Multi-Omics Integration Methods for AI-Based Breast Cancer Molecular Subtypes Classification

Sajid Shah[1], Azurah A Samah[2]*, Siti Zaiton Mohd Hashim[3], Sarahani Harun[4],
Zuraini Binti Ali Shah[5], Farkhana Binti Muchtar[6], Syed Hamid Hussain Madni[7]
Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia [1, 2, 3, 5, 6]
Institute of Systems Biology, Universiti Kebangsaan Malaysia (UKM), 43600 Bangi, Selangor, Malaysia [4]
School of Electronics and Computer Science, University of Southampton Malaysia, 79100 Iskandar Puteri, Johor, Malaysia [7]

***Abstract*—Breast cancer is one of the most life-threatening and heterogeneous diseases. It contains various molecular subtypes, each subtypes have different characteristics, treatment outcomes, and prognosis. The proper integration of multi-omics data, including genomics, epigenomics, transcriptomics, and proteomics, is very important for enhancing the breast cancer molecular subtypes classification accuracy. Despite the increase in high-dimensional multi-omics data, selecting a suitable integration method for multi-omics data in breast cancer molecular subtypes classification still remains a crucial challenge. This study aims to evaluate and compare, and assess the effectiveness of the multi-omics data integration methods, including exploring the advantages, limitations, and highlighting their performance in terms of accuracy, interpretability, scalability, and biological relevance. Our findings indicate that transformer-based integration methods are increasingly adopted in recent studies due to their superior ability to handle high-dimensional heterogeneous data and capture intricate cross-omics relationships while providing interpretable insights. Additionally, we provide a comparative overview of existing models, discuss key trends over the years, and offer actionable guidance for method selection based on dataset characteristics and research objectives. Finally, we suggest future research directions, emphasizing hybrid deep learning frameworks, graph-based models, and attention mechanisms to enhance predictive accuracy and biological interpretability.**

Keywords*—Breast cancer; classification; integration methods; molecular subtypes; multi-omics

I. Introduction

Breast cancer is one of the major life-threatening and most prevalent cancers affecting women worldwide. Despite the advancements in the treatment and screening process, the heterogeneity of breast cancer still presents a major challenge [1]. The heterogeneity of breast cancer is not only morphological but also rooted deeply at a molecular level, which influences disease progression, therapeutic response and prognosis [2]. As a result, the precise and accurate classification of breast cancer into its molecular subtypes has gained significant attention nowadays.

Molecular subtyping of breast cancer, such as Luminal-A, Luminal-B, HER2-Enriched, Basal-Like and Normal-Like, has emerged from the gene expression profiling [3]. These molecular subtypes reflect diverse oncogenic mechanisms and clinical outcomes, necessitating precise and robust classification frameworks [4]. While traditional subtyping relies basically on mono-omics data, recent studies show that the integration of multi-omics data can significantly enhance the accuracy and reliability of breast cancer molecular subtypes classification [5].

Recently, high-throughput technologies such as multi-omics data have become increasingly accessible, including genomics, epigenomics, transcriptomics, and proteomics [6]. Each multi-omics data type captures different aspects of breast tumor heterogeneity, such as genomics, which reveals DNA-level alterations, epigenomics uncovers regulatory modifications, transcriptomics measures gene activity, and proteomics reflects functional outputs [7]. Integrating the diverse multi-omics data types provides a more comprehensive view of breast cancer, allowing for more accurate subtype classification of breast cancer [7].

Accurate and precise multi-omics data integration for breast cancer molecular subtypes classification is very important, as these molecular subtypes are defined through coordinated alteration across the multiple biological layers rather than by individual molecular signals [8]. The effective and accurate integration of multi-omics data variants allows the identification of meaningful biological patterns by integrating various variations in genes, gene expression profiles, protein level interaction, and epigenetic modification that help in the precise classification of breast cancer molecular subtypes [9]. The precise classification of breast cancer molecular subtypes is very important for enhancing the diagnostic accuracy, personalized treatment plans, and prognostic assessment [10]. Inappropriate or poor integration of multi-omics data in breast cancer molecular subtypes classification can lead to loss of important biological information, noise amplification, or even a misleading molecular subtypes classification of breast cancer [11]. Therefore, the selection of an appropriate method for integrating the multi-omics data is required to maintain the biological relevance as well as the data heterogeneity, which is a key challenge in multi-omics breast cancer molecular subtypes classification research.

Despite the rapid increase in multi-omics breast cancer subtypes classification research, there is still a lack of clear guidance regarding the appropriate multi-omics data integration methods for breast cancer molecular subtypes classification. Existing reviews have provided a broad overview of multi-omics integration methods in various domains. However, there is a lack of focus on multi-omics integration methods specifically for breast cancer molecular subtypes classification.

*Corresponding author.

To overcome this issue, this review study introduces a breast-cancer focused analytical evaluation of the multi-omics integration methods and compares the three most common integration methods, including concatenation-based, model-based, and transformation-based integration. By evaluating these integration methods using evaluation criteria, including biological interpretability, scalability, and robustness. This study provides a structured and comparative assessment of the integration methods for multi-omics data in breast cancer molecular subtypes classification.

This study evaluates the integration methods used by existing research, as well as provides actionable insights and methodological guidance for researchers in choosing a suitable multi-omics integration method for breast cancer molecular subtypes classification. The key contributions of this review are summarized as follows:

- To explore the existing multi-omics integration methods for breast cancer molecular subtypes classification.
- To provide a clear comparison and evaluation of integration methods for multi-omics data to classify the breast cancer molecular subtypes.
- To highlight the advantages and disadvantages of the integration methods for the multi-omics data, primarily for breast cancer molecular subtypes classification.
- To suggest future suggestions and directions for developing more accurate and useful integration models.

The rest of the study is organized as follows: Section II emphasizes related works, while Section III focuses on the existing integration methods for multi-omics data in breast cancer molecular subtypes classification. Section IV is about the analysis of multi-omics data integration methods for breast cancer molecular subtypes classification. Section V concentrates on the discussion, while Section VI emphasizes the suggestions and recommendations for future research. Lastly, Section VII is the conclusion of the study.

