

Optimizing the Multi-Omics Data Types and Variant Combinations for Accurate Breast Cancer Molecular Subtypes Classification

Sajid Shah¹, Azurah A Samah^{2*}, Syed Hamid Hussain Madni³,
Sarina Sulaiman⁴, Zuraini Ali Shah⁵, Wong Yee Leng⁶, Aryati Bakri⁷

Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia^{1, 2, 4, 5, 6, 7}

School of Electronics and Computer Science, University of Southampton Malaysia, 79100 Iskandar Puteri, Johor, Malaysia³

Abstract—Breast cancer is a highly heterogeneous disease with Luminal-A, Luminal-B, HER2-Enriched, Basal-Like, and Normal-Like molecular subtypes. Accurate classification of breast cancer molecular subtypes is essential for effective diagnosis, treatment, and planning. In recent years, multi-omics data has been widely used to improve classification performance. However, most of the existing studies focus on various combinations of multi-omics data types and variants without considering their biological relevance and computational effectiveness. This research study aims to systematically analyze, validate, and optimize the combinations of multi-omics data types and variants for accurate breast cancer molecular subtypes classification. The main goal is to identify the most suitable biologically meaningful combinations for improving classification performance. This research study provides the biological rationale for integrating the multi-omics data types and variants, and analyzes the various combinations used by existing studies for breast cancer subtype classification and the reasons behind their selection. Based on this analysis, possible and best-proposed combinations of multi-omics data types and variants are presented for the accurate classification of breast cancer molecular subtypes, based on both biological and computational perspectives. In addition, this research study identifies and recommends reliable public databases that provide multi-omics datasets with verified PAM50 labels for accurate subtype classification. The findings can help researchers design more accurate and reliable classification models by using the best proposed combination of multi-omics data types and variants, and select appropriate datasets with validated subtype labels.

Keywords—Breast cancer; molecular subtypes; multi-omics; data integration; data types and variants; datasets

I. INTRODUCTION

Breast cancer is considered one of the most serious health issues worldwide. According to GLOBOCAN [1], Kim et al. [2], the breast cancer cases recorded 2.6 million new cases and 685,000 deaths in 2024, and breast cancer became the most commonly diagnosed cancer. To treat the patients, doctors need to know the exact type of breast cancer and help them receive the fastest and best treatment, knowing the condition of breast cancer [3].

The genetic alterations and DNA damage can be influenced by estrogen exposure in the cell, that led to breast cancer. DNA defects or cancer-causing genes like BRCA1 and BRCA2 can occasionally be inherited [4]. Consequently, the chance of

developing breast cancer is increased if breast cancer has already appeared in the family. In a healthy person, the immune system attacks the cells with aberrant DNA or abnormal development. When breast cancer patients experience this failure, tumors develop and spread [5].

Researchers have found that breast cancer is not one single disease. It has different molecular subtypes that behave in different ways, as shown in Fig. 1. Based on the heterogeneity, breast cancer is classified into two stratifications, which are histological and molecular [6]. Histological classification is mainly based on morphological study, whereas molecular classification is based on the molecular pattern of breast cancer. Knowing the specific molecular subtype is very important because each subtype responds and plays differently in the process of treatment [6].

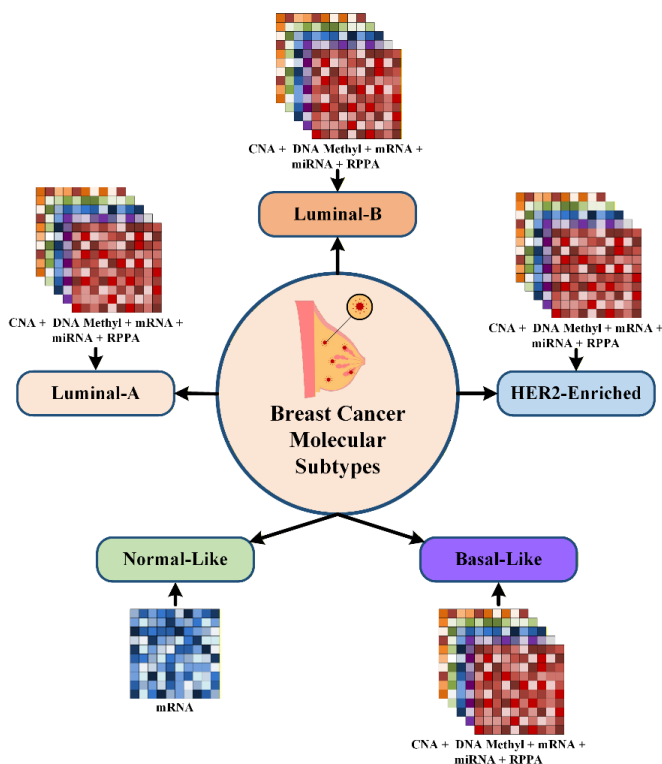


Fig. 1. Breast cancer molecular subtypes

*Corresponding author

Despite all this progress, existing research studies for breast cancer molecular subtypes classification lack accurate combinations of multi-omics data types and variants. The various combinations of data types and variants are used for breast cancer molecular subtypes classification without thinking deeply about the biology behind them, such as how genes, proteins, and chemical marks work together in each subtype. Both biological reasons (for example, which layers best capture hormone pathways or cell growth signals) and computer science reasons (for example, which mixes give stable models and clear results) play an important role in the classification of breast cancer molecular subtypes.

According to Shah et al. [5], multi-omics can improve accuracy, but there are still problems that exist, including HDLSS, limited explainability of AI models, class imbalance, multi-omics integration, and choosing the correct multi-omics data types and variant combinations. Many existing studies select various combinations of multi-omics data types and variants, but they do not explain why those combinations are best for the breast cancer molecular subtypes classification. Also, the majority of the breast cancer classification studies are struggling to identify verified and authentic databases for breast cancer datasets, while most of the studies also find it hard to locate the datasets that are completed with accurate PAM50 subtypes labeling.

Wrong or incomplete classification of breast cancer molecular subtypes due to the inappropriate combination of data types and variants, selection of wrong databases and datasets can lead to bad treatment choices and poorer patient results. Therefore, there is a pressing need to identify the optimal multi-omics data type and variant combinations along with the verified multi-omics databases and datasets for the accurate and reliable classification of breast cancer molecular subtypes. It will help create more reliable tools for doctors and move the field toward truly personalized medicine for breast cancer.

The main contributions of this research study to the field of breast cancer research and multi-omics data analysis are mentioned below:

- It provides a comprehensive investigation of various combinations of multi-omics data types and variants for breast cancer molecular subtype classification based on biological and computational evidence.
- It clearly explains the biological rationale for integrating multi-omics data, showing the biological importance of combining these multi-omics data types and variants necessary for accurate classification.
- It analyzes the existing studies on multi-omics integration for breast cancer subtype classification and highlights the various multi-omics data types and variant combinations.
- It proposes the optimal possible and best combination of multi-omics data types and variants based on both strong biological evidence and computational effectiveness for breast cancer molecular subtypes classification.

