ISSN 2616-5775 Vol. 8, Issue 2: 26-36, DOI: 10.25236/AJCIS.2025.080204

HRAGNN: A cancer subtype identification method using multi-omics data and heterogeneous graph neural networks

Hanwen Bai

School of Software, Henan Polytechnic University, Jiaozuo, 454003, China
212209020006@home.hpu.edu.cn

Abstract: *Cancer is one of the leading causes of mortality worldwide, and its inherent diversity and heterogeneity pose significant challenges in early diagnosis, drug development, and prognosis. Accurate identification of cancer subtypes has therefore become a critical aspect of personalized cancer treatment. With the advancement of omics technologies, multi-omics data offer a more comprehensive understanding of cancer's underlying mechanisms. However, effectively integrating such diverse datasets to identify cancer subtypes remains a significant challenge. In this study, we introduce a novel approach, termed HRAGNN, for cancer subtype identification through the integration of multi-omics data. HRAGNN first constructs an integrated graph and then leverages Relational Attention Mechanism (RAM) and Graph Neural Network (GNN) to capture complex features across these multi-omics layers. Subsequently, the Multi-view Fusion Network (MVFN) is employed to fuse the features derived from the different omics data. We evaluated the performance of HRAGNN on three datasets, comparing it with other existing methods. The experimental results demonstrate that HRAGNN outperforms other approaches in terms of several key evaluation metrics.*

Keywords: *Cancer-subtype classification, Heterogeneous graph neural network, multi-omics integration*

1. Introduction

Cancer is a complex disease driven by genetic mutations in cells, with its development and metastasis involving intricate physiological processes, influenced by a wide array of factors. Cancers can arise in various tissues and organs, presenting with diverse types and stages, each of which directly impacts disease severity and prognosis. Notably, cancer remains one of the most significant threats to patient survival[1]. Its complexity and severity are further compounded by its inherent diversity and heterogeneity[2]. Emerging research has revealed that even within the same cancer type, subtypes can exhibit substantial variations in molecular characteristics and biological behaviors. These differences are critical for the design of personalized treatment regimens[3]. This heterogeneity suggests that cancer types should be viewed not as homogeneous entities, but as a collection of distinct subtypes, each potentially responding differently to therapies and exhibiting divergent clinical outcomes[4]. The goal of cancer subtype classification is to utilize established cancer subtypes as category labels[6], incorporate patient data across various biological levels as sample features, and use intelligent algorithms to train the classification model with the highest degree of fitting. This approach aims to facilitate the precise classification of patients within the same cancer type[5]. With the ongoing progress in high-throughput sequencing technologies, international collaborative initiatives like The Cancer Genome Atlas (TCGA), the International Cancer Genome Consortium (ICGC), and the Cancer Cell Line Encyclopedia (CCLE) have produced and consolidated a vast amount of multi-omics data.[7]. These datasets provide a comprehensive and diverse array of cancer-related insights, resulting in research methods targeting single omics data and cancer subtype identification methods combining multiple omics data.

Cancer subtype identification methods based on single-omics data: In 2018, Guo et al[8]. proposed BCDForest, a deep learning model that trains classifiers using multi-class granular scanning methods combined with an enhancement strategy to emphasize key features. This approach effectively mitigates overfitting and significantly improves classification performance. In 2021, Zhong et al[9]. introduced LACFNForest, a classifier based on the Deep Flexible Neural Forest (DFNForest)[10]. By incorporating a hierarchical broadening ensemble method and an output judgment mechanism, LACFNForest enhanced both classification accuracy and robustness, achieving promising results in the identification of cancer subtypes. The following year, Shen et al[11]. developed DCGN, a model based on single-omics data that integrates a Convolutional Neural Network (CNN) with a Bidirectional Gated Recurrent Unit (BiGRU)[12]. This combination enables the extraction of effective features from high-dimensional data, addresses issues of sample scarcity and sparse features, and improves model accuracy. While these single-omics-based methods for cancer subtype identification have demonstrated efficacy within their respective datasets, the inherent limitations of single-omics data, such as providing only specific levels of information and the inability to capture complex interactions, result in a one-sided perspective.

Cancer subtype identification methods based on multi-omics data: Given the limitations of single-omics data, numerous researchers have developed methods for cancer subtyping based on multi-omics data. In 2019, Xu et al.[13] applied a hierarchical integrated Deep Flexible Neural Forest approach to combine gene expression, miRNA expression, and DNA methylation data for classifying cancer subtypes across various TCGA datasets. The experimental results demonstrated that the combined data improved classification performance by 3.8%–11.5% compared to single-omics data, showcasing robust classification capabilities. In 2021, MOGONET[14] was introduced as an integrated Graph Convolutional Network (GCN) model designed for multi-omics data integration. By utilizing a unified graph structure and multi-layer graph convolution operations, MOGONET effectively captures the interactive features among various omics datasets and extracts deep feature representations from the global graph structure. This approach provides an efficient method for cancer subtype recognition, significantly enhancing both classification accuracy and generalization capabilities. In 2023, Wen and Li[15] leveraged GCN to fuse two types of omics data and perform deep survival prediction analysis. Their experimental results outperformed existing methods, confirming that the fusion of multiple omics features can generate higher-level feature representations. In 2024, to address the challenges of capturing characteristics between heterogeneous data and enhancing dynamic data representation, the model MCRGCN[16] was proposed. MCRGCN employs advanced graph convolution techniques and dynamic representation learning to effectively overcome the heterogeneity and dynamics inherent in graph data, further improving the accuracy and robustness of cancer subtype classification. In the same year, Shen et al[17]. introduced the CAEM-GBDT model, which utilizes a convolutional autoencoder and an attention module to extract features, followed by classification recognition using a Gradient-Enhanced Decision Tree (GBDT). Comparative experiments with various multi-omics and single-omics methods validated the feasibility of the model and underscored the necessity of multi-omics data integration. In recent years, an increasing number of cancer subtype recognition models have adopted multi-omics data as their training inputs. By integrating diverse datasets and leveraging various associations among the data, these models facilitate a more comprehensive understanding of the molecular characteristics and underlying mechanisms of cancer[18]. Consequently, this integration enhances classification performance and provides deeper biological insights.

