BMC Genomics

RESEARCH

Open Access

DeepMoIC: multi-omics data integration via deep graph convolutional networks for cancer subtype classification

Jiecheng Wu[1], Zhaoliang Chen[2], Shunxin Xiao[3], Genggeng Liu[1], Wenjie Wu[4] and Shiping Wang[1*]

Abstract

Background Achieving precise cancer subtype classification is imperative for effective prognosis and treatment. Multi-omics studies, encompassing diverse data modalities, have emerged as powerful tools for unraveling the complexities of cancer. However, owing to the intricacies of biological data, multi-omics datasets generally show variations in data types, scales, and distributions. These intractable problems lead to challenges in exploring intact representations from heterogeneous data, which often result in inaccuracies in multi-omics information analysis.

Results To address the challenges of multi-omics research, our approach DeepMoIC presents a novel framework derived from deep Graph Convolutional Network (GCN). Leveraging autoencoder modules, DeepMoIC extracts compact representations from omics data and incorporates a patient similarity network through the similarity network fusion algorithm. To handle non-Euclidean data and explore high-order omics information effectively, we design a Deep GCN module with two strategies: residual connection and identity mapping. With extracted higher-order representations, our approach consistently outperforms state-of-the-art models on a pan-cancer dataset and 3 cancer subtype datasets.

Conclusion The introduction of Deep GCN shows encouraging performance in terms of supervised multi-omics feature learning, offering promising insights for precision medicine in cancer research. DeepMoIC can potentially be an important tool in the field of cancer subtype classification because of its capacity to handle complex multi-omics data and produce reliable classification findings.

Keywords Multi-omics, Deep graph convolutional network, Supervised learning, Cancer subtype classification

Introduction

Cancer, an extensive spectrum of diseases, can virtually manifest in any organ or tissue within the human body [1]. The identification of cancer subtypes and the prognosis estimation for patients are key aspects of cancer research. Due to the recent rapid progress in high-throughput biomedical technology, diverse types of omics data have been collected with unprecedented levels of detail, encompassing diverse molecular processes such as Copy Number Variation, mRNA expression, and DNA methylation. Although individual omics data can capture specific aspects of biological complexity, a greater comprehension of the complex biological

*Correspondence:
Shiping Wang
shipingwangphd@163.com
[1] College of Computer and Data Science, Fuzhou University, Fuzhou 350108, China
[2] Department of Computer Science, Hong Kong Baptist University, Hong Kong, SAR, China
[3] School of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China
[4] Department of Ophthalmology, Shengli Clinical Medical College of Fujian Medical University, Fujian Provincial Hospital, Fuzhou University Affiliated Provincial Hospital, Fuzhou 350001, China

processes is made possible by the integration of diverse omics data types [2]. In particular, current research has provided compelling evidence that the integration of data from diverse omics technologies considerably enhances the performance of forecasting clinical results compared to using only one type of omics data [3–6]. In light of these advancements, there arises a necessity for innovative integrative analysis methods adept at harnessing the correlation and additional details inherent in multi-omics data.

Navigating the complexities of multi-omics studies is inherently challenging, owing to the diversity of data types, scales, and distributions, which are often characterized by numerous variables and limited samples. Additionally, biological datasets may introduce unwanted complexity and noise, potentially containing errors stemming from measurement inaccuracies or inherent biological variability. To address this challenge, numerous methods and strategies based on deep learning have emerged in recent years to extract meaningful information and integrate diverse omics data into coherent models [7]. With the advancement of personalized medicine, meticulously annotated datasets that provide comprehensive details about sample phenotypes or traits are becoming increasingly accessible. A precise classifier for cancer molecular subtypes is essential for early-stage diagnosis, prognosis, and drug development. Consequently, supervised multi-omics integration models, which can identify disease-related biomarkers and predict outcomes in new samples, are gaining increasing popularity [3]. For instance, Lin et al. [8] employed a type-specific encoding module to extract features from different data types, and combined these features to predict breast cancer subtypes. Poirion et al. [9] utilize an autoencoder for dimensionality reduction to predict survival cancer subtypes with a deep learning framework. Moreover, the predominant focus of current cancer subtype identification methods revolves around unsupervised multi-omics data integration [10–12].

As an effective solution to the integration of multi-omics data, Patient Similarity Network (PSN) was devised to integrate multi-omics data and construct interpretable models [13, 14]. To effectively process non-Euclidean data with PSN, previous studies had utilized Graph Convolutional Network (GCN) [15], which could directly operate on graphs and discover underlying correlations among samples, and has gained popularity in the domain of bioinformatics [6, 16–18]. For example, Dai et al. [19] employed a sample similarity network and a residual GCN for cancer subtype identification. Li et al. [20] designed a multi-omics data fusion method incorporating a two-layer GCN to process the no-Euclidean data similarity network for cancer subtypes classification and analysis.

In spite of the achievements of the aforementioned methods, most GCN-based methods only used shallow structures owing to the over-smoothing issue [21], which results in notorious performance when constructing deep GCN frameworks. Nonetheless, a deep GCN is beneficial to discover remote nodes, and considerable studies have revealed the necessity of propagating information to high-order neighbors in various bioinformatics fields, such as protein-phenotype associations prediction [22], liver cancer diagnosis [23], and protein-ligand binding residue prediction [24]. In particular, owing to the complexity of multi-omics data, shallow GCNs may struggle with the higher-order feature exploration. Thus, with limited samples, adopting deeper networks becomes crucial to the improvement of multi-omics feature learning performance.

In summary, the existing methods, especially those graph-based multi-omics models, generally face the following challenges: (1) Neglecting the *relationships* between different omics data types; (2) Overlooking the *incorporation of patient similarity*; (3) Ignoring the *high-order relationships* between omics data samples. Consequently, to tackle these issues, we propose a multi-omics data integration approach, called DeepMoIC, for cancer subtype classification, as presented in Fig. 1. Initially, autoencoders are employed to extract latent embedding representations from multi-omics data, providing a compact data representation across multiple omics modalities. Subsequently, a PSN is constructed using the similarity network fusion algorithm. To effectively handle non-Euclidean data represented by the latent omics data representation and PSN, we design a Deep GCN for in-depth exploration of high-order information. To handle the challenge of capturing high-order correlation of samples in multi-omics data, we implement two effective strategies: initial residual connection and identity mapping, which facilitate the propagation of omics information to remote neighbors. The proposed method is trained and assessed on 4 benchmark datasets and compared against other state-of-the-art methods in multimodal and multi-omics learning fields. The experimental results indicate that DeepMoIC consistently achieves significant improvements across all datasets, suggesting its potential to provide deeper insights into clinical diagnosis and cancer subtype classification.

Methods

The proposed DeepMoIC comprises three main components. First, multi-omics data are input into autoencoders to extract compact representations. Then, the similarity network fusion method is applied to construct a PSN

