Deciphering Breast Cancer Complexity: Harnessing Spatial Transcriptomics and AI for Personalized Therapeutic Strategies

By

Ayesha Iman

(Registration No: 00000402396)

Department of Sciences

School of Interdisciplinary Engineering & Sciences (SINES)

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(2024)

Deciphering Breast Cancer Complexity: Harnessing Spatial Transcriptomics and AI for Personalized Therapeutic Strategies

By

Ayesha Iman

(Registration No: 00000402396)

A thesis submitted to the National University of Sciences and Technology, Islamabad,

in partial fulfillment of the requirements for the degree of

Master of Science in

Bioinformatics

Supervisor: Dr. Mehak Rafiq

School of Interdisciplinary Engineering & Sciences (SINES)

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

Annex A to NUST Letter No. 0972/102/Exams/Thesis-Cert dated 23 Dec 16.

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Ms. **Ayesha Iman** Registration No. **402396** of **SINES** has been vetted by undersigned, found complete in all aspects as per NUST Statutes/Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature with stamp: _____

DR. MEHAK RAFIQ
Assistant Professor
SINES, National University of Science & Technology
H-12 Islamabad

Name of Supervisor: Dr Mehak Rafiq

Date: 23/8/24.

Signature of HoD with stamp: _____

Dr. Fouzia Malik
HoD Sciences
Professor
SINES NUST
Islamabad

Date: 28-8-2024

Countersign by

Signature (Dean/Principal): _____

Date: 29/08/2024

AUTHOR'S DECLARATION

I Ayesha Iman hereby state that my MS thesis titled "Deciphering Breast Cancer Complexity: Harnessing Spatial Transcriptomics and AI for Personalized Therapeutic Strategies" is my own work and has not been submitted previously by me for taking any degree from National University of Sciences and Technology, Islamabad or anywhere else in the country/ world.

At any time if my statement is found to be incorrect even after I graduate, the university has the right to withdraw my MS degree.

Name of Student: Ayesha Iman

DEDICATION

I dedicate this thesis to my exceptional parents, siblings, friends, and teachers whose unconditional love, support, and guidance led me to this world of accomplishment.

ACKNOWLEDGEMENTS

All praise is for **Almighty Allah**, the ultimate source of all knowledge. By His grace, I have reached this stage of knowledge with the ability to contribute something beneficial to His creation. My deepest respects are to **the Holy Prophet Hazrat Muhammad (PBUH)**, the symbol of guidance and fountain of knowledge.

I earnestly thank my supervisor, **Dr. Mehak Rafiq** for her keen interest, invaluable guidance, encouragement, and continuous support throughout my research journey. I am extremely grateful for her thought-provoking discussions, sound advice, and valuable suggestions. Her mentorship has enabled me to tackle problems more meaningfully and provided me with the resources to pursue my objectives diligently and sincerely.

I am also thankful to my GEC committee members **Dr.Masood ur Rehman Kiyani** and **Dr. Uzma Habib** who guided me throughout my project and offered valuable feedback and suggestions to refine my thesis. Additionally, I acknowledge all the **faculty** members of SINES for their kind assistance at various phases of this research.

My gratitude extends to my colleagues in the Data Analytics lab. Special thanks to **Ariba Abbasi**, **Adeel Nisar**, **Amman Safeer**, and **Talha Zuberi** for their continuous help and feedback at every stage of the research. I am also grateful to my friends **Muneeb Nasir** and **Adnan Tariq** for their unwavering support. Lastly, I am deeply thankful to my parents, siblings and fiancé **Zunair Khalid** for their immense support throughout this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page No.

LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

TNBC	Triple Negative Breast Cancer
HDI	Human Development Index
HER2	human epidermal growth factor receptor
PR	Progestrone Receptor
ER	Estrogen Receptor
TME	Tumor microenviornment
EMT	Epithelialtomesenchymal transition
TILs	tumor-infiltrating lymphocytes
TAMs	tumor-associated macrophages
CAFs	cancer-associated fibroblasts
CAAs	cancer-associated adipocytes
NGS	Next generation Sequencing
PCA	Principal Component Analysis
UMAP	Uniform Manifold Approximation and Projection
DEGs	Differentially Expressed Genes
XGBoost	Extreme Gradient Boosting

ABSTRACT

Breast Cancer, particularly Triple-Negative Breast Cancer (TNBC), remains a formidable challenge due to its aggressive nature, lack of targeted therapies, and poor prognosis. This research addresses the critical need for more accurate prognostic markers by integrating spatial transcriptomics with artificial intelligence (AI) to explore the spatial heterogeneity of gene expression within TNBC tumors. Spatial transcriptomics offers a high-resolution view of the tumor microenvironment, preserving the spatial context of gene expression, while single-cell RNA sequencing (scRNA-seq) provides detailed insights into the cellular composition of the tumors. By identifying and analyzing differentially expressed genes (DEGs) across spatial and single-cell datasets, this study aims to uncover key biomarkers that could serve as therapeutic targets and improve patient outcomes. Machine learning models, including XGBoost and Support Vector Machines (SVM), were employed to develop predictive models for cancer classification, disease staging, and prognosis. These models demonstrated high accuracy, enhancing the understanding of TNBC's complex molecular landscape and supporting the development of personalized treatment strategies. The findings highlight the potential of integrating spatial transcriptomics with AI to revolutionize cancer research, offering new avenues for precision medicine in TNBC.

CHAPTER 1: INTRODUCTION

This chapter includes an overview of breast cancer, with particular attention to Triple Negative Breast Cancer (TNBC), an aggressive subtype, its statistics, and contributing variables. It examines the shortcomings of the existing therapy modalities and emphasizes the necessity for creative fixes. This chapter will describe the issue in more detail and go over possible fixes that could make use of cutting-edge methods like artificial intelligence and spatial transcriptomics.

1.1 Breast Cancer

Breast cancer is the most prevalent cancer diagnosis for women worldwide. It is also the leading cause of cancer-related deaths in humans [1]. Breast cancer is caused by aberrant breast cells that proliferate and develop into tumors. Tumors have the potential to grow throughout the body and become lethal if ignored [2].

1.1.1 Epidemiology

In 2022, 2.3 million women worldwide were diagnosed with breast cancer, resulting in 670,000 deaths. Breast cancer affects women in every country, at any age post-puberty, with incidence rates increasing with age. Global data highlight significant disparities based on human development[2]. In countries with a very high Human Development Index (HDI), 1 in 12 women are diagnosed with breast cancer, and 1 in 71 women dies from it. Conversely, in countries with a low HDI, 1 in 27 women are diagnosed with breast cancer, but 1 in 48 women dies from the disease [2]

1.1.2 Risk Factors

The latest research has provided deeper insights into the various demographic, genetic, reproductive, hormonal, metabolic, and lifestyle factors that can influence an individual's breast cancer risk.

One important non-modifiable risk factor is age, with the incidence of breast cancer rising sharply beyond the age of 50 [3]. Gene mutations such as those in BRCA1 or BRCA2 significantly increase the risk of breast cancer. Additionally, a family history of breast or ovarian cancer in first-degree relatives is a significant risk factor [4]. Lifetime hormonal

exposure, which is increased by early menarche, late menopause, nulliparity, and the use of hormone replacement therapy or contraception, also elevates breast cancer risk. Lifestyle and environmental factors play a role as well; inactivity, obesity (particularly after menopause), alcohol use, and exposure to certain chemicals are all associated with a higher risk of developing breast cancer [5]. (Figure 1.1) illustrates some of the most critical risk factors for breast cancer as identified by the Pink Ribbon organization [6].

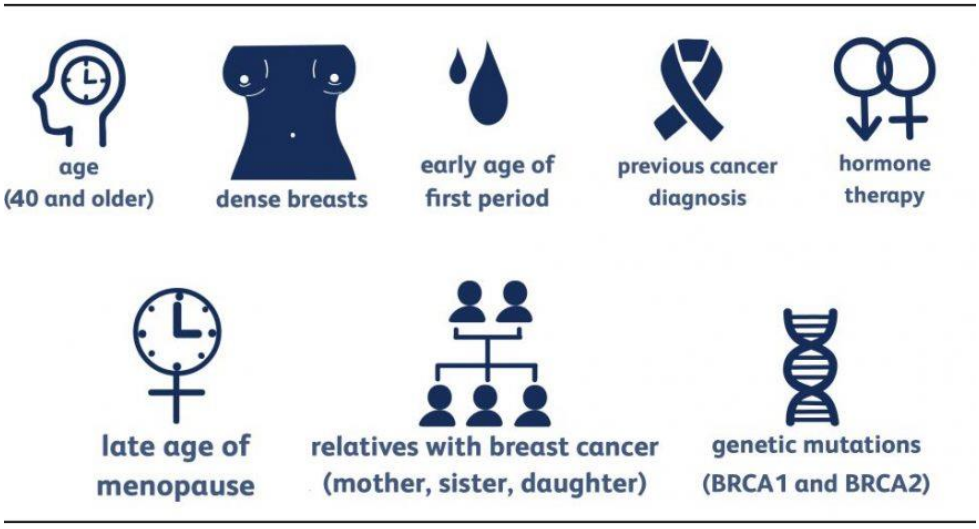

Figure 1.1: Risk factors of breast cancer

1.2 Triple Negative Breast Cancer

Breast cancer encompasses several types, each with distinct characteristics and treatment approaches [7]. The two most common forms are HER2-positive breast cancer, which has an excess of the HER2 protein that promotes cancer cell proliferation, and hormone receptor-positive breast cancer, where cancer cells grow in response to hormones like progesterone or estrogen [8]. Among these, a particularly challenging type is triple-negative breast cancer (TNBC). It is defined by the lack of the three most often targeted biomarkers considered for breast cancer treatment: the human epidermal growth factor receptor (HER2), progesterone receptor (PR), and estrogen receptor (ER). TNBC is more common in younger women and in certain racial groups like African Americans, Hispanics, and Indians [9]. 15% to 20% of cases of breast cancer are TNBC, and these cases typically have a more aggressive clinical course, with poorer evolution occurring within the first 3 to 5 years following diagnosis, early and higher rates of distant recurrences, mainly visceral, and poor survival [10], [11], [12].

1.2.1 Tumour Microenvironment

The tumor microenvironment (TME) has a major impact on the onset, development, proliferation, immune system suppression, angiogenesis, invasion/cell migration, and adverse prognosis of TNBC [13]. The TME is very heterogeneous, with a variety of cellular compositions controlled by several signalling pathways. Cell populations with diverse phenotypic features contribute to the complexity of the TME, and cellular plasticity often facilitates cellular survival and proliferation even in the aftermath of chemotherapy. The TME carves out a space between tumor cells and the surrounding tissues using immune cells and the endothelial system [14].

The complex TME may also trigger the epithelial to mesenchymal transition (EMT), which is the process that leads to the creation of cancer stem cells. Consequently, the development of specific treatment techniques and a knowledge of the TME's intricacy may function as target hallmarks for tumor regression and elimination. Numerous prognostic markers for TME that show immune system suppression have been identified through extensive clinical-pathological reporting. These markers include tumor-infiltrating lymphocytes (TILs), tumor-associated macrophages (TAMs), cancer-associated fibroblasts (CAFs), and cancer-associated adipocytes (CAAs). These tumor-associated markers have a connection to the immune/tumor interactions in TNBC, which support the chemoresistance development process as shown in the (figure 1.2) [15].

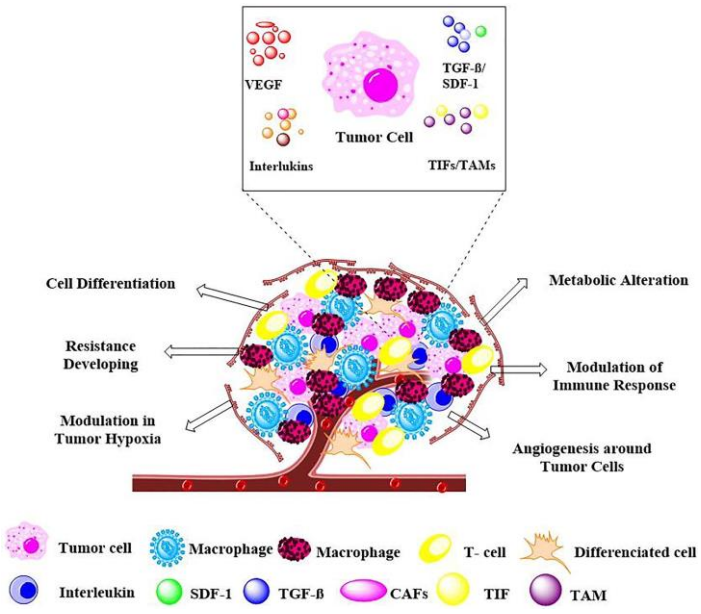

Figure 1.2: Tumor cell interactions with the tumor microenvironment.

The various tumor microenvironments are associated with the development of treatment resistance, immune suppression, metabolic alterations (adaptation to low nutrition), cell differentiation, promotion of hypoxia, and angiogenesis, the formation of new blood vessels. Several variables have been connected to cancer, including transforming growth factor beta (TGF-beta), VEGF, stromal cell-derived factor 1 (SDF-1), tumor-infiltrating lymphocytes (TILs), and tumor-associated macrophages (TAMs).

1.2.2 Treatment and Therapeutic targets

TNBC is a concern for medical professionals and patients alike because, in comparison to other breast cancer subtypes, it has a higher death rate, a worse prognosis, and less available treatment choices [16]. The lack of receptors makes it unresponsive to hormonal and HER2-targeted therapies, posing significant treatment challenges. Consequently, a multifaceted approach is required to manage TNBC effectively [17]. The following paragraphs outline the various treatment options currently available, each tailored to address the unique aspects of this challenging disease. Surgery is a primary treatment option for TNBC, involving procedures such as lumpectomy, which removes the tumor and some surrounding tissue, and mastectomy, which involves the partial or complete removal of one or both breasts [18]. Radiation therapy is another critical component, using high-energy rays to target and kill cancer cells. External beam radiation is the most common form, while brachytherapy, or internal radiation, involves placing radioactive seeds near the tumor site [19].

Chemotherapy remains a cornerstone in the treatment of TNBC due to the lack of specific hormonal or HER2 targets [20]. The conventional chemotherapy regimen typically includes a combination of taxanes and anthracyclines, which are known for their efficacy in inhibiting cancer cell proliferation. Additionally, combinations such as cyclophosphamide, methotrexate, and 5-fluorouracil are employed to target the rapidly dividing cancer cells through various mechanisms. Another effective regimen includes epirubicin and cyclophosphamide followed by paclitaxel, which helps in managing the disease by attacking cancer cells at different phases of their growth cycle [21]. Moreover, platinum-based chemotherapies, specifically carboplatin and cisplatin, have shown significant promise, particularly in patients with BRCA1/2 mutations, due to their ability to induce DNA damage and apoptosis in cancer cells [22].

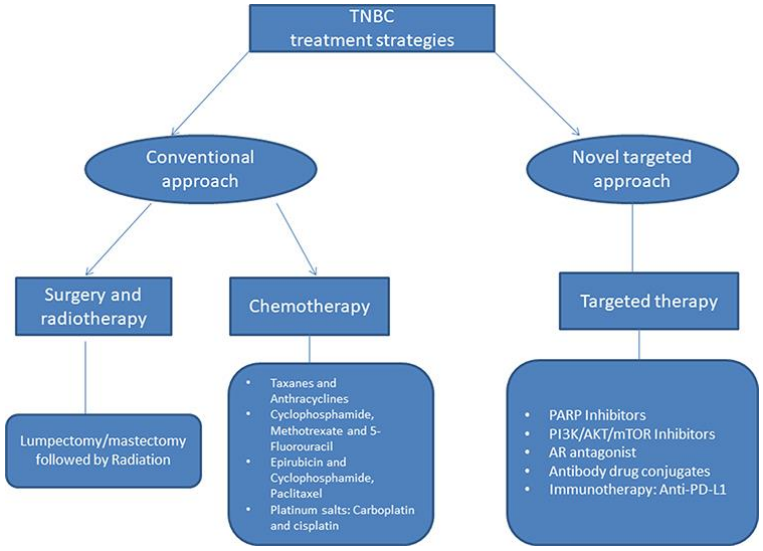

Figure 1.3: Treatment strategies for triple-negative breast cancer

The advent of targeted therapy has provided new avenues for treating TNBC, focusing on specific molecular abnormalities within the cancer cells [21]. PARP inhibitors, for instance, are particularly effective in patients with BRCA mutations, as they exploit the defective DNA repair mechanisms in these cancer cells, leading to cell death [23]. The PI3K/AKT/mTOR pathway, which plays a crucial role in cell growth and survival, is another target, with inhibitors designed to disrupt these signaling processes and curb cancer progression [24]. Additionally, androgen receptor (AR) antagonists are being explored as a potential treatment, given the presence of AR in a subset of TNBC cases. Advanced therapeutic strategies also include antibody-drug conjugates that deliver cytotoxic agents directly to cancer cells, thereby minimizing systemic toxicity[25]. Immunotherapy, particularly anti-PD-L1 agents, has emerged as a promising approach by harnessing the patient's immune system to recognize and destroy cancer cells, offering a new beacon of hope for those battling this aggressive cancer subtype [20].

1.3 Sequencing Techniques

The first technique for sequencing DNA was published over twenty-five years after the structure of DNA was known [26]Sanger sequencing was the first technique which determined the nucleotide sequence of DNA by using chain-terminating dideoxynucleotides [27]. The main limitation of the Sanger Sequencing was low-quality sequences within the first

15–40 bp, and the method's inability to identify single base pair differences in longer segments (e.g., > 900 bp) [28]. One innovative high-throughput technique for thorough transcriptome analysis is RNA sequencing, or RNA-Seq. It can concurrently assess the expression levels of hundreds of genes, offering information on the rules and functional pathways governing biological processes [29]. RNA sequencing is limited by biases introduced during library preparation and amplification, which can affect the accuracy of gene expression measurements. Additionally, the short read lengths can complicate the assembly and quantification of transcripts, particularly for genes with repetitive or complex structures [30].

Then came the Next Generation Sequencing. NGS technologies were released between 2004 and 2006, revolutionizing biomedical research and leading to a significant increase in sequencing data output [31]. But NGS was limited by their short read lengths, which complicates the accurate reconstruction of long DNA sequences and the detection of structural variations. Additionally, the labor-intensive preparation process and challenges in mapping repetitive genomic regions further hinder their efficiency and accuracy [32].