II. Related Works

Multi-omics integration methods are widely explored to enhance disease characterization and subtype classification by integrating various multi-omics data types and analyzing high-dimensional multi-omics data such as genomics, transcriptomics, epigenomics and proteomics. This section reviews the existing related works on the comparison of multi-omics integration methods for cancer subtypes classification and provides the necessary background and context for understanding current methodological trends and limitations.

Acharya and Mukhopadhyay [12] review various network-based methods based on ML for the integration of multi-omics data in precision oncology, such as clustering and factorization. The study also explores the challenges associated with multi-omics data integration, particularly in the context of network-based integration methods. It also determines the challenges in multi-omics data integration in precision oncology. Similarly, Menyhárt and Győrffy [8] discuss algorithmic frameworks and data integration methods for cancer subtypes, disease mechanism and diagnosis. The study sheds light on both single-omics and multi-omics data integration methods, including the bottom-up and top-down integration methods. It also explains various other integration methods, such as multivariate, statistical, network, fusion-based, similarity-based and correlation-based integration. The study also explores the application of multi-omics data integration methods in various multi-omics data types.

In the same way, Adossa et al. [13] introduce recent developments in the single-cell multi-omics and comprehensively review the existing data integration methods. The study particularly focuses on early, intermediate and late data integration methods, including exploring the conceptual principles and main characteristics of each data integration method. Also, the study determines various tools that are used for the integration methods in single-cell multi-omics data. Moreover, Vahabi and Michailidis [14] provide an overview of multi-omics data integration methods with different statistical approaches, focusing on unsupervised learning tasks, including disease onset prediction, biomarker discovery, disease subtyping and module discovery. The study mainly focuses on unsupervised multi-omics data integration methods, particularly regression-based integration methods, clustering and network-based integration methods. It also elaborates on the working of each data integration method in mono-omics data, such as genomics, epigenomics, and transcriptomics.

Additionally, Cai et al. [15] review ML-based multi-omics data integration methods for cancer, specifically early integration, intermediate integration and late integration methods. It mainly reviews the integration methods working process in different multi-omics datasets. The study also explores various multi-omics data integration tools and their applications. Also, the study determines the strength of each integration method in cancer subtypes classification. Moreover, Subramanian et al. [16] review the multi-omics integration methods, particularly network-based, fusion-based, similarity-based and neighborhood-based integration methods. The study also explores the application of these integration methods in the prediction of biomarkers, diagnostics and driver genes for diseases. It also provides an overview of portals for visualization and interpretation of multi-omics datasets.

Furthermore, Heo et al. [17] provide an overview of the rationale and concepts of multi-omics integration methods in cancer research. The study explores multi-omics integration methods and techniques used to help in the integration process and how the multi-omics integration methods are applied in different cancer subtypes classification, cancer pathophysiology, drug target discovery and clinical decision support. It also determines the latest findings and implications in cancer multi-omics studies. Though enough progress has been made and various research has been conducted on multi-omics integration methods across various cancer types and domains, as shown in Table I. However, there is no existing study that focuses on the multi-omics data integration methods for breast cancer molecular subtypes classification. To overcome this issue, this comparison study introduces a breast cancer-specific multi-omics integration methods that evaluate and compares the existing multi-omics integration methods for multi-omics data in breast cancer molecular subtypes classification.

TABLE I. EXISTING RELATED WORKS ON MULTI-OMICS INTEGRATION

Reference	Objective	Integration Methods				Limitations
		Concatenation-based Integration	Model-based Integration	Transformation-based Integration	Others	
Acharya and Mukhopadhyay [12]	Reviews ML-based integration methods for multi-omics data	✗	✗	✗	✓	Only focus on network-based multi-omics integration methods in precision oncology.
Menyhárt and Győrffy [8]	Explore the multi-omics integration methods with application in cancer research	✗	✗	✗	✓	Focus only on multivariate, statistical, network, fusion-based, similarity-based, and correlation-based integration methods for tumor subtypes, prognosis and diagnosis only.
Adossa, et al. [13]	Explore computational strategies for single-cell multi-omics integration methods	✓	✓	✓	✗	Focus on early, intermediate and late integration methods only for single-cell multi-omics data.
Vahabi and Michailidis [14]	Review unsupervised multi-omics data integration methods for various disease subtyping	✗	✗	✗	✓	Focus only on regression-based, clustering and network-based integration methods in different disease subtyping.
Cai, et al. [15]	Explore the ML-based multi-omics integration methods	✓	✓	✓	✗	Focus only on early integration, intermediate integration and late integration methods in general cancer research.
Subramanian, et al. [16]	Review multi-omics data integration methods and their application in the cancer domain	✗	✗	✗	✓	Focus only on network-based, fusion-based, similarity-based and neighborhood-based integration methods in general cancer.
Heo, et al. [17]	Explore multi-omics data integration methods in cancer research.	✗	✗	✗	✓	Focus only on computation integration methods for cancer subtypes.
Current Study	Compare the multi-omics integration methods for breast cancer molecular subtypes classification	✓	✓	✓	✗	Focus specifically on multi-omics integration methods for breast cancer molecular subtypes classification.

III. INTEGRATION METHODS FOR MULTI-OMICS DATA IN BREAST CANCER MOLECULAR SUBTYPES

The integration of multi-omics data is highly significant as it integrates various biological data types and variants, including genomics (Copy Number Alteration (CNA), Copy Number Variation (CNV), Single Nucleotide Polymorphism (SNP)), epigenomics (DNA-Methylation (DNA-Methyl)), transcriptomics (Messenger RNA (mRNA), MicroRNA (miRNA)) and proteomics (Reverse Phase Protein Array (RPPA)). for evaluating and understanding the mechanism and structure of breast cancer disease [17]. While mono-omics data offer limited information about breast cancer [18]. Multi-omics data integration captures the interaction and relationship among breast cancer molecular subtypes at the molecular level and provides a deeper understanding of breast cancer cellular function [19]. With the help of accurate and proper multi-omics data integration, healthcare professionals can explore the aggressiveness and behavior of breast cancer by classifying molecular subtypes, and also provide information about the potential causes behind breast cancer. Proper integration of multi-omics data variants can also assist in understanding the flow from one omics data layer to another omics layer [16].

