2020). *PARP inhibitors: clinical relevance, mechanisms of action and tumor resistance. Front Cell Dev Biol. 2020; 8: 564601*. Epub 2020/10/06. doi: 10.3389/fcell. 2020.564601. PubMed PMID: 33015058.
- Sachdeva, R. K., Bathla, P., Rani, P., Kukreja, V., & Ahuja, R. (2022). A Systematic Method for Breast Cancer Classification using RFE Feature Selection. *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 1673–1676. https://doi.org/10.1109/ICACITE53722.2022.9823464
- Sanz, H., Valim, C., Vegas, E., Oller, J. M., & Reverter, F. (2018). SVM-RFE: selection

and visualization of the most relevant features through non-linear kernels. *BMC Bioinformatics*, *19*, 1–18.
- Seyfried, T. N., & Huysentruyt, L. C. (2013). On the origin of cancer metastasis. *Critical Reviews in Oncogenesis*, *18*(1–2), 43–73. https://doi.org/10.1615/critrevoncog.v18.i1-2.40
- Si, W., Shen, J., Zheng, H., & Fan, W. (2019). The role and mechanisms of action of microRNAs in cancer drug resistance. *Clinical Epigenetics*, *11*, 1–24.
- Statnikov, A., Wang, L., & Aliferis, C. F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, *9*, 1–10.
- Stein, G., Chen, B., Wu, A. S., & Hua, K. A. (2005). Decision tree classifier for network intrusion detection with GA-based feature selection. *Proceedings of the 43rd Annual Southeast Regional Conference-Volume 2*, 136–141.
- Tang, Y., Wang, Y., Kiani, M. F., & Wang, B. (2016). Classification, Treatment Strategy, and Associated Drug Resistance in Breast Cancer. *Clinical Breast Cancer*, *16*(5), 335–343. https://doi.org/10.1016/j.clbc.2016.05.012
- Tian, X., Liu, Z., Niu, B., Zhang, J., Tan, T. K., Lee, S. R., Zhao, Y., Harris, D. C. H., & Zheng, G. (2011). E-cadherin/β-catenin complex and the epithelial barrier. *Journal of Biomedicine & Biotechnology*, *2011*, 567305. https://doi.org/10.1155/2011/567305
- Waks, A. G., & Winer, E. P. (2019). Breast Cancer Treatment: A Review. *JAMA*, *321*(3), 288–300. https://doi.org/10.1001/jama.2018.19323
- Wang, A., An, N., Yang, J., Chen, G., Li, L., & Alterovitz, G. (2017). Wrapper-based gene selection with Markov blanket. *Computers in Biology and Medicine*, *81*, 11–23.
- Waspada, I., Wibowo, A., & Meraz, N. S. (2017). Supervised machine learning model for microrna expression data in cancer. *Jurnal Ilmu Komputer Dan Informasi*, *10*(2), 108–115.
- Wojciechowska, G., Szczerbinski, L., Kretowski, M., Niemira, M., Hady, H. R., & Kretowski, A. (2022). Exploring microRNAs as predictive biomarkers for type 2 diabetes mellitus remission after sleeve gastrectomy: A pilot study. *Obesity*, *30*(2), 435–446.
- Won, K., & Spruck, C. (2020). Triple-negative breast cancer therapy: Current and future perspectives. *International Journal of Oncology*.
- Wu, D., Jennings, C., Terpenny, J., Gao, R. X., & Kumara, S. (2017). A Comparative Study on Machine Learning Algorithms for Smart Manufacturing: Tool Wear Prediction Using Random Forests. *Journal of Manufacturing Science and Engineering*, *139*(7). https://doi.org/10.1115/1.4036350
- Wu, J., & Hicks, C. (2021). Breast Cancer Type Classification Using Machine Learning. In *Journal of Personalized Medicine* (Vol. 11, Issue 2). https://doi.org/10.3390/jpm11020061
- Xu, C., & Jackson, S. A. (2019). Machine learning and complex biological data. In *Genome biology* (Vol. 20, Issue 1, p. 76). https://doi.org/10.1186/s13059-019-1689-0
- Xu, L., Wu, Z., Chen, Y., Zhu, Q., Hamidi, S., & Navab, R. (2014). MicroRNA-21 (miR-21) regulates cellular proliferation, invasion, migration, and apoptosis by targeting PTEN, RECK and Bcl-2 in lung squamous carcinoma, Gejiu City, China. *PloS One*, *9*(8), e103698.
- Xu, P., Wu, Q., Lu, D., Yu, J., Rao, Y., Kou, Z., Fang, G., Liu, W., & Han, H. (2020). A systematic study of critical miRNAs on cells proliferation and apoptosis by the shortest