- It identifies and recommends reliable public databases and datasets that provide complete multi-omics datasets with accurate PAM50 subtype labels for future research.

The rest of the study is organized as follows: Section II emphasizes the breast cancer molecular subtypes, while Section III focuses on the omics, including mono-omics and multi-omics, along with the details of the multi-omics data types and variants. Section IV is about the biological rationale for combining multi-omics data types in breast cancer molecular subtypes. Furthermore, Section V describes multi-omics integration for breast cancer molecular subtype classification. Section VI is about the database and dataset selection. Section VII is future directions and suggestions and lastly, Section VIII is the conclusion of the study.

II. BREAST CANCER MOLECULAR SUBTYPES

Breast cancer is categorized into the five molecular subtypes, including the Luminal-A, Luminal-B, HER2-Enriched, Basal-Like and Normal-Like. Table I shows the description of the breast cancer molecular subtypes based on gene expression features, growth rate and treatment prognosis. Different breast cancer molecular subtypes are identified by unique gene expression patterns that affect on tumors behave and tumors respond to treatment [7]. Each molecular subtype of breast cancer has been elaborated in detail in the subsections below:

A. Luminal-A

Luminal-A is one of the most common molecular subtypes of breast cancer and about 50% to 60% of all breast cancer cases. It is characterized by the higher expression of the ER(+) and PR(+) and it has a HER2(-) status [8]. This tumor has a lower level of proliferation-related genes (measured by Ki-67). Breast cancer patients with Luminal-A have often better prognoses as compared to the other molecular subtypes, including lower recurrence risk and higher survival rates [9].

B. Luminal-B

Luminal-B occur for 15% to 20% of breast cancer molecular subtypes and are often more aggressive. The hormone-positive molecular of Luminal-B is characterized by ER(+), PR(+) and HER2(+) status [8]. As compared to Luminal-A, Luminal-B tumors typically show high proliferative [10]. Also, it shows more aggressive clinical and biological behaviour.

C. HER2-Enriched

HER2 occur 15% to 20% of the breast cancer molecular subtypes. It is characterized by high expression of the HER2(+) and protein along with the ER(-) and PR(-) status [8]. This molecular subtype confers more aggressive clinical and biological behavior.

D. Basal-Like

Basal-Like is also known as Triple-Negative Breast Cancer (TNBC) and is characterized by the absence of ER(-), PR(-) and HER2(-) [8]. This tumor is more aggressive than the Luminal-A and Luminal-B and grows faster. The genes of these tumors are typically found in the basal/myoepithelial cells of the abnormal breast [11]. It affects younger women, particularly those who have BRCA1 gene mutations.

TABLE I. DESCRIPTION OF BREAST CANCER MOLECULAR SUBTYPES

Subtypes	Gene Expression Features	Growth Rate	Prognosis	Percentage
Luminal-A	ER(+), PR(+), HER2(-)	Slow	Normal	About 50% to 60%
Luminal-B	ER(+), PR(+), HER2(+)	Fast	Bad	About 15% to 20%
HER2-Enriched	ER(-), PR(-), HER2(+)	Faster	Worse	About 15% to 20%
Basal-Like	ER(-), PR(-), HER2(-)	Fastest	Worst	About 8% to 37%
Normal-Like	ER(- /+), PR(-/+), HER2 (-/+)	Slow to Moderate	Good	About 5% to 10%

E. Normal-Like

Normal-Like accounts for about 5% to 10% of all breast cancer molecular subtypes. Normal-Like is characterized by ER(-), PR(-) and HER2(-) status [8]. It is a rare molecular subtype that is identified with the help of gene expression profiling that demonstrates similar characteristics to normal breast tissues. Normal-Like typically expresses the genes that are associated with the adipose tissues and other non-epithelial cells that are found in the breast [8].

III. OMICS DATA

Omics data refers to the large-scale biological data that contain various levels of molecular information of a living organism. The term “omics” contains biological information such as genomics (study of DNA and genetic variations), epigenomics (study of heritable changes in gene expression that do not involve changes in the DNA sequence), transcriptomics (study of RNA and gene expression) and proteomics (study of proteins) [12]. Each omics data type provides a unique structure and view of the biological organism [13].

A. Mono-Omics

Mono-omics data is the utilization of a single type of omics data type to study biological systems, such as genomics or epigenomics or transcriptomics. Mono-omics analysis is simpler, easier to process and requires less computational power compared to multi-omics analysis [14]. Mono-omics data types are widely used in previous research and have offered insights into disease mechanisms. For instance, Cristovao, et al. [15], Zeng, et al. [16], Li and Nabavi [17] use mon-omics data, particularly on genomics or transcriptomics data types, to analyze and classify breast cancer molecular subtypes. However, mono-omics data focus on capturing only one layer of biological information, which limits the potential of mono-omics data to fully represent the complexity of heterogeneous diseases such as breast cancer molecular subtypes classification [18].

B. Multi-Omics

Multi-omics data refers to the integration of various data types of omics to provide a detailed and comprehensive structure and view of biological systems [19]. Instead of depending on a single layer of information, such as mono-omics data type, multi-omics combines different molecular data sources at different molecular layers of an organism to better understand the complex and heterogeneous diseases, including the breast cancer molecular subtypes classification [20]. Multi-omics data helps in capturing the important relationships and interactions among different biological processes of an organism that lead to

more reliable and accurate classification of breast cancer molecular subtypes [21].

C. Multi-Omics Data Types

There are four main data types of multi-omics, including genomics, epigenomics, transcriptomics and proteomics. Each multi-omics data type is described in subsections.

1) *Genomics*: It is the study of the complete set of DNA within a cell, including all genes and their functions. It concentrates on the understanding of genetic changes in an organism, such as insertions, deletions and mutations [22]. In breast cancer molecular subtypes, genomics assists in identifying the acquired or inherited genetic alterations that trigger cancer development or influence tumor behavior in the breast [23]. The genetic changes in breast cancer deactivate or activate the oncogenes and tumor suppressor genes that lead to uncontrolled breast cancer cell growth [22]. Overall, genomics provides the foundational layer of biological information for understanding the origin of breast cancer at the DNA level.

2) *Epigenomics*: It is the study of changes in gene activity that do not involve alterations in the DNA sequence itself. Instead, it concentrates on the regulatory mechanisms such as DNA-Methylation, histone modification and chromatin structure changes. The epigenetic modifications can turn genes “on” or “off” any gene without changing the underlying genetic code [24]. In breast cancer, abnormal epigenetic changes lead to the silencing of important tumor suppressor genes or the activation of harmful genes, contributing to breast cancer progression [25]. Epigenomics is important because it contains the information about the environmental and biological factors that can influence gene behavior without changing DNA.