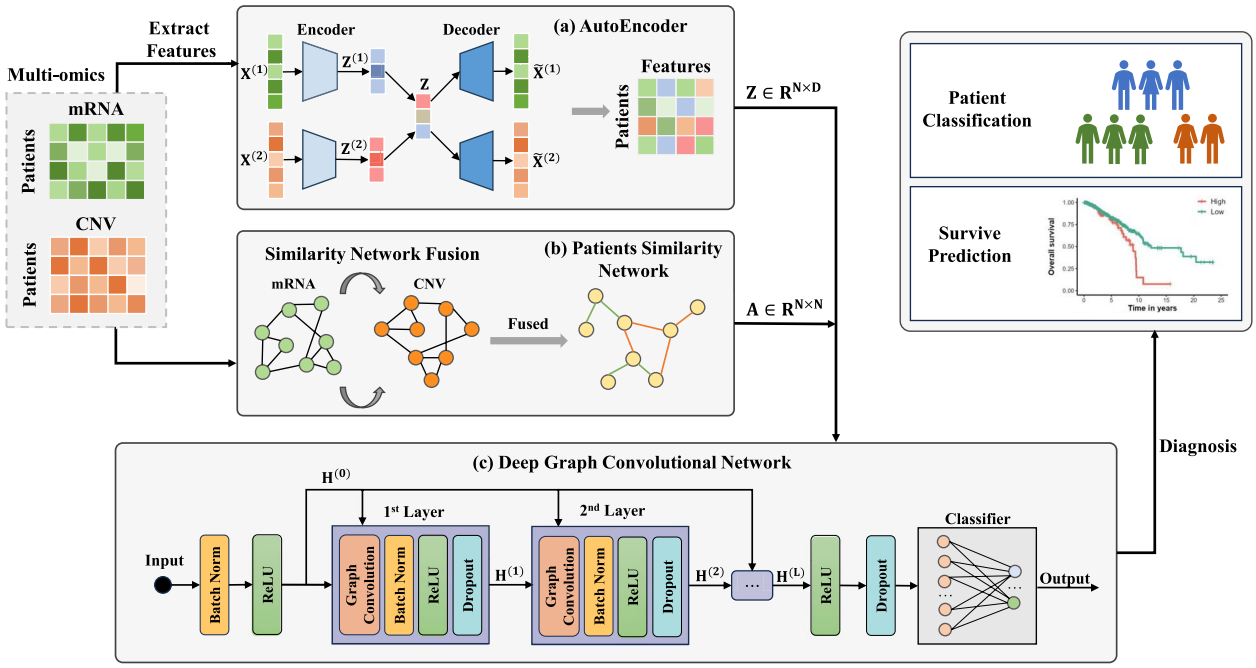


Fig. 1 The overall workflow of DeepMoIC, consisting of two main stages: **a**/**b** Utilizing AutoEncoder to extract features and constructing a patient similarity network (PSN) through the similarity network fusion algorithm, and **c** Deep GCN module to process the PSN and feature matrices for downstream tasks

structure. Finally, the Deep GCN module integrates the feature matrices and PSN for network training and cancer subtype prediction.

Autoencoder architecture

To cope with challenges posed by limited samples and high-dimensional genomic features in multi-omics data analysis, we first utilize a multi-layer autoencoder to reduce data dimensionality and computational cost. In detail, the i-th encoder that learns the compressed representation $\mathbf{Z}_i$ of multi-omics features is defined as

$$\mathbf{Z}_i^{(l)} = f_e\left(\mathbf{Z}_i^{(l-1)}\right) = \sigma\left(\mathbf{W}_i^{\top(l)} Z_i^{(l-1)} + b_i^{(l)}\right), \quad (1)$$

where $\mathbf{Z}_i^{(0)} = \mathbf{X}_i$ that denotes the i-th omics features, $\mathbf{W}_i^{(l)}$ denotes the weight matrix, $b_i^{(l)}$ denotes the bias of the l-th layer and σ is the sigmoid activation function. Subsequently, we employ the decoder layers to learn a reconstructed representation, defined as

$$\tilde{\mathbf{X}}_i^{(l)} = f_d\left(\mathbf{Z}_i^{(l)}\right) = \sigma\left(\mathbf{W}_i^{\top'(l)} \mathbf{Z}_i^{(l)} + b_i^{'(l)}\right), \quad (2)$$

where $\tilde{\mathbf{X}}_i^{(l)}$ represents the rebuilt features. To minimize the reconstruction loss, we employ the Mean Square Error (MSE) loss function to quantify the difference between the rebuilt and the original feature matrices, defined as

$$\mathcal{L}_{MSE}\left(\mathbf{x}_i, \tilde{\mathbf{x}}_i^{(L)}\right) = \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{x}_i - \tilde{\mathbf{x}}_i^{(L)}||^2, \quad (3)$$

where n denotes the count of samples and L denotes the count of layers. Considering that the input data encompasses multiple data types and is represented by different features $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_M$, we assign varied weights to each omics data based on prior knowledge to underscore their impact to the model, with all weights summing up to one. Considering this, the loss function is formulated as

$$\mathcal{L}_{AE} = \sum_{i=1}^{M} \lambda_i \mathcal{L}_{MSE}\left(\mathbf{X}_i, \tilde{\mathbf{X}}_i^{(L)}\right), \quad (4)$$

where M denotes the count of omics types with $\sum_{i=1}^{M} \lambda_i = 1$. Finally, we obtain the unique latent representation extracted from multi-omics data with a weighted integration, i.e., $\mathbf{Z} = \sum_{i=1}^{M} \lambda_i \mathbf{Z}_i^{(L)}$.

Patient similarity network

In typical multi-omics data, establishing direct relationships between samples is challenging due to the diversity of biological information, especially when handling gene expression and protein levels. Existing omics data obtained through independent experiments or measurements often lack inherent relationships between samples. Thus, constructing semantic corrections between

samples is essential for a comprehensive understanding of multi-omics data, especially when bridging the gap between diverse omics data.

In our work, we adopt a Similarity Network Fusion (SNF) algorithm [25] which is designed to build a similarity network among patients based on various types of data. Specifically, for each data type, SNF algorithm first generates the patient similarity matrix and then constructs the corresponding patient adjacency matrix for each omics type. Finally, the algorithm combines different categories of patient similarity matrices to construct a fused graph.

Assume that there are n samples and M types of features (such as mRNA, CNV, and DNA methylation). For the m-th data type, the scaled exponential similarity matrix is computed by

$$\mathbf{S}_{i,j} = \exp\left(-\frac{\theta^2(\mathrm{x_i}, \mathrm{x_j})}{\mu\delta_{\mathrm{i,j}}}\right), \tag{5}$$

where $\theta(x_i, x_j)$ denotes the Euclidean distance within samples x_i and x_j, μ is a hyperparameter, and $\delta_{i,j}$ is employed to address the scaling issue, which is computed by

$$\delta_{i,j} = \frac{mean(\theta(x_i, N_i)) + mean(\theta(x_j, N_j)) + \theta(x_i, x_j)}{3}, \tag{6}$$

where N_i denotes the set of x_i's neighbors and $\mathrm{mean}(\theta(\mathrm{x_i}, \mathrm{N_i}))$ denotes the mean distance from node x_i to each neighbor. To calculate the fused matrix from different omics types, the similarity matrix of all samples is calculated by

$$\mathbf{P}_{i,j} = \begin{cases} \frac{\mathbf{S}_{i,j}}{2\sum_{v\neq i}\mathbf{S}_{i,v}}, & j \neq i, \\ \frac{1}{2}, & j = i. \end{cases} \tag{7}$$

Subsequently, the similarity matrix $\mathbf{K}$ recording the k nearest neighbors is calculated by

$$\mathbf{K}_{i,j} = \begin{cases} \frac{\mathbf{S}_{i,j}}{\sum_{v\in N_i}\mathbf{S}_{i,v}}, & j \in N_i, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Observe that matrix $\mathbf{P}$ encompasses the complete messages regarding the similarity of each sample to all others, while matrix $\mathbf{K}$ only represents the similarity to k most similar samples for each individual and k is set to 20 in our work. In the case of various data types ($M > 2$), different omics similarity matrices are fused by an iterative process, i.e.,

$$\mathbf{P}_{t+1}^{(m)} = \mathbf{K}^{(m)} \times \left(\frac{\sum_{v\neq m}\mathbf{P}_t^{(v)}}{M-1}\right) \times (\mathbf{K}^{(m)})^{\top}, \tag{9}$$

where $\mathbf{P}_t^{(m)}(m = 1, 2, ..., M)$ denotes results at the t-th iteration from the m-th omics data, and the iteration process continues until the algorithm reaches convergence or the specified iteration budget. Finally, the PSN matrix is defined as

$$\mathbf{P} = \frac{\mathbf{P}_t^{(1)} + \mathbf{P}_t^{(2)} + ... + \mathbf{P}_t^{(M)}}{M}. \tag{10}$$

Construction of deep graph convolutional network

After obtaining the compressed intact node feature matrix $\mathbf{Z} \in \mathbb{R}^{n\times d}$ with AE and generating the PSN matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$ through SNF algorithm, we construct a deep graph convolutional network module to effectively process the non-Euclidean data, aiming to learn latent representations. By harnessing the deep GCN, we elevate the capability of multi-omics data to uncover intricate relationships. The multiple layers of deep GCN can facilitate the exploration of high-order connectivity information, empowering the model to capture more nuanced representations and enhance predictive performance. Nevertheless, a primary problem of deep GCN is the over-smoothing issue, where node representations become excessively similar with the growing number of layers. This prevents the model from discovering high-order information in multi-omics data. To extend GCN into a deeper model, we employ two strategies to tackle this challenge: initial residual connection [26] and identity mapping [27]. Then, we design the deep GCN module based on two strategies better to handle multi-omics fusion data for different downstream tasks.