path. *BMC Bioinformatics*, *21*(1), 1–14. https://doi.org/10.1186/s12859-020-03732-x
- Yan, K., & Zhang, D. (2015). Feature selection and analysis on correlated gas sensor data with recursive feature elimination. *Sensors and Actuators B: Chemical*, *212*, 353–363.
- Yu, K., Xie, W., Wang, L., & Li, W. (2021). ILRC: a hybrid biomarker discovery algorithm based on improved L1 regularization and clustering in microarray data. *BMC Bioinformatics*, *22*, 1–19.
- Zeng, Z. L., Zhu, Q., Zhao, Z., Zu, X., & Liu, J. (2021). Magic and mystery of microRNA-32. *Journal of Cellular and Molecular Medicine*, *25*(18), 8588–8601.
- Zhang, C., Sheng, W., Al-Rawe, M., Mohiuddin, T. M., Niebert, M., Zeppernick, F., Meihold-Heerlein, I., & Hussain, A. F. (2022). EpCAM-and EGFR-specific antibody drug conjugates for triple-negative breast cancer treatment. *International Journal of Molecular Sciences*, *23*(11), 6122.
- Zhang, J.-X., Yang, W., Wu, J.-Z., Zhou, C., Liu, S., Shi, H.-B., & Zhou, W.-Z. (2021). MicroRNA-32-5p inhibits epithelial-mesenchymal transition and metastasis in lung adenocarcinoma by targeting SMAD family 3. *Journal of Cancer*, *12*(8), 2258.
- Zhang, J. L., Zheng, H. F., Li, K., & Zhu, Y. P. (2022). miR-495-3p depresses cell proliferation and migration by downregulating HMGB1 in colorectal cancer. *World Journal of Surgical Oncology*, *20*(1), 101.
- Zhang, R.-R., Wang, L.-M., & Shen, J.-J. (2020). Overexpression of miR-32 inhibits the proliferation and metastasis of ovarian cancer cells by targeting BTLA. *European Review for Medical & Pharmacological Sciences*, *24*(9).
- Zhang, X., Jonassen, I., & Goksøyr, A. (2021). Machine learning approaches for biomarker discovery using gene expression data. *Bioinformatics*.
- Zhang, Y., Ding, C., & Li, T. (2008). Gene selection algorithm by combining reliefF and mRMR. *BMC Genomics*, *9*, 1–10.
- Zhavoronkov, A., Vanhaelen, Q., & Oprea, T. I. (2020). Will artificial intelligence for drug discovery impact clinical pharmacology? *Clinical Pharmacology & Therapeutics*, *107*(4), 780–785.
- Zhu, T., Fan, D., Ye, K., Liu, B., Cui, Z., Liu, Z., & Tian, Y. (2020). Role of miRNA-542-5p in the tumorigenesis of osteosarcoma. *FEBS Open Bio*, *10*(4), 627–636.

Appendixes

Appendix 1 Filtered dataset

Appendix 2 Dataset of GSE154255

Appendix 3 Dataset of GSE100453

Appendix 4 Dataset of TNBC and normal samples

Appendix 5 Machine learning with python

العنوان : التعرف على المؤشرات الحيوية المحتملة لتشخيص سرطان الثدي الثلاثي السلبي والتنبؤ به باستخدام الذكاء الاصطناعي.

إعداد: شهد مصطفى يحيى قواسمه.

إشراف : الدكتور يوسف نجاجرة .