3) *Transcriptomics*: It is the study of RNA molecules that are produced from DNA through the process of transcription. It mainly concentrates on the gene expression levels, demonstrating which genes are transcribed actively into RNA under specific conditions [26]. In breast cancer, transcriptomics assists in identifying the differences in gene expression patterns across different cancer molecular subtypes, such as Luminal or Basal-like types [27]. These expression patterns provide important insights into tumor activity, cellular processes, and disease progression.

4) *Proteomics*: Proteomics is the study of proteins, which are the final functional products of genes. It concentrates on the protein abundance, modifications, interactions and structure within biological systems [28]. Proteins carry out most cellular

functions; proteomics provides direct insight into the way cells behave in normal and disease conditions [29]. In breast cancer, abnormal protein interactions or expression indicate tumor metastasis, growth and treatment resistance [27]. Proteomics is important because it reflects the actual functional state of the cell, making it highly valuable for understanding disease mechanisms and improving subtype classification.

Overall, the integration of multi-omics data types is very important for breast cancer molecular subtypes classification. Each omics layer provides complementary and unique information, and combining them helps capture the full complexity of breast cancer biology. By integrating multi-omics data types in an appropriate biological and computational way, researchers can improve classification accuracy, identify more reliable biomarkers, and develop better personalized treatment strategies for different breast cancer subtypes.

D. Multi-Omics Data Variants

In breast cancer research, multi-omics data are further categorized into different variants that represent specific types of molecular measurements within each multi-omics data type. These variants provide more detailed and fine-grained biological information about tumor development, progression, and subtype differences [30]. Instead of treating each omics layer as a single dataset, researchers analyze different variants such as genetic alterations, gene expression levels, epigenetic modifications, and protein activity. Each multi-omics data variant captures a unique aspect of cancer biology, and when combined, they help to build a more complete and accurate model for breast cancer molecular subtype classification [31].

1) *Copy Number Alteration (CNA)*: Copy Number Alteration refers to changes in the number of copies of a particular DNA segment in the genome. These alterations can include gains (amplifications) or losses (deletions) of chromosomal regions. CNA is a key variant of genomics because it directly reflects structural changes in DNA that may lead to cancer development by increasing oncogene activity or deleting tumor suppressor genes [32].

2) *Copy Number Variation (CNV)*: Copy Number Variation is a form of genomic structural variation where sections of the genome vary in copy number between individuals. CNVs can influence gene dosage and are strongly associated with cancer susceptibility and tumor progression. In breast cancer, CNVs help explain variability in gene expression and subtype-specific genomic instability [32].

3) *Single Nucleotide Polymorphism (SNP)*: Single Nucleotide Polymorphisms are variations at a single nucleotide position in DNA. SNPs are the most common type of genetic variation and can affect how genes function or how proteins are produced. In breast cancer genomics, SNPs are used to identify genetic risk factors and subtype-specific mutations [33].

4) *DNA Methylation*: DNA methylation is an epigenomic modification where methyl groups are added to DNA, usually affecting gene expression without changing the DNA sequence. It plays a major role in gene silencing and activation [34]. In breast cancer, abnormal DNA methylation patterns are linked to

tumor suppression loss and subtype-specific regulatory changes [35].

5) *Messenger RNA Expression (mRNA)*: mRNA represents the transcriptomic level of gene activity, measuring how actively genes are being transcribed into RNA. It reflects real-time cellular function and is widely used for defining breast cancer intrinsic subtypes, such as Luminal and Basal-Like. mRNA profiles provide strong signals for tumor classification and biological interpretation [36].

6) *MicroRNA Expression miRNA*: miRNAs are small non-coding RNA molecules that regulate gene expression by binding to target mRNA and preventing protein production. miRNA is considered a regulatory transcriptomic variant that influences cancer progression by controlling multiple genes simultaneously [37]. In breast cancer, miRNA expression patterns are associated with subtype differences and tumor aggressiveness [38].

7) *Reverse Phase Protein Array (RPPA)*: Reverse Phase Protein Array is a proteomics-based technique used to measure protein expression and activation levels, including post-translational modifications such as phosphorylation. RPPA provides functional information about signaling pathways in breast cancer. It reflects the final stage of gene expression and is essential for understanding how molecular changes translate into cellular behavior and treatment response [39].

IV. BIOLOGICAL RATIONALE FOR COMBINING MULTI-OMICS DATA TYPES IN BREAST CANCER MOLECULAR SUBTYPES

Breast cancer is a highly heterogeneous disease with various molecular subtypes. Each subtype shows different genetic mutations, gene expression patterns, epigenetic regulation, and protein activity [23]. Mono-omics data types are not sufficient to fully explain tumor behavior because cancer development is driven by multiple biological mechanisms acting together across different omics levels. For instance, Basal-like tumors show very high TP53 mutation rates, while Luminal tumors show frequent PIK3CA and GATA3 mutations, and HER2-enriched tumors are strongly associated with ERBB2 amplification [39]. These differences clearly indicate that breast cancer cannot be fully understood using only one type of omics data.

A. Relationship of Multi-Omics Data Types and Variants

Genomics, epigenomics, transcriptomics, and proteomics multi-omics data types are closely associated and work together in order to describe the detailed and comprehensive view of a full biological system of a cell. Multi-omics data types are not independent; instead, they form an organized, sequenced, and layered biological flow where each multi-omic layer influences the next [40]. Genomics represents the foundation because DNA contains all the genetic information required for the development and function of a cell. However, the activity of a gene is not determined by DNA alone. Epigenomics acts as a regulatory layer on top of genomics by controlling whether specific genes are turned on or off through mechanisms. The epigenetic changes directly influence transcriptomics because only genes that are activated at the epigenomic level are transcribed into RNA. Transcriptomics then reflects the real-time gene expression profile of a cell, showing which genes are

actively producing RNA molecules under certain conditions. The RNA is further translated into proteins, making proteomics the final functional layer of the biological system. Proteins carry out most cellular activities, including signaling, metabolism, and structural functions, and therefore represent the actual biological outcome of genomic instructions [12]. The relationship among the multi-omics data types is highly interconnected, meaning that changes in one layer directly lead to effects in others, as shown in Fig. 2. Because of the strong interdependencies, integrating multi-omics data is very important for understanding heterogeneous and complex diseases like breast cancer. The interconnected biological flow enables researchers to capture not only individual molecular changes but also the dynamic relationships between different molecular layers, leading to more accurate and biologically meaningful cancer subtype classification [39].

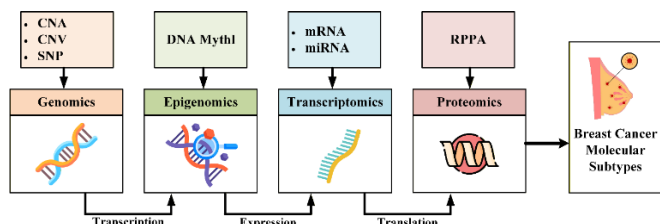


Fig. 2. Relationship among the multi-omics data types and variants.