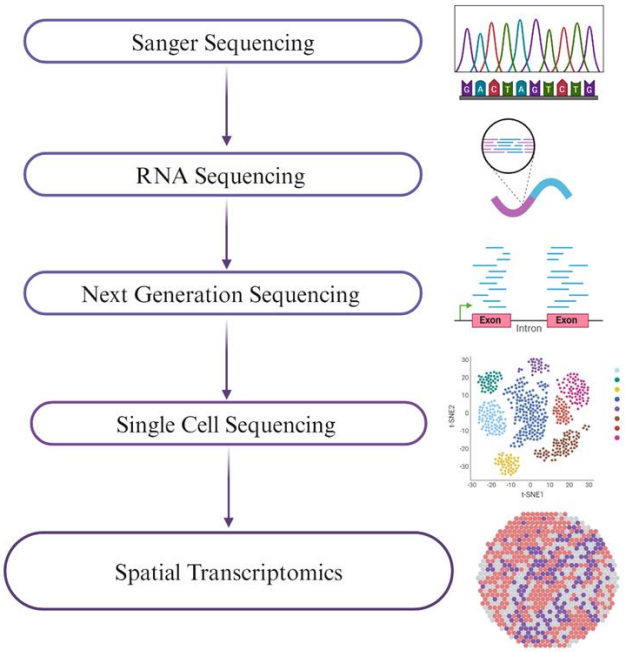

Figure 1.4: Sequencing technologies

Single cell RNA sequencing is a new-generation sequencing technology based on nanopore equipment after second-generation sequencing [33]. Using scRNA-seq, it has been possible to

demonstrate in recent years the heterogeneity of cells, cell dynamic differentiation processes, tumor prognosis, treatment, and other aspects of cancer [34] Accurately characterizing the expression of genes in the microenvironment and the tumor, as well as knowing the degree of transcriptome heterogeneity within the tumor entity, may assist determine more effective molecular targets for prognosis and treatment [35]. scRNA-seq also helps with personalized therapy because of its ability to identify cell subsets and biomarkers with potential treatments[36]. Precisely single-cell RNA sequencing (scRNA-Seq) allows for the detailed examination of the cellular composition and heterogeneity within tumors by profiling gene expression in thousands of individual cells. Building on this, spatial transcriptomics maps these cellular profiles within their original tissue context, revealing how different cell types are organized and interact spatially within the tumor microenvironment [36].

1.3.1 Spatial Transcriptomics

In recent years, transcriptome research has revolutionized cancer diagnosis and treatment [37]. Novel developments in single-cell RNA sequencing (scRNA-seq) technology have yielded a wealth of information regarding the cellular makeup of malignancies [38]. scRNA-seq analysis of cell transcriptomes can be done at high throughput, however tissue processing eliminates the transcriptomes' spatial context. On the other hand, techniques such as immunohistochemistry (IHC) and in situ hybridization yield excellent spatial resolution, but they often need pre-selection of targets, which limits their suitability for high-throughput exploratory studies. Subsequent investigations have demonstrated that most tumor-associated cell states are intricate and challenging to define using a small number of surface receptors or marker genes [39][39], which presents both a technical and financial challenge to targeted spatial methods [40]. Here comes spatial trancriptomics, a cutting-edge technology that allows for the comprehensive analysis of gene expression patterns while preserving the spatial context within tissue samples. It provides an intricate window into how tissue microenvironments relate to or influence gene expression [41]. With spatial transcriptomics, we can now profile averaged transcriptomes in complex cellular soups or in relative isolation using Sc RNA-seq, both of which obliterate geographical information. Instead, by completing the picture of how the position of a cell inside a tissue can impact its gene expression, we can close the gaps concerning complex biological problems [42].

1.3.1.1 Methodologies of Spatial Transcriptomics

Spatial transcriptomics encompasses several methodologies, each with its own unique approach to preserving and analyzing spatial gene expression. The first method, **Sequencing-based**, involves placing tissue sections on spatially barcoded arrays to capture mRNA, constructing libraries, and sequencing the captured RNA to map spatial gene expression, exemplified by 10X Genomics Visium. The second method, **Probe-based**, uses barcoded RNA probes or RNA scope probes for tissue staining, followed by imaging and selection of regions of interest (ROIs). UV cleavage is then used to collect barcodes for each ROI, leading to data analysis and visualization, as seen with GeoMx. The third method, **Imaging-based**, involves immunostaining or probe hybridization/labeling, microscopic imaging, and visualization of the tissue, followed by data analysis, illustrated by CosMx SMI. The fourth method, **Image-guided spatially resolved**, includes immunostaining, microscopic imaging and visualization, and photo-selection of target cells or ROIs, followed by single-cell sorting and sequencing. This comprehensive approach allows for detailed data analysis and visualization, demonstrated by Spatially annotated FUNseq.

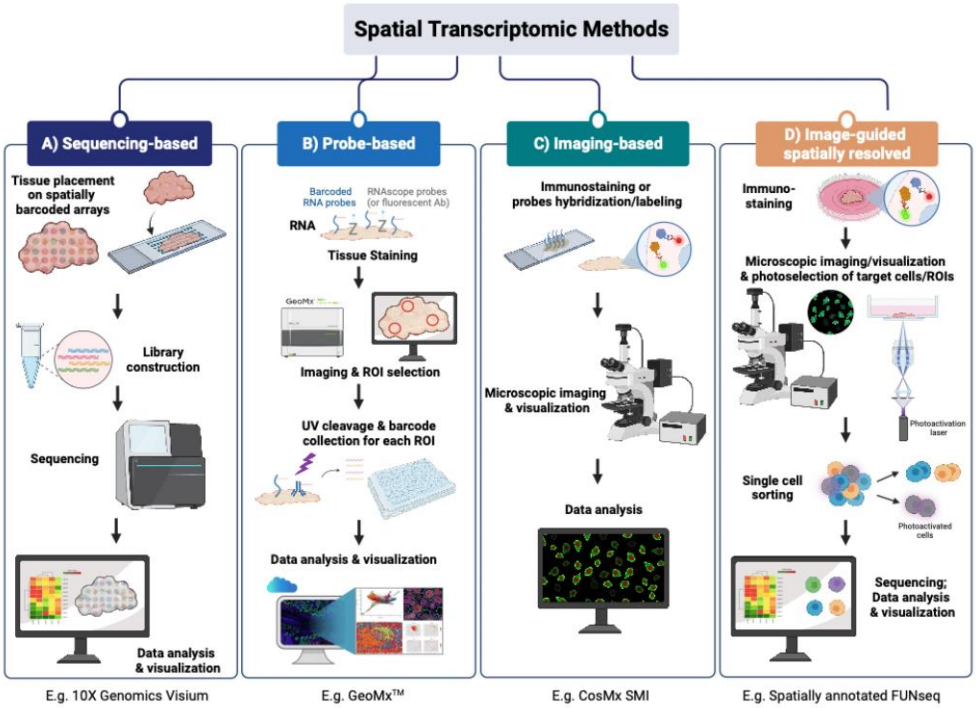

Figure 1.5: (A) Sequencing-based, (B) Probe-based, (C) Imaging-based, and (D) Image-guided methods.

Each type of spatial transcriptomics method offers unique advantages and is suited for different applications, depending on the required resolution, throughput, and specific research goals. These methodologies collectively enhance our ability to understand the spatial organization of gene expression within complex tissues, providing critical insights into cellular function and disease mechanisms [43].

1.3.1.2 Computational Tools and Softwares

Scanpy and Squidpy are essential computational tools for analyzing spatial transcriptomics data, offering powerful capabilities to researchers [44]. Scanpy, a Python-based package, is widely used for processing and visualizing single-cell gene expression data, including spatial transcriptomics. It provides a comprehensive range of functions, from basic preprocessing and normalization to advanced tasks like clustering and differential expression analysis. This tool is particularly useful for integrating spatial information with gene expression data, allowing researchers to visualize and interpret complex patterns within tissues [45]. Squidpy, built on top of Scanpy, extends these capabilities specifically for spatial transcriptomics. It incorporates spatial statistics and image processing techniques to analyze spatial dependencies and cellular interactions within tissues. Squidpy enables spatially aware clustering, neighborhood analysis, and integration of high-resolution tissue images with gene expression data, providing a detailed view of tissue architecture and cellular microenvironments. Together, these tools enhance the analytical power of spatial transcriptomics studies, helping researchers uncover the biological mechanisms underlying tissue function and disease [46].

In recent years, machine learning and deep learning techniques have become popular for analyzing spatial transcriptomics data due to their ability to handle large and complex datasets [47]. By integrating spatial transcriptomics with genomic, proteomic, and clinical data, machine learning models create a comprehensive view of the tumor microenvironment [48]. This approach helps identify new biomarkers and therapeutic targets, improves our understanding of tumor heterogeneity, and enhances the precision of personalized medicine. Despite these advancements, there is still a challenge in cancer research, particularly for triple-negative breast cancer (TNBC). Traditional bulk or single-cell studies often overlook the spatial expression patterns within tissues, ignoring the gene expression variations within the tumor microenvironment. This oversight highlights the need for research that considers these spatial patterns to develop more effective and personalized treatment strategies.

1.4 Problem Statement

Previous research on prognostic markers for TNBC has predominantly focused on bulk or single-cell analyses, neglecting the crucial spatial distribution of gene expression within tumor tissues. This significant research gap fails to account for the inherent heterogeneity of gene expression, which is vital for accurate prognostication. As a result, current prognostic markers lack the necessary consistency across different regions of the tumor, leading to potential unreliability in clinical applications.

1.5 Proposed Solution

To address the research gap in prognostic markers for *T*NBC, we propose an integrated approach combining scRNA-seq and spatial transcriptomics to capture the spatial heterogeneity of gene expression within tumors. This will enable the identification of key marker genes that reflect the tumor's complexity. Using these markers, we will develop a predictive machine learning model for TNBC disease staging and a prognostic model for survival outcomes. These models will leverage algorithms such as Random Forest, K-Nearest Neighbors, Multi-Layer Perceptron, and Support Vector Machine, ensuring robustness and accuracy. These models will account for the tumor's spatial heterogeneity, leading to more accurate and reliable prognostication and aiding in the personalized treatment of TNBC patients.

1.6 Objectives

- Conduct single-cell analysis and spatial transcriptomics to identify marker genes in Triple-Negative Breast Cancer.
- Develop and validate a predictive machine learning model for disease staging in TNBC.
- Create a prognostic model for survival outcomes in TNBC patients.

CHAPTER 2: REVIEW OF LITERATURE

2.1 Overview

In this chapter, we will explore significant advancements in breast cancer research, covering genomic and transcriptomic studies, as well as proteomics and metabolomics approaches. We will discuss the integration of these fields through comprehensive omics analysis. Additionally, we will examine the impact of AI and machine learning in cancer research, focusing on predictive and classification models and their role in multi-omics analysis. Finally, we will address the challenges and current solutions in integrating spatial transcriptomics with AI, highlighting its potential in enhancing breast cancer research.

2.2 Advances in Breast Cancer Research

In recent years, the field of breast cancer research and treatment has seen remarkable advancements, driven by cutting-edge technologies and innovative approaches. Researchers have made significant strides in understanding the molecular underpinnings of breast cancer, utilizing genomic, proteomics and transcriptomic studies to uncover critical insights.

2.2.1 Proteomics and Genomics Studies

Functional biological parameters do not always match gene expression characteristics because there can be differences between RNA levels and protein levels. Functional proteomics analysis is therefore employed as additional information, and the combination of transcriptome and genomic data facilitates the identification of novel targets [49]. Proteomics research has revealed differentially expressed proteins that may be targets for therapy and prognostic indicators, greatly advancing our understanding of TNBC. The sensitivity and accuracy of protein detection have been improved by recent developments in mass spectrometry-based proteomics, such as tandem mass tag (TMT) labeling and isobaric tags for relative and absolute quantitation (iTRAQ), which allow for thorough comparative analyses of protein expression across TNBC samples. Additionally, new molecular targets and important regulatory networks involved in TNBC have been identified through the integration of proteomics with transcriptomics and genomes data using cutting-edge bioinformatics methods. The field of functional proteomics has yielded valuable insights into protein-protein interactions and post-

translational changes, emphasizing their involvement in the progression of tumor necrosis factor-α and resistance to treatment. However, significant limitations exist within proteomics studies, particularly in the context of TNBC. One major challenge is detecting low-abundance proteins, which may play critical roles in cancer progression but are often overshadowed by more abundant proteins. Additionally, the complexity of the proteome, with its vast dynamic range and extensive post-translational modifications, poses difficulties in comprehensive protein analysis [50], [51].

It is well known that genomic mutations and genetic heterogeneity are more prevalent in breast cancer and other malignancies. A long-term solution has not been possible with traditional oncology techniques. The use of targeted therapies has proven crucial in managing the intricacy and resistance linked to breast cancer. But as genomic technology has advanced, our knowledge of the genetic makeup of breast cancer has changed, creating new opportunities for the development of more effective anti-cancer treatments [52]. Studies utilizing genomic technologies have identified key driver genes for breast cancer, including TP53, PIK3CA, MYC, PTEN, and BRCA1/2. In the context of TNBC, particular emphasis has been placed on mutations in BRCA1/2 and other significant genes. These findings underscore the genetic diversity and complexity inherent in TNBC. However, there are notable limitations in current genomic approaches, including difficulties in interpreting the clinical significance of numerous mutations and the necessity for more comprehensive, integrated analyses. Such integrative efforts are essential to fully elucidate the genomic intricacies and heterogeneity of breast cancer [50].

2.2.2 Multi-Omics Approaches in Breast Cancer Research

High-throughput technologies have rapidly advanced over the past few decades, enabling a variety of genetic investigations at the cellular and tissue levels. Furthermore, the extensive gathering of gene expression data and DNA methylation profiles has been made possible by highly developed genome screening technologies like whole exome sequencing (WES) and whole genome sequencing (WGS) [53] Single-cell technology allows for new scientific insights into cytological characteristics and gene activity at the cellular level [54]. Furthermore, the development of mass spectrometry techniques has allowed for the highly accurate detection of vast quantities of proteins and metabolites. Proteomics technologies are

moving approaching single-cell resolution and can detect nearly all human proteins. Still, a single platform is not enough to uncover a strong correlation with cancer driver mutations or to unravel the intricacy that underlies cancer genomes. As a result, there is a growing endeavour to create data-driven computational and mathematical techniques for analysing high-dimensional datasets that come from various innovative analytical platform [55].

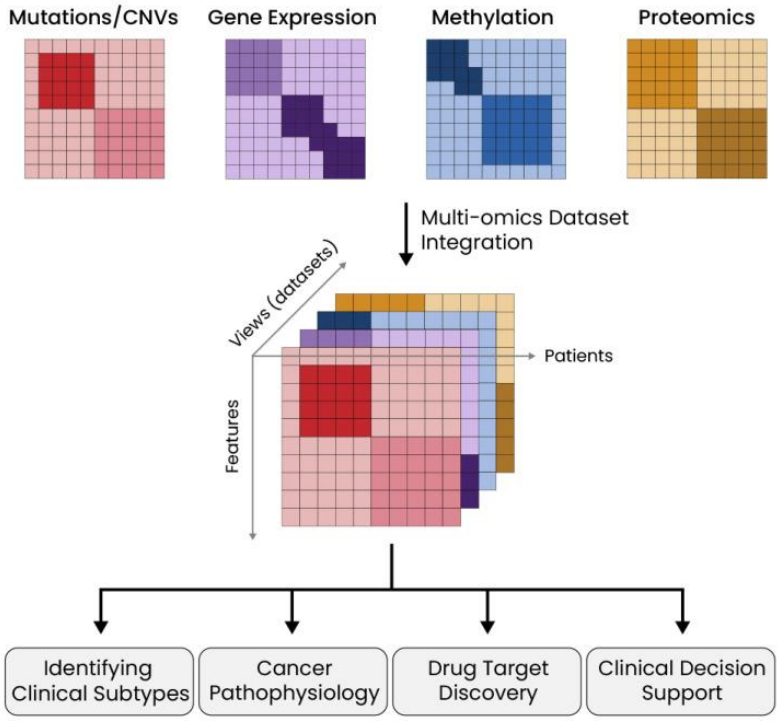

Figure 2.1: The integration of multiple omics datasets—Mutations/CNVs, Gene Expression, Methylation, and Proteomics—into a comprehensive analysis platform.

Multi-omics approaches have been developed to combine various patient-generated omics datasets to identify consistent and maintained genetic or clinical features across multiple datasets (Figure 2.1). Multi-omics research aims to uncover patient subgroups and biological aspects underpinning cancer pathophysiology to overcome the present complications generated by genetic and phenotypic heterogeneity that hamper our understanding of cancer genesis and progression [56].

RNA sequencing (RNA-seq) is a pivotal component of multi-omics approaches in studying TNBC. This technology allows for the comprehensive analysis of the transcriptome, enabling researchers to identify differentially expressed genes, alternative splicing events, and novel transcripts that may play critical roles in TNBC development and progression. In TNBC studies, RNA-seq has been employed to compare the gene expression profiles of TNBC tumors

with adjacent normal tissues and other breast cancer subtypes, such as estrogen receptor-positive (ER+) and HER2-positive (HER2+) tumors. This comparison helps to pinpoint TNBC-specific genes and pathways that are significantly dysregulated, providing insights into the unique molecular characteristics of TNBC. For instance, RNA-seq data integration from multiple cohorts, including The Cancer Genome Atlas (TCGA), has led to the identification of key genes involved in mammary gland morphogenesis and hormone-related pathways, which are crucial for understanding the aggressive nature of TNBC and its poor prognosis [57], [58], [59], [60]. Despite its powerful capabilities, RNA sequencing in TNBC research faces several limitations [61]. One of the primary challenges is the inherent complexity and heterogeneity of TNBC tumors, which can lead to variability in gene expression profiles and complicate the interpretation of RNA-seq data [62]. Additionally, the high cost and technical demands of RNA-seq limit its widespread application, particularly in large-scale clinical studies [63]. Another significant limitation is the potential for technical biases during sample preparation, sequencing, and data analysis, which can affect the accuracy and reproducibility of the results [64].

Advancements in single-cell analysis have greatly improved our understanding of TNBC, offering solutions to the limitations of traditional RNA sequencing. This technique involves isolating individual cells from a tumor sample and examining their gene expression. This allows scientists to identify different cell types within the tumor, such as cancer stem cells, immune cells, and stromal cells, all of which play unique roles in tumor growth and progression. By mapping these cellular landscapes, researchers can pinpoint specific cell populations that drive malignancy and resistance to treatment. Studies have shown that scRNA-seq can reveal the detailed structure of tumor subclones, track changes during disease progression, and uncover new biomarkers for targeted therapies. However, despite its significant impact, single-cell analysis in TNBC has limitations like high costs, extensive computational needs, and technical challenges related to separating and integrating data from individual cells. Additionally, while scRNA-seq is excellent at capturing the diversity of gene expression, it may miss important spatial context and protein-level variations essential for a complete tumor profile. Overcoming these challenges will require continued innovation and collaborative efforts to fully harness single-cell technologies in the fight against TNBC [62], [63], [64].

Spatial transcriptomics is an advanced molecular technique that allows the mapping of gene expression in tissue sections while preserving the spatial context [65] . This approach overcomes the limitations of traditional bulk and single-cell RNA sequencing by maintaining the physical location of each cell within the tissue architecture. By integrating spatial information with transcriptomic data, spatial transcriptomics provides a comprehensive understanding of the cellular microenvironment, cell-to-cell interactions, and the spatial organization of gene expression patterns within complex tissues [66]. This technique is particularly valuable in cancer research, where the spatial heterogeneity of tumors plays a critical role in disease progression, metastasis, and therapeutic response. In TNBC, spatial transcriptomics has provided pivotal insights into the tumor's complex cellular landscape and its interaction with the surrounding microenvironment. TNBC is characterized by a lack of hormone receptors and HER2 expression, making it more challenging to treat with conventional therapies. Spatial transcriptomics has revealed the presence of distinct cellular niches within TNBC, including areas of high immune infiltration and regions dominated by cancer stem-like cells [67].