Incorporating the initial residual connection guarantees that, despite stacking multiple layers, the eventual representation of each node keeps the information from the input layer in a fraction of α. The message-passing operation is formulated as

$$\mathbf{H}^{(l+1)} = \sigma\left((1-\alpha)\tilde{\mathbf{L}}\mathbf{H}^{(l)} + \alpha\mathbf{H}^{(0)}\right), \tag{11}$$

where α is a hyperparameter. Recall that $\tilde{\mathbf{L}} = \tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}$, where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ represents the adjacency matrix with additional self-connections, $\tilde{\mathbf{D}}$ is the degree matrix of $\tilde{\mathbf{A}}$ and initially $\mathbf{A} = \mathbf{P}$.

While the initial residual connection provides partial relief from over-smoothing, the decline in performance persists as the model deepens. To address this issue, the identity matrix is introduced to the weight matrix, and the message-passing operation is formulated as

$$\mathbf{H}^{(l+1)} = \sigma\left(\left((1-\alpha)\tilde{\mathbf{L}}\mathbf{H}^{(l)} + \alpha\mathbf{H}^{(0)}\right)\left((1-\beta_l)\mathbf{I} + \beta_l\mathbf{W}_d^{(l)}\right)\right), \tag{12}$$

where $\beta_l = log\left(\frac{\lambda}{l}+1\right) \approx \frac{\lambda}{l}$ when $\frac{\lambda}{l}$ is small enough, and λ is a hyperparameter. Note that $\mathbf{H}^{(l)}$ is the outcome of the previous layer and $\mathbf{H}^{(0)} = \mathbf{Z}$. $\sigma(\cdot)$ is the activation function and $\mathbf{W}_d^{(l)}$ is a learnable weight matrix for the l-th layer. The factor β is strategically set to make sure that the decay of the weight matrix increases adaptively with the stacking of more layers.

Subsequently, we employ the aforementioned basic layer to construct a deep GCN, where each layer is structured in the following sequence: *Graph Convolution* → *Batch Normalization* → *ReLU* → *Dropout*. Herein, *Batch Normalization* [28] is employed to improve the stability of the model training, which standardizes inputs within each mini-batch and not only accelerates the convergence speed but also alleviates the issues of gradient vanishing and exploding.

The outcome of the last layer of deep GCN will be fed into a linear classifier, which will then be fed into the softmax layer to obtain the classification probability distribution for the loss function computation. To quantify the difference between the predicted results and the ground truth, we employ the cross-entropy loss function, i.e.,

$$\mathcal{L} = -\frac{1}{q}\sum_{i=1}^{q}\sum_{c=1}^{C} y_{ic} log(p_{ic}), \tag{13}$$

where q is the count of training samples and c is the count of classes. If $y_{ic} = 1$, the ground truth of the i-th node is c, and p_{ic} represents the predicted confidence of the i-th node belonging to the c-th class. The entire procedure is summed up in Algorithm 1.

Algorithm 1 Training Algorithm of DeepMoIC

Input: Multi-omics data $\mathcal{X} = \mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_m$; The ground truth label of train samples $\mathcal{Y}_{train}$; The number of hidden layers L and hyperparameters α and λ for deep GCN.
Output: The predicted label of test samples $\mathcal{Y}_{test}$.
1: Calculate fused feature matrix $\mathbf{Z} \in \mathbb{R}^{n\times d}$ using the autoencoder architecture;
2: Construct the PSN matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$ with the SNF method;
3: Initialize $\mathbf{H}^{(0)} = \mathbf{Z}$, $\mathbf{A} = \mathbf{P}$;
4: **while** not convergent **do**
5: Compute the hidden representations $\{\mathbf{H}^{(l)}\}_{l=1}^{L}$ by Eq. (12);
6: Compute the loss value $\mathcal{L}$ by Eq. (13);
7: Update trainable parameters $\{\mathbf{W}_d^{(l)}\}_{l=1}^{L}$ with back propagation;
8: **end while**
9: **Return** The predicted label $\mathcal{Y}_{test}$.

Results

Data preparation

For the task of pan-cancer subtype classification, we leverage the TCGA Pan-cancer dataset [29], which integrates RNA-seq and Copy Number Variation (CNV) dataset. The CNV dataset via GISTIC2 method consists of 10,845 samples, while the batch effect normalized RNA-seq dataset encompasses 11,060 samples. After filtering out missing data, the eventual TCGA Pan-cancer dataset comprises 9,664 samples from 28 distinct subtypes.

Table 1 Statistics of multi-omics datasets

Datasets	Samples	Features	Subtypes
TCGA	9,664	gene expression (17,944)	28
		CNV(17,944)	
BRCA	511	mRNA (19,580)	4
		CNV (19,273)	
		RPPA (223)	
KIPAN	707	meth (2,000)	3
		mRNA (2,000)	
		miRNA (472)	
LGG	524	meth (2,000)	2
		mRNA (2,000)	
		miRNA (548)	

For the specific task of recognizing cancer subtypes, we use 3 cancer subtype datasets. The BRCA dataset is used for breast invasive carcinoma PAM50 subtype classification, which comprises 3 omics data types: mRNA, CNV, and Reverse-Phase Protein Array (RPPA), and encompasses 511 samples from 4 subtypes: Luminal A, Luminal B, TNBC, and HER2(+). The KIPAN dataset is used for kidney cancer type classification, which comprises 3 omics data types: DNA methylation, miRNA, and mRNA, and encompasses 707 samples from 3 subtypes: KICH, KIRC, and KIRP. The LGG dataset is used for grade classification in glioma, which comprises 3 omics data types: DNA methylation, miRNA, and mRNA, and encompasses 524 samples from 2 subtypes: Grade 2 and Grade 3. To facilitate survival prediction tasks on the BRCA dataset, we also retrieve clinical information from the GDC Data Portal (https://portal.gdc.cancer.gov/). In the experiment, 60% of the samples are randomly selected as the training set and the rest as the test set. The detailed dataset statistics are provided in Table 1.

Cancer subtype classification

Several experiments are conducted to evaluate the performance and efficacy of the proposed DeepMoIC. We compare our proposed method with 9 methods, including classical machine learning methods and state-of-the-art deep learning methods. Specifically, Supporting Vector Machine (SVM), Random Forest (RF), and *K*-Nearest Neighbor (KNN) are single-view baselines. Multi-view techniques have been demonstrated

to possess the advantages of uncovering cross-talk patterns and capturing the heterogeneity of samples in multi-omics data mining [30], attributed to which we compare them with two multi-view methods, i.e., CoGCN [31] and ERL-MVSC [32]. Finally, some recently proposed multi-omics data analysis approaches, including DeepMO [8], MOGONET [3], MoGCN [20] and Moanna [33], are also compared in our experiments.

