frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 04 January 2023
DOI 10.3389/fgene.2022.1090394

OPEN ACCESS

EDITED BY
Dominik Heider,
University of Marburg, Germany

REVIEWED BY
Hans A. Kestler,
University of Ulm, Germany
Andreas Holzinger,
Medical University Graz, Austria

*CORRESPONDENCE
Jili Hu,
hujili@ahtcm.edu.cn

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 05 November 2022
ACCEPTED 09 December 2022
PUBLISHED 04 January 2023

CITATION
Liu C, Duan Y, Zhou Q, Wang Y, Gao Y, Kan H and Hu J (2023), A classification method of gastric cancer subtype based on residual graph convolution network. *Front. Genet.* 13:1090394. doi: 10.3389/fgene.2022.1090394

A classification method of gastric cancer subtype based on residual graph convolution network

Can Liu[1,2], Yuchen Duan[1], Qingqing Zhou[1], Yongkang Wang[1,2], Yong Gao[1,2], Hongxing Kan[1,2] and Jili Hu[1,2]*

[1]School of Medical Informatics Engineering, Anhui University of Chinese Medicine, Hefei, Anhui, China, [2]Anhui Computer Application Research Institute of Chinese Medicine, China Academy of Chinese Medical Sciences, Hefei, Anhui, China

Background: Clinical diagnosis and treatment of tumors are greatly complicated by their heterogeneity, and the subtype classification of cancer frequently plays a significant role in the subsequent treatment of tumors. Presently, the majority of studies rely far too heavily on gene expression data, omitting the enormous power of multi-omics fusion data and the potential for patient similarities.

Method: In this study, we created a gastric cancer subtype classification model called RRGCN based on residual graph convolutional network (GCN) using multi-omics fusion data and patient similarity network. Given the multi-omics data's high dimensionality, we built an artificial neural network Autoencoder (AE) to reduce the dimensionality of the data and extract hidden layer features. The model is then built using the feature data. In addition, we computed the correlation between patients using the Pearson correlation coefficient, and this relationship between patients forms the edge of the graph structure. Four graph convolutional network layers and two residual networks with skip connections make up RRGCN, which reduces the amount of information lost during transmission between layers and prevents model degradation.

Results: The results show that RRGCN significantly outperforms other classification methods with an accuracy as high as 0.87 when compared to four other traditional machine learning methods and deep learning models.

Conclusion: In terms of subtype classification, RRGCN excels in all areas and has the potential to offer fresh perspectives on disease mechanisms and disease progression. It has the potential to be used for a broader range of disorders and to aid in clinical diagnosis.

KEYWORDS

multi-omics, autoencoder, patient similarity network, residual graph convolutional network, classification

1 Introduction

Gastric cancer (GC) is a highly aggressive cancer with significant heterogeneity in terms of cell types, states, and subpopulation distribution in the immune microenvironment (Shao et al., 2021; Kim et al., 2022). According to the epidemiological survey (Ferlay et al., 2021), the incidence of GC is the fifth highest among tumor diseases worldwide, and the mortality rate is the third highest among tumor deaths (Wang et al., 2021; Dong et al., 2021). Studies have shown that several variables, including genetics, the immune system, lifestyle choices, and psychological factors, can affect the development and occurrence of tumors (Shin et al., 2022). Multiple pathological processes at various levels and dimensions, including the genome, transcriptome, and proteome, are involved in complex diseases like cancer (Menyhárt and Győrffy, 2021).

With the advancement of high-throughput sequencing and omics technology, researchers progressively understood the limits of employing a single omics (Sun et al., 2019; Jia et al., 2022). To better understand the essence of the disease, it is required to undertake a joint analysis of various types of data, get more comprehensive information, construct a perfect body regulatory network, and thoroughly investigate the regulation and causal relationships between molecules (Tao et al., 2020). Consequently, one of the areas of research that is now quite active is the integration of multi-omics data for cancer subtyping (Lindskrog et al., 2021; Sivadas et al., 2022). The biological information contained in multi-omics data is critical for disease diagnosis and treatment. However, due to its huge scale, high dimension, high noise, and strong heterogeneity, data is difficult to handle and analyze, posing significant obstacles to cancer typing (Duan et al., 2021; Picard et al., 2021).

The Graph Convolutional Network (GCN) (Kipf and Welling, 2017) is a convolutional neural network that was built in recent years that can directly act on graphs and use their structural information, and it is gaining popularity in the field of bioinformatics (Zhang et al., 2021). It can identify unlabeled nodes and categorize them using both the node's feature vector and network topology data (Li et al., 2022).