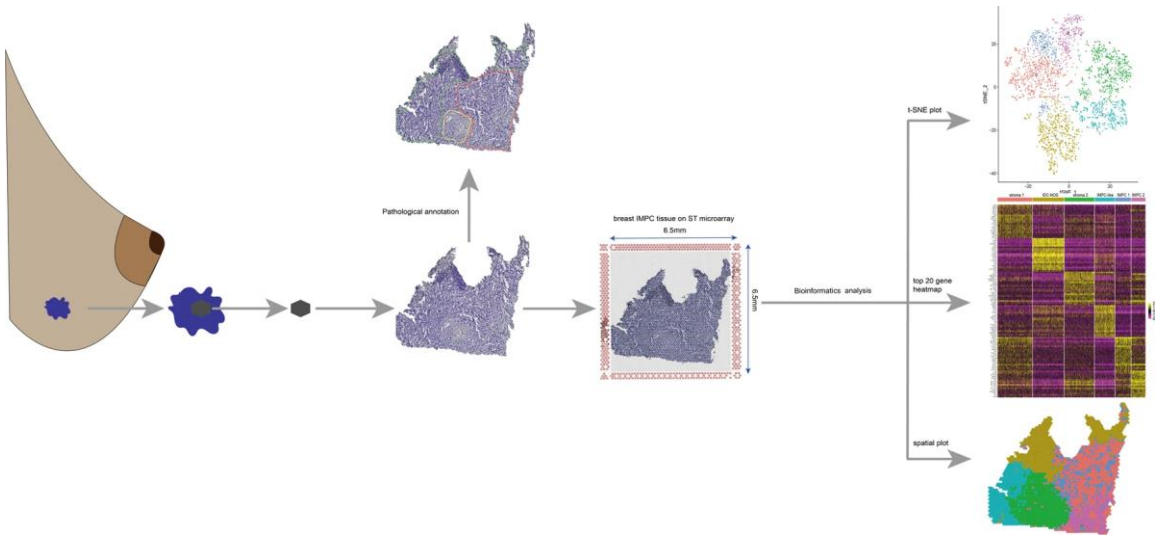

Figure 2.2: Outline of the spatial transcriptomics workflow, beginning with tissue sampling and pathological annotation, followed by spatial mapping and sequencing, and concluding with bioinformatics analysis for cellular composition and gene expression profiling.

The image (Figure 2.2) illustrates the workflow of spatial transcriptomics applied to breast cancer research. It begins with the collection and sectioning of a breast cancer tissue sample. Pathologists then annotate the tissue sections to identify regions of interest. The annotated tissue is placed onto a specialized spatial transcriptomics microarray, which captures

spatial gene expression data. This data undergoes bioinformatics analysis, generating visual representations such as t-SNE plots, heatmaps of the top genes, and spatial plots that map gene expression back onto the tissue sections. This process provides a detailed view of the spatial distribution of gene expression within the tumor, revealing its molecular landscape and heterogeneity [68]. These insights have highlighted potential therapeutic targets and biomarkers specific to certain tumor regions. Furthermore, spatial transcriptomics has been used to study the spatial dynamics of immune evasion mechanisms in TNBC, offering new avenues for immunotherapy [69]. Spatial transcriptomics is a relatively new and emerging technique in the field of molecular biology, and consequently, there is still a paucity of research utilizing this method. Previous research on prognostic markers for TNBC has predominantly focused on bulk or single-cell analyses, neglecting the crucial spatial distribution of gene expression within tumor tissues. This significant research gap fails to account for the inherent heterogeneity of gene expression, which is vital for accurate prognostication. As a result, current prognostic markers lack the necessary consistency across different regions of the tumor, leading to potential unreliability in clinical applications. To address these limitations, further studies and explorations employing spatial transcriptomics are imperative. This approach can provide a more comprehensive understanding of TNBC's complex cellular landscape, uncovering spatial patterns of gene expression that are essential for developing reliable prognostic markers and effective targeted therapies.

2.3 Artificial Intelligence and Machine Learning in Breast Cancer Research

Artificial Intelligence (AI) is transforming various fields, from finance and transportation to education and healthcare. Its ability to analyze large datasets, recognize patterns, and make predictions has made it an invaluable tool in many domains [70]. In biomedical research, AI has brought significant advancements, helping scientists understand complex biological systems and discover new treatments. AI techniques, such as machine learning and deep learning, are used to analyze genetic data, identify disease markers, and develop personalized medicine approaches. These technologies have accelerated drug discovery, improved diagnostic accuracy, and optimized clinical trials, making biomedical research more efficient and effective. In cancer research, AI has been particularly impactful. By analyzing vast amounts of genomic, transcriptomic, and proteomic data, AI can uncover patterns and correlations that traditional methods might lack. This has led to better understanding of cancer progression, metastasis, and treatment resistance [71]. TNBC, a

challenging subtype lacking hormone receptors and HER2 expression, AI has been crucial. AI-powered imaging techniques enhance early diagnosis, while machine learning models predict disease outcomes and response to treatments [72]. Additionally, AI-driven analyses of gene expression and mutation data have identified new biomarkers and therapeutic targets, offering hope for more effective and personalized treatments for TNBC [73].

2.3.1 Machine Learning Models for Cancer Classification and Prediction

Machine learning models are transforming many aspects of our lives, from everyday tasks like personalized movie recommendations to complex ones like autonomous vehicles navigating city streets. These models excel at analyzing large amounts of data, finding patterns, and making accurate predictions, proving their versatility in many real-time applications.

2.3.1.1 Classification Models

Classification models are essential in machine learning for sorting data into specific categories. In single-cell and spatial transcriptomics, these models help classify different cell types based on their gene expression profiles. Common techniques include support vector machines (SVMs) and decision trees. SVMs find the best boundary that separates different classes in high-dimensional space [74], while decision trees split the data into branches to make decisions based on certain features. These models are crucial for identifying cell states and subpopulations within tissues, helping to understand cellular diversity and disease mechanisms [75].

2.3.1.2 Prediction Models

Prediction models are used to forecast future outcomes based on past data. In cancer research, these models can predict disease progression, treatment responses, and patient survival. Techniques like regression analysis and ensemble methods, such as random forests, are used. Regression models predict continuous outcomes, such as the expression levels of specific genes, while ensemble methods combine multiple models to improve prediction accuracy and reliability. These prediction models integrate genetic, proteomic, and clinical data to provide comprehensive insights into disease dynamics, aiding personalized medicine approaches [76], [77] .

2.3.1.3 Neural Networks

Neural networks, especially deep learning models, have revolutionized the analysis of complex biological data. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are widely used in spatial transcriptomics. CNNs are effective for image-based data, capturing spatial details in tissue samples to identify gene expression patterns. RNNs are suitable for sequential data and can capture changes in gene expression over time. Deep learning models handle high-dimensional data well, automatically learning important features without manual input. This makes them invaluable for tasks like cell type identification, spatial pattern recognition, and integrating multi-omics data [78][79].

2.3.1.4 Integration and Impact

By combining classification models, prediction models, and neural networks, researchers can analyze single-cell and spatial transcriptomics data with great precision. These machine learning techniques provide a complete view of the tumor environment, uncovering new biomarkers and therapeutic targets. This integrated approach enhances the understanding of tumor diversity and improves the precision of personalized medicine. The growing capabilities of AI and machine learning continue to expand what is possible in spatial transcriptomics, offering new insights into diseases like breast cancer and paving the way for innovative treatments.

2.4 Integration of Spatial Transcriptomics and Artificial Intelligence

The integration of spatial transcriptomics and artificial intelligence (AI) represents a groundbreaking advancement in biomedical research. By combining high-resolution spatial gene expression data with powerful AI algorithms, researchers can analyze and interpret complex biological patterns with unprecedented precision.

2.4.1 Current Approaches and Solution

A very recent study applied spatial transcriptomics to investigate the heterogeneity of triple-negative breast cancer (TNBC) tumors. Utilizing a spatial transcriptomics platform, researchers captured spatially resolved gene expression patterns within tumor tissues, complemented by digital pathology tools for histomorphological analysis. Open-source libraries and proprietary software facilitated image processing and feature extraction. The study employed clustering algorithms to identify distinct TNBC ecotypes based on gene expression profiles, which were validated using external datasets. Predictive models incorporating supervised learning algorithms were developed to correlate spatial and molecular features with clinical outcomes. The results delineated multiple TNBC ecotypes, each associated with specific clinical outcomes and therapeutic responses, underscoring the heterogeneity of TNBC and its implications for personalized treatment strategies. This integration of spatial transcriptomics and machine learning offers promising avenues for enhancing precision medicine in cancer care [80].

Another study leveraged digital image analysis and machine learning to predict the response to neoadjuvant chemotherapy (NAC) in triple-negative breast cancer (TNBC) patients. The researchers utilized whole-slide imaging (WSI) combined with machine learning techniques to classify histological components of the tumor microenvironment (TME). They processed and analyzed the images using software such as QuPath for digital pathology and Python libraries including OpenCV and scikit-image for image preprocessing. Feature extraction was performed using techniques like gray-level co-occurrence matrix (GLCM), Gabor filters, and local binary patterns (LBP). The machine learning models used for classification included linear support vector machine (SVM), radial basis function SVM (rbfSVM), and ensemble tree methods implemented with the RUSBoost algorithm. The study employed a stratified eightfold cross-validation strategy to ensure robust model performance and reliability. The researchers further modeled spatial relationships within the TME using graph-based approaches to characterize tissue states and interactions.

The results revealed that the machine learning models effectively predicted NAC response by analyzing spatial and histological features of the TME. This approach demonstrated the potential of combining digital pathology with advanced machine learning

techniques to improve predictive accuracy and guide personalized treatment strategies in TNBC. This study underscores the value of integrating digital image analysis and machine learning to enhance predictive modeling in cancer treatment [81].

2.4.2 Challenges

Spatial transcriptomics technologies face several significant challenges, particularly regarding resolution and sensitivity. Achieving high spatial resolution while maintaining the sensitivity required for accurate gene expression detection is difficult, which can result in the loss of critical spatial information necessary for mapping gene expression patterns within the tumor microenvironment [82]. Additionally, the data generated by spatial transcriptomics are highly complex, combining high-dimensional gene expression data with spatial coordinates. This complexity necessitates sophisticated computational tools and expertise, making the integration of spatial data with transcriptomic profiles to extract meaningful biological insights a formidable task[83]. The relatively small size of spatial transcriptomics datasets due to the high cost and technical demands also poses a significant challenge. Small datasets can lead to overfitting in machine learning models and reduce the statistical power of findings, complicating the generalization of results across larger populations [84].

Integrating AI and machine learning with spatial transcriptomics adds another layer of complexity. One major challenge is the interpretability of machine learning models, particularly deep learning approaches, which often function as "black boxes" where the decision-making process is not transparent. This lack of interpretability can be a barrier in clinical applications where understanding the rationale behind predictions is crucial for gaining trust and ensuring patient safety [85]. Effective integration also requires combining computational predictions with existing biological knowledge, necessitating interdisciplinary collaboration between data scientists, biologists, and clinicians to ensure that AI models are biologically plausible and clinically relevant. Moreover, integration demands substantial computational resources and infrastructure, posing a barrier for many research institutions. Standardizing data formats and analysis pipelines to facilitate data sharing and reproducibility across different studies and platforms is another significant challenge, as variability in data processing methods can lead to inconsistent results [86].

2.4.3 Further Direction and Innovation

The future of integrating spatial transcriptomics and AI in TNBC research is set to transform personalized medicine and treatment strategies. One key innovation is combining spatial transcriptomics with other types of data, like proteomics and metabolomics, to gain a comprehensive understanding of the tumor environment and molecular interactions. Advanced AI models, especially those using deep learning and graph-based methods, will enhance the resolution and interpretation of spatial data, aiding in the precise identification of tumor subtypes and treatment targets. Improvements in data standardization and sharing will facilitate collaboration, leading to larger, more diverse datasets that improve the reliability of AI models. Innovations in single-cell analysis and spatially resolved transcriptomics will further refine our ability to map cellular diversity and dynamics within TNBC tumors, paving the way for targeted treatments and early diagnosis. By addressing current challenges and utilizing these advancements, researchers aim to significantly enhance TNBC patient outcomes through more personalized and effective therapies.

CHAPTER 3: METHODOLOGY

In this chapter, a detailed methodology was employed to integrate spatial transcriptomics with artificial intelligence for breast cancer analysis. The study began with data acquisition, including spatial transcriptomics, imaging, and clinical data. Single-cell analysis was then performed to explore gene expression patterns, followed by differential expression analysis. Machine learning models were developed to predict patient outcomes. This comprehensive approach yielded profound insights into the complex landscape of tumor heterogeneity and the cellular patterns that contribute to breast cancer progression. Furthermore, the analysis identified potential therapeutic targets, offering valuable directions for personalized treatment strategies.

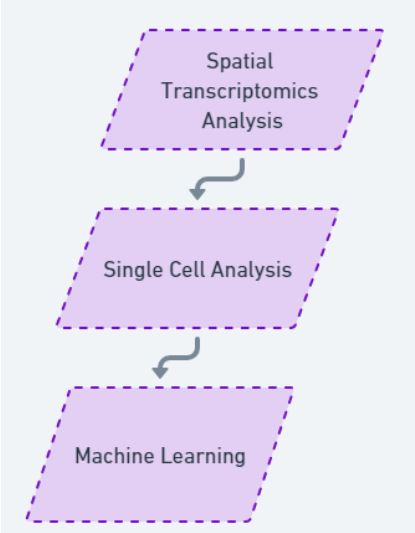

Figure3.1: General workflow from spatial transcriptomics to machine learning

3.1 Data Description and Acquisition

The data of TNBC for spatial transcriptomics analysis was retrieved from GEO datasets under the accession number GSE210616. This dataset consisted of 22 patients out of which 15 were African-American and 7 who were Caucasian. 2 sections from each patient were taken (except for patient 19) which made them 43 samples. 28 tissue sections representing 14 primary TNBC tumors were subject to spatial transcriptomics using the 10x Genomics Visium platform, followed by high throughput sequencing.

Table 3.1: Dataset for spatial transcriptomics analysis

Database	Gene Expression Omnibus
Platform	10X Genomics Visium
Accession Number	GSE210616
Cancer Type	Triple Negative
No. of Patients	22
Tissue Section per Patient	2
No. of Samples	43
Ethnicity	African-American
Country	USA

For the single-cell analysis of normal breast tissue, Data was sourced from two primary databases. First, data was retrieved from the Genotype Tissue Expression (GTEx) database, focusing on breast tissue samples from three individuals without breast cancer.

Figure 3.2: Dataset for single cell analysis of healthy individuals

Second, data was accessed from the GEO database under accession number GSE161529. This dataset included various types of breast cancer as well as samples from healthy individuals without breast cancer. From this dataset, 13 normal breast tissue samples were specifically selected for analysis.

3.2 Spatial Transcriptomics Analysis

For the spatial transcriptomics analysis, the process began by downloading the processed files, which included both the normalized gene expression data and the corresponding image files with mapped spatial spots. These files provided a comprehensive dataset that integrated spatial information with gene expression profiles, essential for understanding the spatial organization of the tissue.

3.2.1 Data Collection and Processing:

For the spatial transcriptomics data, a tool called the ***Loupe Browser***, provided by 10x Genomics, was used. Processed data files for each sample were first downloaded in a format called cloupe, which is compatible with the Loupe Browser. These files contained important information, such as the locations of different spots on the tissue images, clusters of these spots, and the associated gene expression data.

3.2.2 Visualization

Each sample was opened in the Loupe Browser to view the tissue images with overlaid spots. These spots represent specific regions of the tissue that were analyzed during sequencing. The browser also allowed us to see clusters of these spots using UMAP projections, which helped us understand the differences in gene expression across different regions of the tissue.

3.2.3 Extracting Gene Expression Data

For each sample, the gene expression data linked to the spatial spots was downloaded. This data was saved in CSV files, which were later used for further analysis. By doing this, specific regions of the tissue were connected with their gene expression profiles, facilitating easier analysis and interpretation of the biological information.

3.2.4 Differentially Expressed Genes (DEG) Analysis

For the analysis of differentially expressed genes (DEGs), Python was used within the Visual Studio Code (VS Code) environment. To begin, the data was organized by creating a

dedicated directory named ***gene_expression_data,*** where all 43 CSV files containing the gene expression data for each sample were stored.

Next, differential expression analysis was performed using a Python script that iterated through each of the 43 files. The analysis identified genes that were differentially expressed between conditions, applying a p-value threshold of less than 0.05 and a log2 fold change threshold of greater than 1 (or less than -1) to determine significance. Only genes meeting this criterion were considered differentially expressed, ensuring that the results were both statistically and biologically meaningful and reliable. After identifying these DEGs, the results were saved into new CSV files, which were stored in a separate directory called ***DEG_results***. Each file was named according to its corresponding sample, making it easy to track and reference the results.

3.2.5 Selection and Consolidation of Top DEG's

Following the initial differential expression analysis, the next step was to identify the most significant genes across all clusters and samples. Utilizing the DEG_results directory, the top 20 differentially expressed genes from each cluster within each sample were selected. For these selected genes, key metrics including their log2 fold change values, p-values, and adjusted p-values were extracted, which are essential for understanding both the significance and the extent of gene expression changes.

To accomplish this, a Python script was employed that systematically processed each file within the DEG_results directory. The script first filtered the genes based on a p-value threshold of less than 0.05, ensuring that only statistically significant genes were considered. It then sorted the remaining genes by their adjusted p-values, prioritizing those with the smallest values, which typically indicate higher statistical significance after correcting for multiple comparisons. From this sorted list, the top 20 genes were selected for each cluster within each sample and saved in a file.

3.2.6 Identification of Common DEG's Across Samples

After compiling the significant genes from each sample into a file, the focus shifted to identifying the most consistently significant genes across all samples. The top 20 genes that were common across all samples were identified by selecting those that consistently appeared and demonstrated significant differential expression. This resulted in a file containing the top

20 common genes identified across all samples. These 20 common genes, highlighted for their consistency and significance, were recognized as key biomarkers or potential therapeutic targets, offering valuable insights for further biological interpretation and exploration.

3.3 Single Cell Analysis

Single-cell RNA sequencing (scRNA-seq) data was obtained from the Gene Expression Omnibus (GEO) for 13 healthy individual's samples and from the Genotype-Tissue Expression (GTEx) project for 3 normal samples. The data included **features**, **barcodes**, and **matrix** files, which were used to construct the gene expression matrices necessary for downstream analysis. The **Scanpy** library was used for the analysis of single cell.

3.3.1 Preprocessing

The data was loaded into an **AnnData** object, which is a standard format for handling large single-cell datasets. For each dataset, we wrote a function to read the matrix, genes, and barcodes files, and subsequently construct the AnnData object. Specifically, the matrix file was loaded and transposed to ensure that the rows corresponded to cells and the columns to genes. The genes and barcodes were then assigned as variable names **(var_names)** and observation names **(obs_names)** in the AnnData object, respectively. This process was repeated separately for the GEO and GTEx samples, resulting in two distinct AnnData objects: one for the GEO samples and one for the GTEx samples.

Once both datasets were loaded, they were concatenated into a single AnnData object to facilitate joint analysis. During this concatenation, a new key was introduced to distinguish between the two datasets, enabling subsequent analyses to account for potential batch effects. To make sure that the variable (feature) names are unique, Scanpy uses the ***"adata.var_names_make_unique()"*** function. If required, it updates the variable names to make them unique after determining whether there are any duplicates.