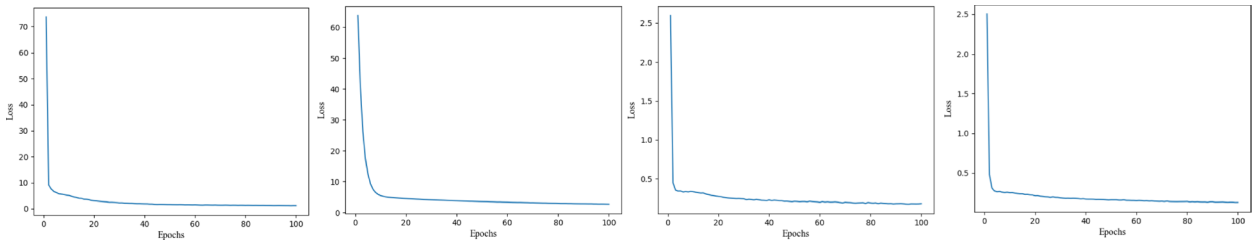
For single-view methods, the multi-omics features are combined to construct an integrated feature matrix. For a fair comparison, all compared methods are tested by their default settings. For DeepMoIC, the hidden dimensions of the autoencoders are set to 1,000 for the TCGA and 100 for the other datasets, because the TCGA dataset has more samples and we ought to keep more features for a comprehensive description. The count of layers of autoencoder is set to 1. During the training of the autoencoder, the training epoch is set to 100, as shown in Fig. 2, which eventually converges after 20 epochs across all datasets. During the training of the deep GCN module, the learning rate is configured as 0.001, and the training epochs are set to 300. Through the gird search, the hyperparameters α and λ are both set to 0.5 on all datasets. The hidden unit dimensions of deep GCN are set to 512 for the TCGA and 64 for the other datasets.

For all methods, every experiment is conducted 5 times, and we utilize the mean results along with the standard deviation as the eventual result. Table 2 provides the performance comparison of the proposed method and other baselines, which is assessed by Accuracy, F1 score, Precision, and Recall. From experimental results, we can observe that DeepMoIC performs superior in the cancer subtype classification task on all datasets.

By comparing the classification performance of these datasets, we can observe that some existing multi-omics data analysis methods obtain undesired performance with a small sample size, such as on the BRCA dataset. However, on the TCGA dataset with a larger sample size, these methods fail to achieve satisfactory performance and even may perform unfavorably compared to some single-view methods. This further demonstrates the robustness and scalability of our proposed method. Besides, MoGCN is characterized as shallow GCN models that exhibit inferior performance compared to DeepMoIC. Notably, the performance of MoGCN on the TCGA dataset significantly declines compared with that on the smaller BRCA dataset. This observation underscores the superior capability of deep GCN in extracting intricate relationships within multi-omics data. It is plausible that the limitations of DeepMO and Moanna, which

Table 2 Performance (mean% ± std%) comparison of all compared algorithms

Methods	Accuracy	F1 score	Precision	Recall	Accuracy	F1 score	Precision	Recall
Dataset	**TCGA**				**BRCA**			
SVM	81.55(0.00)	73.88(0.00)	76.15(0.00)	73.51(0.00)	87.01(0.00)	84.32(0.00)	90.59(0.00)	81.26(0.00)
RF	81.11(0.26)	70.32(0.44)	66.66(0.73)	74.41(0.37)	84.16(1.09)	77.90(0.84)	70.08(1.36)	87.71(0.58)
KNN	64.79(0.00)	47.26(0.00)	40.85(0.00)	56.06(0.00)	81.82(0.00)	77.45(0.00)	68.40(0.00)	89.27(0.00)
CoGCN	71.80(0.60)	64.51(0.63)	68.24(1.51)	63.88(0.62)	78.63(1.57)	74.31(2.11)	76.96(2.47)	72.66(2.03)
ERL-MVSC	80.03(0.16)	77.75(0.46)	77.64(0.67)	77.86(0.50)	87.51(0.74)	86.38(1.7)	89.73(2.36)	83.28(1.24)
DeepMO	72.43(4.77)	73.02(2.65)	77.08(1.33)	73.48(4.35)	86.05(1.76)	84.98(1.54)	83.90(2.06)	86.94(1.13)
MOGONET	67.03(0.69)	52.89(0.56)	53.83(0.84)	60.17(0.15)	77.76(2.22)	74.85(3.72)	81.26(1.24)	74.03(3.05)
MoGCN	72.06(0.79)	52.28(2.29)	54.89(2.33)	59.31(1.90)	89.37(0.78)	87.51(0.95)	89.54(0.65)	85.98(1.09)
Moanna	80.72(0.42)	74.49(0.79)	77.03(0.37)	73.80(0.92)	90.73(0.78)	88.02(1.28)	89.43(1.31)	87.13(1.18)
DeepMoIC	**84.28(0.16)**	**81.01(0.35)**	**81.97(0.17)**	**81.09(0.49)**	**92.98(0.66)**	**91.52(0.71)**	**92.22(0.56)**	**90.88(0.84)**
Dataset	**KIPAN**				**LGG**			
SVM	93.43(0.00)	93.91(0.00)	93.94(0.00)	93.88(0.00)	68.75(0.00)	67.82(0.00)	69.45(0.00)	68.09(0.00)
RF	93.90(0.74)	88.15(1.44)	88.18(1.45)	88.13(1.42)	72.66(2.62)	60.04(2.33)	59.80(2.22)	60.28(2.22)
KNN	91.55(0.00)	83.66(0.00)	84.44(0.00)	84.05(0.00)	63.29(0.00)	53.67(0.00)	52.85(0.00)	54.53(0.00)
CoGCN	94.61(0.65)	93.33(0.64)	93.92(0.50)	92.78(0.79)	68.86(2.29)	68.67(2.51)	69.59(1.66)	69.07(2.16)
ERL-MVSC	94.13(1.36)	92.51(1.73)	91.06(1.79)	94.02(1.98)	70.67(2.56)	71.92(2.21)	72.39(2.03)	71.46(2.40)
DeepMO	95.11(1.00)	93.56(1.44)	92.53(1.97)	94.91(1.62)	72.00(2.53)	71.83(2.55)	72.01(2.60)	71.81(2.55)
MOGONET	94.84(0.42)	93.37(0.85)	95.16(0.36)	91.91(1.17)	65.71(3.20)	65.03(3.84)	68.45(1.38)	66.71(2.70)
MoGCN	92.08(0.36)	92.47(0.41)	92.41(0.31)	92.67(0.65)	67.14(2.61)	66.41(3.90)	67.86(1.40)	66.90(2.98)
Moanna	93.78(0.62)	93.73(0.80)	93.51(0.68)	94.16(1.01)	69.33(1.69)	68.50(1.97)	70.49(1.80)	68.84(1.76)
DeepMoIC	**96.25(0.17)**	**95.36(0.13)**	**95.39(0.14)**	**95.35(0.12)**	**73.24(1.39)**	**73.18(1.41)**	**73.21(1.39)**	**73.17(1.42)**



(a) TCGA (b) BRCA (c) KIPAN (d) LGG

Fig. 2 The training loss curves of the autoencoder

solely rely on feature attributes, and MOGONET, which neglects the exploration of relationships between different histological data during the construction of the similarity network, contribute to their comparatively poor performance.

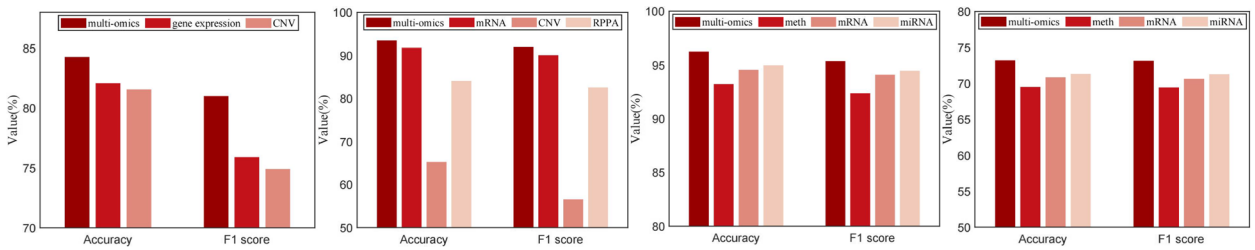
As presented in Fig. 3, we also compare the classification performance of single-omic and multi-omics data using the DeepMoIC, we can observe that the performance of utilizing single-omics data is inferior to that of utilizing multi-omics data across all datasets. This proves that research on cancer subtype classification can benefit from integrating multi-omics data to account for multiple perspectives. By leveraging diverse biological data types, we can gain a deeper and more accurate understanding of cancer subtypes, leading to improved classification accuracy and robustness.