مشرف مشارك: الدكتور رشيد جيوسي.

ملخص:

يُعد سرطان الثدي الثلاثي السلبي(TNBC) واحدًا من أكثر أنواع السرطان العدوانية، ويرتبط بأعلى معدلات الوفيات. يتميز سرطان الثدي الثلاثي السلبي بعدم وجود مستقبلات الاستروجين والبروجستيرون و عامل النمو البشري. تلعب الجزيئات الصغيرة من الحمض النووي الريبي (MicroRNA) دورًا رئيسيًا في التعبير الجيني من خلال تفاعلها مع جزيئات الحمض النووي الريبي الرسول. MicroRNAقد تعمل كمؤشرات حيوية لتشخيص سرطان الثدي الثلاثي السلبي والتنبؤ بتوقعات المرض. الهدف الرئيسي من الدراسة هو تحليل التعبير الجيني لجين MicroRNA عن طريق استخلاص مستويات التعبيرالجيني والبيانات السريرية لمرضى سرطان الثدي الثلاثي السلبي من قاعدة بيانات التعبير الجيني (GEO)لاكتشاف المؤشرات الحيوية لهذا النوع من السرطان باستخدام تعلم الآلة. تشمل أهداف الدراسة استخدام الذكاء الاصطناعي وتعلم الآلة لتحديد المؤشرات الحيوية المحتملة للتشخيص والتنبؤ لسرطان الثدي الثلاثي السلبي. بالإضافة إلى ذلك، بناء نموذج بإجراء تحليل تلوي، يليه تحليل التعبير التفاضلي، الذي يكشف عن الارتباطات ذات الأهمية الإحصائية بين التوقيعات الجينية المتعددة. تتضمن هذه الدراسة أيضًا مقارنة ملفات التعبيرالجيني بين أنسجة سرطان الثدي الثلاثي السلبي والأنسجة الطبيعية، وكذلك بين أنسجة سرطان الثدي الثلاثي السلبي وأنسجة السرطان غير الثلاثي السلبي. في النهاية، تم بناء نموذج تعلم الآلة باستخدام طريقة اختيار ميزات مهجنة لاختيار المؤشرات الحيوية. شاركت الدراسة في دمج مجموعتين من البيانات، مما نتج عنه مجموعة بيانات مدمجة تحتوي على 4577 MicroRNA. في مرحلة بناء نموذج تعلم الآلة، تم استخدام مجموعة من طرق اختيار الميزات لتحديد التوقيعات الحيوية التي تميز سرطان الثدي الثلاثي السلبي عن الأنسجة الطبيعية. يشمل ذلك طريقة الالتفاف باستخدام الاستبعاد التكراري للميزات، جنبًا إلى جنب مع طرق مدمجة باستخدام الغابات العشوائية وآلة المتجه الداعم. كشفت دراستنا عن اختلافات كبيرة في التعبير الجيني بين أنسجة سرطان الثدي الثلاثي السلبي والأنسجة الطبيعية. في المقابل، لم يختلف التعبيرالجيني بشكل كبير بين أنسجة سرطان الثدي الثلاثي السلبي وأنسجة غير الثلاثي السلبي. علاوة على ذلك، تُظهر الدراسة أن استخدام الاستبعاد التكراري للميزات وآلة المتجه الداعم والغابات العشوائية كخوارزمية اختيار ميزات هجينة لملفات التعبيرأو مجموعات بيانات مماثلة تحتوي على عدد كبير من الميزات مقارنة بعدد العينات يمكن أن يزيل الميزات الزائدة بشكل فعال، ويحدد المؤشرات الحيوية ذات الصلة التشخيصية، ويحافظ على دقة تصنيف عالية. وأخيرًا، حددت الدراسة أن(miR-32-5P) يمكن أن يُستخدم كمؤشر حيوي محتمل لتشخيص سرطان الثدي الثلاثي السلبي، وأن التعبير العالي له يرتبط ارتباطًا كبيرًا بزيادة البقاء على قيد الحياة بشكل عام لدى مرضى سرطان الثدي الثلاثي السلبي.