One of the main tasks of the multi-omics data in breast cancer is the early classification of molecular subtypes, for instance, classifying the breast cancer based on molecular level, such as Luminal-A, Luminal-B, HER2-Enriched, Basal-Like and Normal-Like [20]. Breast cancer patients' treatment depends on their specific molecular subtype. By properly utilizing the multi-omics data integration methods and considering various levels of the multi-omics data variants, breast cancer molecular subtypes can be accurately classified at the molecular level so that the patients can get better treatment on time [20].

To gain the full potential of multi-omics data for accurate classification of breast cancer molecular subtypes, effective integration methods are important to capture the behavior and aggressiveness of breast cancer based on its molecular level through the integration of various multi-omics data variants. Typically, the three most common and prominent methods are used for multi-omics data integration, including the concatenation-based integration, model-based integration and transformation-based integration, which are discussed in subsections.

A. Concatenation-Based Integration (CBI)

Concatenation-Based Integration (CBI), also known as early integration, is one of the most direct integration methods, as shown in Fig. 1. It directly combines various multi-omics data variants and creates a joint matrix. The CBI method usually does not require comprehensive feature engineering or extensive preprocessing, as all omics features are added to a large input matrix. The joint matrix is delivered into the model for the

classification tasks. In this integration method, a joint matrix is formed by stacking the feature vectors together from various datasets, which creates a single and larger feature vector for each dataset.

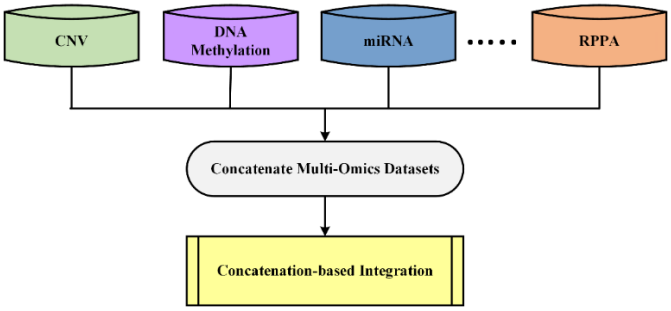

Fig. 1. Concatenation-based integration method.

For the breast cancer molecular subtypes classification, Cristovao et al. [21] use CBI for integrating mRNA and miRNA to train the Feed Forward Neural Network (FFNN) model for classifying breast cancer molecular subtypes. CBI assists the models in learning omics information at once and also captures the non-linear pattern. However, it leads the model to high-dimensional input and affects the model's performance. It indicates that CBI struggles in modelling and capturing the complex relationships and interactions among various omics variants and also leads to reduced interpretability and potential overfitting when labeled data are limited. In the same way, Zeng et al. [22] utilize CNN5 and integrate CNV and mRNA using CBI, where multi-omics data are merged at the input feature level and processed jointly by the CNN5. According to the study, CBI is straightforward, efficient, and avoids complex pre-processing or feature transformation and it is suitable for a lightweight DL model. It affects learning because concatenated data emphasize dominant subtypes (e.g., Luminal A). Hence, CBI omics specific detail can be lost, and also high dimensionality challenges can occur and cannot explicitly capture complex relationships between omics variants.

Furthermore, Rakshit et al. [23] apply SVM and integrate DNA-Methyl, mRNA and miRNA using CBI into a single large dataset. The integrated data is directly processed using Stacked Autoencoder (SAE) to reduce dimensionality before being fed into the classification model. CBI is straightforward and allows unified processing of high-dimensional multi-omics data, but it loses omics-specific signals, introduces noise from irrelevant features, and does not provide a solution to weigh the contributions of different multi-omics data variants.

B. Model-Based Integration (MBI)

Model-Based Integration (MBI), also known as intermediate integration, is a more organized, structured and adaptive integration method as shown in Fig. 2. In the model-based integration method, all multi-omics data variants are first processed by their module and converted into an intermediate form before a final model is created using various intermediate models. Hence, the integration happens to determine the various multi-omics data associated with a specific disease.

For the breast cancer molecular subtypes classification, Lin, et al. [24] propose DeepMO and integrate mRNA, DNA-Methyl and CNV through MBI and their own feature extraction subnetwork, and then fuse the learned high-level patterns into a single integrated representation for classification. The MBI method allows the model to capture complex cross-omics relationships more effectively than the CBI method. However, MBI is sensitive to class imbalance, relies on default hyperparameters, and has higher computational demands due to large feature sets. Similarly, Choi and Chae [25] propose moBRCA-net and integrate DNA-Methyl and miRNA by MBI through separate self-attention modules to learn the importance of individual features. The high-level data are then fused into a single joint representation, which is used by a shared classification network to classify subtypes. MBI preserves biological relationships between omics variants, reduces dimensionality, and enhances interpretability. Hence, MBI relies on feature selection, misses some cross-omics dependencies, and has higher computational requirements.

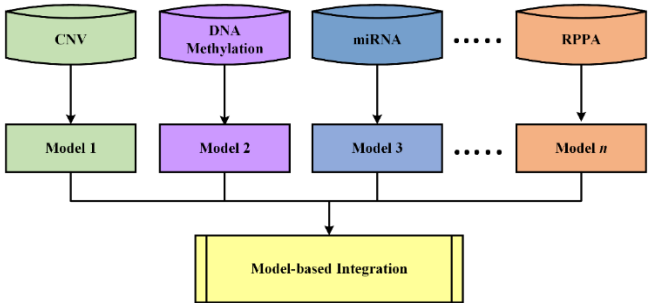

Fig. 2. Model-based integration method.

In the same way, Lupat et al. [26] develop MOANNA and integrate Gene-Exp, CNV and Somatic Mutation through MBI using a semi-supervised AE, transforming the high-dimensional input into a 64-dimensional latent representation. This integrated feature vector is then used by multiple supervised classifiers to classify breast cancer molecular cancer subtypes. By learning compact and informative features, MBI improves classification performance compared to other methods. However, it requires large training datasets and may still face challenges with high-dimensional data. Moreover, Guo et al. [27] propose AGCN and integrate CNV, DNA-Methyl and mRNA through MBI into Protein-Protein Interaction (PPI) network. MBI captures complex cross-omics relationships and gene-level structural information before classification, producing a joint representation for breast cancer molecular subtype classification. Hence, MBI depends on accurate graph construction, which is computationally intensive and less flexible for non-graph data.