In cancer research, gene expression data reflects the transcriptional activity of genes, thereby revealing the molecular characteristics of cancer and playing a fundamental role in understanding tumor biology. DNA methylation involves the addition of methyl groups to gene promoter regions, which suppresses gene transcription and consequently reduces gene expression. This epigenetic modification identifies genes that are silenced or activated due to abnormal methylation patterns, thereby influencing overall gene expression levels. MicroRNAs (miRNAs) regulate gene expression by binding to the messenger RNA (mRNA) of target genes, inhibiting translation or promoting mRNA degradation. This establishes a direct regulatory relationship between miRNAs and gene expression. Abnormal miRNA expression can lead to the dysregulation of cancer-related genes. Additionally, miRNA genes themselves may be suppressed or activated through methylation[19], further modulating the expression of cancer-related genes.

However, in existing methods, most models consider only one type of connection, either inter-group or intra-group connections[20], and predominantly employ a single aggregation method during model integration. In multi-omics data, the identification of cancer subtypes involves not only features such as gene expression but also encompasses complex heterogeneous data, including gene mutations, protein interactions, and epigenetic information, which may exhibit highly dynamic and irregular relationships[21], To accurately capture the nonlinear and dynamic relationships within the data, we adopt a graph structure that is more suitable for describing the complex interactions inherent in biomedical data[22-24]. This approach facilitates more comprehensive biological information mining[25].

To address the challenges associated with extracting effective cancer subtype information from high-dimensional multi-omics data, we present HRAGNN, a novel hybrid fusion method for cancer subtype identification that integrates multi-omics data using a Heterogeneous Relational Attention Graph Neural Network (HRAGNN). HRAGNN introduces a Multi-Relation Attention Mechanism, employing multiple Relation-Attention heads to independently capture diverse relational features between and within sample groups. This multi-head approach enables the model to effectively learn and represent the complex relationships inherent in multi-omics data. Initially, we construct a similarity network by calculating the cosine similarity across different omics datasets. Subsequently, the similarity graph for each sample's omics data is input into the Heterogeneous Relational Attention Network (HRAN), which extracts multi-omics features enriched with information from each omics layer. These features are then processed by a three-layer GNN to learn and capture complex relationships and structural information among nodes and their neighbors within the graph on a global scale. Finally, the multi-view omics data were integrated and classified using an MVFN[14] architecture augmented with gating units, enabling precise determination of the cancer subtype. We applied HRAGNN to datasets of invasive breast cancer (BRCA) and glioblastoma multiforme (GBM), integrating gene expression, miRNA expression, and DNA methylation data. Experimental results demonstrate that HRAGNN outperforms other integrated multi-omics data classification methods, underscoring its effectiveness in accurately identifying cancer subtypes.

2. Results

Framework of HRAGNN: We introduce HRAGNN for cancer subtype identification, which integrates three multi-omics datasets: gene expression, miRNA expression, and DNA methylation data. HRAGNN operates through five key steps: (i) Data preprocessing: The method first reduces noise and standardizes the data to enhance data quality; (ii) Constructing similarity graph: A similarity graph is created for each omics dataset, and these graphs are subsequently combined to form a final similarity graph; (iii) Feature extraction: The similarity graph is input into a relational attention network, followed by processing through a Residual Graph Neural Network (RGNN) to extract features; (iv) Feature fusion: Multi-omics features are fused using MVFN; (v) Identification: Finally, the fused features are processed through a SoftMax layer to determine the class probabilities for each sample. A flowchart of HRAGNN is presented in Fig. 1.