3.3.2 Quality Control

After loading the single-cell RNA sequencing (scRNA-seq) data, a quality control (QC) process was performed to ensure the integrity and reliability of the data. This step is crucial for filtering out low-quality cells that could skew the results of downstream analyses.

Mitochondrial genes, identified by their **"MT-"** prefix, were analyzed to calculate their expression as a percentage of total cellular RNA, with high levels indicating potentially stressed or apoptotic cells suitable for exclusion. Ribosomal genes, marked by **"RPS"** and **"RPL"** prefixes, were quantified to assess the translational state of cells, as abnormal expression might signal cellular stress or technical artifacts. Additionally, hemoglobin genes were identified by specific naming patterns to detect possible contamination from blood cells, with elevated expression in non-erythroid cells flagged for further investigation.

In this analysis, we utilized the **`sc.pp.calculate_qc_metrics`** function to compute quality control metrics for each cell. The **`log1p=True`** parameter was employed to apply a log-transformation to these metrics, which helps stabilize variance and improve the interpretability of the data. The results were directly stored in the `AnnData` object, allowing us to flag cells with abnormal gene expression patterns for further scrutiny.

3.3.3 Visualization of Quality Control Metrics

After calculating the quality control (QC) metrics, visualization of these metrics was done to gain insights into the distribution and variability of the data across all cells. Visualization plays a crucial role in the QC process as it allows us to identify potential outliers and anomalies that might indicate issues such as low-quality cells, doublets, or other technical artifacts. Specifically, the focused was on three key metrics: the number of genes expressed per cell **(`n_genes_by_counts`),** the total counts per cell **(`total_counts`)**, and the percentage of counts in mitochondrial genes **(`pct_counts_mt`).** Each of these metrics provides critical information about cell quality.

To visualize these QC metrics, violin plots using the **`sc.pl.violin`** function from the Scanpy library were employed. Violin plots were chosen because they effectively display both the distribution and density of the data, making it easier to spot trends and outliers.

A scatter plot using **'sc.pl.scatter'** was made to examine the relationship between total RNA counts per cell (`total_counts`) and the number of genes expressed per cell (`n_genes_by_counts`). This plot helped identify potential doublets and low-quality cells. Additionally, we colored the points based on the percentage of mitochondrial gene expression **(`pct_counts_mt`),** allowing us to visually assess cells with high mitochondrial content.

3.3.4 Filtering Cells and Genes

As part of quality control workflow, filtering criteria was applied to both cells and genes to enhance the overall quality of our dataset. The first step was to filter out cells with fewer than 100 detected genes **(min_genes=100)**, a step designed to exclude low-quality cells. Following this, the genes that were not expressed in at least 3 cells **(min_cells=3)** were filtered out**,** thereby removing genes with sparse expression that are likely to contribute more noise than meaningful biological insight.

3.3.5 Normalization

After filtering cells and genes to ensure that our dataset included only high-quality data points, with the crucial step was of normalization. Normalization was necessary to account for variations in sequencing depth across individual cells, which could otherwise introduce technical biases into our analysis. To preserve the original data for potential downstream comparisons, the data was saved in the raw count data in a separate layer of the AnnData object.

Following this, a logarithmic transformation to the scaled data was applied using the log1p method, which adds one to each expression value before applying the logarithms. The normalization and log transformation were carried out using the Scanpy library's **sc.pp.normalize_total** and **sc.pp.log1p** functions, respectively.

3.3.6 Identification of Highly Variable Genes

Highly variable gene (HVGs) discovery using single-cell RNA sequencing (scRNA-seq) enables the identification of genes that significantly influence cell-to-cell variation in a homogeneous cell population, like a population of embryonic stem. This step deals with the top 2000 genes that were present in each sample and performs further analysis using the function ***sc.pp.highly_variable_genes()***. To display the mean relationship and gene expression dispersion, an illustration was created for each sample to highlight the genes that demonstrated greater variability among cells.

3.3.7 Dimensionality Reduction

After highly variable genes were identified, PCA was first applied to the normalized data using the ***`sc.tl.pca`*** function from the Scanpy library. This function computes the principal components and orders them according to the amount of variance they explain. To determine

how many PCs to retain for further analysis, the variance explained by each component was examined using the ***`sc.pl.pca_variance_ratio`*** function. This plot helps in deciding the appropriate number of PCs to use in downstream analysis, ensuring that enough components are included to capture the essential structure of the data while avoiding overfitting by including too many.

3.3.8 Nearest neighbor graph construction and visualization

Following dimensionality reduction, the next step was to construct the nearest neighbor graph, which is crucial for understanding the relationships between cells in the dataset. In the analysis, the ***`sc.pp.neighbors`*** function from the Scanpy library was utilized to construct the neighborhood graph. To visualize the high-dimensional data in a more interpretable form, Uniform Manifold Approximation and Projection (UMAP) was then applied using the ***`sc.tl.umap`*** function. UMAP is a widely used technique that projects high-dimensional data into a two-dimensional space, preserving the local and global structure of the data as much as possible. The resulting UMAP plot was colored based on specific metadata, such as sample origin, to visually assess the distribution and separation of different cell populations.

3.3.9 Clustering

In the final stage of analysis, the focus was on identifying distinct cell populations within the dataset by performing clustering, which is a key step in single-cell RNA sequencing (scRNA-seq) analysis. Leiden clustering was implemented using the ***`sc.tl.leiden`*** function from the Scanpy library. This function operates on the nearest neighbor graph constructed in the previous step, grouping cells into clusters based on their local neighborhood structure. The ***"igraph"*** implementation was specified, and the number of iterations was set to 2 to ensure a robust clustering solution. After clustering, the results were visualized using UMAP, with each cell colored according to its assigned cluster.

3.3.10 Differential Gene Expression Analysis

Following the clustering step, the focus shifted to identifying differentially expressed genes (DEGs) within each cluster. To accomplish this, differential expression analysis was performed using the Wilcoxon rank-sum test, a non-parametric test that compares gene expression between each cluster and all other clusters in the dataset.

This analysis was implemented using the ***`sc.tl.rank_genes_groups`*** function from the Scanpy library. This function ranks genes by their differential expression across clusters, aiding in the identification of the most distinct and biologically relevant markers for each group. By specifying the clustering result ***(`leiden_res_0.50`)*** as the grouping variable, DEGs were identified in the context of the previously determined clusters. The Wilcoxon test was chosen for its robustness in handling the typically non-normal distribution of gene expression data in single-cell RNA sequencing.

3.3.11 Identification and Cross-Analysis of Common Genes Between Spatial Transcriptomics and Single-Cell Data

From the spatial transcriptomics analysis, the top 20 genes that were consistently expressed across all samples were identified. To further explore the significance of these genes, their expression was analyzed within the differentially expressed gene set from the single-cell RNA sequencing (scRNA-seq) data.

3.4 Machine Learning

In this study, machine learning was applied to integrate spatial transcriptomics and single-cell RNA sequencing data by identifying key differentially expressed genes (DEGs) associated with breast cancer progression. Spatial transcriptomics data provided spatially resolved gene expression profiles, while single-cell RNA-seq data offered high-resolution insights into individual cellular phenotypes within the tumor microenvironment. Various machine learning models, including Xtreme Gradient Boosting and Support Vector Machines, were trained using DEGs and spatial features to predict breast cancer outcomes. The models were evaluated using accuracy, AUC-ROC, and confusion matrix, revealing critical genes and spatial patterns linked to disease progression.

3.4.1 Normal vs Cancer Classification Model

For the first model, classification was performed to differentiate between healthy individuals and cancerous patients.

3.4.1.1: Data Preparation

The initial step in the process involved preparing the data. For the cancer samples, the gene expression data from the Top_20_Common_Genes.csv file, derived

from the spatial transcriptomics analysis, was utilized. For the normal samples, data from the ***common_genes_expression_scRNAseq.csv*** file, obtained from the single-cell RNA sequencing analysis, was incorporated. The gene expression data from both files was then combined into a single comprehensive dataset. This combined dataset was saved in a file named ***Combined_Genes_Data_Normal_vs_Cancer.csv***, which was subsequently used for training the classification model.

3.4.1.2: Data Splitting

After compiling the gene expression data into the Combined_Genes_Data_Normal_vs_Cancer.csv file, the next step involved splitting the dataset into training and testing subsets. This split was performed to ensure that the model could be trained on one portion of the data while being evaluated on a separate, unseen portion. Typically, 80% of the data was allocated for training, while the remaining 20% was reserved for testing. This approach helped to assess the model's ability to generalize to new, unseen data.

3.4.1.3 Feature Selection:

Before training the classification model, it was crucial to select the most relevant features (gene expression values) to ensure they contributed effectively to the classification process. Feature selection involved choosing the most informative genes, particularly those with significant Log2 fold changes, to serve as predictors in the model. This step was essential for improving the model's performance by focusing on the most critical variables and reducing the complexity of the data.

In addition to selecting the features, a new column named Type was added to the dataset to label each sample as either "normal" or "cancer." This column served as the target variable **(y)** for the model. The selected features, represented by their gene names and corresponding Log2 fold change values, were used as the predictor variables **(x).**

3.4.1.4 Model Training

Following feature selection and data preparation, the Xtreme Gradient Boosting (XGBoost) classification model was employed to train on the cancer and normal gene expression data. XGBoost was chosen for its effectiveness in handling complex datasets, particularly in the context of distinguishing between cancerous and non-cancerous samples. The model was configured with key hyperparameters to optimize performance: **n_estimators=100** determined the number of boosting rounds, **learning_rate=0.1** controlled the step size to prevent overfitting, **max_depth=5** set the maximum depth of each tree to capture feature interactions, subsample=0.8 and colsample_bytree=0.8 introduced randomness to reduce overfitting, and objective='**binary:logistic**' specified the binary classification nature of the task.

3.4.1.5 Cross Validation

For the XGBoost model, a cross-validation process combined with grid search was applied to fine-tune the model and ensure its effectiveness in predicting cancer stages. The data was split into five folds, with the model being trained on four folds and validated on the fifth. This 5-fold cross-validation process was repeated across 16 different candidate configurations, leading to a total of 80 fits.

3.4.1.6 Model Evaluation

After training, the model's performance was evaluated using the testing subset. Key performance metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC), were calculated to assess the model's effectiveness in correctly classifying the samples.

3.4.2: Stage Prediction Model

3.4.2.1: Data Preparation

For the stage prediction model, the process began by creating a comprehensive metadata file using cancer data. This involved merging two key datasets: the first was

the file containing expression data for the top 20 common genes identified across all cancer patients, and the second was a supplementary file from a published study [87] that used the same dataset. The supplementary file included vital survival analysis data for the cancer patients, such as disease progression and survival times. By merging these two files, a unified metadata file was creasted that combined both molecular and clinical data, forming the basis for training the stage prediction model.

3.4.2.2: Data Splitting

After preparing the metadata file, the next step involved splitting the dataset into training and testing sets to ensure the model's performance could be properly evaluated. The data was typically split into two parts: 80% of the data was allocated for training the model, while the remaining 20% was set aside for testing. This split allowed the model to learn from most of the data while being tested on an unseen portion, providing a reliable measure of its ability to generalize to new, unseen data.

3.4.2.3: Feature Selection

For the feature selection process in the stage prediction model, a combination of molecular and clinical features was carefully selected to optimize the model's predictive power. The features included gene names and their corresponding Log2 fold change values, which provided insights into the differential expression of genes. In addition to these molecular features, several key clinical variables were incorporated: days to recurrence, days to death, days to maximum encounter, relapse-free survival, and overall survival days. These clinical metrics were chosen for their relevance in capturing the progression and outcomes of cancer in patients. The target variable (y) in our model was the cancer stage, which the model aimed to predict based on the selected features.

3.4.2.4: Model Training

Following feature selection, the stage prediction model was trained using a Support Vector Machine (SVM) classifier. The SVM model was implemented within a pipeline that included two key components: a **StandardScaler** for feature scaling and the **SVC (Support Vector Classification)** algorithm. The StandardScaler was applied

first to standardize the feature data, ensuring that all features had a mean of zero and a standard deviation of one. This step was crucial for the SVM, as it is sensitive to the scale of input features.

The SVM classifier was configured with a **radial basis function (RBF)** kernel, which is effective for capturing non-linear relationships in the data. Hyperparameter tuning was performed to optimize the model's performance. The parameters tuned included the regularization parameter C, which controls the trade-off between maximizing the margin and minimizing classification errors, with values tested at **0.1, 1, 10, and 100**. Additionally, the gamma parameter, which defines the influence of a single training example, was tuned with values of 1, 0.1, 0.01, and 0.001. These hyperparameters were explored in combination to identify the best configuration for the SVM model.

3.4.2.5: Cross Validation

After the initial model training, cross-validation combined with grid search was performed to fine-tune the stage prediction model and ensure its robustness. Specifically, for the SVM model, we conducted a 5-fold cross-validation, where the data was split into five folds. The model was trained on four folds and validated on the fifth, with this process repeated five times so that each fold served as the validation set once. We explored 12 different candidate configurations for the SVM model, resulting in a total of 60 fits.

3.4.2.6: Model Evaluation

After training, the model’s performance was evaluated using the testing subset. Key performance metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC), were calculated to assess the model’s effectiveness in correctly classifying the samples.

3.4.3 Prognosis Prediction Model

3.4.2.1 Data Preparation

For the prognosis prediction model, the same metadata file used in the stage prediction model was utilized. A new column named ***Prognosis*** was introduced to categorize the patients based on their survival time. Patients with a survival time of 0 to 5 years were labeled as 0, indicating a poor prognosis. Those with a survival time of 5 to 10 years were categorized as 1, indicating a poor prognosis. Finally, patients with a survival time of over 10 years were labeled as 2, indicating a good prognosis. This categorization allowed us to stratify the patients based on their long-term outcomes, providing a foundation for the prognosis prediction model.

3.4.2.2: Data Splitting

After preparing the metadata file, the next step involved splitting the dataset into training and testing sets to ensure the model's performance could be properly evaluated. The data was typically split into two parts: 80% of the data was allocated for training the model, while the remaining 20% was set aside for testing. This split allowed the model to learn from most of the data while being tested on an unseen portion, providing a reliable measure of its ability to generalize to new, unseen data.

3.4.2.3: Feature Selection:

For the feature selection process in the prognosis prediction model, a combination of molecular and clinical features was carefully selected to optimize the model's predictive power. The features included gene names and their corresponding Log2 fold change values, which provided insights into the differential expression of genes. In addition to these molecular features, several key clinical variables were incorporated: days to recurrence, days to death, days to maximum encounter, relapse-free survival, stage, and overall survival days. These clinical metrics were chosen for their relevance in capturing the progression and outcomes of cancer in patients. The target variable (y) in our model was the cancer progression, which the model aimed to predict based on the selected features.

3.4.2.4: Model Training

For training the prognosis prediction model, the **Support Vector Regression** (SVR) technique was applied from the SVM family. The SVR model was implemented within a pipeline that included two main components: a StandardScaler for feature scaling and the SVR algorithm as the regressor. The StandardScaler was applied to standardize the feature data, ensuring that all features had a consistent scale, which is crucial for the performance of SVR.

3.4.2.5: Cross Validation

After the initial model training, cross-validation combined with grid search was performed to fine-tune the stage prediction model and ensure its robustness. Specifically, for the Support Vector Regression (SVR) model, a 5-fold cross-validation was conducted, where the data was split into five folds. The model was trained on four folds and validated on the fifth, with this process repeated five times so that each fold served as the validation set once. We explored 12 different candidate configurations for the SVRmodel, resulting in a total of 80 fits.

3.4.2.6: Model Evaluation

Following the training of the Support Vector Regression (SVR) model for prognosis prediction, its performance was evaluated using key metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and the R^2 score. MSE and MAE were used to measure the accuracy of the model's predictions by quantifying the average error magnitude, with lower values indicating better performance. The R^2 score assessed how well the model captured the variance in the survival time data, with scores closer to 1 reflecting a strong predictive capability.

In this methodology, spatial transcriptomics was integrated with artificial intelligence to analyze TNBC. The approach began with the acquisition of data from GEO datasets and single-cell RNA sequencing, followed by the application of the 10x Genomics Visium platform to map gene expression across tissue sections. Differential expression analysis was conducted, and highly variable genes were identified and visualized using advanced techniques like

UMAP. Subsequently, machine learning models, including XGBoost and SVM, were developed to classify cancer stages, predict patient prognosis, and identify key genes associated with cancer progression, offering insights for personalized treatment strategies.

CHAPTER 4: RESULTS AND DISCUSSION

In this chapter, the findings deduced from an extensive evaluation of spatial transcriptomics, scRNA-seq, and machine learning approaches will be discussed. The findings of the study will be analyzed and discussed in reference to the available literature, with the aim of establishing how the current study relates to the existing literature on breast cancer. The results are compared with previous works to emphasize the advancements made using spatially resolved measurements of gene expression and cellular heterogeneity for breast cancer progression. Additionally, future trends of **machine** learning models will be described, with a special focus on their abilities to find biomarkers for prognosis and enhance patients' lives.

4.1 Spatial Transcriptomics Analysis

The processed files were opened using the Loupe Browser, enabling us to visualize the tissue samples with overlaid spots representing regions of gene expression analysis.

4.1.1 Visualization

The images were visualized in the Loupe Browser, where the tissue sections were overlaid with spots corresponding to regions analyzed for gene expression. The browser provided a spatial projection that revealed distinct clusters within these spots, each represented by different colors. These clusters reflect the spatial heterogeneity of gene expression across the tissue sample. The spatial projection image for sample 1 is shown in figure 4.1.

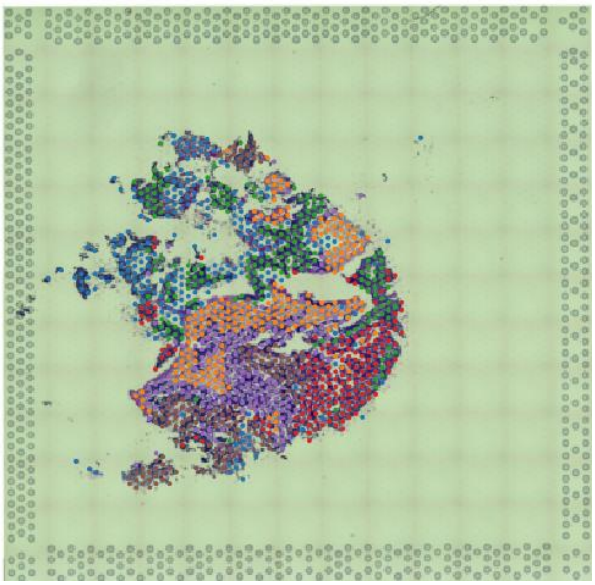

Figure 4.1: Tumor microenvironment clusters representing different cell types.