B. Biological Dependency of Multi-Omics

One of the most important biological reasons for integrating multi-omics data is the strong dependency between multi-omics data types. DNA mutations can influence epigenetic regulation, which then affects gene expression and ultimately protein production. For example, TP53 mutations in Basal-like tumors lead to uncontrolled cell proliferation, while PIK3CA mutations in Luminal tumors activate downstream signaling pathways affecting both RNA expression and protein signaling [39]. Similarly, epigenetic silencing of tumor suppressor genes can reduce RNA and protein expression even when DNA remains unchanged. This cascading biological relationship shows that each omics layer only provides partial information, but together they describe the complete disease mechanism [39].

C. Importance of Multi-Omics Integration for Breast Cancer Molecular Subtypes Classification

Because breast cancer subtypes are defined by combined genetic, epigenetic, transcriptional, and proteomic features, integrating all these data types is necessary for accurate classification. TCGA multi-platform studies have shown that basal-like tumors are consistently distinct across all omics layers, while luminal and HER2 subtypes show overlapping

patterns across different data types [41]. This proves that single-omics approaches fail to capture subtype complexity. Multi-omics integration allows researchers to identify more robust biomarkers, detect hidden subtype relationships, and improve classification accuracy by combining complementary biological signals [42].

In summary, breast cancer is driven by interconnected molecular mechanisms across multiple biological layers. Genomics explains structural changes, epigenomics explains regulatory control, transcriptomics reflects gene activity, and proteomics captures functional outcomes. However, none of these layers alone can fully represent tumor behavior. Therefore, integrating the genomics, epigenomics, transcriptomics, and proteomics multi-omics data is biologically necessary to capture the complete molecular architecture of breast cancer and improve the accuracy of molecular subtype classification.

V. MULTI-OMICS INTEGRATION FOR BREAST CANCER MOLECULAR SUBTYPE CLASSIFICATION

The successful integration of multi-omics variants provides deep insights into the molecular alterations across various regulatory levels of breast cancer. The correct selection of multi-omics data types and variants is very important for breast cancer molecular subtype classification and gives us a full picture of the molecular landscape of breast cancer. This helps in identifying subtypes and finding possible biomarkers for personalised treatment plans [31]. There are various combinations of multi-omics data types and variants that are used by existing research studies for performing the breast cancer molecular subtypes classification, as shown in Fig. 3.

A. Combinations of Multi-Omics Data Types for Breast Cancer Molecular Subtypes Classification

Multi-omics data types integration has become very important in classifying the breast cancer molecular subtype. By combining genomics, epigenomics, transcriptomics, and proteomics multi-omics data types, researchers can capture the heterogeneous and complex structure and mechanism of breast cancer. As compared to mono-omics integration, multi-omics integration improves the ability to identify subtypes of specific biomarkers, discover the interactions and relationships among various multi-omics data types and variants, and enhance the overall classification and prediction performance.

Existing studies have used various combinations of multi-omics data types, as shown in Table II, to improve the breast cancer molecular subtypes classification. The existing studies use combinations of two or three data types to classify the breast cancer molecular subtypes.

TABLE II. MULTI-OMICS DATA TYPES USED BY THE EXISTING STUDIES

No of Multi-Omics Data Types	Genomics	Epigenomics	Transcriptomics	Proteomics	References
2	✓		✓		[15-17]
2		✓	✓		[43-47]
3	✓	✓	✓		[48-55]
3	✓	✓		✓	[7]
3	✓		✓	✓	[56]
4	✓	✓	✓	✓	Proposed

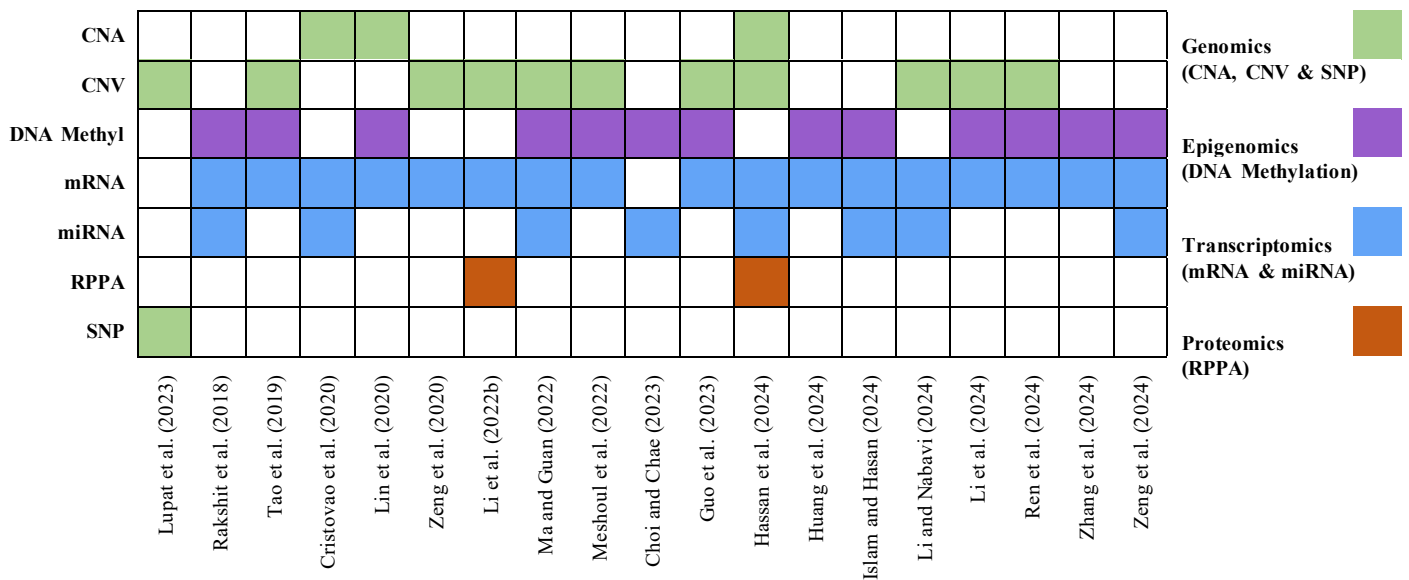


Fig. 3. Multi-omics data types and variants used by the existing studies.

Each combination of multi-omics data types provides unique insights; however, the existing studies that integrated three data types generally show significant performance. Based on the detailed analysis, it is suggested that the integration of four main multi-omics data types tends to provide a more detailed molecular structure, which can lead to more reliable and accurate classification of breast cancer molecular subtypes.

B. Combinations of Multi-Omics Variants for Breast Cancer Molecular Subtypes Classification

With the rapid growth of multi-omics data, Copy Number Variation (CNV), Single Nucleotide Polymorphism (SNP), DNA methylation, mRNA expression, miRNA expression, and protein abundance (RPPA), multi-omics variants have become available. The proper and correct integration of multi-omics variants allows for a more robust and reliable classification of breast cancer molecular subtypes by capturing the important information in each omics data type [42]. Researchers use

various combinations of the multi-omics variants to improve the classification performance in breast cancer molecular subtypes classification.