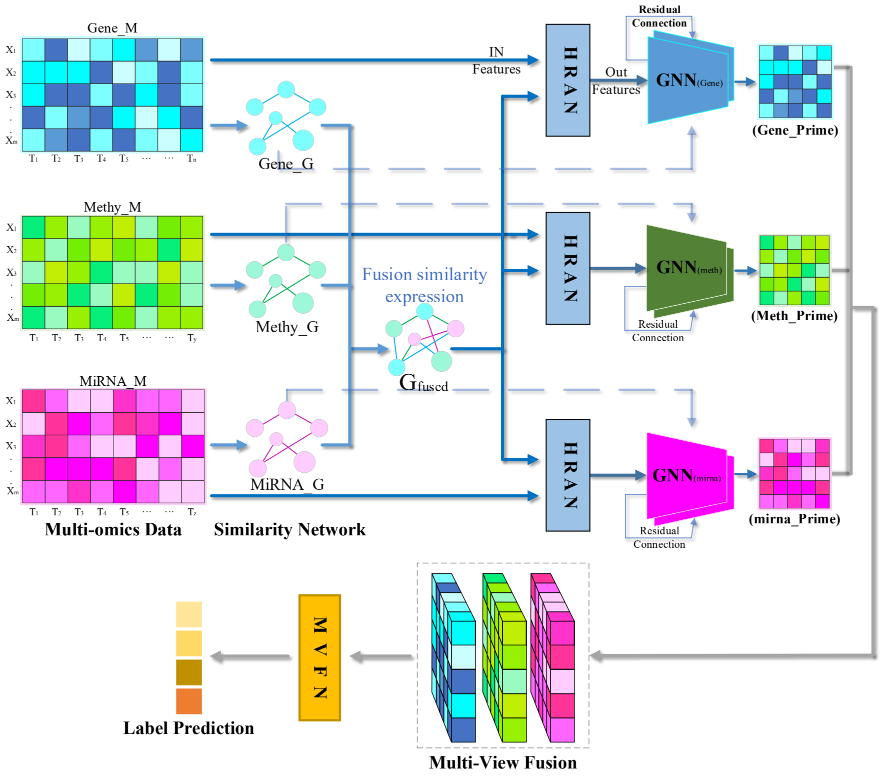

Figure 1. The framework of HRAGNN

Datasets: To rigorously evaluate the performance of the Hierarchical Relational Attention Graph Neural Network (HRAGNN) model, we conducted analyses on a comprehensive cohort comprising 646 cases of invasive breast carcinoma (BRCA) and 248 instances of glioblastoma multiforme (GBM). Each sample was characterized by three distinct multi-omics datasets: transcriptomic gene expression profiles, miRNA expression data, and epigenomic DNA methylation patterns. Both the BRCA and GBM datasets were stratified into four molecularly defined subtypes, with their nomenclature and detailed characteristics delineated in Table 1. The datasets were partitioned into training and testing subsets using a 4:1 ratio. The HRAGNN model was trained over 1,500 epochs, with performance metrics meticulously recorded at every 30-epoch interval to monitor model convergence and mitigate the risk of overfitting. The Invasive Breast Cancer (BRCA) and Glioblastoma Multiforme (GBM) cancer samples used in this study are publicly available on the TCGA website (https://cancergenome.nih.gov).

Table 1 Statistical information of two cancer datasets

Dataset	Subtype	Number of features	Number of samples
BRCA	Luminal A, Luminal B, HER2, Negative.	18514 (Gene) 5000 (Meth) 534 (miRNA)	646
GBM	Classical, Proneural, Mesenchymal, Neural.	12042 (Gene) 5000 (Meth) 534 (miRNA)	248

Ablation experiment: In this study, we evaluated the performance of cancer subtype classification using four external performance metrics: Precision, Accuracy, Recall, and F1 score, to assess the effectiveness of the algorithm. All methods were tested and compared across five different randomly generated training sets to ensure a robust performance evaluation.

Comparative test of polymerization methods: In this study, we compared three distinct aggregation methods: sum aggregation, mean aggregation, and maximum aggregation, when processing graph-structured data within the context of graph neural networks (GNNs). The performance results, summarized in Table 2, reveal notable differences across these methods. Specifically, in the BRCA dataset, sum aggregation achieved an accuracy of 0.861 and an F1 score of 0.851, while maximum aggregation resulted in 0.808 accuracy and 0.757 F1 score. In contrast, mean aggregation demonstrated superior performance with an accuracy of 0.885 and an F1 score of 0.876. Similarly, in the GBM dataset, sum aggregation yielded an accuracy of 0.860 and an F1 score of 0.849, whereas mean aggregation outperformed with an accuracy of 0.881 and an F1 score of 0.871. Maximum aggregation again showed the weakest results, with an accuracy of 0.800 and an F1 score of 0.771. These results suggest that mean aggregation consistently outperforms sum aggregation across both datasets, highlighting its capacity to better capture the underlying patterns in the data. Notably, the accuracy values for sum and mean aggregation in the BRCA dataset were 0.904 and 0.902, respectively, while in the GBM dataset, the corresponding values were 0.889 and 0.895. These findings underscore the importance of aggregation methods in optimizing model performance. They also suggest that mean aggregation enhances the model's ability to generalize, yielding higher accuracy and more stable results across diverse cancer subtypes. Overall, this analysis reinforces the effectiveness of the model proposed in this study, particularly in accurately identifying and distinguishing between specific cancer categories.