Further, Li and Nabavi [28] utilize GNN and CNV, mRNA, and miRNA into a single supra-graph using MBI, where nodes represent genes and edges capture intra- and inter-omics interactions. GNN layers learn hidden representations across the entire graph, and combined with features from a parallel network to form a final integrated vector for classification. The MBI method captures complex biological relationships and improves representation. Hence, it depends on predefined graph connections and requires high computational effort. Furthermore, Tao et al. [29] propose SMOMKL and integrate CNV, DNA-Methyl and mRNA using MBI and transform into

a kernel matrix representing patient similarities. These kernels are combined within the MKL model using learned weights, allowing the integration to occur at the similarity level rather than by raw data concatenation. MBI effectively handles high-dimensionality and improves predictive performance; however, it can be computationally intensive and depends on proper kernel selection.

Moreover, Ren et al. [30] propose MVGNN and integrate DNA-Methyl, mRNA and miRNA using MBI, which is first separately processed through individual GCN branches to extract omics-specific patterns. These patterns are then fused using a multi-view attention mechanism to create a single integrated representation that captures complex cross-omics relationships. MBI effectively models inter-omics correlations and highlights important features. Hence, it requires high-quality graphs and can be computationally demanding. Similarly, Zeng et al. [31] utilize DiffRS-net and use MBI for integrating DNA-Methyl mRNA and miRNA separately to select important features, rather than concatenating raw data directly. The correlated features are fused within a DL using an attention mechanism, which assigns higher importance to more informative features before classification. MBI reduces data dimensionality and better captures cross-omics relationships. Hence, it mainly focuses on linear correlations and requires careful parameter tuning.

Islam et al. [32] apply DNN and use MBI for integrating CAN, Gene-Exp and mRNA. In the main classification model, each omics variant is first processed separately through its own neural network branch to learn high-level features, and these learned features are then merged at an intermediate stage before making the final classification. MBI design allows the model to capture omics-specific patterns first and then learn complementary information from both data variants together. The integrated model achieved better classification performance than models using only a single omics variant. However, it requires a relatively large amount of data and may face challenges related to high dimensionality. In the same way, Li et al. [33] introduce CautionGCN and apply MBI for integrating CNV, DNA-Methyl and RNAseq. By using MBI, each omics variant is first processed through a causal multi-head AE to extract meaningful and low-dimensional features rather than directly concatenating raw data. SNF is then applied to integrate cross-omics relationships at the network level, and a graph convolutional network jointly learns from the fused similarity network and the extracted features to perform classification. MBI design helps capture complex multi-omics interactions and improves robustness to high-dimensional and imbalanced data. Hence, it introduces higher computational complexity and depends on accurate similarity modeling.

C. Transformation-Based Integration (TBI)

Transformation-Based Integration (TBI) is also known as the late integration method. In the TBI method, each multi-omics data layer is transformed into a representable, comparable latent space, such as a graph, kernel similarity matrices, or core matrix. Each multi-omics dataset is transformed into an intermediate form before integrating all the data into a joint transformation. Once the data is transformed, all the multi-omics data layers are integrated into the model using a specific technique, as shown in Fig. 3.

For the breast cancer molecular subtypes classification, Ma and Guan [34] present MOCSC and apply TBI for integrating mRNA, miRNA, DNA-Methyl and CNV. By using TBI, each omics variant is processed separately, where features are first learned using SSDAE and then used by an individual NN to produce separate classification. These omics-specific classifications are finally combined using a VCDN to generate the final classification result. TBI-based fusion improves robustness by integrating information from multiple omics views, but it depends on the quality of each omics-specific model. Similarly, Meshoul et al. [35] present ET and integrate CNV, DNA-Methyl and RNA through TBI, and select features separately for each omics type before combining them. The study showed that the TBI improves the feature relevance and explainability, but it increases complexity and depends heavily on the quality of feature selection. In the same way, Huang et al. [18] propose DSCCN and integrate DNA-Methyl and mRNA using TBI, where each variant is processed separately through its own network to generate independent classification, which are combined at the decision level to produce the final classification. TBI method captures complementary information from each omics type and avoids problems with high-dimensional concatenation. Hence, it may be less effective for underrepresented subtypes and could benefit from additional data or extensions to other omics types.

Further, Zhang et al. [36] propose AET-net and integrate Gene-Exp and DNA-Methyl. Each omics variant is first processed separately and then fused into a shared latent representation using an AE, capturing essential features from both omics variants. This fused feature vector is then processed by a classifier for final classification.

TBI reduces dimensionality, handles high-dimensionality of the data effectively, and models complex dependencies. Hence, it requires careful tuning and overfitting with small datasets. Lastly, Li et al. [37] present MoGCN and integrate CNV, RNAseq and RPPA using TBI. The multi-omics data is first processed through a separate AE to learn a shared latent representation. It then applies SNF to create a unified PSN, and finally, a GCN performs classification using both the joint feature matrix and the fused network. TBI captures nonlinear relationships, improves stability, and achieves high accuracy, though it requires substantial computation and may be sensitive to noise in high-dimensional data.

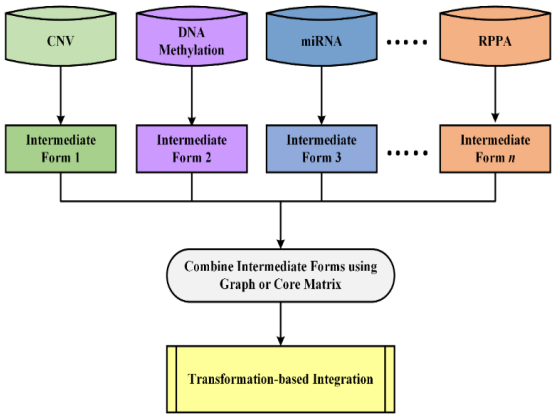

Fig. 3. Transformation-based integration method.