The clear separation of clusters indicates the presence of different microenvironments within the tissue, each potentially representing various cell types or biological states. This kind of spatial resolution is crucial for understanding the complex architecture of breast cancer tissues, as it helps identify regions with specific gene expression profiles that may be linked to distinct functional roles within the tumor microenvironment [88] each sample, the gene expression data associated with the spatial spots was downloaded and saved in CSV files. This data included the gene names, their corresponding log2 fold-change values, and p-values. After identifying these DEGs, we saved the results into new CSV files, which were stored in a separate directory called ***DEG_results***. Each file was named according to its corresponding sample, making it easy to track and reference the results. Utilizing the DEG_results directory, the top 20 differentially expressed genes from each cluster within each sample were selected. From this sorted list, the top 20 genes were selected for each cluster within each sample. Finally, all the selected genes from each sample were compiled into a single comprehensive file named ***Top_DEGs_Summary.csv.*** The top 20 genes that were common across all samples were then identified, selecting those that consistently appeared and demonstrated significant differential expression. The file containing the top 20 common genes identified across all samples was named ***Top_20_Common_Genes.csv.***

4.1.2 Top 20 common Genes

The 20 commonly identified genes were ***MALAT1, IGKC, ACTB, B2M, EEF1A1, FTH1, FTL, GAPDH, HLA-, MT-ATP6, MT-CO1, MT-CO2, MT-CO3, MT-CYB, MT-ND3, MT-ND4, RPL13, RPL37, RPL41, RPLP1***. These genes represent a mix of housekeeping genes, ribosomal proteins, and mitochondrial genes, which are typically involved in fundamental cellular processes such as protein synthesis, energy production, and immune responses. The consistent expression of these genes across various spatial spots suggests their essential role in maintaining basic cellular functions within the breast cancer tissue microenvironment. For instance, genes like GAPDH and ACTB are well-known housekeeping genes that serve as critical controls in gene expression studies due to their stable expression [89]. The identification of these genes reinforces their role in the fundamental cellular architecture of both malignant and non-malignant cells within the tissue [90].

The significance of mitochondrial genes in breast cancer is further highlighted by studies[91], [92] showing that mutations or dysregulation of these genes can lead to increased reactive oxygen species (ROS) production, which contributes to genetic instability, promoting tumor progression and metastasis. Furthermore, mitochondrial DNA (mtDNA) mutations, particularly in genes like ***MT-CO1*** and ***MT-ATP6***, have been associated with more aggressive forms of breast cancer and poorer patient outcomes [93].

The ribosomal genes identified in common genes, including ***RPL13, RPL37, RPL41,*** and ***RPLP1***, play a significant role in breast cancer progression by enhancing the translational capacity of cancer cells[94]. Ribosomal proteins are not merely structural components of the ribosome but are actively involved in regulating cell proliferation, apoptosis, and stress responses—processes that are often hijacked during oncogenesis[95]. Overexpression of ribosomal genes like ***RPL13*** and ***RPLP1*** has been associated with increased protein synthesis, which supports the rapid growth and survival of tumor cells[96]. Additionally, studies have shown that alterations in these ribosomal proteins can influence oncogenic pathways, such as the PI3K/AKT/mTOR pathway, further promoting tumorigenesis [97].

4.1.3 Marker Genes

In this study, MALAT1 and IGKC acted as significant marker genes which have important functions in breast cancer cells. In breast cancer, MALAT1 is overexpressed and is involved in several oncogenous functions such as leading to the change in the type of splicing

pattern, transcription and cell migration. Research has revealed that MALAT1 can bind to splicing factors including SRSF1 that directly promotes production of pro-tumorigenic splice variants that in sprout cancer cells promote mere survival and invasion [98]. Moreover, growing evidence has shown that MALAT1 mRNA can activate pathways such as mTOR, which is just as well in line with the ability of MALAT1 in enhancing the malignancy of breast cancer cells [99]. Since the upregulation of MALAT1 levels has been linked with poor prognosis, the authors suggest that it can be used as prognostic biomarker and that it can be targeted for treatment of breast cancer [100].

TILs are associated with the immunoglobulin produced by B cells and known as IGKC [101]. IGKC has been identified to be overexpressed in various cancers and its high expression has been used to predict better prognosis in breast cancer presumably because of efficient anti-tumor immunity. Research has shown that there is a positive relationship between IGKC expression and survival rates in woman with breast cancer, especially those forms of the disease in which immune activity plays a key role in treatment and containment [102].

4.2 Single Cell Analysis

For the single-cell analysis, the features, barcodes, and matrix files from both datasets were stored into an **AnnData** variable. This AnnData object contained the key data elements, including the number of observations (cells) and the number of variables (genes). In this section, the focus will be on explaining the results derived from the analysis using one representative sample from the dataset.

4.2.1 Preprocessing

The genes and barcodes were then assigned as variable names ***(var_names)*** and observation names ***(obs_names)*** in the AnnData object, respectively. In case of 1 sample that is named as ***N-253*** the adata is given as

Table 4.1: Number of variables and observation in anndata object

Sample	n_obs	n_var
N_253	4966	33538

4.2.2 Quality Control

After the data was stored in the AnnData object, quality control was performed, during which the mitochondrial genes count was identified for N-253.

Table 4.2: Total counts and pct_counts_mt for N-253

total_counts	pct_counts_mt
3587	8.001115
1628	2.641278
7536	2.295648
10366	3.935944

In the analysis, the `***sc.pp.calculate_qc_metrics***` function was used to compute quality control metrics for each cell. A log-transformation was applied to these metrics by setting the `log1p=True` parameter. The results were stored directly in the `***AnnData***` object, enabling the identification and flagging of cells with abnormal gene expression patterns for further investigation. To visualize the distribution of these metrics, a histogram was plotted for sample N-253.

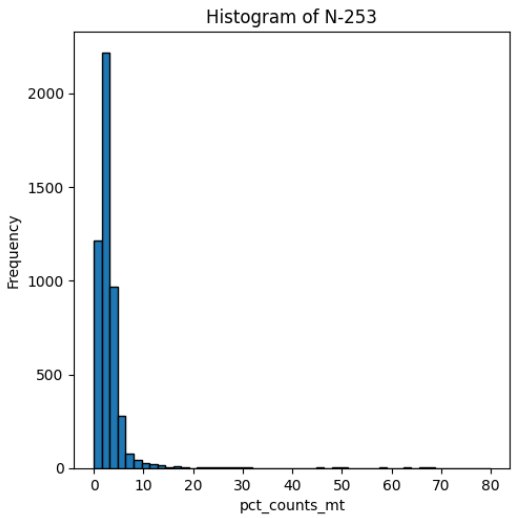

Figure 4.2: Histogram of pct_counts_mt

The histogram for sample N-253 shows a sharp peak at very low pct_counts_mt (around 0-20%) values, indicating that most cells in this sample have a low mitochondrial gene expression, with the frequency rapidly decreasing as the pct_counts_mt increases. The histogram suggests that most cells are likely healthy, with minimal mitochondrial gene expression.

4.2.3 Filtering

As part of the quality control workflow, filtering criteria were applied to both cells and genes to improve the overall quality of the dataset. The filtered genes were visualized using a violin plot. For sample N_253, the violin plot before filtering shows a total of 4,966 cells.

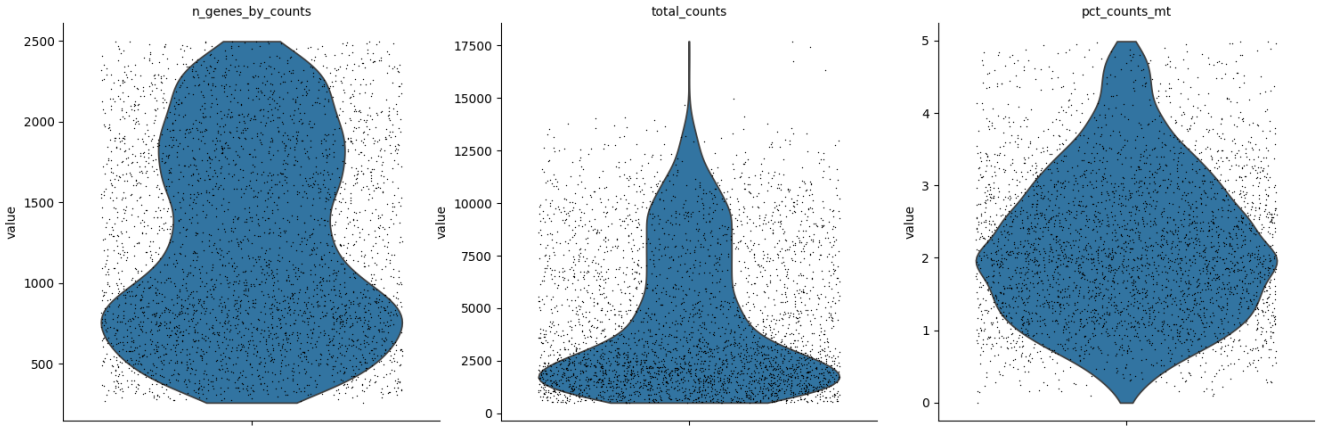

Figure 4.3: Violin Plot before filtering

After filtering, the number of cells was reduced to **3,407**. Violin plots were used to visualize these changes, providing a clear comparison of the cell distributions before and after filtering. This visualization aids in assessing the effectiveness of the filtering process and ensuring that the dataset's quality is improved by removing cells with abnormal gene expression patterns.

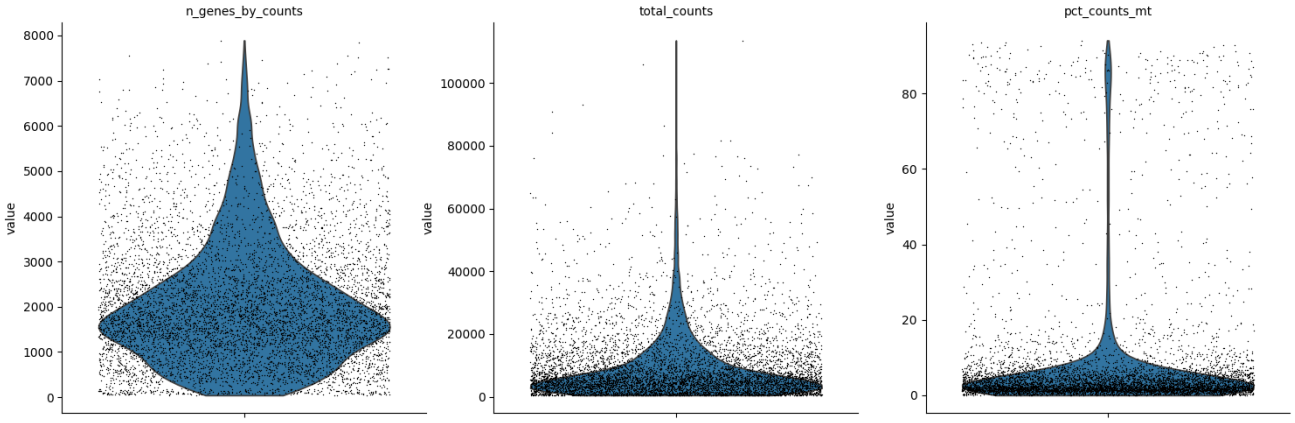

Figure 4.4: Violin Plot after filtering

Scatter plot visualization of sample N-253 was also performed, providing important insights into the quality and distribution of the cells within this sample. The plot revealed that most cells in N-253 clustered within a range of high total gene counts and a substantial number of detected genes, with low levels of mitochondrial gene expression (pct_counts_mt). This pattern suggests that the cells in this sample are predominantly healthy and exhibit high data quality.

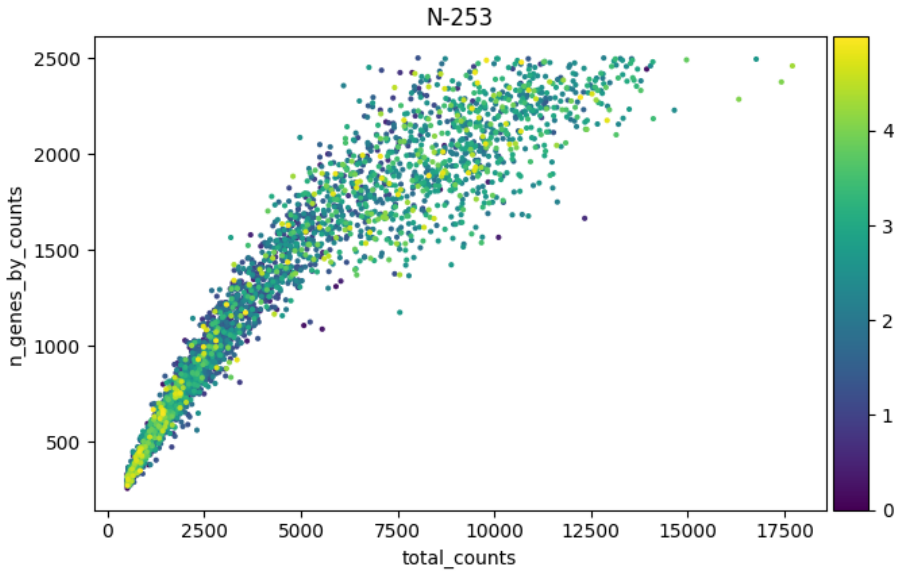

Figure 4.5: Scatter plot after filtering

4.2.4 Normalization

After filtering for high-quality cells and genes, normalization was performed to account for variations in sequencing depth using the ***`sc.pp.normalize_total`*** function, followed by a logarithmic transformation with ***`sc.pp.log1p`.*** To preserve the original data, the raw counts were saved in a separate layer of the AnnData object. The results of these transformations were stored in the `adata` variable for further analysis.

4.2.5 Highly Variable Genes

Highly variable genes with a threshold of the top 2000 genes were identified and then visualized using the ***`sc.pp.highly_variable_genes()`*** function. These genes, which demonstrated greater variability among cells, were then visualized to display the relationship between mean expression and gene dispersion for each sample.

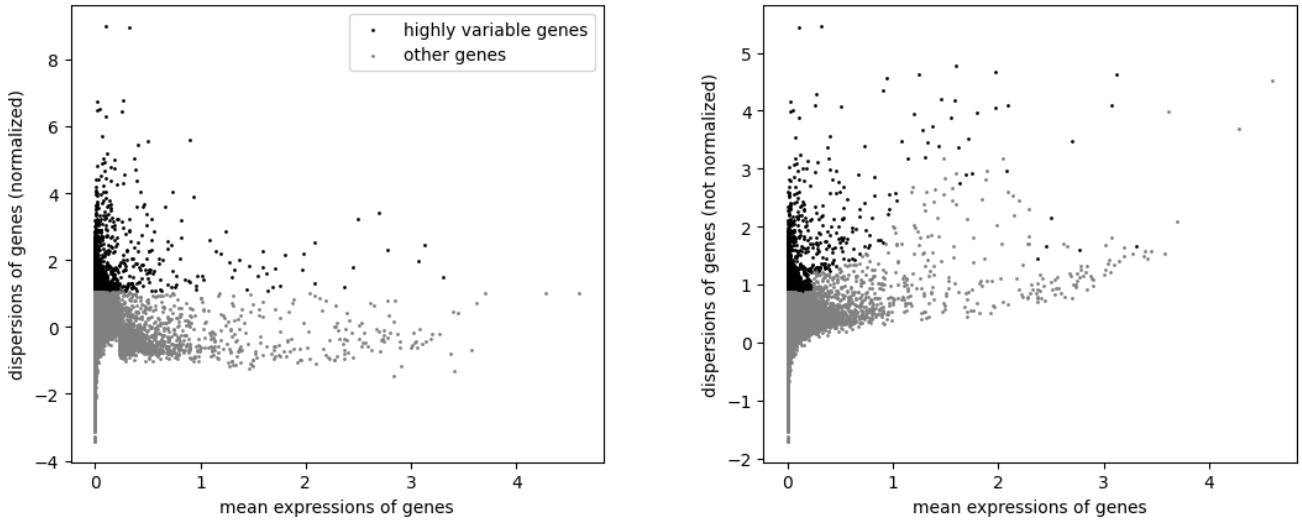

Figure 4.6: Scatter plot of highly variable genes

The two graphs illustrate the identification of highly variable genes (HVGs) by comparing mean gene expression with gene expression dispersion, both before and after normalization. The left graph shows the normalized data, where black dots represent HVGs with high dispersion relative to their mean expression, highlighting genes with significant variability across cells. In contrast, the right graph presents the same analysis on non-normalized data, displaying a similar pattern but without the preprocessing adjustments. Together, these graphs validate the selection of HVGs, demonstrating consistent gene variability patterns regardless of normalization, thereby ensuring the robustness of the identified HVGs for downstream analyses.

4.2.6 Dimensionality Reduction

After highly variable genes were identified, PCA was first applied to the normalized data using the ***`sc.tl.pca`*** function from the Scanpy library. This function computes the principal components and orders them according to the amount of variance they explain. To determine how many PCs to retain for further analysis, the variance explained by each component was examined using the ***`sc.pl.pca_variance_ratio`*** function.

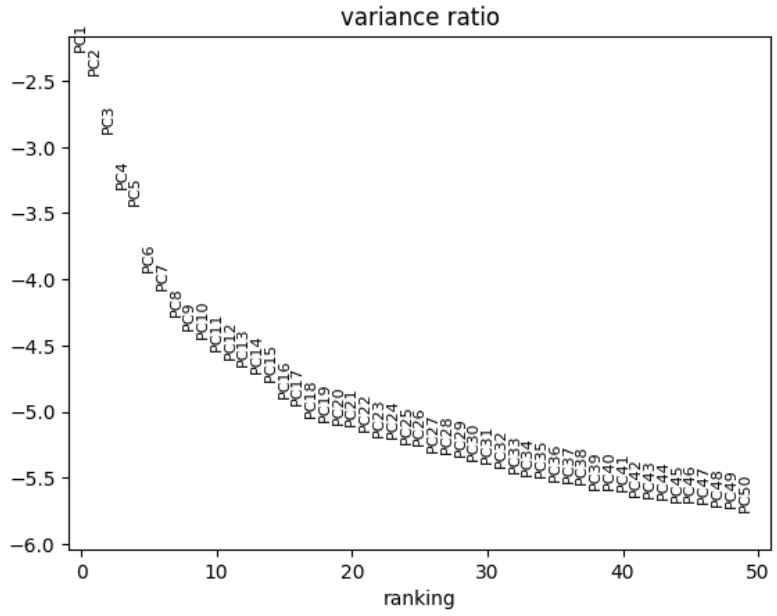

Figure 4.7: Scree plot for PCA

The scree plot visualizes the explained variance ratio for the principal components (PCs) from a Principal Component Analysis (PCA). The x-axis ranks the PCs, while the y-axis shows the variance ratio, likely log transformed. The plot reveals that the first few PCs, particularly PC1 and PC2, capture the most significant variance in the dataset, with the variance ratio steadily decreasing as the ranking increases. An "elbow" is observed around PC6 to PC10, suggesting that these initial components contain the most meaningful variation, while subsequent components contribute less. This helps determine the optimal number of PCs to retain for further analysis, balancing dimensionality reduction with information preservation.

Moreover, the scatter plot was made to provide a detailed visualization of cell distribution across the first four principal components (PC1-PC4) obtained through Principal Component Analysis (PCA), with additional insights into the influence of mitochondrial gene expression (`pct_counts_mt`).