Table 2 Comparative test results of polymerization methods

	Methods	Precision	Recall	F1 Score	Accuracy
BRCA	SUM	0.904	0.813	0.851	0.861
	MAX	0.785	0.737	0.757	0.808
	MEAN	0.902	0.856	0.876	0.885
GBM	SUM	0.889	0.836	0.849	0.860
	MAX	0.822	0.756	0.771	0.800
	MEAN	0.895	0.863	0.871	0.881

Ablation experiment of AN module: to evaluate the impact of the attention network on cancer subtype classification, we implemented three models: a graph neural network (GNN), an attention graph neural network (AGNN), and a relational attention graph neural network (RAGNN), along with their heterogeneous counterparts, HRAGNN and HRAGNN, corresponding to AGNN and RAGNN, respectively. As shown in Table 3, the accuracy of GNNs on the BRCA and GBM datasets reached 0.808 and 0.800, respectively, while AGNN achieved accuracy rates of 0.829 and 0.840. In comparison, RAGNN exhibited an improved accuracy of 0.846 for BRCA and 0.823 for GBM. The accuracy of AGNN on these datasets was 0.838 and 0.823 for BRCA and GBM, respectively, while the performance of RAGNN was 0.868 and 0.848, respectively. When integrating heterogeneous correlation representations, HRAGNN and HRAGNN showed a performance improvement of 1–4% in terms of accuracy and other evaluation metrics, compared to AGNN and RAGNN. These results suggest that, for the BRCA and GBM datasets, the relational attention network outperforms the attention network and is better suited for the model proposed in this study. Furthermore, the inclusion of heterogeneous correlation representations led to higher model performance. However, to more accurately assess whether these improvements were due to the fusion of multiple omics data representations, we conducted the following ablation study.

Table 3 Experimental results of AN module ablation

	Methods	Precision	Recall	F1 Score	Accuracy
BRCA	GNNs	0.829	0.711	0.721	0.808
	AGNN	0.868	0.774	0.808	0.838
	RAGNN	0.866	0.783	0.813	0.846
	HAGNN	0.872	0.792	0.821	0.854
	HRAGNN	0.902	0.856	0.876	0.885
GBM	GNNs	0.840	0.751	0.765	0.800
	AGNN	0.848	0.779	0.795	0.823
	RAGNN	0.853	0.785	0.802	0.823
	HAGNN	0.865	0.802	0.814	0.842
	HRAGNN	0.895	0.863	0.871	0.881

Ablation experiment of fusion heterogeneous adjacency matrix: According to the results presented in Table 3, it is evident that when heterogeneous fusion correlation representations are utilized, the model's performance exceeds that achieved by convolutional relational attention graphs alone. To verify whether the fusion of heterogeneous correlation representations genuinely improves model performance and whether there are potential similarities between different omics samples with the same label, we conducted a two-part ablation study. In this experiment, we compared the results of models without fusion of correlation representations and with fusion of two correlation representations. The results, as shown in Table 4, indicate that, for the BRCA dataset, the accuracy for the three experiments were 0.846, 0.869, and 0.885, and the corresponding F1 scores were 0.813, 0.853, and 0.876, respectively. For the GBM dataset, the accuracy values were 0.823, 0.863, and 0.881, and the F1 scores were 0.802, 0.849, and 0.871, respectively. When no similarity information is shared between the omics datasets, the model's performance is the lowest. However, when relevant representations of multiple omics data are fused, the model's performance improves significantly, with a similar percentage improvement across both datasets. Based on these findings, we conclude that when samples share the same label, there are inherent similarities between different omics data samples. Furthermore, the model's ability to select and extract features can be enhanced by fusing the correlation representations of multiple omics data.

Table 4 Experimental results of fusion heterogeneous adjacency matrix ablation

	Methods	Precision	Recall	F1 Score	Accuracy
BRCA	RAGNN	0.866	0.783	0.813	0.846
	Bis_HRAGNN	0.889	0.831	0.853	0.869
	HRAGNN	0.902	0.856	0.876	0.885
GBM	RAGNN	0.853	0.785	0.802	0.823
	Bis_HRAGNN	0.873	0.835	0.849	0.863
	HRAGNN	0.895	0.863	0.871	0.881

Comparative experiment: To assess the classification performance of the HRAGNN model, we compared it with five other multi-omics data classification algorithms: K-nearest neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF), Fully Connected Neural Network (NN), MCRGCN, and MOGONET. The performance of these six classifiers, including HRAGNN, was evaluated using the same dataset. The experimental results are shown in Table 5 and Table 6.

Table 5 Performance comparison between each model on the BRCA dataset

Methods	Precision	Recall	F1 Score	Accuracy
NN	0.541	0.552	0.544	0.746
KNN	0.515	0.388	0.385	0.639
SVM	0.812	0.754	0.773	0.824
RF	0.808	0.637	0.631	0.645
MOGONET	0.822	0.748	0.782	0.831
MCRGCN	0.833	0.785	0.831	0.849
HRAGNN	0.902	0.856	0.876	0.885

Table 6 Performance comparison between each model on the GBM dataset

Methods	Precision	Recall	F1 Score	Accuracy
NN	0.707	0.650	0.661	0.701
KNN	0.672	0.637	0.631	0.647
SVM	0.803	0.792	0.787	0.801
RF	0.716	0.602	0.615	0.660
MOGONET	0.827	0.723	0.729	0.783
MCRGCN	0.833	0.828	0.821	0.834
HRAGNN	0.895	0.863	0.871	0.881