TABLE II. EXISTING STUDIES ON MULTI-OMICS DATA INTEGRATION METHODS FOR BREAST CANCER MOLECULAR SUBTYPES CLASSIFICATION

References	Objective	Proposed Models	Integration Methods	Accuracy	Performance Metrics	Advantages of Integration Method	Limitations of Integration Method
Cristovao, et al. [21]	Breast Cancer Molecular Subtypes Classification	•LR & FFNN	CBI	94.00%	• Accuracy • Precision • Recall	• Effectively handles high-dimensional multi-omics data. • Captures non-linear relationships across omics variants.	• Lacks interpretability in latent representations. • Sensitive to hyperparameter tuning. • May overfit with limited labeled data.
Lin, et al. [24]	Breast Cancer Molecular Subtypes Classification	• DeepMO	MBI	78.20%	• Accuracy • AUC	• Outperforms single-omics-based integration methods. • Handles high-dimensional data effectively.	• Sensitive to class imbalance. • Computational complexity from large feature sets.
Choi and Chae [25]	Breast Cancer Molecular Subtypes Classification	• moBRCA-net	MBI	90.90%	• Accuracy • F1-Score • Precision • Recall	• Enhances performance by preserving biological relationships between omics variants • Learn feature importance effectively. • Improving interpretability and subtype distinction.	• Relies on feature selection to manage high dimensionality and computational costs. • Potentially missing some omics dependencies. • Does not explicitly model cross-omics dependencies, which could be extended in future work.
Islam, et al. [32]	Breast Cancer Molecular Subtypes Classification	• DNN	MBI	79.20%	• Accuracy	• Effectively handles high integration through feature learning • Improve subtype classification over.	• Can exacerbate the curse of dimensionality. • Requires large training data to avoid overfitting.
Lupat, et al. [26]	Breast Cancer Molecular Subtypes Classification	• MOANNA	MBI	85.60%	• F1-Score • Precision • Recall	• Handles high-dimensional multi-omics effectively. • Improves generalization.	• Potential overfitting in large datasets. • Reconstruction loss may not align perfectly with classification.
Ma and Guan [34]	Breast Cancer Molecular Subtypes Classification	• MOCSC	TBI	95.00%	• Accuracy • AUROC • Macro F1 Value • Weighted F1 Value	• Avoids data inconsistency by fusing at the decision level. • Captures cross-omics correlations effectively.	• Dependent on individual classifier quality. • May propagate errors from weak omics models.
Meshoul, et al. [35]	Breast Cancer Molecular Subtypes Classification	• ET	TBI	84.50%	• Accuracy • Accuracy • AUROC • F1-Score • Macro F1 Value • Precision • Recall • ROC_AUC • Weighted F1 Value	• Improves feature relevance across omics. • Enhances explainability. • flexible with early/late schemes for different data handling.	• A multi-stage process increases complexity. • Dependent on the feature selection quality. • Computationally expensive. • Does not work well with smaller datasets.
Zeng, et al. [22]	Breast Cancer Molecular Subtypes Classification	• CNN5	CBI	90.02%	• F1-Score • Precision • Recall	• Handles imbalanced data via weighted loss; direct learning on combined features.	• Potential loss of omics-specific patterns. • Exacerbates dimensionality issues. • Lacks explicit handling of inter-omics correlations.
Guo, et al. [27]	Breast Cancer Molecular Subtypes Classification	• AGCN Variants: SEGCN & cAGCN	MBI	89.42%	• Accuracy • AUC • MCC	• Effectively captures relational dependencies between genes across omics layers. • Enables biologically meaningful fusion by	• Relies on accurate graph construction, which may introduce bias if prior knowledge is incomplete. • Computationally intensive for large-scale graphs;

References	Objective	Proposed Models	Integration Methods	Accuracy	Performance Metrics	Advantages of Integration Method	Limitations of Integration Method
						incorporating prior structural relationships.	• Lack of flexibility in handling non-graphical data representations.
Huang, et al. [18]	Breast Cancer Molecular Subtypes Classification	• DSCCN	TBI	90.60%	• Accuracy • AUC • F1-Score	• Successfully identifies inter-omics associations to capture complementary information and reduce data heterogeneity. • Outperforms concatenation-based, ensemble-based, and knowledge-driven methods.	• Data imbalance in breast cancer datasets reduces accuracy for minority subtypes. • Future extensions are needed for non-coding omics and data augmentation techniques.
Li, et al. [37]	Breast Cancer Molecular Subtypes Classification	• MoGCN	TBI	89.80%	• Accuracy • F1-Score	• Captures nonlinear relationships and improves stability effectively.	• Relies on unsupervised AE and SNF. • Intensity for high-dimensional data. • Potential noise interference in multi-omics integration.
Li, et al. [33]	Breast Cancer Molecular Subtypes Classification	• CautionGCN	MBI	89.18%	• Accuracy • F1-Score • Precision • Recall	• Reduces bias and enhances robustness to imbalanced data. • Captures multi-omics interactions via graphs.	• Complexity in causal inference • May introduce assumptions. • Dependent on accurate similarity networks. • Computationally demanding.
Li and Nabavi[28]	Breast Cancer Molecular Subtypes Classification	• Multi-Omic GNN	MBI	86.40%	• Accuracy • F1-Score	• Incorporates biological knowledge into graph structures for better representation	• Dependent on predefined graph connections, which may miss unknown relations. • Higher complexity in graph construction. • Potential bias from incomplete biological knowledge in edges.
Tao, et al. [29]	Breast Cancer Molecular Subtypes Classification	• SMOMKL	MBI	87.00%	• Accuracy • AUC	• Learns optimal kernel combinations for omics. • Improves predictive power through kernel fusion.	• Kernel selection is critical and may miss non-linear interactions. • Computationally expensive for large kernels.
Zhang, et al. [36]	Breast Cancer Molecular Subtypes Classification	• AET-net	TBI	90.00%	• Accuracy • F1-Score	• Reduces dimensionality while capturing complex dependencies through Transformation attention, leading to effective multi-omics integration.	• Sensitive to hyperparameter tuning, which can affect integration quality. • Potential for overfitting in smaller datasets.
Ren, et al. [30]	Breast Cancer Molecular Subtypes Classification	• MVGNN	MBI	91.80%	• Accuracy • AUC • F1-Score • Precision • Sensitivity • Specificity	• Captures inter-omics correlations via graphs; attention weights important features. • Handles heterogeneous data structures.	• Relies on graph quality and similarity networks • Computational overhead in multi-view fusion.
Zeng, et al. [31]	Breast Cancer Molecular Subtypes Classification	• DiffRS-net	MBI	91.30%	• Accuracy • F1-Score • Precision • Recall	• Detects correlations across multiple omics variants. • Reduces dimensionality while preserving associations • Enhances interpretability.	• Assumes linear relationships, may miss non-linear interactions. • Sparsity tuning is sensitive. • Lacks dynamic weighting of views.