Application to survive analysis

In this subsection, survival prediction experiments are performed to further validate the effectiveness of the proposed DeepMoIC. First, we use the trained model to predict the test set of BRCA to get subtype classification results. Then, utilizing the lifelines package in Python, we plotted Kaplan-Meier survival curves for two subtypes of the test set, Luminal A and Luminal B, which help describe and compare the survival probabilities of different groups over time. Log-rank tests are employed to assess distinctions among the survival curves, and log-rank *P*-values are calculated to indicate the statistical significance of the observed differences between the survival curves. Differences are deemed statistically significant at *P*-values < 0.05, and the smaller the *P*-value, indicating that the results have greater statistical significance. In Fig. 4, the survival prediction results of all compared algorithms are presented. The performances of DeepMO and Moanna are even worse than SVM and RF, which is further evidence that one cannot rely solely on feature attributes to study complex multi-omics data. The *P*-value of 4.8e−03 reveals the better subtype prediction results of DeepMoIC compared with other methods, highlighting the robust performance and general applicability of DeepMoIC.

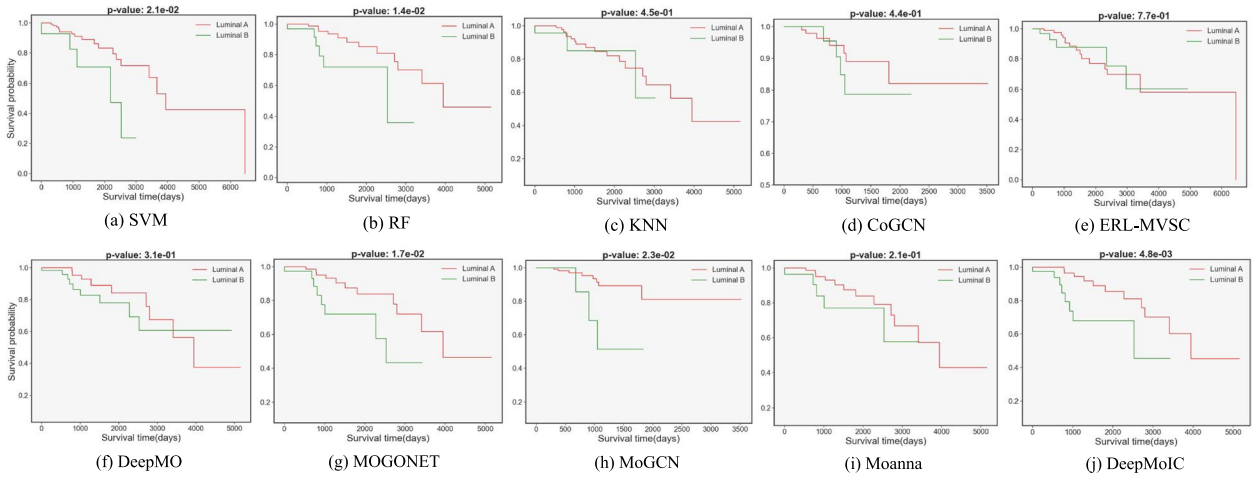
Identification of significant biomarkers by AE

In this part, we use the autoencoder to extract the important gene of mRNA data of KIPAN at the transcriptome level, and enrichment analysis is performed. Specifically, we train AE for 100 epochs up to convergence, extracting 100 genes with the highest score every 10 epochs. The score was calculated by multiplying the sum of the absolute values of the weights of the first encoder layer by the standard deviation of each raw feature, ultimately resulting in a total of 135 genes. Figure 5 presents the Biological Process (BP), Molecular Function (MF), and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotations using DAVID ($p < 0.05$) [34, 35]. For biological processes, the genes are involved in cell proliferation, migration, adhesion, transcription regulation, and so on, which are crucial



(a) TCGA (b) BRCA (c) KIPAN (d) LGG

Fig. 3 The accuracy and F1 score comparison of DeepMoIC with single-omic and multi-omic data on all datasets



(a) SVM (b) RF (c) KNN (d) CoGCN (e) ERL-MVSC

(f) DeepMO (g) MOGONET (h) MoGCN (i) Moanna (j) DeepMoIC

Fig. 4 Survive prediction of all compared algorithms on BRCA dataset

for cancer progression and metastasis. For instance, genes such as EGLN3, JAG1, and JUP were highlighted in the regulation of cell proliferation, indicating their role in tumor growth. Additionally, the BP analysis identifies key pathways like the Wnt signaling pathway, which is critical in driving cancer development [36]. In terms of molecular functions, the analysis revealed significant associations with sequence-specific DNA binding, growth factor binding, and receptor binding, which are critical in cancer cell signaling and communication. The KEGG pathway enrichment analysis showed that these biomarkers are significantly enriched in cancer-related pathways, including pathways in cancer, the p53 signaling pathway, cell adhesion molecules, and so on. These pathways play crucial roles in oncogenic processes and cancer progression. For example, the p53 signaling pathway is essential for regulating cell cycle and apoptosis, processes often dysregulated in cancer, and its role in kidney cancer is particularly critical for tumor suppression and genomic stability [37]. JUP appeared in several key pathways, including the Wnt signaling pathway and cell adhesion molecules, emphasizing its multifunctional role in cancer progression. These findings demonstrate that the AE effectively captures significant biomarkers that are crucial for understanding cancer development and progression. The enrichment of biomarkers in significant BP, MF, and KEGG validates the capability of AE to extract meaningful features for cancer subtype classification. This provides valuable insights for further research and potential therapeutic strategies.

Parameter sensitivity analysis

To explore the impact of depth of deep GCN, we design several models with diverse numbers of layers, and the compared results are displayed in Fig. 6. We can find that as the count of layers increases, the performance rises gradually. This highlights the suitability of deep GCN for fitting omics data, outperforming shallow GCN in cancer subtype classification. When each dataset reaches a certain number of layers, metrics such as Accuracy and F1-Score achieve optimal performance. Nevertheless, as the count of layers keeps increasing, the performance gradually decreases. This is probably because although DeepMoIC overcomes over-smoothing to some extent, having more layers does not necessarily lead to better performance. Moreover, because graph networks are built with multi-omics integration data with significant differences in node connectivity, graph sparsity, information propagation path lengths, and so on, different datasets require specific numbers of layers in deep GCN to achieve optimal performance. Therefore, we need to choose different layers for different multi-omics datasets. In the experiment, the number of layers of DeepMoIC is set to 8 for TCGA and LGG, 20 for BRCA, and 16 for KIPAN.

For another important parameter α of the deep GCN, which controls the initial residual ratio. We compared the performance of α from 0.1 to 1.0 to verify the effectiveness of the initial residual. In Fig. 7, we can observe that when increases from 0.1 to 0.5, the classification performance on all datasets gradually increases. At $\alpha = 0.5$, the model achieves optimal or near-optimal performance across all datasets. This suggests that a balanced contribution from both the initial features and the learned embedding representation is crucial for maximizing model performance. However, after 0.5, for datasets like LGG, which have relatively poor data quality, adding too much initial feature information will lead to a decrease in performance. Other datasets can maintain a relatively stable state but may not achieve the best results. Therefore, we finally set α to 0.5 to achieve a more robust and stable result.