As shown in Table 5 and Table 6, we evaluated the classification performance of multiple models on the BRCA and GBM datasets, using key metrics: accuracy, recall, F1 score, and precision. Among all the models, HRAGNN consistently demonstrated superior performance. For the BRCA dataset, HRAGNN achieved an accuracy of 0.902, recall of 0.856, F1 score of 0.876, and precision of 0.885. On the GBM dataset, HRAGNN reached an accuracy of 0.895, recall of 0.863, F1 score of 0.871, and precision of 0.881. In comparison to MCRGCN, HRAGNN outperformed by 7.2%, 7.5%, and 3.6% in recall, F1 score, and accuracy, respectively, in the BRCA dataset. When compared to MOGONET, HRAGNN showed improvements of 10.8%, 9.4%, and 5.4% in these metrics. In the GBM dataset, HRAGNN improved recall, F1 score, and accuracy by 3.5%, 0.5%, and 4.7%, respectively, over MCRGCN, and by 1.4%, 14.2%, and 9.8% over MOGONET. The lower performance of MOGONET can be attributed to its failure to capture interactions between different omics features, whereas MCRGCN's performance is also inferior to HRAGNN, highlighting the importance of incorporating correlations across multi-omics data. By fusing multi-omics data, HRAGNN leverages sample correlation to provide valuable feature associations, facilitating the extraction of deeper feature correlations and enhancing the relevance of the features identified by graph convolutional networks. These results confirm that HRAGNN, utilizing heterogeneous graph convolutional networks, is highly effective in fusing multi-omics data for accurate cancer subtype classification.

3. Methods

Data preprocessing: Data preprocessing is crucial for improving data quality by reducing noise and ensuring standardization. In this study, the HRAGNN model uses three matrices as input: *Gene_M*$\in R^{m\times t}$, *Meth_M*$\in R^{m\times e}$ and *MiRNA_M*$\in R^{m\times w}$, where *m* represents the number of samples, *t* is the number of genes, *e* is the number of methylation sites, and *w* is the number of miRNA molecules. *a) Gene_M* represents gene expression data, where each row corresponds to a sample, and each column corresponds to a gene. The element *Gene_M*[i, j] represents the expression level of the j-th gene in the i-th sample. *b) Meth_M* represents DNA methylation data, with each row representing a sample and each column representing a methylation site. The element *Meth_M*[i, j] indicates the methylation level at the j-th site in the i-th sample. *c) MiRNA_M* contains miRNA expression data, where each row corresponds to a sample and each column corresponds to an individual miRNA molecule. The element *MiRNA_M*[i, j] represents the expression level of the j-th miRNA in the i-th sample. During preprocessing, an initial filtering step removes any column where all the values are zero across the samples. Subsequently, for each matrix column, the element values are extracted and sorted in ascending order. Based on this ordered sequence, the lower (Q1) and upper (Q3) quartiles are calculated. Any element below Q1 is truncated to Q1, while any element exceeding Q3 is capped at Q3, thus correcting for outliers and extreme values. Finally, each column undergoes a standardization procedure, ensuring that the data across features are scaled uniformly. This normalization step mitigates dimensionality differences between features, thereby enhancing the model's training performance and generalizability.Through the implementation of these standardization and outlier correction techniques, the quality and comparability of the data are significantly improved, providing more precise and robust input for subsequent deep learning model training.

Constructing similarity graph: In this step, HRAGNN constructs three distinct similarity graphs based on the matrices *Gene_M*, *Meth_M* and *MiRNA_M****.*** For the *Gene_M* matrix, HRAGNN first constructs a similarity graph, denoted as *Gene_G*, where each vertex represents a sample, and each edge between two vertices encodes the similarity between those samples. Specifically, the similarity between the i-th and j-th samples is computed using the cosine similarity between their respective rows in *Gene_M* (*Gene_Mi* and *Gene_Mj*). Once the *Gene_G* graph is constructed, HRAGNN retains only the top 10 edges with the highest weights for each vertex, setting all other edges' weights to zero. This procedure ensures that only the most significant relationships between samples are preserved. The same methodology is applied to construct the *Methy_G* and *MiRNA_G* graphs from the *Meth_M* and *MiRNA_M* matrices, respectively. For the fusion of the three similarity graphs (*Gene_G*, *Meth_G*, and *MiRNA_G*), HRAGNN adopts a novel weighted fusion approach, which incorporates the intrinsic characteristics of each graph and adaptively adjusts the contribution of each graph to the final fused graph *G_fused*. Instead of directly combining the adjacency matrices, the fusion process utilizes a graph attention mechanism to learn the optimal weights for each edge in the fused graph. This mechanism allows for the dynamic adjustment of the influence of each individual similarity graph based on the specific relationships between samples across different modalities. And the fusion of the three graphs is expressed as:

$$G_{fused} = \alpha \times G_{Gene}(i,j) + \beta \times G_{Methy}(i,j) + \gamma \times G_{MiRNA}(i,j) \quad (1)$$

Where α, β, γ are the attention weights dynamically learned through the graph attention network, satisfying α+β+γ=1. These weights are adaptively updated during training based on the task-specific loss function, ensuring that the fused graph reflects the most relevant information from all three modalities. The resulting *G_fused* graph, in which the vertices represent samples, provides a more robust and context-aware representation of the relationships between samples, integrating the complementary information from gene expression, DNA methylation, and miRNA expression data. Previous studies have demonstrated that node representation learning relies on the principle that sample nodes with identical labels exhibit similar feature patterns[26]. In this section, we extend this concept by simultaneously constructing both inter-group and intra-group connections across multiple omics data types. This approach enables the model to effectively capture and exploit the underlying relationships and latent features among the various omics data modalities during training, thus enhancing its ability to learn comprehensive and integrated representations.