Table III presents the various combinations of multi-omics variants that are utilized by the existing studies for the successful classification of breast cancer subtypes. Existing studies use combinations of two, three, four, or five multi-omics data variants for the classification of molecular breast cancer subtypes. Overall, the integration of multi-omics variants empowers researchers to explore and capture the complex molecular interactions to understand the mechanism and underlying structure of breast cancer molecular subtypes. Each multi-omics variant combination provides unique information about the breast cancer subtypes on the molecular level. Based on the detailed analysis, it is suggested that integrating the five multi-omics variants can enhance breast cancer molecular subtypes classification accuracy and performance.

TABLE III. MULTI-OMICS DATA VARIANTS USED BY THE EXISTING STUDIES

No of Multi-Omics Data Variants	CNA	CNV	SNP	DNA-M	mRNA	miRNA	RPPA	References
2		✓	✓					[57]
2		✓			✓			[16, 46]
2				✓	✓			[51]
3		✓		✓	✓			[49, 52-54]
3		✓		✓		✓		[48, 55]
3	✓				✓	✓		[15, 17]
3	✓				✓		✓	[7]
3				✓	✓	✓		[43]
4		✓		✓	✓	✓		[50]
5	✓	✓		✓	✓	✓		[56]
5		✓		✓	✓	✓	✓	Proposed

Notably, models incorporating the accurate multi-omics data types and variants tend to achieve superior classification performance and emphasize the value of multi-omics data fusion in breast cancer molecular subtypes analysis.

C. Proposed Combinations of Multi-Omics Variants for Breast Cancer Molecular Subtypes Classification Based on Biological Rationale

The reliable and appropriate integration of multi-omics data types and variants plays an important role in enhancing the accuracy and robustness of breast cancer molecular subtype classification. Based on biological evidence and insights from existing computational classification models, various combinations of multi-omics data variants are utilized for capturing the heterogeneous and complex view of breast cancer, as mentioned in Section V-A and Section V-B. These combinations are selected to represent different layers of biological information, including genomic alterations, epigenetic regulation, gene expression, and protein activity, which collectively define the breast cancer behavior and subtype characteristics.

From a biological perspective, the combination of genomics, epigenomics, transcriptomics, and proteomics multi-omics data types, along with their desired variants, allows for a comprehensive representation of breast cancer structure, behavior, and progression across multiple molecular levels.

Based on the biological rationale and computational models for breast cancer molecular subtypes classification, the combination of CNV, DNA Methylation, mRNA, and RPPA multi-omics data variants, which integrates genomic structure, epigenetic control, gene expression, and protein activity, or the combination of CNV, DNA methylation, miRNA, and RPPA multi-omics data variants, which replaces mRNA with miRNA, resulting in a combination of, focusing more on regulatory transcriptomic mechanisms. These possible combinations have demonstrated effectiveness in capturing subtype-specific patterns and improving classification performance compared to mono-omics analysis. However, these combinations are still incomplete and can be further explored by adding more multi-omics variants.

To address these limitations, this research study proposes an enhanced and more comprehensive combination that integrates CNV, DNA Methylation, mRNA, miRNA, and RPPA multi-omics data variants. The proposed combination is based on the biological justification because it captures all major layers of molecular regulation, starting from DNA-level structural variations (CNV), followed by epigenetic regulation (DNA-Methylation), gene expression (mRNA), gene regulation (miRNA) and finally protein-level functional outcomes (RPPA), as shown in Fig. 4. By including both mRNA and miRNA, the model benefits from a more complete representation of transcriptomic activity, including both expression and regulatory mechanisms. The inclusion of RPPA further strengthens the model by incorporating protein-level validation of upstream molecular changes.

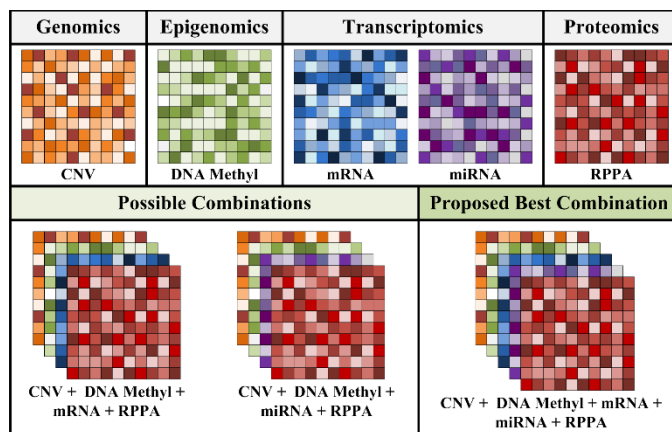


Fig. 4. Possible and proposed best combinations of multi-omics variants for breast cancer molecular subtypes classification.

From a computational perspective, integrating these complementary data variants can improve feature diversity, reduce information loss, and enhance the ability of machine learning and deep learning models to learn complex patterns associated with different breast cancer subtypes. The proposed comprehensive combination not only improves classification accuracy but also increases model robustness and biological interpretability. Therefore, the proposed combination of multi-omics data types and variants is considered the most suitable and effective strategy for accurate breast cancer molecular subtype classification.

VI. DATABASE AND DATASET SELECTION

There are various databases used for breast cancer research, as mentioned in Table IV, each with its own characteristics, quality, multi-omics types, and variants. Compared to other commonly used databases such as CPTAC, NCBI, GEO, and METABRIC, the TCGA and UCSC Xena databases are among the most widely used and reliable sources for breast cancer multi-omics datasets. TCGA (The Cancer Genome Atlas) is the original and one of the most comprehensive cancer research projects, led by the National Cancer Institute (NCI), which profiled more than 20,000 primary tumors across 33 different cancer types, including breast cancer. It provides a complete range of multi-omics data types and their variants, making it highly suitable for molecular subtype classification studies. The official repository for TCGA data is the Genomic Data Commons (GDC), where both raw and processed datasets are available. On the other hand, UCSC Xena is another powerful data visualization and exploration platform that serves as a user-friendly portal for accessing breast cancer datasets. UCSC Xena re-processes and organizes multi-omics data into easy-to-use formats, allowing researchers to efficiently explore, integrate, and download datasets across all major omics types and variants. Due to their comprehensive coverage of multi-omics data types and variants, as shown in Table IV, standardized processing, and availability of clinically annotated information such as PAM50 subtype labels, TCGA, and UCSC Xena are considered the most suitable and commonly used databases for breast cancer multi-omics research and classification tasks.