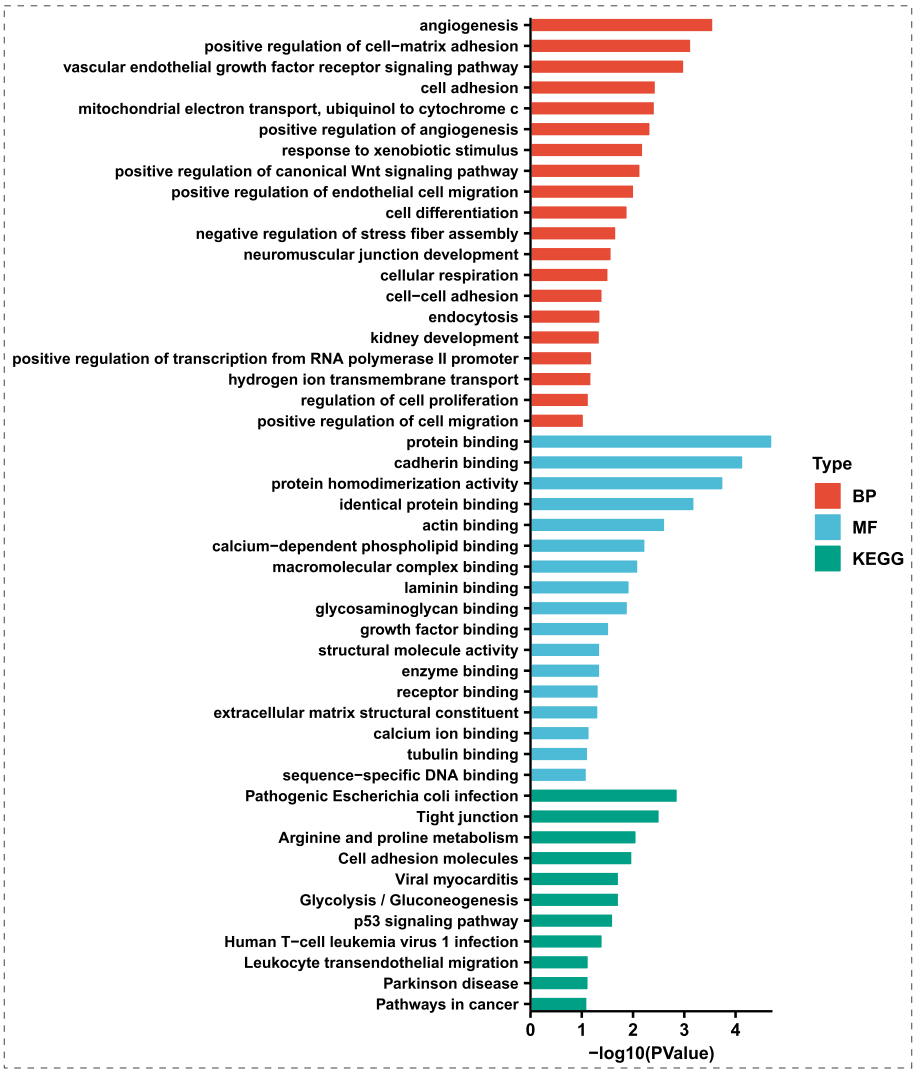


Fig. 5 Biological Process, Molecular Function, and KEGG pathway annotations of mRNA data from KIPAN

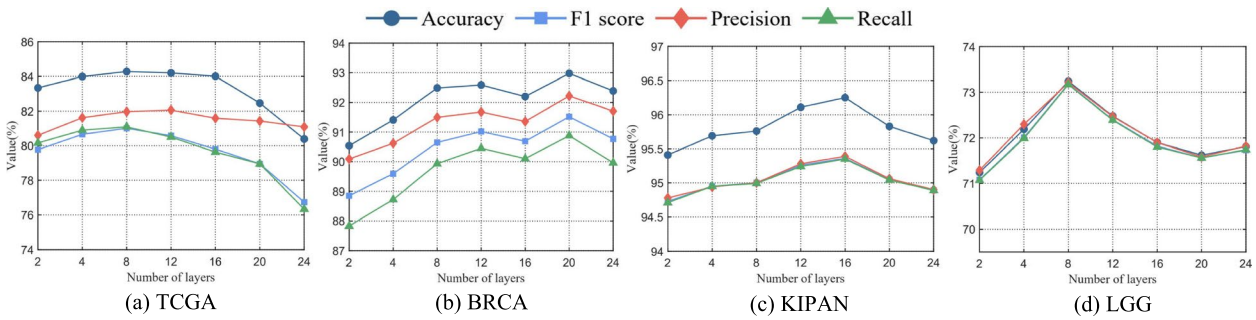


Fig. 6 Classifiaction performance comparison of DeepMolC with different numbers of layers on all datasets

For the similarity matrix **K** in the PSN construction process, which records the k nearest neighbors. We compare different k to evaluate the impact of the number of neighbors on the performance of the method. As shown in Fig. 8, when k=10, we can see that the proposed method cannot achieve good results on all datasets,

because it cannot capture enough information from similar neighbors. When k=30, except for KIPAN, the performance of other datasets has declined compared with k=20, because too much neighbor information may also bring more noise in the information aggregation process and affect the performance. So in the end we choose k as 20 to achieve a more balanced result.

Effect of initial residual and identity mapping

As depicted in Fig. 9, we perform an ablation experiment on all datasets, focusing on two strategies for constructing deep GCN. The results reveal that the over-smoothing issue can be partially mitigated with the inclusion of initial residuals, and performance is sustained even with an increased number of layers, though it does not surpass the performance achieved using both strategies. In contrast, utilizing only identity mapping leads to solid performance in shallow layers but exhibits a rapid decline as the count of layers increases. Similar performance degradation is observed in the absence of either policy, with a notable drop occurring as the count of layers increases. Thus, the optimal approach involves employing a combination of both strategies for improved performance.

Effect of patient similarity network

A key module of DeepMoIC is the construction of the PSN matrix, which is a pivotal component that enables the model to harness information from neighborhood patients. To validate the impact of the PSN matrix on cancer diagnosis, we conduct an ablation study. Two experiments are designed: one with the PSN matrix as input, whereas the other using the identity matrix as input. Experimental results in Fig. 10 reveal that the model trained with the PSN outperforms its counterpart trained without it. This emphasizes the crucial role of the PSN matrix within the DeepMoIC framework. By incorporating information from neighborhood patients, our proposed method exhibits enhanced classification results. This stresses the effectiveness of utilizing the PSN information to enhance the robustness of cancer diagnosis.

Discussion

Multi-omics research is becoming increasingly relevant in cancer research due to its potential to provide a comprehensive view of the biological processes involved in cancer development and progression. By integrating data from several omics types, such as genomics, transcriptomics, and proteomics, it can unveil complicated connections and identify crucial biomarkers that single-omics techniques may overlook. To address obstacles in multi-omics research, we propose a novel multi-omics data integration approach named DeepMoIC, specifically designed for cancer subtype classification. The