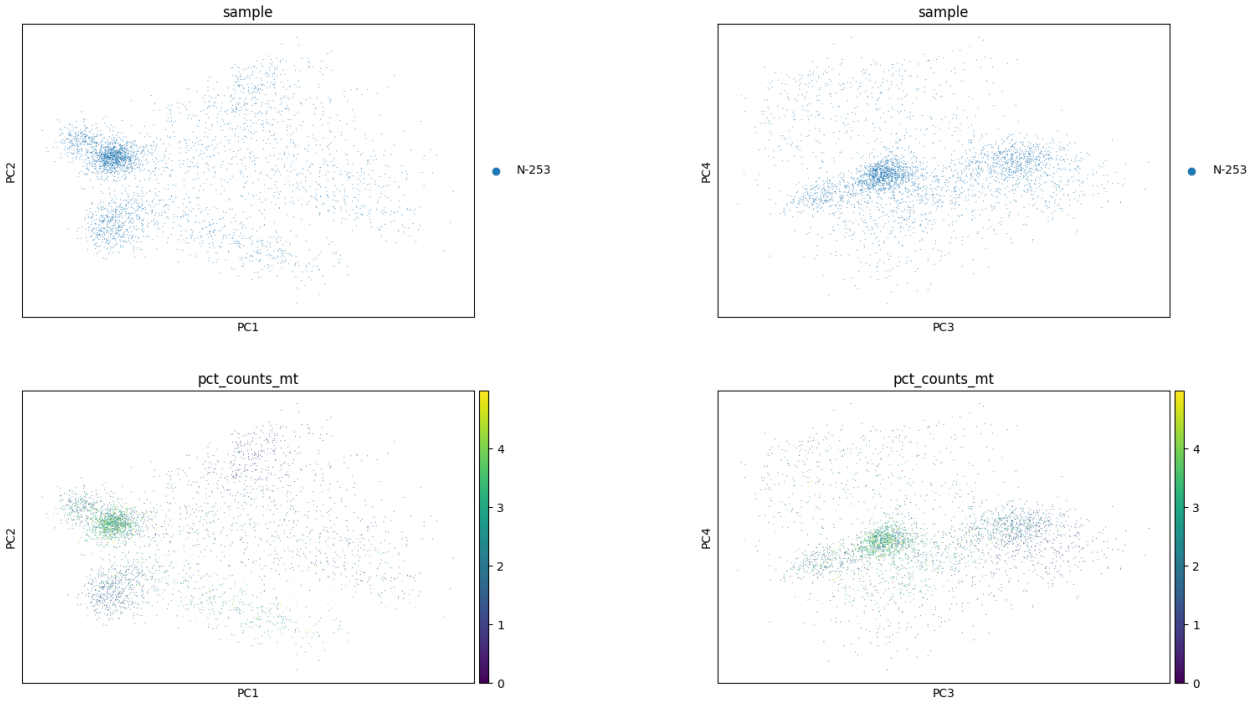

Figure 4.8: PCA plots of cells colored by sample and mitochondrial gene expression (pct_counts_mt).

The top two plots (PC1 vs. PC2 and PC3 vs. PC4) illustrate the main sources of variability in the dataset, highlighting how cells cluster based on shared features. The bottom plots integrate a color gradient representing `pct_counts_mt`, revealing how mitochondrial content correlates with these principal components. This combined analysis helps identify patterns or clusters in the data, offering a comprehensive understanding of both the structural variation within the dataset and the potential impact of mitochondrial gene expression on this variability.

4.2.7 Nearest neighbour graph construction and visualization

After performing dimensionality reduction, the nearest neighbor graph was constructed to better understand the relationships between cells in the dataset. The ***sc.pp.neighbors*** function from the Scanpy library was used to build this neighborhood graph. To make the high-dimensional data more interpretable, Uniform Manifold Approximation and Projection (UMAP) was applied using the ***`sc.tl.umap`*** function for visualization.

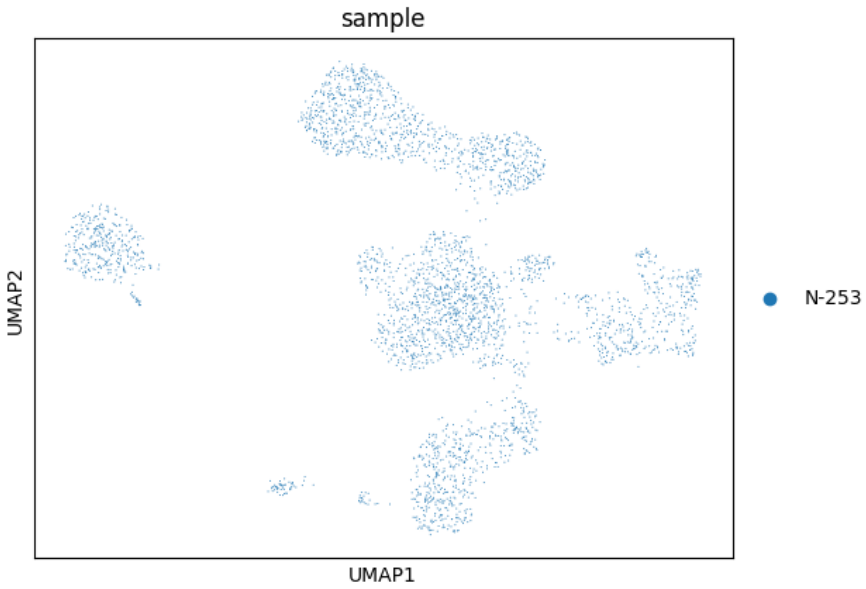

Figure 4.9: UMAP plot of cells colored by sample

The UMAP plot for sample N-253 reveals distinct clusters of cells, indicating the presence of multiple cell populations within the sample. These clusters represent groups of cells with similar gene expression profiles, suggesting potential biological heterogeneity or different cell types. The clear separation between clusters highlights the effectiveness of UMAP in capturing and visualizing the underlying structure of the data, providing valuable insights into the cellular diversity present in the dataset.

4.2.8 Clustering

For the final stages of this analysis clustering was done for all the samples. The UMAP plot for sample N-253 displayed, labeled "leiden," illustrates the clustering of cells based on the Leiden algorithm, a community detection method used to identify distinct groups within the dataset.

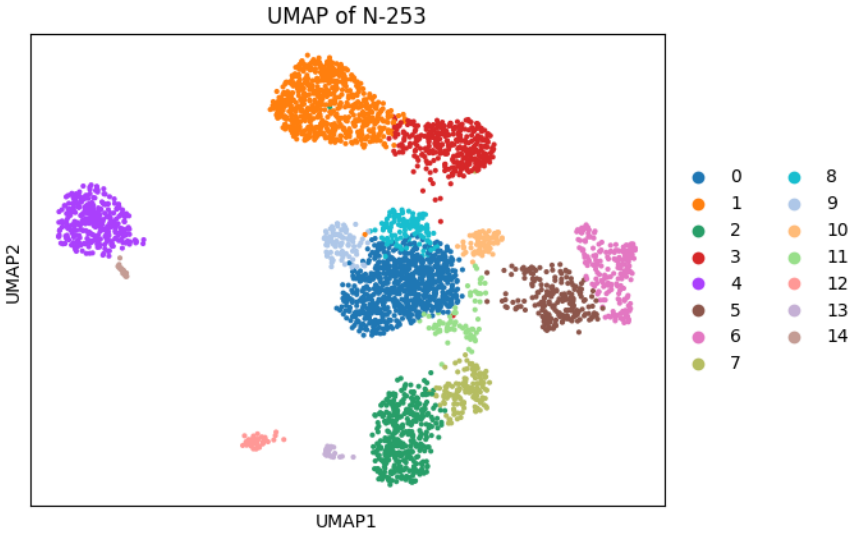

Figure 4.10: UMAP plot of N-253 cells showing distinct clusters labelled by colour.

Each colour represents a different cluster, with the numbers corresponding to specific clusters identified by the algorithm. The plot shows that the cells in N-253 are organized into several well-defined clusters, suggesting the presence of multiple distinct cell populations or states within the sample.

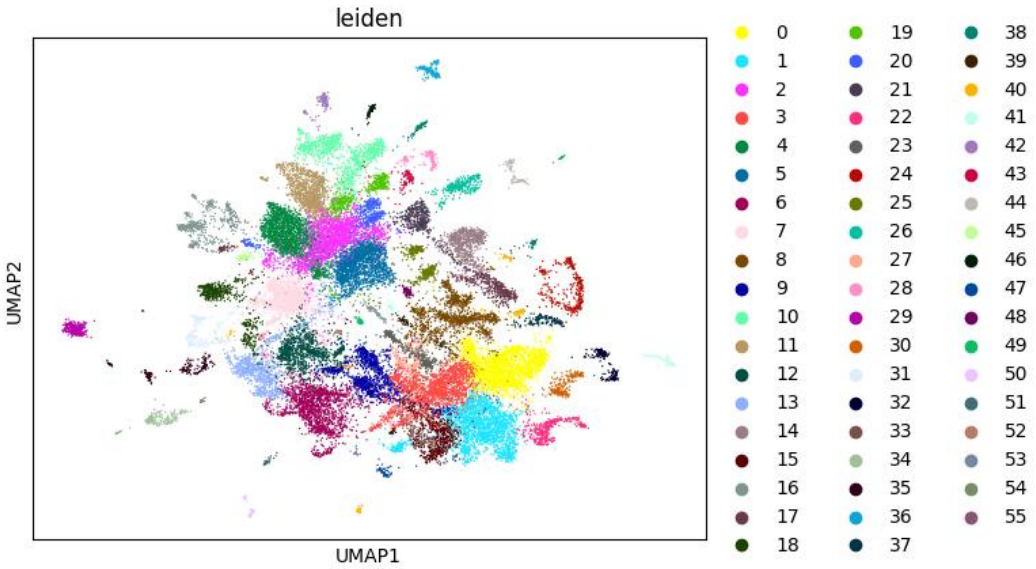

Figure 4.11: UMAP Plot showing clustering of all samples.

The anndata was then concatenated for all the samples and clustering was done. The presence of multiple, well-separated clusters indicates the existence of distinct cell populations or subtypes within the sample, which have been effectively captured and distinguished by the

Leiden clustering method. This visualization is crucial in the context of your results, as it demonstrates the cellular heterogeneity within the dataset and provides a clear overview of the different cell types or states present, which can be further analysed for biological insights.

4.2.9 Differential Gene Expression Analysis

After identifying the differentially expressed genes (DEGs), the next step was to store these results for further analysis and interpretation. To facilitate this, the DEG results from all 16 samples were stored into a single CSV file named ***differential_expression_results.csv***. These differentially expressed genes were used in comparison with spatial analysis and for machine learning.

4.2.10 Cross-Analysis of Common Genes between ST and Single-Cell Data

The top 20 common genes identified from the spatial transcriptomics data were then extracted within the single-cell data, along with their expression values. This information was compiled into a file named ***`common_genes_expression_scRNAseq.csv`,*** which serves as a crucial dataset for subsequent machine learning applications. By ensuring consistency across both data modalities, this file enables us to investigate the roles these genes play in various contexts. This file was further analysed in the machine learning model.

4.3 Machine Learning

In machine learning analysis, three models were employed for the classification and prediction of Triple-Negative Breast Cancer (TNBC). The results of these models are detailed below, providing insights into their performance and predictive accuracy in identifying TNBC.

4.3.1 Normal vs Cancer Classification Model

For the classification of normal versus cancer patients, the XGBoost classification model was applied. XGBoost is a highly effective machine learning algorithm used in cancer data analysis for its ability to handle complex, high-dimensional data, such as gene expression and mutation profiles[98]. It excels in making accurate predictions for cancer outcomes, like patient survival or treatment response, by capturing intricate patterns in the data. XGBoost's robustness against overfitting and efficiency in processing large datasets make it a valuable tool in developing personalized cancer treatments and improving diagnostic accuracy [99]. The model's performance was evaluated using various metrics, including accuracy, precision,

recall, F1-score, and AUC-ROC, to confirm its effectiveness in correctly classifying the samples.

Table 4.3: Classification report of XGBoost

Model: XGBoost	**0**	**1**
Accuracy	0.85	0.85
Precision	0.84	0.86
Recall	0.86	0.84
F1 Score	0.85	0.85

The classification report for the XGBoost model indicates a balanced performance across both classes, with an overall accuracy of 85%. The precision and recall for both classes are closely matched, each at 0.84 or 0.86, resulting in an F1-score of 0.85 for both classes, suggesting the model is equally effective at minimizing false positives and false negatives.

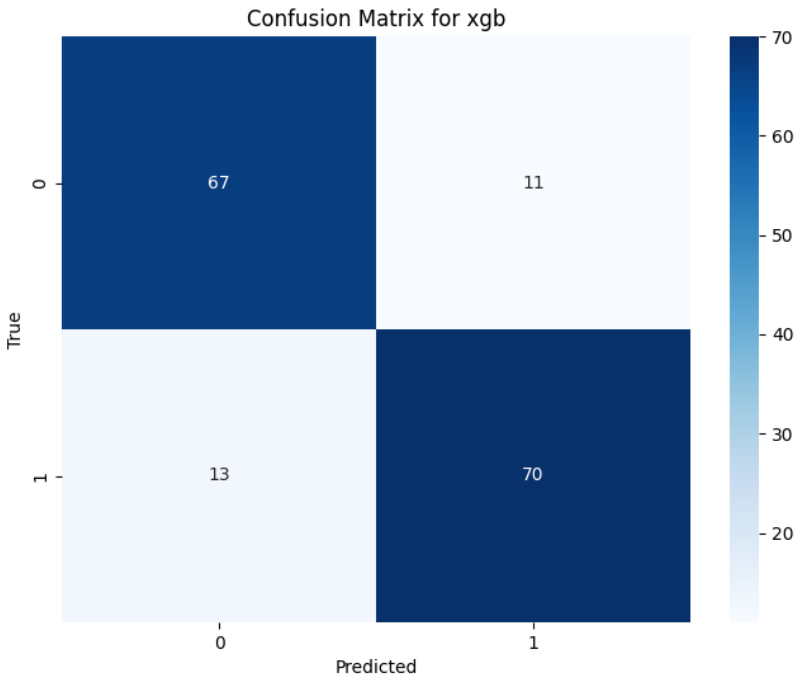

Figure 4.12: Confusion matrix of Extreme gradient Boosting

The confusion matrix shows that the XGBoost model correctly identified 67 out of 78 instances for class 0 and 70 out of 83 instances for class 1, with 11 false positives and 13 false negatives. The high number of true positives and true negatives indicates that the model effectively distinguishes between the two classes, though there is some room for improvement in minimizing the 11 false positives and 13 false negatives. Overall, the matrix reflects a strong classification performance with balanced accuracy across both classes.

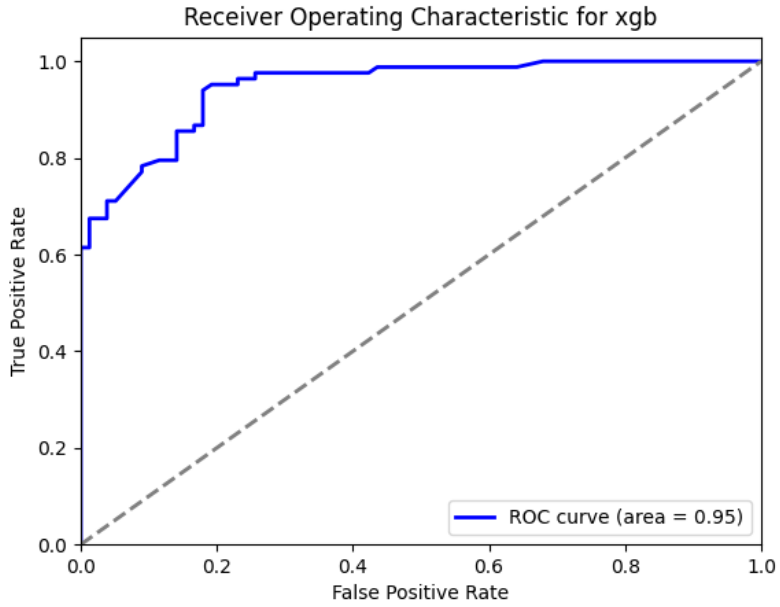

Figure 4.13: ROC curve of Extreme Gradient Boosting

The ROC curve for the XGBoost model shows an AUC of 0.95, indicating excellent classification performance. The curve's proximity to the top left corner reflects a high true positive rate and a low false positive rate, demonstrating the model's strong ability to accurately distinguish between the two classes.

4.3.2 Stage Prediction Model

The same procedure was followed to prepare the stage prediction model, where a single metadata file was prepared combining molecular and clinical data; the data was divided into training and test dataset; relevant molecular and clinical features were then selected and an SVM classifier was trained. Support Vector Classification (SVC) is an algorithm of the

supervised machine learning type mainly used in classification problems. It is based on Support Vector Machines (SVMs) which aims at finding the perfect hyperplane that can adequately take in high number of classes for data points [100]. Because of this, SVC is particularly effective in problems with high dimensions, for example, gene expression data in cancer study.

SVC has been extensively applied in recent studies for the prediction and classification of breast cancer, underlining its effectiveness in this domain. For instance, a study by Zheng et al. demonstrated that an SVM-based model, using an RBF kernel, outperformed other machine learning models in predicting breast cancer stages with high accuracy and robustness [100]. Similarly, Chen et al. employed an SVC model to classify breast cancer subtypes, finding that it provided superior predictive performance due to its ability to handle the high-dimensional gene expression data effectively [101].

Hyperparameter tuning and cross validation was done to fine-tune the SVC model, and evaluated its performance using accuracy, precision, recall, F1-score, and AUC-ROC.

Table 4.4: Classification report of SVC

Model: SVC	**0**	**1**	**2**	**3**
Accuracy	0.83	0.83	0.83	0.83
Precision	0.86	0.77	0.83	0.92
Recall	0.89	0.88	0.62	0.75
F1 Score	0.88	0.82	0.71	0.83

The SVC model's performance across four different classes, labeled 0 through 3, demonstrates consistent accuracy of 83% across all classes, indicating that the model correctly classified 83% of the instances overall. Precision values vary, with class 3 achieving the highest precision at 92%, suggesting that most of the positive predictions for this class were correct. However, class 1 has the lowest precision at 77%, indicating a higher rate of false positives for this class. Recall, which measures the model's ability to correctly identify actual positives, is highest for class 0 at 89%, showing the model's effectiveness in detecting true positives for this class. In contrast, recall is lowest for class 2 at 62%, indicating some difficulty in correctly

identifying all true positives in this class. The F1 Score, which balances precision and recall, is relatively high for most classes, with class 0 achieving the highest F1 Score of 88%, reflecting the model's strong overall performance in this category. Class 2, however, has the lowest F1 Score at 71%, suggesting that while the model is generally effective, it has room for improvement in balancing precision and recall for this class.