Extracting features: *a)* Construct relational attention network: Upon constructing the fused similarity graph G_{fused} for the samples, HRAGNN integrates G_{fused} alongside the multi-omics data matrices *Gene_M*, *Meth_M* and *MiRNA_M* into Relational Attention Network (RAN) to elucidate the intricate relationships among samples. Initially, HRAGNN inputs G_{fused} and *Gene_M* into the RAN. We define two pivotal parameters: *W* and *a*. Here *W* represents a linear transformation weight matrix that maps input features to a higher-dimensional feature space, while *a* is a learnable parameter vector utilized to compute attention coefficients. In the *Gene_M* matrix, each row $h_i=\{d_1, d_2, d_3, d_4, \dots, d_n\}$ denotes the feature vector of the i-th sample, After applying the linear transformation, the transformed feature vector is represented as $Wh_i=\{ D_1,D_2, D_3, D_4, \dots, D_n \}$. To ascertain the attention score between any pair of samples *i* and *j*, the transformed feature vectors Wh_i and Wh_j are concatenated and subsequently passed through a LeakyReLU activation function following a linear combination with *a.* This process yields the raw attention coefficient e_{ij} defined as:

$$e_{ij} = LeakyReLU(a^T \times [Wh_i||Wh_j]) \tag{2}$$

Where "||" denotes the concatenation of vectors. Subsequently, attention coefficients corresponding to non-connected sample pairs in G_{fused} are filtered by setting $e`_{ij}$ to $e_{ij}$ if $G_{fused}$ (i, j) > 0, and to negative infinity (−∞) otherwise. Formally, the filtered attention coefficients $e`_{ij}$ are expressed as:

$$e'_{ij} = \begin{cases} e_{ij}, & if\ G_{fused_{ij}} > 0 \\ -\infty, & otherwise \end{cases} \tag{3}$$

This filtering mechanism ensures that only pertinent edges, as delineated by G_{fused}, contribute to the attention computation. The normalized attention coefficients $a`_{ij}$ are then obtained by applying the SoftMax function over the neighborhood N_i of each node *i*:

$$a'_{ij} = \frac{\exp(e'_{ij})}{\sum_{k\in N_i}\exp(e'_{ik})} \tag{4}$$

Where N_i denotes the set of neighboring nodes of node *i*. Finally, Finally, the intermediate feature representation *Gene_H* is derived by aggregating the transformed features W_h weighted by the normalized attention coefficients. This procedure is similarly applied to the *Meth_M* and *MiRNA_M* matrices, resulting in the intermediate feature representations *Meth_H* and *MiRNA_H*, respectively. Through this multi-omics integration facilitated by the Relational Attention Network, HRAGNN effectively captures and models the complex relationships among samples, leveraging the complementary information inherent in gene expression, DNA methylation, and miRNA expression data. b) Construct residual GNN networkIn this section, HRAGNN constructs three residual Graph Neural Networks (GNNs) utilizing the intermediate feature matrices: *Gene_H*, *Meth_H* and *MiRNA_H* derived previously, alongside their corresponding multi-omics similarity graphs: *Gene_G*, *Meth_G* and *MiRNA_G*. Linear Transformation and Regularization: Focusing on *Gene_H*, initiates the process by defining a weight matrix *W* and applying a linear transformation to *Gene_H* to obtain *Wh*, where Wh_i represents product of *W* and the feature vector of the i-th sample:

$$Wh_i = W \times h_i \tag{5}$$

To mitigate the risk of overfitting, L1 Regularization is imposed on *Wh*:

$$\mathcal{L}_{L1} = \lambda \sum |W_{ij}| \tag{6}$$

where λ is the regularization coefficient. Aggregation Strategies: HRAGNN employs three distinct aggregation strategies: mean aggregation, summation aggregation, and maximum aggregation, to process the similarity graphs. In the context of *Gene_G*, the following notations are used:*a)* d_i: Degree of the i-th node (i.e., the number of nodes connected to node *i*). *b)* X_i: Feature vector of the i-th node. *c)* $Gene\_G_{ij}$: Connection weight between nodes *i* and *j* in *Gene_G.* For this study, mean aggregation is selected as the aggregation method. The output for the i-th node is computed as:

$$output_i = LeakyReLU\left(\frac{1}{d_i}\sum_{j\in N(i)} Gene\_G_{ij} \times (X_j W) + b\right) \tag{7}$$

Where N_i denotes the set of nodes adjacent to node *i*. *b* is a bias vector. Residual Connections and Activation: To further prevent the model from overfitting, a residual connection is incorporated after the first GNN layer, linking it directly to the input of the final layer. This residual connection facilitates the flow of information and enhances the stability of the network during training. Subsequently, the Sigmoid activation function scales the output to the range [0, 1], resulting in the primary feature matrix *Gene_prime:*

$$Gene\_prime = \sigma(output) \tag{8}$$

where represents the Sigmoid function. The same procedure is applied to the intermediate feature matrices *Meth_H* and *MiRNA_H*, in conjunction with their respective similarity graphs *Meth_G* and *MiRNA_G*. Consequently, HRAGNN generates three primary feature matrices: *Gene_prime*, *Meth_prime* and *MiRNA_prime*.