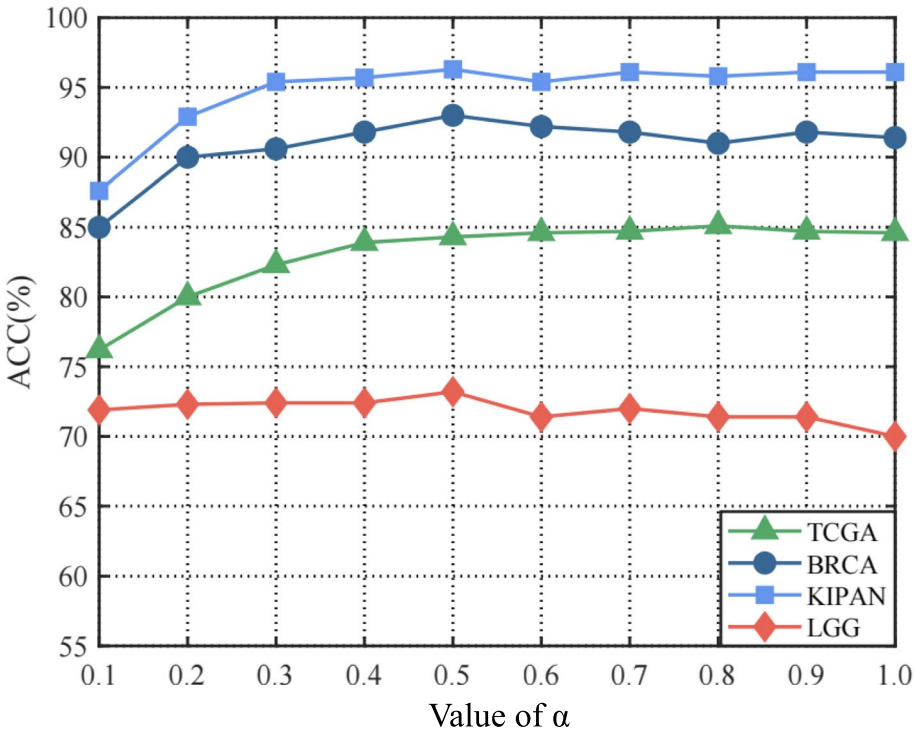


Fig. 7 The classification accuracy of the proposed method w.r.t. parameter α

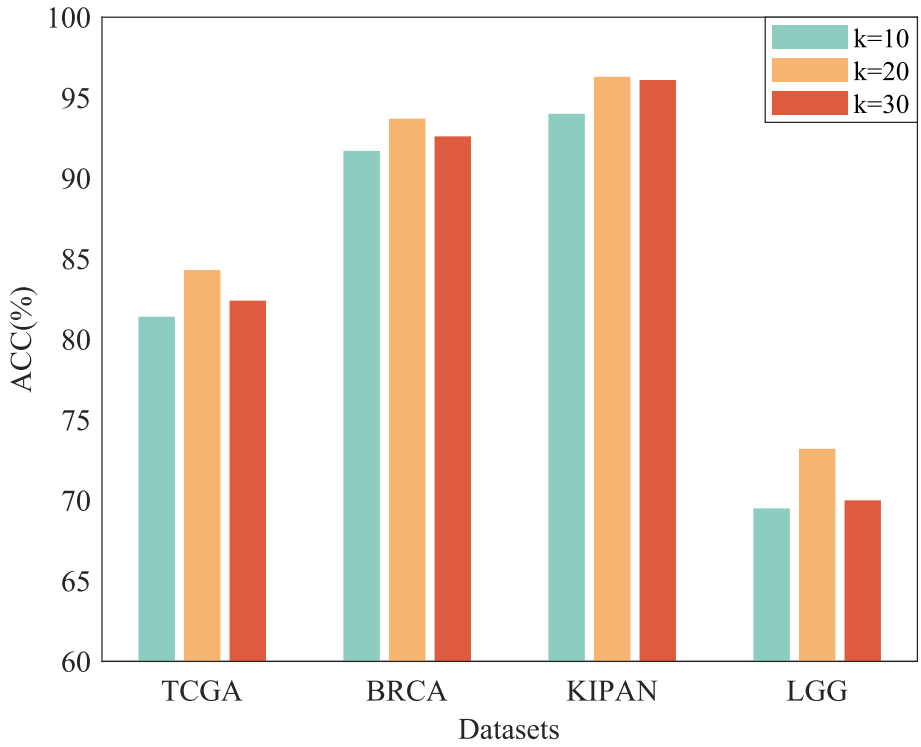


Fig. 8 The classification accuracy of the proposed method w.r.t. parameter *k*

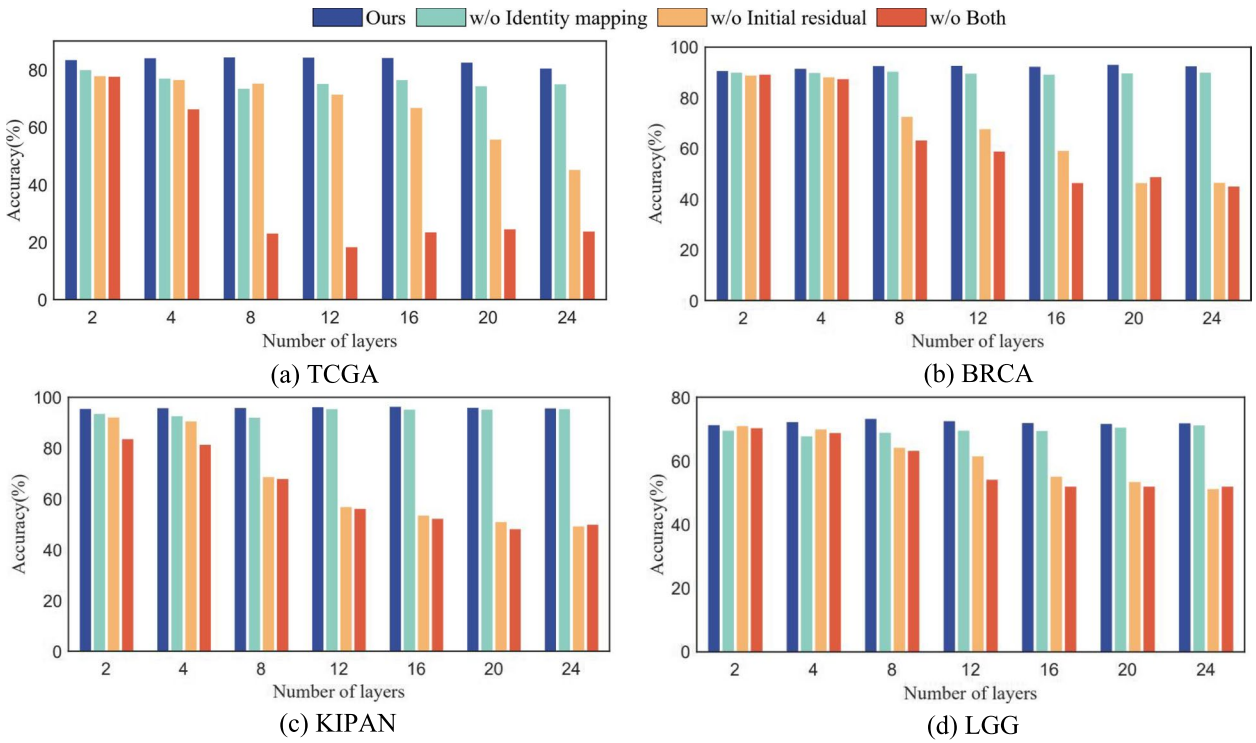


Fig. 9 Ablation study on initial residual connection and identity mapping of deep GCN module

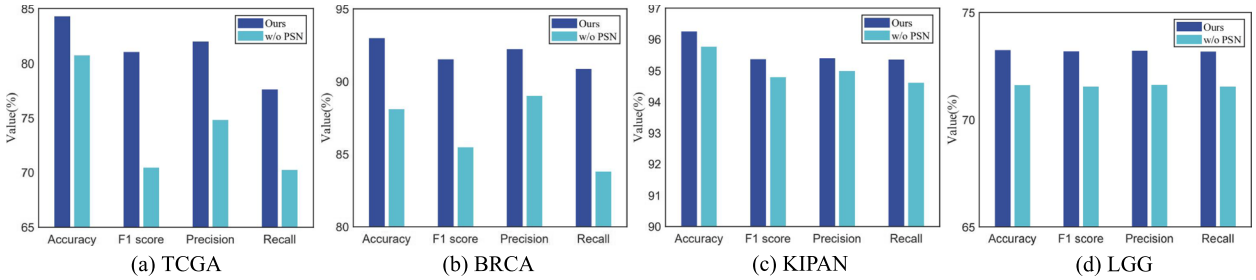


Fig. 10 Ablation study w.r.t. the PSN matrix

proposed DeepMoIC leverages deep GCN based on two novel strategies to improve performance and robustness. Firstly, we apply the AE to extract compact and meaningful representations from high-dimensional omics data. This procedure minimizes noise and improves feature quality, facilitating more accurate downstream analysis. Additionally, we find that AE can capture important biomarkers, significantly improving the interpretability of DeepMoIC by providing insights into the biological significance of the extracted features, as well as indicating that the reduced feature matrix is highly informative and improves classification performance. Furthermore, DeepMoIC incorporates a patient similarity network (PSN) into the model, improving comprehension of patient relationships using multi-omics data. The PSN captures complex interactions across different omics types, allowing the model to identify patient groups with similar molecular characteristics. This relational insight is essential for accurate cancer subtype classification. The experiment also demonstrates that including the PSN considerably improves classification performance.