Table IV presents a widely used databases and repositories for breast cancer molecular subtype classification datasets. It shows that multi-omics data, particularly genomics, epigenomics, transcriptomics, and proteomics, can be effectively obtained from platforms such as Linkedomics-TCGA-BRCA, UCSC-XENA-BRCA, and GDC-TCGA-BRCA. While other repositories like CPTAC, NCBI, GEO, and METABRIC also provide valuable breast cancer datasets, they have certain limitations, such as missing omics layers (e.g., proteomics or epigenomics), restricted cohorts, or static data

availability. Table V describes the multi-omics GDC-TCGA-BRCA dataset used for breast cancer molecular subtype classification. This dataset is downloaded from the GDC-TCGA repository and carefully preprocessed to ensure accurate subtype labeling using PAM50 standards. Since labeling inconsistencies are a common challenge in existing studies, all samples were manually validated and assigned standardized labels to improve reliability. The dataset includes multiple multi-omics data variants along with their respective feature dimensions and sample sizes.

TABLE IV. COMPARISON OF USED MULTI-OMICS DATASETS FOR BREAST CANCER MOLECULAR SUBTYPES CLASSIFICATION

Datasets and Repositories	Genomics	Epigenomics	Transcriptomics	Proteomics	Limits
Linkedomics (Linkedomics-TCGA-BRCA) https://linkedomics.org/data_download/TCGA-BRCA/	Y	Y	Y	Y	N/A
University of California, Santa Cruz (UCSC-XENA-BRCA) https://xena.ucsc.edu	Y	Y	Y	Y	N/A
Genomic Data Commons – (GDC-TCGA-BRCA) https://portal.gdc.cancer.gov/projects/TCGA-BRCA	Y	Y	Y	Y	N/A
Clinical Proteomic Tumor Analysis Consortium (CPTAC) https://cptac-data-portal.georgetown.edu/cptacPublic/	N	N	N	Y	Only the proteomics cohort from TCGA
National Center for Biotechnology Information (NCBI) https://www.ncbi.nlm.nih.gov/datasets/	Y	Y	Y	N	Proteomics missing
Gene Expression Omnibus (GEO) https://www.ncbi.nlm.nih.gov/geo/	Y	Y	Y	N	A cohort from NCBI missing proteomics Data
Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) http://molonc.bccrc.ca/aparicio-lab/research/metabric/	Y	N	Y	N	Static data & genomics and proteomics are missing

TABLE V. DESCRIPTION OF MULTI-OMICS GDC-TCGA-BRCA DATASET FOR BREAST CANCER MOLECULAR SUBTYPES CLASSIFICATION

Multi-Omics Datatypes	Multi-Omics Variants	Original Features	Original Sample Size	Label Sample Size
Genomics	CNV	19569	759	754
Epigenomics	DNA-Methylation	19050	759	754
Transcriptomics	mRNA	18207	759	754
	miRNA	369	759	754

TABLE VI. DESCRIPTION OF MULTI-OMICS LINKEDOMICS-TCGA-BRCA DATASET FOR BREAST CANCER MOLECULAR SUBTYPES CLASSIFICATION

Multi-Omics Datatypes	Multi-Omics Variants	Original Features	Original Sample Size	Label Sample Size
Genomics	CNV	24776	1080	1039
Epigenomics	DNA-Methylation	20106	783	772
Transcriptomics	mRNA	20155	1093	1051
	miRNA	823	755	740
Proteomics	RPPA	175	887	858

TABLE VII. DESCRIPTION OF MULTI-OMICS UCSC-XENA-BRCA DATASET FOR BREAST CANCER MOLECULAR SUBTYPES CLASSIFICATION

Multi-Omics Datatypes	Multi-Omics Variants	Original Features	Original Sample Size	Label Sample Size
Genomics	CNV	60624	1082	1021
Epigenomics	DNA-Methylation	486428	893	774
Transcriptomics	mRNA	60661	1226	1052
	miRNA	1882	1202	1036
Proteomics	RPPA	488	919	853

Table VI outlines the multi-omics LINKEDOMICS-TCGA-BRCA dataset, which was obtained from the Linkedomics platform. Similar to the GDC dataset, this dataset was also rigorously preprocessed, and PAM50 labels were assigned and validated to ensure consistency and accuracy in breast cancer molecular subtype classification. It contains a more comprehensive set of omics data, including genomics, epigenomics, transcriptomics, and proteomics (RPPA), making it particularly suitable for integrative multi-omics analysis. Finally, Table VII presents the multi-omics UCSC-XENA-BRCA dataset, which was collected from the UCSC Xena repository. This dataset was also preprocessed following the same standardized pipeline, where PAM50 labels were carefully validated and assigned to address common labeling challenges in breast cancer datasets. It includes extensive high-dimensional data across genomics, epigenomics, transcriptomics, and proteomics, further supporting robust multi-omics-based classification. Collectively, these datasets provide a strong foundation for developing and evaluating accurate breast cancer molecular subtype classification models.

In addition, the PAM50-validated labels generated and refined in this research study for all the datasets in Table V, Table VI, and Table VII are publicly available to support reproducibility and future research. The breast cancer molecular subtypes PAM50-labels can be downloaded from the GitHub repository, enabling other researchers to utilize standardized and reliable data for breast cancer molecular subtypes classification.

VII. SUGGESTIONS AND FUTURE DIRECTIONS

This research study shows that the best combination of multi-omics data types and variants for breast cancer molecular subtypes classification is CNV + DNA methylation + mRNA + miRNA + RPPA because it covers all important biological layers from DNA changes to protein activity. Future research should evaluate the proposed best combination with more advanced machine learning and deep learning models to determine whether classification accuracy and robustness can be improved further. Researchers can also explore additional possible and optimized combinations of multi-omics data types and variants to strengthen predictive performance and improve generalization capability.

In addition, the optimized multi-omics combinations identified and validated in this research study should be further tested through practical implementation and real-world validation using a breast cancer molecular subtype classification experiment. Practical simulation using actual multi-omics breast cancer datasets and clinical classification environments is important to confirm the reliability, scalability, and effectiveness of the identified combinations in real diagnostic scenarios.

It is also important to evaluate the proposed models on other cancer-related datasets and in clinical settings to assess their adaptability and clinical usefulness. Future work should further focus on improving model interpretability so that researchers and healthcare professionals can better understand the classification decisions. Moreover, challenges such as missing data handling, high dimensionality, data imbalance, and computational complexity should also be addressed in future studies. Furthermore, future studies should also investigate the

computational cost and dimensionality challenges associated with integrating multiple high-dimensional multi-omics variants simultaneously. The integration of CNV, DNA methylation, mRNA, miRNA, and RPPA data can significantly increase data dimensionality, storage requirements, computational complexity, and model training time, particularly when implemented using standard computing resources. Therefore, future work should focus on developing computationally efficient frameworks, dimensionality reduction strategies, optimized feature selection methods, and scalable architectures to improve practical implementation feasibility and resource efficiency for large-scale multi-omics classification systems.

By following these directions, researchers can develop more reliable, interpretable, and accurate multi-omics classification systems for breast cancer molecular subtypes, which may contribute to better personalized treatment strategies for breast cancer patients as well as patients affected by other cancer types.