Fusing features: In this section, we introduce the Multi-View Fusion Neural Network (MVFN), a neural network model designed to perform fusion processing of multiple omics data views and generate final prediction outputs. Let *C* denote the number of cancer subtypes (classes), *V* represent the number of omics data categories. The input dimension In_{dim} is defined as V^C. The MVFN architecture employs gating units to dynamically modulate the contribution of each omics view during the feature fusion process. It comprises a sequential model consisting of two fully connected (dense) layers: 1) First Fully Connected Layer: Maintains both input and output dimensions at In_{dim}. 2) Second Fully Connected Layer: Maps the dimension from In_{dim} to the number of classes *C*. Activation and Feature Scaling: Initially, HRAGNN applies the Sigmoid activation function to each primary feature in the input list *in_list*[i], where $i \in \{1,2,3\}$ corresponds to *Gene_Prime*, *Meth_Prime*, and *MiRNA_Prime* respectively. This ensures that the input values are scaled between 0 and 1:

$$in\_list[i] \;=\; \sigma(x) \;=\; \frac{1}{1+e^{-x}} \tag{9}$$

Fusion of Multi-View Features: The fusion process initiates by deploying gating units on the features of each view, thereby dynamically modulating their contributions based on learned weights. For each view *i*, the gated feature is determined as:

$$gated\_feature_i = in\_list[i] \times gate_i \tag{10}$$

where $gate_i$ denotes the output of the gating unit corresponding to view *i*. These gating units consist of a linear transformation followed by a sigmoid activation function, enabling the model to assign a scalar weight between 0 and 1 to each feature, effectively controlling its influence in the fusion process. Subsequently, the gated features are integrated using high-order tensor multiplication to capture complex interactions across multiple views. Specifically, the outer product of the first two gated features is computed and reshaped to dimensions [-1, C^2, 1], where *C* represents the number of cancer subtypes (classes). For datasets comprising more than two views, this process is iteratively applied within a loop. In each iteration, the intermediate tensor *x* is matrix-multiplied with the gated features of the newly introduced view *in_list*[i], and the tensor dimensions are accordingly adjusted to accommodate the increasing number of views:

$$x = reshape(matmul(x, unsqueeze(in\_list[i], 1)), (-1, C^{i+1}, 1)) \tag{11}$$

Where *matmul* denotes matrix multiplication and *unsqueeze* adds a new dimension to facilitate appropriate matrix operations. Feature Vector Formation: After processing all views, the multi-view features are reshaped into one-dimensional vectors to satisfy the input dimensional requirements of the model's fully connected layers:

$$mvfn\_feat = reshape(x, (-1, C^V)) \tag{12}$$

Through these operations, the MVFN effectively integrates features from multiple omics views, producing a comprehensive composite feature representation. This fused feature vector is then passed through the sequential fully connected layers to generate the final prediction output. The final output dimension is [*batch_size*, *C*], corresponding to the prediction scores for each cancer subtype.

Identifying: Upon integrating and processing the multi-omics features, HRAGNN generates the fused view representations, which are subsequently transformed into probability distributions over the cancer subtypes using the SoftMax activation function. The SoftMax function is defined as:

$$softmax(Z_i) = \frac{e^{Z_j}}{\sum_{j-1}^{K} e^{Z_j}} \tag{13}$$

Where Z_i is the raw score for class *i*, e^{Zi} represents the exponentiation of Z_i, and K is the total number of cancer subtypes. This formulation ensures that the output probabilities are normalized, summing to one across all classes. The resultant probability distribution is then compared with the true labels to compute the final model performance. Specifically, the cross-entropy loss function is employed to quantify the discrepancy between the predicted probabilities and the actual labels. The cross-entropy loss is expressed as:

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \cdot \log(p_i) \tag{14}$$

Where C denotes the total number of classes, y_i is the true label for class *i* (typically represented as a one-hot encoded vector), p_i is the predicted probability for class *i.* To enhance the training stability and prevent the learning rate from diminishing too rapidly, HRAGNN utilizes the AdaBound optimization algorithm. AdaBound dynamically adjusts the learning rate by setting adaptive bounds based on the gradients. The learning rate η_t at iteration t is updated according to the following rule:

$$\eta t = min(\eta_{max}, max(\eta_{min}, \eta_{base} \times \frac{1}{\sqrt{v_t}+\epsilon})) \tag{15}$$

Where, η_{max} and η_{min} are the upper and lower bounds of the learning rate respectively, η_{base} is is the initial learning rate, v_t represents the second moment estimate of the gradient at iteration t, ϵ is a small constant added for numerical stability. By employing AdaBound, HRAGNN effectively regulates the learning rate within predefined boundaries, thereby improving the convergence behavior and overall training stability.

Parameter settings: Residual Graph Neural Network Architecture: In this study, In this study, the Residual Graph Neural Network is architected with a layer configuration of [in_features,512,512,256] This structure comprises an input layer followed by two hidden layers with 512 neurons each, and an output layer with 256 neurons. Regularization Techniques: To prevent overfitting and improve the generalization capability of the model, regularization strategies are employed post each GNN layer, and the dropout rate is 0.3.

Code availability: The source code and datasets of this work can be downloaded from GitHub (https://github.com/1book1/HRAGNN-model).