Finally, we design a deep GCN module based on initial residual connection and identity mapping for the cancer subtype classification. This module is critical for extracting high-order features from multi-omics data, which are required to understand complex relationships between multiple omics layers. High-order features enable the model to detect subtle patterns and associations that shallow architectures might miss, leading to more precise and significant biological discoveries. The experiments reveal that deeper GCN architectures outperform shallower ones in classification tasks, up to an optimal depth, which varies across datasets. For example, the optimal depth for the TCGA is 8, while for the BRCA is 20. This variation is likely due to the different complexities and characteristics of each dataset. Currently, we determine the optimal number of layers through extensive experimentation. While this approach has proven effective, future research could explore more systematic and theoretically grounded methods for selecting the optimal number of layers, tailored to the specific complexities of various datasets. To overcome the over-smoothing problem, we employ two strategies: initial residual connections and identity mapping. These strategies enable the deep GCN module to leverage the benefits of deep architectures while preserving input data integrity. This balance leads to superior performance in classification tasks, as evidenced by our experiments. The ability to capture and preserve high-order features improves the accuracy and robustness of cancer subtype classification, highlighting the efficacy of our multi-omics integration strategy.

Conclusion

The proposed DeepMoIC method leverages deep Graph Convolutional Network and effectively addresses challenges in multi-omics studies. By efficiently extracting compact omics representations with autoencoder and integrating patient similarity networks, DeepMoIC significantly enhances the understanding of cancer, leading to improved performance in cancer subtype classification. The results demonstrate that DeepMoIC consistently outperforms all compared methods across all datasets, potentially setting a new benchmark for supervised multi-omics integration and offering enhanced precision medicine insights in cancer research.

Acknowledgements

We thank the authors of the cited references for their excellent work.

Authors' contributions

JW and SW conceptualized the study; All authors conceived and designed the final study; GL and WW provided several valuable suggestions in the development of the algorithm. JW and SX completed data processing and programming implementation; JW and ZC drafted the manuscript; All authors reviewed and approved the final version of the manuscript.

Funding

This work was supported by the National Natural Science Foundation of China under Grant U21A20472 and 62276065, the National Key Research and Development Plan of China under Grant 2021YFB3600503, the Medical Innovation Program of Fujian Province under Grant 2023CXA001 and the Natural Science Foundation of Fujian Province under Grant 2022J011006.

Data availability

The TCGA Pan-cancer dataset was obtained from the GitHub repository https://github.com/Xiaoshunxin/MPK-GNN. The BRCA dataset was obtained from the GitHub repository https://github.com/Lifoof/MoGCN. The KIPAN and LGG datasets were downloaded from https://drive.google.com/drive/folders/1_tJ2ekhTmWp7ZcRVjUVGx0cqGMRKEhNo [5].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 17 June 2024 Accepted: 2 December 2024
Published online: 18 December 2024

References

1. Garraway LA, Lander ES. Lessons from the cancer genome. Cell. 2013;153(1):17–37.
2. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18:1–15.
3. Wang T, Shao W, Huang Z, Tang H, Zhang J, Ding Z, et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nat Commun. 2021;12(1):3445.
4. Moon S, Lee H. MOMA: a multi-task attention learning algorithm for multi-omics data interpretation and classification. Bioinformatics. 2022;38(8):2287–96.
5. Lu Y, Peng R, Dong L, Xia K, Wu R, Xu S, et al. Multiomics dynamic learning enables personalized diagnosis and prognosis for pancancer and cancer subtypes. Brief Bioinform. 2023;24(6):bbad378.
6. Li B, Nabavi S. A multimodal graph neural network framework for cancer molecular subtype classification. BMC Bioinformatics. 2024;25(1):27.
7. Picard M, Scott-Boyer MP, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021;19:3735–46.
8. Lin Y, Zhang W, Cao H, Li G, Du W. Classifying breast cancer subtypes using deep neural networks based on multi-omics data. Genes. 2020;11(8):888.
9. Poirion OB, Jing Z, Chaudhary K, Huang S, Garmire LX. DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. Genome Med. 2021;13:1–15.
10. Tini G, Marchetti L, Priami C, Scott-Boyer MP. Multi-omics integration—a comparison of unsupervised clustering methodologies. Brief Bioinform. 2019;20(4):1269–79.
11. Yang H, Chen R, Li D, Wang Z. Subtype-GAN: a deep learning approach for integrative cancer subtyping of multi-omics data. Bioinformatics. 2021;37(16):2231–7.
12. Yang B, Yang Y, Wang M, Su X. MRGCN: cancer subtyping with multi-reconstruction graph convolutional network using full and partial multi-omics dataset. Bioinformatics. 2023;39(6):btad353.
13. Pai S, Bader GD. Patient similarity networks for precision medicine. J Mol Biol. 2018;430(18):2924–38.
14. Liu C, Duan Y, Zhou Q, Wang Y, Gao Y, Kan H, et al. A classification method of gastric cancer subtype based on residual graph convolution network. Front Genet. 2023;13:1090394.
15. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In: Proceedings of the 5th International Conference on Learning Representations, Toulon, France. 2017. pp. 1–14.
16. Zhang XM, Liang L, Liu L, Tang MJ. Graph neural networks and their current applications in bioinformatics. Front Genet. 2021;12:690049.
17. Wen G, Li L. FGCNSurv: dually fused graph convolutional network for multi-omics survival prediction. Bioinformatics. 2023;39(8):btad472.
18. Sun X, Jia X, Lu Z, Tang J, Li M. Drug repositioning with adaptive graph convolutional networks. Bioinformatics. 2024;40(1):btad748.
19. Dai W, Yue W, Peng W, Fu X, Liu L, Liu L. Identifying cancer subtypes using a residual graph convolution model on a sample similarity network. Genes. 2021;13(1):65.
20. Li X, Ma J, Leng L, Han M, Li M, He F, et al. MoGCN: a multi-omics integration method based on graph convolutional network for cancer subtype analysis. Front Genet. 2022;13:806842.
21. Li Q, Han Z, Wu XM. Deeper insights into graph convolutional networks for semi-supervised learning. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI Press; 2018. p. 3538–45.
22. Liu L, Mamitsuka H, Zhu S. HPODNets: deep graph convolutional networks for predicting human protein-phenotype associations. Bioinformatics. 2022;38(3):799–808.
23. Zhang G, Peng Z, Yan C, Wang J, Luo J, Luo H. A novel liver cancer diagnosis method based on patient similarity network and DenseGCN. Sci Rep. 2022;12(1):6797.
24. Wang W, Sun B, Yu M, Wu S, Liu D, Zhang H, et al. GraphPLBR: Protein-Ligand Binding Residue Prediction With Deep Graph Convolution Network. IEEE/ACM Trans Comput Biol Bioinforma. 2023;20(3):2223–32.
25. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11(3):333–7.
26. Gasteiger J, Bojchevski A, Günnemann S. Predict then Propagate: Graph Neural Networks Meet Personalized PageRank. In: Proceedings of the 7th International Conference on Learning Representations. New Orleans; 2019. pp. 1–15.
27. Chen M, Wei Z, Huang Z, Ding B, Li Y. Simple and deep graph convolutional networks. In: Proceedings of the 37th International Conference on Machine Learning. PMLR; 2020. pp. 1725–1735.
28. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on Machine Learning. Lille: JMLR; 2015. pp. 448–56.
29. Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, Kamath A, et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol. 2020;38(6):675–8.
30. Xiao S, Lin H, Wang C, Wang S, Rajapakse JC. Graph neural networks with multiple prior knowledge for multi-omics data analysis. IEEE J Biomed Health Inform. 2023;27(9):4591–600.
31. Li S, Li WT, Wang W. Co-GCN for multi-view semi-supervised learning. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI Press; 2020. pp. 4691–8.
32. Huang A, Wang Z, Zheng Y, Zhao T, Lin CW. Embedding regularizer learning for multi-view semi-supervised classification. IEEE Trans Image Process. 2021;30:6997–7011.
33. Lupat R, Perera R, Loi S, Li J. Moanna: multi-omics autoencoder-based neural network algorithm for predicting breast cancer subtypes. IEEE Access. 2023;11:10912–24.
34. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57.
35. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res. 2022;50(W1):W216–21.
36. Duchartre Y, Kim YM, Kahn M. The Wnt signaling pathway in cancer. Crit Rev Oncol/Hematol. 2016;99:141–9.
37. Marei HE, Althani A, Afifi N, Hasan A, Caceci T, Pozzoli G, et al. p53 signaling in cancer progression and therapy. Cancer Cell Int. 2021;21(1):703.

Publisher's Note