VIII. CONCLUSION

This research study investigated various combinations of multi-omics data types and variants for accurate breast cancer molecular subtype classification. It explained omics data, including mono-omics, multi-omics data types and variants, and also examined the biological reasons for combining the multi-omics data types, particularly genomics, epigenomics, transcriptomics, and proteomics, along with the justification of computational modelling studies that have performed breast cancer molecular subtypes classification using various combinations of multi-omics data. According to our observation, the majority of existing research does not choose combinations carefully based on both biology and computer science. To fill this gap, we proposed the best combination of CNV + DNA methylation + mRNA + miRNA + RPPA multi-omics variants to classify the breast cancer molecular subtypes because it captures all important layers of cancer biology and works well with classification models. Furthermore, this research study also recommended TCGA and UCSC Xena as the most suitable databases because they contain all the required multi-omics data types and variants with correct PAM50 subtype labels. This research study gives clear guidelines for researchers to build better, more accurate, and biologically meaningful classification models. Overall, the findings will help improve early diagnosis, personalized treatment, and patient outcomes in breast cancer.

ACKNOWLEDGMENT

This study was supported by the Fundamental Research Grant Scheme (FRGS/1/2023/ICT02/UTM/03/1) from the Malaysian Ministry of Higher Education.

DATA AVAILABILITY

GitHub	Repository	Link:
		https://github.com/sajidshah232/Breast-Cancer-Molecular-Subtypes-PAM50-Labels.git

REFERENCES

- [1] GLOBOCAN, "Global Cancer Observatory " 2025.
- [2] J. Kim et al., "Global patterns and trends in breast cancer incidence and mortality across 185 countries," *Nature Medicine*, pp. 1-9, 2025.

- [3] M. Akram, M. Iqbal, M. Daniyal, and A. U. Khan, "Awareness and current knowledge of breast cancer," *Biological research*, vol. 50, no. 1, p. 33, 2017.
- [4] A. R. Venkitaraman, "How do mutations affecting the breast cancer genes BRCA1 and BRCA2 cause cancer susceptibility?," *DNA repair*, vol. 81, p. 102668, 2019.
- [5] S. Shah, A. A. Samah, S. H. H. Madni, S. Z. M. Hashim, and M. Faheem, "Recent advancements in artificial intelligence-driven breast cancer molecular subtypes classification using multi-omics: A comprehensive review," *Engineering Applications of Artificial Intelligence*, vol. 170, p. 114237, 2026.
- [6] R. G. do Nascimento and K. M. Otoni, "Histological and molecular classification of breast cancer: what do we know?," *Mastology*, vol. 30, pp. 1-8, 2020.
- [7] X. Li et al., "MoGCN: a multi-omics integration method based on graph convolutional network for cancer subtype analysis," *Frontiers in Genetics*, vol. 13, p. 806842, 2022.
- [8] O. Yersal and S. Barutca, "Biological subtypes of breast cancer: Prognostic and therapeutic implications," *World journal of clinical oncology*, vol. 5, no. 3, p. 412, 2014.
- [9] J. J. Gao and S. M. Swain, "Luminal A breast cancer and molecular assays: a review," *The oncologist*, vol. 23, no. 5, pp. 556-565, 2018.
- [10] F. Ades et al., "Luminal B breast cancer: molecular characterization, clinical management, and future perspectives," *Journal of clinical oncology*, vol. 32, no. 25, pp. 2794-2803, 2014.
- [11] J. Y. So, J. Ohm, S. Lipkowitz, and L. Yang, "Triple negative breast cancer (TNBC): Non-genetic tumor heterogeneity and immune microenvironment: Emerging treatment options," *Pharmacology & therapeutics*, vol. 237, p. 108253, 2022.
- [12] U. Srivastava, S. Kanchan, M. Keshri, M. K. Gupta, and S. Singh, "Types of omics data: genomics, metagenomics, epigenomics, transcriptomics, proteomics, metabolomics, and phenomics," in *Integrative omics*: Elsevier, 2024, pp. 13-34.
- [13] S. M. Meystre, R. Gouripeddi, and A. V. Alekseyenko, "Molecular, Genetic, and Other Omics Data," in *Clinical Research Informatics*: Springer, 2023, pp. 309-328.
- [14] J. Lee, D. Y. Hyeon, and D. Hwang, "Single-cell multiomics: technologies and data analysis methods," *Experimental & Molecular Medicine*, vol. 52, no. 9, pp. 1428-1442, 2020.
- [15] F. Cristovao et al., "Investigating deep learning based breast cancer subtyping using pan-cancer and multi-omic data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 1, pp. 121-134, 2020.
- [16] J. Zeng, H. Cai, and T. Akutsu, "Breast cancer subtype by imbalanced omics data through a deep learning fusion model," in *Proceedings of the 2020 10th International Conference on Bioscience, Biochemistry and Bioinformatics*, 2020, pp. 78-83.
- [17] B. Li and S. Nabavi, "A multimodal graph neural network framework for cancer molecular subtype classification," *BMC bioinformatics*, vol. 25, no. 1, p. 27, 2024.
- [18] A. Peng, X. Mao, J. Zhong, S. Fan, and Y. Hu, "Single-cell multi-omics and its prospective application in cancer biology," *Proteomics*, vol. 20, no. 13, p. 1900271, 2020.
- [19] S. Graw et al., "Multi-omics data integration considerations and study design for biological systems and disease," *Molecular omics*, vol. 17, no. 2, pp. 170-185, 2021.
- [20] S. Shah et al., "Multi-Omics Integration Methods for AI-Based Breast Cancer Molecular Subtypes Classification," *International Journal of Advanced Computer Science & Applications*, vol. 17, no. 1, 2026.
- [21] G. T. Jung, K.-P. Kim, and K. Kim, "How to interpret and integrate multi-omics data at systems level," *Animal cells and systems*, vol. 24, no. 1, pp. 1-7, 2020.
- [22] A. M. Lesk, *Introduction to genomics*. Oxford University Press, 2017.
- [23] U. Testa, G. Castelli, and E. Pelosi, "Breast cancer: a molecularly heterogeneous disease needing subtype-specific treatments," *Medical Sciences*, vol. 8, no. 1, p. 18, 2020.
- [24] K. Gagnidze and D. W. Pfaff, "Epigenetic mechanisms: DNA methylation and histone protein modification," in *Neuroscience in the 21st century: from basic to clinical*: Springer, 2022, pp. 2677-2716.
- [25] R. Mathur et al., "Epigenetic factors in breast cancer therapy," *Frontiers in Genetics*, vol. 13, p. 886487, 2022.
- [26] J. Rajawat, "Transcriptomics," in *Omics Approaches, Technologies And Applications: Integrative Approaches For Understanding OMICS Data*: Springer, 2019, pp. 39-56.
- [27] Z. Zhu, L. Jiang, and X. Ding, "Advancing breast cancer heterogeneity analysis: insights from genomics, transcriptomics and proteomics at bulk and single-cell levels," *Cancers*, vol. 15, no. 16, p. 4164, 2023.
- [28] C. Monti, M. Zilocchi, I. Colugnati, and T. Alberio, "Proteomics turns functional," *Journal of proteomics*, vol. 198, pp. 36-44, 2019.
- [29] T. D. Veenstra, "Omics in systems biology: current progress and future outlook," *Proteomics*, vol. 21, no. 3-4, p. 2000235, 2021.
- [30] D. Lee, Y. Park, and S. Kim, "Towards multi-omics characterization of tumor heterogeneity: a comprehensive review of statistical and machine learning approaches," *Briefings in bioinformatics*, vol. 22, no. 3, p. bbaa188, 2021.
- [31] M. M. Ortiz and E. R. Andreckek, "Molecular characterization and landscape of breast cancer models from a multi-omics perspective," *Journal of mammary gland biology and neoplasia*, vol. 28, no. 1, p. 12, 2023.
- [32] G. Hovhannisyann, T. Harutyunyan, R. Aroutiounian, and T. Liehr, "DNA copy number variations as markers of mutagenic impact," *International journal of molecular sciences*, vol. 20, no. 19, p. 4723, 2019.
- [33] K. S. Allemailem et al., "Single nucleotide polymorphisms (SNPs) in prostate cancer: its implications in diagnostics and therapeutics," *American journal of translational research*, vol. 13, no. 4, p. 3868, 2021.
- [34] G. A. Dhar, S. Saha, P. Mitra, and R. Nag Chaudhuri, "DNA methylation and regulation of gene expression: Guardian of our health," *The Nucleus*, vol. 64, no. 3, pp. 259-270, 2021.
- [35] K. Holm et al., "Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns," *Breast cancer research*, vol. 12, no. 3, p. R36, 2010.
- [36] A. Szymiczek, A. Lone, and M. R. Akbari, "Molecular intrinsic versus clinical subtyping in breast cancer: A comprehensive review," *Clinical genetics*, vol. 99, no. 5, pp. 613-637, 2021.
- [37] K. Pierouli et al., "Long non-coding RNAs and microRNAs as regulators of stress in cancer," *Molecular medicine reports*, vol. 26, no. 6, p. 361, 2022.
- [38] K. Kalecky, R. Modisette, S. Pena, Y.-R. Cho, and J. Taube, "Integrative analysis of breast cancer profiles in TCGA by TNBC subgrouping reveals novel microRNA-specific clusters, including miR-17-92a, distinguishing basal-like 1 and basal-like 2 TNBC subtypes," *BMC cancer*, vol. 20, no. 1, p. 141, 2020.
- [39] D. Koboldt, R. Fulton, M. McLellan, H. Schmidt, J. Kalicki-Veizer, and J. McMichael, "Comprehensive molecular portraits of human breast tumours. *Nature [Internet]*. 2012; 490: 61-70," ed, 2021.
- [40] R. Duan et al., "Evaluation and comparison of multi-omics data integration methods for cancer subtyping," *PLoS computational biology*, vol. 17, no. 8, p. e1009224, 2021.
- [41] P. Goel and N. Shaikh, "Uncovering Clinically Relevant Breast Cancer Subtypes Biomarkers Using Integrative Bioinformatics and Machine Learning Approaches," *Biomarkers*, no. just-accepted, pp. 1-12, 2026.
- [42] O. Menyhárt and B. Györfy, "Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis," *Computational and structural biotechnology journal*, vol. 19, pp. 949-960, 2021.
- [43] S. Rakshit, I. Saha, S. S. Chakraborty, and D. Plewczynski, "Deep learning for integrated analysis of breast cancer subtype specific multi-omics data," in *TENCON 2018-2018 IEEE region 10 conference*, 2018, pp. 1917-1922: IEEE.
- [44] Y. Huang, P. Zeng, and C. Zhong, "Classifying breast cancer subtypes on multi-omics data via sparse canonical correlation analysis and deep learning," *BMC bioinformatics*, vol. 25, no. 1, p. 132, 2024.