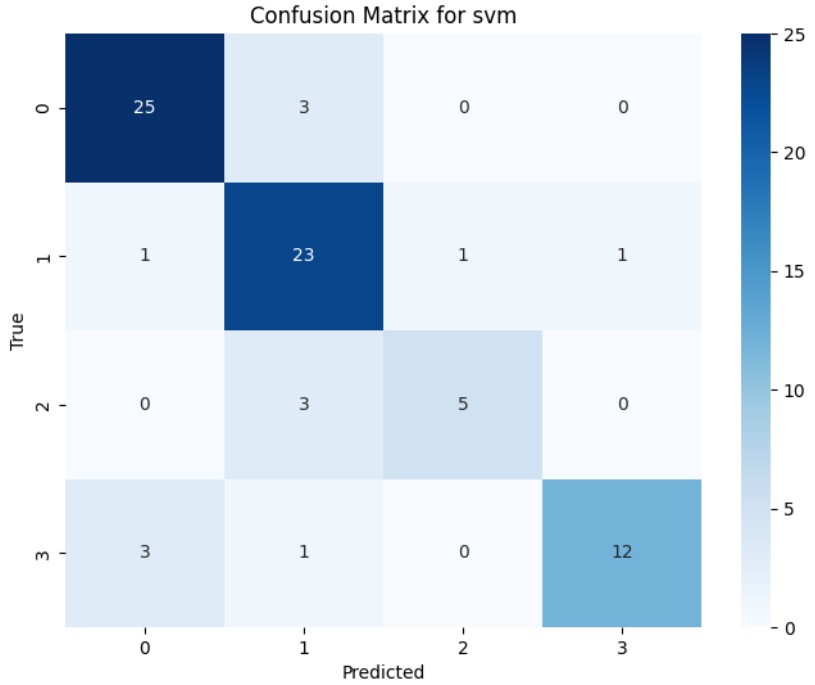

Figure 4.14: Confusion Matrix for Support Vector Classifier

The confusion matrix for the SVM model demonstrates strong performance in classifying instances of classes 0 and 1, with 25 and 23 correct classifications, respectively, though there is some confusion between these two classes, with a few instances of class 0 being misclassified as class 1 and vice versa. Class 2 has the fewest correct classifications (5) and shows some misclassification as class 1, indicating a challenge in distinguishing between these classes. Class 3 is generally well classified with 12 correct predictions, but a few instances were incorrectly classified as class 0, suggesting some feature overlap. Overall, the SVM model performs well, but there is room for improvement in differentiating between certain closely related classes.

4.3.3 Prognosis Prediction Model

The prognosis prediction model was developed using the same comprehensive metadata file as the stage prediction model, with an additional column introduced to categorize patients based on their survival times into three prognosis categories: bad (0-5 years), poor (5-10 years), and good (over 10 years). For prognosis prediction **Support Vector Regression** was chosen. Like SVM, SVR operates by finding a hyperplane in a high-dimensional space that best fits the data while maintaining the margin of tolerance (epsilon) around the hyperplane. The objective of SVR was to minimize the error by ensuring that as many data points as possible lie within this margin, while also controlling model complexity to avoid overfitting through a regularization parameter C [100]. Recent studies have demonstrated the effectiveness of SVR in predicting breast cancer outcomes. For example, researchers have used SVR to predict survival rates based on gene expression data, demonstrating superior performance compared to traditional statistical methods[101]. Other studies have applied SVR to predict the likelihood of metastasis and response to therapy, highlighting its robustness in handling complex biological data [102].

Table 4.5: Evaluation metric of SVR

Mean Squared Error:	**0.014836081926522578**
Mean Absolute Error:	0.09239555821271597
R^2 Score	0.9765976866888868

The evaluation metrics for the Support Vector Machine (SVM) model indicate strong predictive performance. The Mean Squared Error (MSE) of 0.0148 reflects a very low average squared difference between the predicted and actual values, suggesting high accuracy. The Mean Absolute Error (MAE) of 0.0924 further supports this by indicating that the average magnitude of errors in the predictions is small, reinforcing the model's precision. Additionally, the R-squared (R^2) score of 0.9766 demonstrates that approximately 97.66% of the variance in the data is explained by the model, underscoring its robustness and reliability in capturing the underlying patterns in the dataset.

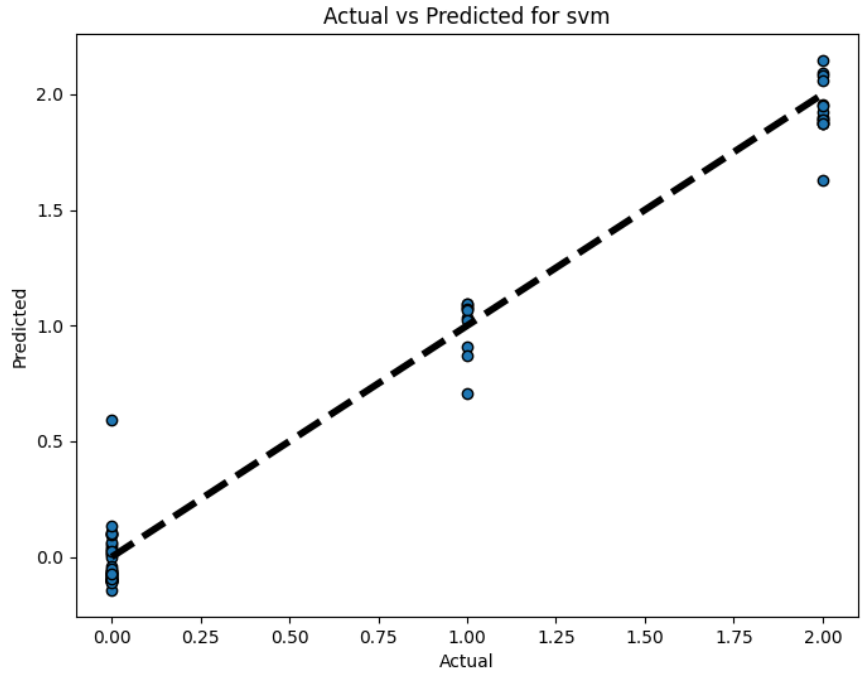

Figure 4.15: Scatter plot of Support Vector Regressor

The plot shows the actual versus predicted values for an SVM model, with most points closely aligning along the diagonal line, indicating high prediction accuracy. Points on the line represent perfect predictions, while deviations indicate errors. Overall, the model performs well, with only minor discrepancies between actual and predicted values. This visual aligns with the strong R^2 score previously mentioned.

In this chapter, the results obtained from the analysis of spatial transcriptomics, single-cell RNA sequencing (scRNA-seq), and machine learning approaches were thoroughly evaluated and discussed. The findings were compared with existing literature to contextualize the advancements made in understanding breast cancer progression, particularly using spatially resolved gene expression and cellular heterogeneity. The visualization of tumor microenvironments, identification of top differentially expressed genes, and the roles of key marker genes were explored. Additionally, machine learning models, including XGBoost for cancer classification and Support Vector Machines for stage and prognosis prediction, were developed and assessed for their predictive accuracy.

CHAPTER 5: CONCLUSION AND FUTURE RECOMMENDATIONS

Breast cancer is the most common but also one of the most lethal diseases affecting women worldwide. Many researchers have focused on understanding its molecular underpinnings and developing therapeutic strategies. Traditional approaches, including genomics, proteomics, and transcriptomics, have provided valuable insights into the disease's complexity. However, due to advancements in technology, spatial transcriptomics is emerging as a powerful new tool in the field, and very little work has been done on it compared to other techniques.

This study provides a comprehensive integration of spatial transcriptomics and artificial intelligence to unravel the complexity of triple-negative breast cancer (TNBC). By combining single-cell RNA sequencing with spatial transcriptomics, we were able to capture the spatial heterogeneity of gene expression within TNBC tumors, which is crucial for identifying reliable prognostic markers and therapeutic targets. Notably, we identified two marker genes that show promise for use in personalized medicine and as therapeutic targets, potentially improving the precision and effectiveness of treatment strategies. The machine learning models we developed for cancer classification, disease staging, and prognosis prediction demonstrate the potential of spatially resolved data in improving patient outcomes. These models predict outcomes with a high degree of accuracy, which could be highly beneficial in real-life applications, offering clinicians better tools for decision-making in treatment planning.

This study also has limitations that should be acknowledged. Firstly, the dataset we used is relatively small, which may introduce bias and affect the generalizability of our findings. Additionally, we focused exclusively on one type of breast cancer, namely triple-negative breast cancer (TNBC), which limits the broader applicability of our results to other breast cancer subtypes. The machine learning models we developed, while promising, may lack robustness due to the limited data and could benefit from further validation. Cross-validating our models' performance on larger and more diverse datasets would be essential to ensure their reliability and effectiveness.

For future research, it is recommended to expand the study to include a larger and more diverse dataset encompassing multiple breast cancer subtypes, which would enhance the

robustness and generalizability of the machine learning models. Moreover, integrating additional omics data, such as proteomics and metabolomics, with spatial transcriptomics could provide a more comprehensive understanding of the tumor microenvironment and its impact on disease progression. Finally, collaboration with clinical experts to apply and test these models in real-world settings could bridge the gap between research and clinical application, ultimately leading to more effective and personalized treatment strategies for breast cancer patients.

REFERENCES

[1] K. Morton Cuthrell and N. Tzenios, “Article no. IRJO.100349 Review Article Cuthrell and Tzenios,” 2023. [Online]. Available: https://www.sdiarticle5.com/review-history/100349

[2] “World Health Organization.”

[3] O. Nicolis, D. De Los Angeles, and C. Taramasco, “A contemporary review of breast cancer risk factors and the role of artificial intelligence,” Front Oncol, vol. 14, p. 1356014, Apr. 2024, doi: 10.3389/FONC.2024.1356014/BIBTEX.

[4] “Breast Cancer Risk Factors- CDC.”

[5] S. Łukasiewicz, M. Czeczelewski, A. Forma, J. Baj, R. Sitarz, and A. Stanisławek, “Breast Cancer—Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies—An Updated Review,” Cancers (Basel), vol. 13, no. 17, Sep. 2021, doi: 10.3390/CANCERS13174287.

[6] “Risk Factors -Pink Ribbon Pakistan.”

[7] Q. Jiao et al., “The latest progress in research on triple negative breast cancer (TNBC): risk factors, possible therapeutic targets and prognostic markers,” J Thorac Dis, vol. 6, no. 9, p. 1329, 2014, doi: 10.3978/J.ISSN.2072-1439.2014.08.13.

[8] J. Wu and C. Hicks, “Breast Cancer Type Classification Using Machine Learning,” Journal of Personalized Medicine 2021, Vol. 11, Page 61, vol. 11, no. 2, p. 61, Jan. 2021, doi: 10.3390/JPM11020061.

[9] “New study finds triple-negative breast cancer tumors with an increase in immune cells have lower risk of recurrence after surgery - Mayo Clinic News Network.” Accessed: Jul. 16, 2024. [Online]. Available: https://newsnetwork.mayoclinic.org/discussion/new-study-finds-triple-negative-breast-cancer-tumors-with-an-increase-in-immune-cells-have-lower-risk-of-recurrence-after-surgery/

[10] C. K. Anders, V. Abramson, T. Tan, and R. Dent, “The Evolution of Triple-Negative Breast Cancer: From Biology to Novel Therapeutics,” American Society of Clinical Oncology Educational Book, no. 36, pp. 34–42, May 2016, doi: 10.1200/EDBK_159135/ASSET/IMAGES/LARGE/EDBK_159135-BOX1.JPEG.

[11] C. M. Perou et al., “Molecular portraits of human breast tumours,” Nature 2000 406:6797, vol. 406, no. 6797, pp. 747–752, Aug. 2000, doi: 10.1038/35021093.

[12] W. D. Foulkes, I. E. Smith, and J. S. Reis-Filho, “Triple-Negative Breast Cancer,” New England Journal of Medicine, vol. 363, no. 20, pp. 1938–1948, Nov. 2010, doi: 10.1056/NEJMRA1001389.

[13] H. Kumar et al., “A review of biological targets and therapeutic approaches in the management of triple-negative breast cancer,” J Adv Res, vol. 54, pp. 271–292, Dec. 2023, doi: 10.1016/J.JARE.2023.02.005.

[14] M. Shibata and M. O. Hoque, “Targeting Cancer Stem Cells: A Strategy for Effective Eradication of Cancer,” Cancers 2019, Vol. 11, Page 732, vol. 11, no. 5, p. 732, May 2019, doi: 10.3390/CANCERS11050732.

[15] H. Kumar et al., “A review of biological targets and therapeutic approaches in the management of triple-negative breast cancer,” J Adv Res, vol. 54, pp. 271–292, Dec. 2023, doi: 10.1016/J.JARE.2023.02.005.

[16] M. Maqbool, F. Bekele, and G. Fekadu, “Treatment Strategies Against Triple-Negative Breast Cancer: An Updated Review,” Breast Cancer: Targets and Therapy, vol. 14, p. 15, 2022, doi: 10.2147/BCTT.S348060.

[17] M. Mustafa et al., “Molecular pathways and therapeutic targets linked to triple-negative breast cancer (TNBC),” Molecular and Cellular Biochemistry 2023 479:4, vol. 479, no. 4, pp. 895–913, May 2023, doi: 10.1007/S11010-023-04772-6.

[18] Q. Jiao et al., “The latest progress in research on triple negative breast cancer (TNBC): risk factors, possible therapeutic targets and prognostic markers,” J Thorac Dis, vol. 6, no. 9, p. 1329, 2014, doi: 10.3978/J.ISSN.2072-1439.2014.08.13.

[19] F. Sarno et al., “Triple Negative Breast Cancer Treatment Options and Limitations: Future Outlook,” Pharmaceutics 2023, Vol. 15, Page 1796, vol. 15, no. 7, p. 1796, Jun. 2023, doi: 10.3390/PHARMACEUTICS15071796.

[20] A. Mandapati and K. E. Lukong, “Triple negative breast cancer: approved treatment options and their mechanisms of action,” J Cancer Res Clin Oncol, vol. 149, no. 7, pp. 3701–3719, Jul. 2023, doi: 10.1007/S00432-022-04189-6/FIGURES/5.

[21] S. Downs-Canner and E. A. Mittendorf, “Preoperative Immunotherapy Combined with Chemotherapy for Triple-Negative Breast Cancer: Perspective on the KEYNOTE-522 Study,” Ann Surg Oncol, vol. 30, no. 6, pp. 3166–3169, Jun. 2023, doi: 10.1245/S10434-023-13267-Z/FIGURES/1.

[22] J. Ogier du Terrail et al., “Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer,” Nature Medicine 2023 29:1, vol. 29, no. 1, pp. 135–146, Jan. 2023, doi: 10.1038/s41591-022-02155-w.

[23] Y. Tai et al., “FLT1 activation in cancer cells promotes PARP-inhibitor resistance in breast cancer,” EMBO Mol Med, 2024, doi: 10.1038/S44321-024-00094-2/SUPPL_FILE/44321_2024_94_MOESM10_ESM.PDF.

[24] Y. Peng, Y. Wang, C. Zhou, W. Mei, and C. Zeng, “PI3K/Akt/mTOR Pathway and Its Role in Cancer Therapeutics: Are We Making Headway?” Front Oncol, vol. 12, p. 819128, Mar. 2022, doi: 10.3389/FONC.2022.819128/BIBTEX.

[25] E. A. Kolyvas, C. Caldas, K. Kelly, and S. S. Ahmad, “Androgen receptor function and targeted therapeutics across breast cancer subtypes,” Breast Cancer Research 2022 24:1, vol. 24, no. 1, pp. 1–15, Nov. 2022, doi: 10.1186/S13058-022-01574-4.

[26] J. D. Watson and F. H. C. Crick, “Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid,” Nature 1953 171:4356, vol. 171, no. 4356, pp. 737–738, Apr. 1953, doi: 10.1038/171737a0.

[27] A. K. Gupta and U. D. Gupta, “Next generation sequencing and its applications,” Animal Biotechnology: Models in Discovery and Translation, pp. 395–421, Jan. 2020, doi: 10.1016/B978-0-12-811710-1.00018-5.

[28] B. M. Crossley et al., “Guidelines for Sanger sequencing and molecular assay monitoring,” Journal of Veterinary Diagnostic Investigation, vol. 32, no. 6, pp. 767–775, Nov. 2020, doi: 10.1177/1040638720905833/ASSET/IMAGES/LARGE/10.1177_1040638720905833-FIG1.JPEG.

[29] Z. Khatoon, B. Figler, H. Zhang, and F. Cheng, “Introduction to RNA-Seq and its Applications to Drug Discovery and Development,” Drug Dev Res, vol. 75, no. 5, pp. 324–330, Aug. 2014, doi: 10.1002/DDR.21215.

[30] K. R. Kukurba and S. B. Montgomery, “RNA Sequencing and Analysis,” Cold Spring Harb Protoc, vol. 2015, no. 11, p. pdb. top084970, Nov. 2015, doi: 10.1101/PDB.TOP084970.

[31] E. R. Mardis, “Next-generation sequencing platforms,” Annual Review of Analytical Chemistry, vol. 6, no. Volume 6, 2013, pp. 287–303, Jun. 2013, doi: 10.1146/ANNUREV-ANCHEM-062012-092628/CITE/REFWORKS.

[32] T. Hu, N. Chitnis, D. Monos, and A. Dinh, “Next-generation sequencing technologies: An overview,” Hum Immunol, vol. 82, no. 11, pp. 801–811, Nov. 2021, doi: 10.1016/J.HUMIMM.2021.02.012.

[33] L. Ren et al., “Single cell RNA sequencing for breast cancer: present and future,” Cell Death Discovery 2021 7:1, vol. 7, no. 1, pp. 1–11, May 2021, doi: 10.1038/s41420-021-00485-1.

[34] L. Alberti-Servera et al., “Single-cell RNA sequencing reveals developmental heterogeneity among early lymphoid progenitors,” EMBO J, vol. 36, no. 24, pp. 3619–3633, Dec. 2017, doi: 10.15252/EMBJ.201797105/SUPPL_FILE/EMBJ201797105-SUP-0008-DATASETEV7.ZIP.

[35] K. T. Kim et al., “Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells,” Genome Biol, vol. 16, no. 1, pp. 1–15, Jun. 2015, doi: 10.1186/S13059-015-0692-3/FIGURES/5.

[36] W. Chung et al., “Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer,” Nature Communications 2017 8:1, vol. 8, no. 1, pp. 1–12, May 2017, doi: 10.1038/ncomms15081.

[37] A. Andersson et al., “Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions,” Nature Communications 2021 12:1, vol. 12, no. 1, pp. 1–14, Oct. 2021, doi: 10.1038/s41467-021-26271-2.

[38] L. N. G. Castro, I. Tirosh, and M. L. Suvà, “Decoding Cancer Biology One Cell at a Time,” Cancer Discov, vol. 11, no. 4, pp. 960–970, Apr. 2021, doi: 10.1158/2159-8290.CD-20-1376.

[39] X. Ren, L. Zhang, Y. Zhang, Z. Li, N. Siemers, and Z. Zhang, “Insights Gained from Single-Cell Analysis of Immune Cells in the Tumor Microenvironment,” Annu Rev Immunol, vol. 39, no. Volume 39, 2021, pp. 583–609, Apr. 2021, doi: 10.1146/ANNUREV-IMMUNOL-110519-071134/CITE/REFWORKS.

[40] M. Asp, J. Bergenstråhle, and J. Lundeberg, “Spatially Resolved Transcriptomes—Next Generation Tools for Tissue Exploration,” BioEssays, vol. 42, no. 10, p. 1900221, Oct. 2020, doi: 10.1002/BIES.201900221.

[41] M. Rossi and D. C. Radisky, “Multiplex Digital Spatial Profiling in Breast Cancer Research: State-of-the-Art Technologies and Applications across the Translational Science Spectrum,” Cancers 2024, Vol. 16, Page 1615, vol. 16, no. 9, p. 1615, Apr. 2024, doi: 10.3390/CANCERS16091615.

[42] P. R. Harrison, D. Conkie, J. Paul, and K. Jones, “Localisation of cellular globin messenger RNA by in situ hybridisation to complementary DNA,” FEBS Lett, vol. 32, no. 1, pp. 109–112, May 1973, doi: 10.1016/0014-5793(73)80749-5.