4. Conclusion

The key innovation of HRAGNN lies in its ability to construct a fusion correlation network by capturing the inter-omics relationships, allowing the graph neural network to extract deeper, more meaningful features for classification. Unlike existing methods, HRAGNN assigns learnable weight parameters to the correlation networks of each omics type and subsequently applies weighted aggregation during the fusion process. This approach ensures that the similarity networks of the various omics data samples are trained in a coordinated manner, leading to enhanced classification performance. The effectiveness of HRAGNN is validated through ablation studies on heterogeneous relational attention networks and comparison with established models such as MOGONET and MCRGCN. While HRAGNN shows promising results, it has limitations, particularly in unsupervised clustering tasks. Addressing these challenges, such as improving model adaptability and scalability in unsupervised settings, will be the focus of future research.

References

[1] K. S. et al. Advances in cancer immunotherapy 2019 – latest trends. J. Exp. Clin. Cancer Res. 38, 268; 10.1186/s13046-019-1266-0 (2019).

[2] Alexander et al. Assessment of the molecular heterogeneity of E-Cadherin expression in invasive lobular breast cancer. Cancers 14, 295; 10.3390/cancers14020295 (2022).

[3] Ding, H. & Luo, J. MAMnet: detecting and genotyping deletions and insertions based on long reads and a deep learning approach. Brief. Bioinform. 23, bbac195; 10.1093/bib/bbac195 (2022).

[4] Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. Cell 144, 646–674 (2011).

[5] Luo, J. et al. BreakNet: detecting deletions using long reads and a deep learning approach. BMC Bioinformatics 22, 577; 10.1186/s12859-021-04499-5 (2021).

[6] S. J. P. et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. 41, 26 (2023).

[7] Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics data integration, interpretation, and its application. Bioinformatics Biol. Insights 14; 10.1177/1177932219899051 (2020).

[8] Guo, Y. et al. BCDForest: a boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data. BMC Bioinformatics 19, Suppl. 5, 118; 10.1186/s12859-018-2095-4 (2018).

[9] Zhong, L., Meng, Q., Chen, Y., Du, L. & Wu, P. A laminar augmented cascading flexible neural forest model for classification of cancer subtypes based on gene expression data. BMC Bioinformatics 22, 475; 10.1186/s12859-021-04391-2 (2021).

[10] Xu, J. et al. A novel deep flexible neural forest model for classification of cancer subtypes based on gene expression data. IEEE Access 7, 22086–22095; 10.1109/ACCESS.2019.2898723 (2019).

[11] Shen, J. et al. Deep learning approach for cancer subtype classification using high-dimensional gene expression data. BMC Bioinformatics 23, 1–17; 10.1186/s12859-022-04980-9 (2022).

[12] Cho, K., Bengio, Y. et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1724–1734. Association for Computational Linguistics (2014).

[13] Xu, J. et al. A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data. BMC Bioinformatics 20, 527; 10.1186/s12859-019-3116-7 (2019).

[14] Wang, T. et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nature Communications 12, 3445 (2021).

[15] Wen, G. & Li, L. FGCNSurv: dually fused graph convolutional network for multi-omics survival prediction. Bioinformatics 39, 1–9; 10.1093/bioinformatics/btad472 (2023).

[16] Chen, F. et al. Supervised graph contrastive learning for cancer subtype identification through multi-omics data integration. Health Information Science and Systems 12, 12 (2024).

[17] Shen, J., Guo, X., Bai, H. & Luo, J. CAEM-GBDT: a cancer subtype identifying method using multi-omics data and convolutional autoencoder network. Frontiers in Bioinformatics 4, art. no. 1403826; 10.3389/fbinf.2024.1403826 (2024).

[18] Gao, R., Luo, J., Ding, H. & Zhai, H. INSnet: a method for detecting insertions based on deep learning network. BMC Bioinformatics 24, 80; 10.1186/s12859-023-05216-0 (2023).

[19] Fuso, A., Raia, T., Orticello, M., & Lucarelli, M. The complex interplay between DNA methylation and miRNAs in gene expression regulation. Biochimie, 173, 12-16; 10.1016/j.biochi.2020.02.006 (2020).

[20] Li, Z. Q. et al. Multi-omics analysis of five species of milk and specific composition links within each species. Food Chemistry 457, art. no. 140028; 10.1016/j.foodchem.2024.140028 (2024).

[21] Subramanian, S., Verma, S., Kumar, A., & Jere, A. K. Multi-omics data integration, interpretation, and its application. Bioinformatics and Biology Insights 14; 10.1177/1177932219899051 (2020).

[22] Su, X. et al. A comprehensive survey on community detection with deep learning. IEEE Transactions on Neural Networks and Learning Systems 35, 4, 4682–4702; 10.1109/TNNLS.2021.3137396 (2024).

[23] Hamilton, W., Ying, R., & Leskovec, J. Inductive representation learning on large graphs. In Proceedings of the Neural Information Processing Systems (2017).

[24] Mahmud, M. et al. Deep learning in mining biological data. Cognitive Computation 13, 1–33; 10.1007/s12559-020-09773-x (2021).

[25] Maghsoudi, Z., Nguyen, H., Tavakkoli, A., & Nguyen, T. A comprehensive survey of the approaches for pathway analysis using multi-omics data integration. Briefings in Bioinformatics 23, 6, art. no. bbac435; 10.1093/bib/bbac435 (2022).

[26] Ma, Y. et al. Is homophily a necessity for graph neural networks? arXiv preprint arXiv:2106.06134 (2021).