References	Objective	Proposed Models	Integration Methods	Accuracy	Performance Metrics	Advantages of Integration Method	Limitations of Integration Method
Rakshit, et al. [23]	Breast Cancer Molecular Subtypes Classification	• SVM	CBI	93.50%	• Accuracy	• Straightforward combination of heterogeneous data • Enables unified processing in a deep model • Reduces the preprocessing complexity.	• May miss omics-specific signals in high-dimensional space. • Increases the risk of noise from irrelevant features. • Lacks a mechanism to weigh different omics contributions.

Table II shows the comparison of existing studies on multi-omics integration, highlighting the utilization of integration methods from a CBI to MBI and TBI. It summarizes the objectives, proposed models for the classification, the type of integration method used, and their reported performance metrics. and highlighting the advantages and limitations of each integration method used in breast cancer molecular subtypes classification.

IV. Analysis of Multi-Omics Data Integration Methods for Breast Cancer Molecular Subtypes Classification

All three multi-omics data integration methods are used by the existing studies for breast cancer molecular subtypes classification. Each integration method has its own characteristics, strengths and limits, as shown in Table III.

Table III provides a practical guideline for selecting a suitable integration method for multi-omics data in breast cancer molecular subtypes classification, based on aspects such as data type, sample size, interpretability, and computational requirements from the observation of the existing studies. According to the findings from the existing studies, CBI is appropriate for both mono-omics and multi-omics data, particularly when the dataset is small to medium in size. CBI is straightforward to implement, computationally light, and highly interpretable, making it ideal when simplicity and clarity are important. It works best for combining a limited number of omics variants, where the primary goal is to preserve interpretability and generate easily understandable insights. While MBI is specifically designed for medium to large multi-omics datasets and is appropriate for medium interpretability and computational complexity. Lastly, TBI provides a balance between predictive performance and interpretability, handling high-dimensional omics variants more effectively. It is especially designed for large multi-omics datasets and is suitable for high-dimensional data. It can easily capture more complex interactions and relationships among different omics data types and variants, and is more appropriate for a graph-based architecture.

TABLE III. Comparison of Multi-Omics Data Integration Methods for Breast Cancer Molecular Subtypes Classification

Integration Methods	Datasets Suitability	Interpretability	Computational Complexity	Strenghts	Weaknesses
Concatenation-Based Integration (CBI): Combines all omics data into one large input matrix without much preprocessing.	Small to Medium	High	Low	• Simple, fast and straightforward. • No preprocessing is required as all omics features are integrated into a large input matrix. • Uses a full feature set. • Handle all omics data types in a single joint matrix. • Produces continuous data, which makes it easier for various ML models to analyze.	• Ignores inter-omics relationships. • Increasing the number of features together increases dimensionality. • Sensitive to noise/missing data, high-dimensional.
Model-Based Integration (MBI): Learns features separately using sub-models, then merges the learned representation	Medium to Large	Medium	Medium	• Captures omics-specific relationships and interactions.	• Complex to design. • Cannot handle heterogeneous data. • Perform feature selection before integrating omics data, which can lead to the loss of important features. • Require careful hypermeter tuning for each sub-model.
Transformation-Based Integration (TBI): Converts omics data into a shared latent space (e.g., graphs), then integrates.	Large	Low to Medium	High	• Suitable for high-dimensional data. • Capture more complex interactions and relationships among different omics data types. • Most common integration data method. • Easily deal with data heterogeneity. • Suitable for graph-based architecture. • Include the statistical frameworks, such as core-based integration and graph-based semi-supervised learning, for developing a model.	• Transformation can cause information loss, depending on design quality. • Challenges in transforming data into an intermediate form. • Kernel methods are computationally extensive.

Overall, Table III serves as a decision-support guideline for researchers, guiding them in choosing the most appropriate integration methods for omics data, particularly in breast cancer subtypes classification based on the specific characteristics of their dataset, research objectives, and available computational resources.

V. Discussion

Existing studies have increasingly explored CBI, MBI, and TBI methods for multi-omics data analysis in breast cancer molecular subtypes classification. Each integration method has its own strengths and limitations in capturing inter-omics relationships, managing data dimensionality, and improving classification performance.

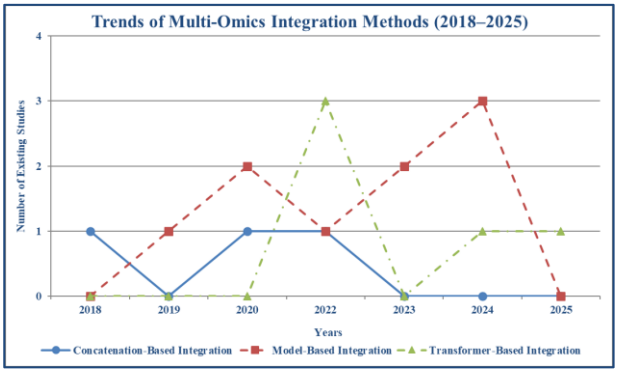

Fig. 4. Trends of multi-omics integration methods in breast cancer molecular subtypes classification.

Fig. 4 shows trends in the use of multi-omics integration methods for breast cancer molecular subtype classification from 2018 to 2025. Based on the graph, it is clear that early studies predominantly use CBI methods due to their simplicity and ease of implementation, while MBI methods became increasingly popular from 2019 onward, reflecting their ability to capture inter-omics interactions and improve predictive performance. Additionally, TBI emerged in 2022 and shows a growing adoption trend, highlighting its superior capacity to handle high-dimensional heterogeneous data, model complex relationships across omics layers, and provide attention-driven interpretability. This finding shows a clear shift in the field toward leveraging deep learning architectures for more accurate and biologically informative breast cancer molecular subtype classification, indicating that TBI methods are becoming a preferred approach for multi-omics integration in recent studies.