[43] T. Y. Chen, L. You, J. A. U. Hardillo, and M. P. Chien, “Spatial Transcriptomic Technologies,” Cells 2023, Vol. 12, Page 2042, vol. 12, no. 16, p. 2042, Aug. 2023, doi: 10.3390/CELLS12162042.

[44] V. Svensson, S. A. Teichmann, and O. Stegle, “SpatialDE: identification of spatially variable genes,” Nat Methods, vol. 15, no. 5, pp. 343–346, Apr. 2018, doi: 10.1038/NMETH.4636.

[45] F. A. Wolf, P. Angerer, and F. J. Theis, “SCANPY: Large-scale single-cell gene expression data analysis,” Genome Biol, vol. 19, no. 1, pp. 1–5, Feb. 2018, doi: 10.1186/S13059-017-1382-0/FIGURES/1.

[46] G. Palla et al., “Squidpy: a scalable framework for spatial omics analysis,” Nature Methods 2022 19:2, vol. 19, no. 2, pp. 171–178, Jan. 2022, doi: 10.1038/s41592-021-01358-2.

[47] Y. Li, S. Stanojevic, and L. X. Garmire, “Emerging artificial intelligence applications in Spatial Transcriptomics analysis,” Comput Struct Biotechnol J, vol. 20, pp. 2895–2908, Jan. 2022, doi: 10.1016/J.CSBJ.2022.05.056.

[48] T. Liu et al., “Artificial intelligence-enabled spatially resolved transcriptomics reveal spatial tissue organization of multiple tumors,” Tumor Discovery 2024, 3(1), 2049, vol. 3, no. 1, p. 2049, Mar. 2024, doi: 10.36922/TD.2049.

[49] R. Akbani et al., “A pan-cancer proteomic perspective on The Cancer Genome Atlas,” Nature Communications 2014 5:1, vol. 5, no. 1, pp. 1–15, May 2014, doi: 10.1038/ncomms4887.

[50] J. Cuzick et al., “Prognostic value of a combined estrogen receptor, progesterone receptor, Ki-67, and human epidermal growth factor receptor 2 immunohistochemical score and comparison with the genomic health recurrence score in early breast cancer,” Journal of Clinical Oncology, vol. 29, no. 32, pp. 4273–4278, Nov. 2011, doi: 10.1200/JCO.2010.31.2835/ASSET/IMAGES/ZLJ9991014930005.JPEG.

[51] Y. Li, X. Kong, Z. Wang, and L. Xuan, “Recent advances of transcriptomics and proteomics in triple-negative breast cancer prognosis assessment,” J Cell Mol Med, vol. 26, no. 5, pp. 1351–1362, Mar. 2022, doi: 10.1111/JCMM.17124.

[52] J. Sukumar, K. Gast, D. Quiroga, M. Lustberg, and N. Williams, “Triple-negative breast cancer: promising prognostic biomarkers currently in development,” Expert Rev Anticancer Ther, vol. 21, no. 2, p. 135, 2021, doi: 10.1080/14737140.2021.1840984.

[53] R. M. Neve et al., “A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes,” Cancer Cell, vol. 10, no. 6, pp. 515–527, Dec. 2006, doi: 10.1016/j.ccr.2006.10.008.

[54] S. Lee, J. Kim, and J. E. Park, “Single-Cell Toolkits Opening a New Era for Cell Engineering,” Mol Cells, vol. 44, no. 3, pp. 127–135, Mar. 2021, doi: 10.14348/MOLCELLS.2021.0002.

[55] “Mapping Cellular Coordinates through Advances in Spatial Transcriptomics Technology,” Mol Cells, vol. 43, no. 7, pp. 591–599, Jul. 2020, doi: 10.14348/MOLCELLS.2020.0020.

[56] Y. J. Heo, C. Hwa, G. H. Lee, J. M. Park, and J. Y. An “Integrative Multi-Omics Approaches in Cancer Research: From Biological Networks to Clinical Subtypes,” Mol Cells, vol. 44, no. 7, p. 433, Jul. 2021, doi: 10.14348/MOLCELLS.2021.0042.

[57] F. Chen et al., “RNA-seq analysis identified hormone-related genes associated with prognosis of triple negative breast cancer,” J Biomed Res, vol. 34, no. 2, p. 129, Apr. 2020, doi: 10.7555/JBR.34.20190111.

[58] C. Robert, “A decade of immune-checkpoint inhibitors in cancer therapy,” Nature Communications 2020 11:1, vol. 11, no. 1, pp. 1–3, Jul. 2020, doi: 10.1038/s41467-020-17670-y.

[59] C. Denkert, C. Liedtke, A. Tutt, and G. von Minckwitz, “Molecular alterations in triple-negative breast cancer—the road to new treatment strategies,” The Lancet, vol. 389, no. 10087, pp. 2430–2442, Jun. 2017, doi: 10.1016/S0140-6736(16)32454-0.

[60] C. C. Chen et al., “Shisa3 is associated with prolonged survival through promoting β-catenin degradation in lung cancer,” Am J Respir Crit Care Med, vol. 190, no. 4, pp. 433–444, Aug. 2014, doi: 10.1164/RCCM.201312-2256OC/SUPPL_FILE/DISCLOSURES.PDF.

[61] F. Zhao et al., “Single-cell and bulk RNA sequencing analysis of B cell marker genes in TNBC TME landscape and immunotherapy,” Front Immunol, vol. 14, p. 1245514, Dec. 2023, doi: 10.3389/FIMMU.2023.1245514/BIBTEX.

[62] D. Tierno et al., “Next-Generation Sequencing and Triple-Negative Breast Cancer: Insights and Applications,” International Journal of Molecular Sciences 2023, Vol. 24, Page 9688, vol. 24, no. 11, p. 9688, Jun. 2023, doi: 10.3390/IJMS24119688.

[63] M. Li, T. Yan, M. Wang, Y. Cai, and Y. Wei, “Advances in Single-Cell Sequencing Technology and Its Applications in Triple-Negative Breast Cancer,” Breast Cancer : Targets and Therapy, vol. 14, p. 465, Dec. 2022, doi: 10.2147/BCTT.S388534.

[64] H. Chen, F. Ye, and G. Guo, “Revolutionizing immunology with single-cell RNA sequencing,” Cellular & Molecular Immunology 2019 16:3, vol. 16, no. 3, pp. 242–249, Feb. 2019, doi: 10.1038/s41423-019-0214-4.

[65] M. Karaayvaz et al., “Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq,” Nature Communications 2018 9:1, vol. 9, no. 1, pp. 1–10, Sep. 2018, doi: 10.1038/s41467-018-06052-0.

[66] C. Kim et al., “Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing,” Cell, vol. 173, no. 4, pp. 879-893.e13, May 2018, doi: 10.1016/J.CELL.2018.03.041.

[67] S. Z. Wu et al., “A single-cell and spatially resolved atlas of human breast cancers,” Nature Genetics 2021 53:9, vol. 53, no. 9, pp. 1334–1347, Sep. 2021, doi: 10.1038/s41588-021-00911-1.

[68] J. Lv et al., “Spatial transcriptomics reveals gene expression characteristics in invasive micropapillary carcinoma of the breast,” Cell Death & Disease 2021 12:12, vol. 12, no. 12, pp. 1–11, Nov. 2021, doi: 10.1038/s41419-021-04380-6.

[69] P. L. Ståhl et al., “Visualization and analysis of gene expression in tissue sections by spatial transcriptomics,” Science, vol. 353, no. 6294, pp. 78–82, Jul. 2016, doi: 10.1126/SCIENCE.AAF2403.

[70] E. J. Topol, “High-performance medicine: the convergence of human and artificial intelligence,” Nature Medicine 2019 25:1, vol. 25, no. 1, pp. 44–56, Jan. 2019, doi: 10.1038/s41591-018-0300-7.

[71] G. Litjens et al., “A survey on deep learning in medical image analysis,” Med Image Anal, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/J.MEDIA.2017.07.005.

[72] J. A. Cruz and D. S. Wishart, “Applications of machine learning in cancer prediction and prognosis,” Cancer Inform, vol. 2, pp. 59–77, Jan. 2006, doi: 10.1177/117693510600200030/ASSET/IMAGES/LARGE/10.1177_117693510600200030-FIG5.JPEG.

[73] N. G. Thaker et al., “The Role of Artificial Intelligence in Early Cancer Detection: Exploring Early Clinical Applications,” https://home.liebertpub.com/aipo, vol. 1, no. 2, pp. 91–105, Apr. 2024, doi: 10.1089/AIPO.2023.0011.

[74] M. Zubair, S. Wang, and N. Ali, “Advanced Approaches to Breast Cancer Classification and Diagnosis,” Front Pharmacol, vol. 11, p. 632079, Feb. 2021, doi: 10.3389/FPHAR.2020.632079/BIBTEX.

[75] M. U. Ghani, T. M. Alam, and F. H. Jaskani, “Comparison of Classification Models for Early Prediction of Breast Cancer,” 3rd International Conference on Innovative Computing, ICIC 2019, Nov. 2019, doi: 10.1109/ICIC48496.2019.8966691.

[76] T. Anothaisintawee, Y. Teerawattananon, C. Wiratkapun, V. Kasamesup, and A. Thakkinstian, “Risk prediction models of breast cancer: a systematic review of model performances,” Springer, vol. 133, no. 1, pp. 1–10, May 2012, doi: 10.1007/s10549-011-1853-z.

[77] S. M. Domchek, A. Eisen, K. Calzone, J. Stopfer, A. Blackwood, and B. L. Weber, “Application of Breast Cancer Risk Prediction Models in Clinical Practice,” https://doi.org/10.1200/JCO.2003.07.007, vol. 21, no. 4, pp. 593–601, Sep. 2016, doi: 10.1200/JCO.2003.07.007.

[78] R. R. Janghel, A. Shukla, R. Tiwari, and R. Kala, “Breast Cancer diagnosis using Artificial Neural Network models,” Proceedings - 3rd International Conference on Information Sciences and Interaction Sciences, ICIS 2010, pp. 89–94, 2010, doi: 10.1109/ICICIS.2010.5534716.

[79] A. Bhardwaj and A. Tiwari, “Breast cancer diagnosis using Genetically Optimized Neural Network model,” Expert Syst Appl, vol. 42, no. 10, pp. 4611–4620, Jun. 2015, doi: 10.1016/J.ESWA.2015.01.065.

[80] X. Wang, D. Venet, D. Larsimont, L. Stenbeck, F. Dupont, and A. J. Garcia, “Spatial transcriptomics reveals substantial heterogeneity in triple negative breast cancer with potential clinical implications,” Feb. 2024, doi: 10.21203/RS.3.RS-3921508/V1.

[81] T. B. Fisher et al., “Digital image analysis and machine learning-assisted prediction of neoadjuvant chemotherapy response in triple-negative breast cancer,” Breast Cancer Research, vol. 26, no. 1, pp. 1–13, Dec. 2024, doi: 10.1186/S13058-023-01752-Y/TABLES/3.

[82] I. Wall et al., “Abstract 6604: Spatial transcriptomics delineates tumor heterogeneity in NACT triple-negative breast cancer,” Cancer Res, vol. 84, no. 6_Supplement, pp. 6604–6604, Mar. 2024, doi: 10.1158/1538-7445.AM2024-6604.

[83] R. Y. Lee, C. W. Ng, M. P. Rajapakse, N. Ang, J. P. S. Yeong, and M. C. Lau, “The promise and challenge of spatial omics in dissecting tumour microenvironment and the role of AI,” Front Oncol, vol. 13, p. 1172314, May 2023, doi: 10.3389/FONC.2023.1172314/BIBTEX.

[84] L. Zhang et al., “Clinical and translational values of spatial transcriptomics,” Signal Transduction and Targeted Therapy 2022 7:1, vol. 7, no. 1, pp. 1–17, Apr. 2022, doi: 10.1038/s41392-022-00960-w.

[85] X. Wang, D. Venet, D. Larsimont, L. Stenbeck, F. Dupont, and A. J. Garcia, “Spatial transcriptomics reveals substantial heterogeneity in triple negative breast cancer with potential clinical implications,” Feb. 2024, doi: 10.21203/RS.3.RS-3921508/V1.

[86] J. Cao et al., “Spatial Transcriptomics: A Powerful Tool in Disease Understanding and Drug Discovery,” Theranostics, vol. 14, no. 7, p. 2946, 2024, doi: 10.7150/THNO.95908.

[87] L. Moses and L. Pachter, “Museum of spatial transcriptomics,” Nature Methods 2022 19:5, vol. 19, no. 5, pp. 534–546, Mar. 2022, doi: 10.1038/s41592-022-01409-2.

[88] A. J. Cornish et al., “The genomic landscape of 2,023 colorectal cancers,” Nature 2024, vol. 26, pp. 1–10, Aug. 2024, doi: 10.1038/s41586-024-07747-9.

[89] L. Han, G. Lei, Z. Chen, Y. Zhang, C. Huang, and W. Chen, “IGF2BP2 Regulates MALAT1 by Serving as an N6-Methyladenosine Reader to Promote NSCLC Proliferation,” Front Mol Biosci, vol. 8, Jan. 2022, doi: 10.3389/FMOLB.2021.780089.

[90] Y. Zhou et al., “The m6A reader IGF2BP2 regulates glycolytic metabolism and mediates histone lactylation to enhance hepatic stellate cell activation and liver fibrosis,” Cell Death Dis, vol. 15, no. 3, Mar. 2024, doi: 10.1038/S41419-024-06509-9.

[91] M. J. A. Weerts, S. Sleijfer, and J. W. M. Martens, “The role of mitochondrial DNA in breast tumors,” Drug Discov Today, vol. 24, no. 5, pp. 1202–1208, May 2019, doi: 10.1016/J.DRUDIS.2019.03.019.

[92] Q. Zhang, Z. Liang, Y. Gao, M. Teng, and L. Niu, “Differentially expressed mitochondrial genes in breast cancer cells: Potential new targets for anti-cancer therapies,” Gene, vol. 596, pp. 45–52, Jan. 2017, doi: 10.1016/J.GENE.2016.10.005.

[93] L. P. Jayasekera et al., “Mitochondrial genome in sporadic breast cancer: A case control study and a proteomic analysis in a Sinhalese cohort from Sri Lanka,” PLoS One, vol. 18, no. 2, p. e0281620, Feb. 2023, doi: 10.1371/JOURNAL.PONE.0281620.

[94] Z. Lin, R. Peng, Y. Sun, L. Zhang, and Z. Zhang, “Identification of ribosomal protein family in triple-negative breast cancer by bioinformatics analysis,” Biosci Rep, vol. 41, no. 1, Jan. 2021, doi: 10.1042/BSR20200869/227258.

[95] C. M. Harold, A. F. Buhagiar, Y. Cheng, and S. J. Baserga, “Ribosomal RNA Transcription Regulation in Breast Cancer,” Genes 2021, Vol. 12, Page 502, vol. 12, no. 4, p. 502, Mar. 2021, doi: 10.3390/GENES12040502.

[96] J. Chen, X. Qian, Y. He, X. Han, and Y. Pan, “Novel key genes in triple-negative breast cancer identified by weighted gene co-expression network analysis,” J Cell Biochem, vol. 120, no. 10, pp. 16900–16912, Oct. 2019, doi: 10.1002/JCB.28948.

[97] X. Lin, L. Guo, X. Lin, Y. Wang, and G. Zhang, “Expression and prognosis analysis of mitochondrial ribosomal protein family in breast cancer,” Scientific Reports 2022 12:1, vol. 12, no. 1, pp. 1–13, Jun. 2022, doi: 10.1038/s41598-022-14724-7.

[98] X. Tian and G. Xu, “Clinical value of lncRNA MALAT1 as a prognostic marker in human cancer: systematic review and meta-analysis,” BMJ Open, vol. 5, no. 9, p. e008653, Sep. 2015, doi: 10.1136/BMJOPEN-2015-008653.

[99] B. Goyal, S. R. M. Yadav, N. Awasthee, S. Gupta, A. B. Kunnumakkara, and S. C. Gupta, “Diagnostic, prognostic, and therapeutic significance of long non-coding RNA MALAT1 in cancer,” Biochimica et Biophysica Acta (BBA) - Reviews on Cancer, vol. 1875, no. 2, p. 188502, Apr. 2021, doi: 10.1016/J.BBCAN.2021.188502.

[100] H. T. Zheng et al., “High expression of lncRNA MALAT1 suggests a biomarker of poor prognosis in colorectal cancer,” Int J Clin Exp Pathol, vol. 7, no. 6, p. 3174, 2014, Accessed: Aug. 18, 2024. [Online]. Available: /pmc/articles/PMC4097248/

[101] M. Schmidt et al., “Prognostic impact of immunoglobulin kappa c (Igkc) in early breast cancer,” Cancers (Basel), vol. 13, no. 14, p. 3626, Jul. 2021, doi: 10.3390/CANCERS13143626/S1.

[102] M. Lohr et al., “The prognostic relevance of tumour-infiltrating plasma cells and immunoglobulin kappa C indicates an important role of the humoral immune response in non-small cell lung cancer,” Cancer Lett, vol. 333, no. 2, pp. 222–228, Jun. 2013, doi: 10.1016/J.CANLET.2013.01.036.

[103] X. Y. Liew, N. Hameed, and J. Clos, “An investigation of XGBoost-based algorithm for breast cancer classification,” Machine Learning with Applications, vol. 6, p. 100154, Dec. 2021, doi: 10.1016/J.MLWA.2021.100154.

[104] S. Kabiraj et al., “Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm,” 2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020, Jul. 2020, doi: 10.1109/ICCCNT49239.2020.9225451.

[105] V. Jakkula, “Tutorial on Support Vector Machine (SVM)”.

[106] T. Sarkodie-Gyan, W. Caesarendra, M. Ebrahim, A. Ahmed, H. Sedky, and S. Mesbah, “Accuracy Assessment of Machine Learning Algorithms Used to Predict Breast Cancer,” Data 2023, Vol. 8, Page 35, vol. 8, no. 2, p. 35, Feb. 2023, doi: 10.3390/DATA8020035.

[107] B. Sahoo, Z. Pinnix, S. Sims, and A. Zelikovsky, “Identifying Biomarkers Using Support Vector Machine to Understand the Racial Disparity in Triple-Negative Breast Cancer,” Journal of Computational Biology, vol. 30, no. 4, pp. 502–517, Apr. 2023, doi: 10.1089/CMB.2022.0422/ASSET/IMAGES/CMB.2022.0422_FIGURE6.JPG.

[108] S. Goli, H. Mahjub, J. Faradmal, H. Mashayekhi, and A. R. Soltanian, “Survival Prediction and Feature Selection in Patients with Breast Cancer Using Support Vector Regression,” Comput Math Methods Med, vol. 2016, no. 1, p. 2157984, Jan. 2016, doi: 10.1155/2016/2157984.

[109] R. Shafique et al., “Breast Cancer Prediction Using Fine Needle Aspiration Features and Upsampling with Supervised Machine Learning,” Cancers 2023, Vol. 15, Page 681, vol. 15, no. 3, p. 681, Jan. 2023, doi: 10.3390/CANCERS15030681.

[110] R. Thalib, M. A. Bakar, and N. Larasati, “Early diagnosis breast cancer by using hybrid machine learning and advanced Lanczos algorithm,” AIP Conf Proc, vol. 2738, no. 1, Jun. 2023, doi: 10.1063/5.0140165/2894244.