- [45] S. Islam and M. N. Hasan, "Personalized graph feature-based multi-omics data integration for cancer subtype identification," arXiv preprint arXiv:2408.08832, 2024.
- [46] Q. Zhang et al., "AET-net: A framework for subtype classification based on the multi-omics data of breast cancer," *Molecular & Cellular Biomechanics*, vol. 21, no. 4, pp. 785-785, 2024.
- [47] P. Zeng, C. Huang, and Y. Huang, "DiffRS-net: A Novel Framework for Classifying Breast Cancer Subtypes on Multi-Omics Data," *Applied Sciences*, vol. 14, no. 7, p. 2728, 2024.
- [48] M. Tao et al., "Classifying breast cancer subtypes using multiple kernel learning based on omics data," *Genes*, vol. 10, no. 3, p. 200, 2019.
- [49] Y. Lin, W. Zhang, H. Cao, G. Li, and W. Du, "Classifying breast cancer subtypes using deep neural networks based on multi-omics data," *Genes*, vol. 11, no. 8, p. 888, 2020.
- [50] Y. Ma and J. Guan, "MOCS: a multi-omics data-based framework for cancer subtype classification," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2022, pp. 2853-2859: IEEE.
- [51] J. M. Choi and H. Chae, "moBRCA-net: a breast cancer subtype classification framework based on multi-omics attention neural networks," *BMC bioinformatics*, vol. 24, no. 1, p. 169, 2023.
- [52] H. Guo, X. Lv, Y. Li, and M. Li, "Attention-based GCN integrates multi-omics data for breast cancer subtype classification and patient-specific gene marker identification," *Briefings in Functional Genomics*, vol. 22, no. 5, pp. 463-474, 2023.
- [53] N. Li, F. Yang, Q. Zhang, Q. Li, X. Zhang, and J. Teng, "CautionGCN: Cancer Subtype Classification by Developing Causal Multi-Head Autoencoder and Graph Convolutional Network," in *2024 IEEE International Conference on Medical Artificial Intelligence (MedAI)*, 2024, pp. 626-634: IEEE.
- [54] S. Meshoul, A. Batouche, H. Shaiba, and S. AlBinali, "Explainable multi-class classification based on integrative feature selection for breast cancer subtyping," *Mathematics*, vol. 10, no. 22, p. 4271, 2022.
- [55] Y. Ren et al., "Classifying breast cancer using multi-view graph neural network based on multi-omics data," *Frontiers in Genetics*, vol. 15, p. 1363896, 2024.
- [56] A. M. Hassan, S. M. Naeem, M. A. Eldosoky, and M. S. Mabrouk, "Multi-omics-based Machine Learning for the Subtype Classification of Breast Cancer," *Arabian Journal for Science and Engineering*, pp. 1-14, 2024.
- [57] R. Lupat, R. Perera, S. Loi, and J. Li, "Moanna: multi-omics autoencoder-based neural network algorithm for predicting breast cancer subtypes," *Ieee Access*, vol. 11, pp. 10912-10924, 2023.