Fig. 5 demonstrates the comparison of CBI, MBI, and TBI integration methods for multi-omics data in breast cancer molecular subtypes classification based on the accuracy of the models. For CBI, the minimum accuracy 90.20% is shown by CNN5 with CBI. While SVM shows a medium accuracy of 93.50% with CBI, and FFNN achieves a maximum accuracy of 94.00% with CBI. Models with CBI show stable and strong accuracy, but their performance mainly depends on integrating different multi-omics variants into a joint matrix directly, which may increase data dimensionality and complexity. For MBI, the performance varies more widely. DeepMo and DNN with MBI shows minimum accuracy of 78.20%, indicating weaker performance. Models such as MOANNA, MoGNN, and SMOMKL with MBI show medium accuracy of 85.60%, 86.40% and 87.00% demonstrate more good performance. The models, including MVGNN, DiffRS-net, and moBRCA-net with MBI, achieve significant accuracy of 91.80%, 91.30%, and 90.90%, showing that MBI can perform well when advanced models are used. For TBI, MoGCN attained 89.80% with TBI represents the minimum accuracy, while AET-net and DSCCN with TBI achieve accuracy of 90.00% and 90.60% show medium performance. While the best results are achieved by ET and MOCSC, with TBI attaining an accuracy of 95.00%, which is the maximum accuracy in the entire comparison.

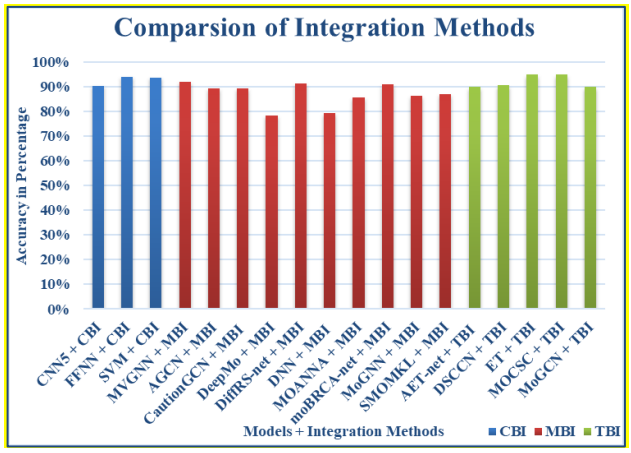

Fig. 5. Comparison of multi-omics integration methods in breast cancer molecular subtypes classification.

Overall, TBI is the strongest and most effective integration method, as it consistently achieves high accuracy and includes the best-performing models. This indicates that transforming multi-omics data into a shared representation before integration is more effective for breast cancer molecular subtypes classification than direct concatenation or model-level integration.

As per the analysis of the existing studies, the CBI method is widely used due to its simplicity and ease of implementation. However, this integration method completely ignores the complex interaction and relationships among the multi-omics data types, which leads to the potential loss of the most relevant information. Additionally, the CBI method suffers from high dimensionality, increases computational complexity, overfitting risks, and is sensitive to missing values or noise. Most of the existing studies highlight that the concatenation-based method is easy to implement. However, it does not fully capture the heterogeneity and hierarchical nature of the multi-omics data. Furthermore, another main disadvantage of this method is that it frequently increases the number of features in the joint matrix, which further increases the dimensionality of the multi-omics data.

MBI provides a more adaptive and structured approach. It allows capturing the omics-specific interactions and relationships, thereby offering more detailed biological insights. However, the model-based integration method is relatively complex to design and also computationally expensive, particularly when dealing with high-dimensional multi-omics data. Furthermore, domain-specific modeling and hypermeter tuning are required for model-based integration, which further increases the proposed model complexity.

Lastly, TBI has been less frequently used in the existing studies, particularly for breast cancer molecular subtypes classification. But it has several advantages, including the potential to handle high-dimensional heterogeneous data types, capture both inter-omics and intra-omics interaction and relationships and provide more structured insights through the graph and attention mechanism. Although the TBI is not used frequently in breast cancer molecular subtypes classification according to our observation, it has gained significant attention in other cancer domains because of showing superior performance in modeling long-range dependencies and dynamic relationships between features. TBI allows modular processing of each multi-omics data type while also enabling meaningful cross-model interaction, which makes it suitable for integrative analysis.

While the CBI method remains popular due to its simplicity, the MBI offers enhanced interpretability and biological relevance, and the TBI method has great potential for next-generation multi-omics data analysis. As the multi-omics research continues to evolve, combining the strengths of the multi-omics integration methods could lead to a more robust and accurate classification of breast cancer molecular subtypes, eventually contributing to improve the diagnosis, prognosis and personalized treatment strategies.

VI. Suggestions and Future Directions

Based on the analysis of existing multi-omics integration methods for breast cancer molecular subtype classification, several opportunities for future research exist. While current methods, including CBI, MBI, and TBI, have demonstrated significant progress, challenges remain in balancing accuracy, interpretability, scalability, and robustness to incomplete data. The following section outlines key directions and practical suggestions to advance the development of more effective and reliable multi-omics integration strategies.

A. Development of Hybrid Integration Methods

Future studies should explore hybrid approaches that combine the strengths of CBI, MBI, and TBI integration methods. Such integration methods could leverage the simplicity and efficiency of CBI, the interpretability of MBI, and the high-dimensional, attention-driven capabilities of TBI. Hybrid models may provide improved classification accuracy while maintaining biological interpretability and scalability.

B. Enhancing TBI Models

TBI has shown great promise in capturing complex intra- and inter-omics relationships. Future work should focus on optimizing these methods for efficiency and robustness, including techniques to handle class imbalance, reduce computational costs, and incorporate multi-modal biological knowledge. This could enable TBI methods to generalize better across diverse multi-omics datasets and multiple omics variants.

C. Robust Integration for Missing or Partial Omics Data

Many current integration methods assume complete multi-omics data variants, which is often not feasible in real-world studies. Future research should develop integration methods, such as multi-view transformers or imputation-enhanced models, to robustly integrate incomplete or partially missing omics data. This will increase the applicability of multi-omics integration in clinical and large-scale cohort studies.

D. Focus on Interpretability and Biological Insight

While TBI and MBI methods provide high accuracy, their interpretability can still be limited. Future research should emphasize methods that not only classify breast cancer molecular subtypes accurately but also provide interpretable outputs, highlighting key molecular features, pathways, and cross-omics interactions. Attention mechanisms, feature selection, or graph-based interpretability techniques can support this goal.

E. Scalability to High-Dimensional Multi-Omics Datasets

As multi-omics datasets grow in size and complexity, there is a need for methods that can scale efficiently without sacrificing performance. Future studies should focus on dimensionality reduction, computational optimizations, and adaptive integration methods that can handle large, heterogeneous datasets while maintaining predictive accuracy.

VII. Conclusion

In terms of multi-omics data, the integration methods include concatenation-based, model-based, and transformation-based integration. These integration methods allow the discovery of hidden patterns in the multi-omics data. This study evaluates the multi-omics data integration methods for classifying the breast cancer molecular subtypes. The concatenation-based integration is straightforward, fast, and simple, and it also does not require preprocessing. However, CBI ignores the inter-omics interaction and also increases the number of features in the joint matrix. Model-based integration captures the interaction and relationship among multi-omics data types. However, MBI is more complex to design and it also cannot handle the high-dimensional heterogeneous data. The transformation-based integration appears to be a more promising alternative, particularly for the multi-omics data to classify the breast cancer molecular subtypes, as it not only handles the high-dimensional data heterogeneity effectively but also captures more complex interactions. Each method has distinct advantages and is suitable for different research scenarios. Simple or concatenation-based models are recommended for small datasets; transformation-based or model-based integration methods support high interpretability. Furthermore, transformation-based integration methods also handle large heterogeneous datasets, and transformation-based multi-view models are effective for datasets with missing modalities.

Overall, the findings show that each multi-omics data integration method for breast cancer molecular subtypes classification provides various advantages and limitations. However, the selection of the multi-omics integration methods totally varies on the combination of multi-omics data types, variants and nature of the proposed model for the classification of breast cancer molecular subtypes. Hence, these insights provide practical guidance for selecting appropriate multi-omics integration methods based on study objectives, data characteristics, and computational constraints, offering structured guidelines for future AI-driven research in breast cancer subtype classification.

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.

REFERENCES

[1] A. G. Waks and E. P. Winer, "Breast cancer treatment: a review," Jama, vol. 321, no. 3, pp. 288-300, 2019.

[2] M. Akram, M. Iqbal, M. Daniyal, and A. U. Khan, "Awareness and current knowledge of breast cancer," Biological research, vol. 50, pp. 1-23, 2017.

[3] S. Ochoa, G. de Anda-Jáuregui, and E. Hernández-Lemus, "Multi-omic regulation of the PAM50 gene signature in breast cancer molecular subtypes," Frontiers in oncology, vol. 10, p. 845, 2020.

[4] R. R. Bastien et al., "PAM50 breast cancer subtyping by RT-qPCR and concordance with standard clinical molecular markers," BMC medical genomics, vol. 5, pp. 1-12, 2012.

[5] C. Shi et al., "Multi-omics integration analysis: tools and applications in environmental toxicology," Environmental Pollution, p. 124675, 2024.

[6] X. Dai and L. Shen, "Advances and trends in omics technology development," Frontiers in Medicine, vol. 9, p. 911861, 2022.

[7] M. K. Bharti, D. Chandra, R. Siddique, K. Ranjan, and P. Kumar, "Recent advancement in high-throughput "omics" technologies," in Current omics advancement in plant abiotic stress biology: Elsevier, 2024, pp. 343-355.

[8] O. Menyhárt and B. Győrffy, "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.

[9] C. Zhang et al., "Advancing precision and personalized breast cancer treatment through multi-omics technologies," American Journal of Cancer Research, vol. 14, no. 12, p. 5614, 2024.

[10] X. Zhang, "Molecular classification of breast cancer: relevance and challenges," Archives of Pathology & Laboratory Medicine, vol. 147, no. 1, pp. 46-51, 2023.

[11] S. Yang, Z. Wang, C. Wang, C. Li, and B. Wang, "Comparative evaluation of machine learning models for subtyping triple-negative breast cancer: a deep learning-based multi-omics data integration approach," Journal of Cancer, vol. 15, no. 12, p. 3943, 2024.

[12] D. Acharya and A. Mukhopadhyay, "A comprehensive review of machine learning techniques for multi-omics data integration: challenges and applications in precision oncology," Briefings in functional genomics, vol. 23, no. 5, pp. 549-560, 2024.

[13] N. Adossa, S. Khan, K. T. Rytkönen, and L. L. Elo, "Computational strategies for single-cell multi-omics integration," Computational and Structural Biotechnology Journal, vol. 19, pp. 2588-2596, 2021.

[14] N. Vahabi and G. Michailidis, "Unsupervised multi-omics data integration methods: a comprehensive review," Frontiers in genetics, vol. 13, p. 854752, 2022.

[15] Z. Cai, R. C. Poulos, J. Liu, and Q. Zhong, "Machine learning for multi-omics data integration in cancer," Iscience, vol. 25, no. 2, 2022.

[16] I. Subramanian, S. Verma, S. Kumar, A. Jere, and K. Anamika, "Multi-omics data integration, interpretation, and its application," Bioinformatics and biology insights, vol. 14, p. 1177932219899051, 2020.

[17] 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," Molecules and cells, vol. 44, no. 7, pp. 433-443, 2021.

[18] 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.

[19] C. Rossi et al., "Breast cancer in the era of integrating "Omics" approaches," Oncogenesis, vol. 11, no. 1, p. 17, 2022.

[20] 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.

[21] 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.

[22] 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.

[23] S. Rakshit, I. Saha, S. S. Chakraborty, and D. Plewczyski, "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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] M. Tao et al., "Classifying breast cancer subtypes using multiple kernel learning based on omics data," Genes, vol. 10, no. 3, p. 200, 2019.

[30] 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.

[31] 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.

[32] M. M. Islam, S. Huang, R. Ajwad, C. Chi, Y. Wang, and P. Hu, "An integrative deep learning framework for classifying molecular subtypes of breast cancer," Computational and structural biotechnology journal, vol. 18, pp. 2185-2199, 2020.

[33] 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.

[34] Y. Ma and J. Guan, "MOCSC: 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.

[35] 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.

[36] 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.

[37] 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.