Kim et al., (2021) proposes an analytical framework named DrugGCN based on gene expression data for predicting drug responses using graph convolutional networks (GCNs). Baul et al., (2022) offers omicsGAT, a graph attention network (GAT) model that blends graph learning with attention processes for cancer subtype identification based on RNA-seq data. By allocating various attention coefficients to nearby samples, the multi-head attention mechanism can more successfully protect the connection between them. However, such experimental results are neither applicable nor interpretable when only one set of omic data is considered. According to studies (Sun and Hu, 2016; Xu et al., 2019), different forms of data have complementarities, and multi-omics can fuse the rich information in each type of data to facilitate categorization. Li et al., (2022) developed a multi-omics ensemble model, MoGCN, with two-layer graph convolutional networks for the classification and analysis of cancer subtypes. Ramirez et al., (2020) constructed a graph convolutional neural network for classifying tumor and non-tumor samples based on unstructured gene expression data. Unfortunately, as depth increases, graph convolutional networks suffer from vanishing gradients and over-smoothing, which significantly reduces model accuracy. Zhang et al., (2022) proposes a new method for detecting liver cancer using a fusion similarity network, denoising autoencoder, and dense graph convolutional neural network. Liang et al., (2021) proposed a Consensus Guided Graph Autoencoder (CGGA) to identify cancer subtypes and bring fresh insights into the treatment of patients with diverse subtypes. Wang et al., (2021) introduces a unique multi-omics integrative approach called the Multi-Omics Graph Convolutional Networks (MOGONET), which is utilized for biomedical classification and can find key biomarkers from various omics data sources. Finally, Dai et al., (2021) combined GCN with a residual network to build a cancer subtype classification model, named ERGCN, which performed well on three different TCGA cancer types, presenting a new method for precision cancer treatment.

Therefore, we integrated multi-omics data and designed a model RRGCN based on graph convolution for GC subtype classification. High-dimensional multi-omics data is integrated into low-dimensional space using an artificial neural network autoencoder (AE) to extract hidden layer characteristics. The Patient Similarity Network (PSN) combines the network topology generated by each data type and analyzes the links between patients using the Pearson correlation coefficient (Benesty et al., 2009). The fused network can collect information from multiple data sources that are both shared and complementary. Two residual networks with skip connections are merged with four GCN layers to collect feature matrices and patient similarity correlations to discover and classify GC subtypes, and the classification results are finally output by softmax. The results of the comparison with random forest (RF), support vector machine (SVM), MoGCN, and ERGCN reveal that RRGCN has the best performance. The classification accuracy of the GC subtype is 0.87, the AUC value is 0.98, and the values of other indicators of RRGCN are also the highest when compared to other methods. We believe that RRGCN can provide new and unique insights into the identification, classification, and clinical diagnosis of GC subtypes.

2 Materials and methods

Proposed method

We designed a GC subtype classification model, namely RRGCN, which is based on the residual graph convolutional

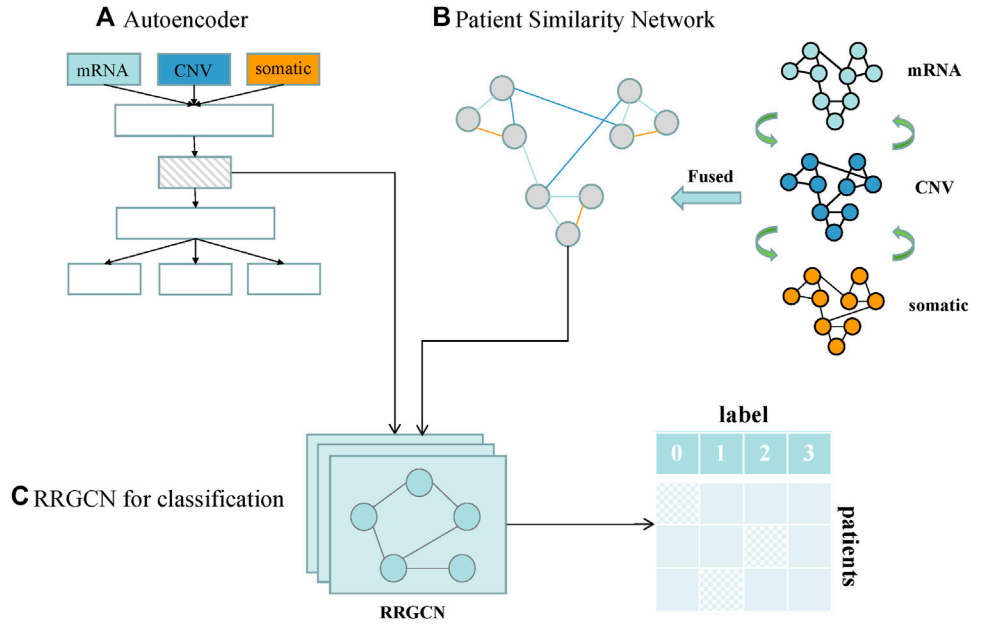

FIGURE 1
Workflow of RRGCN. **(A)** Features extracted by AE network. **(B)** Patient similarity network constructed by SNF algorithm. **(C)** RRGCN for GC subtype classification.

network. The input consists of the multi-omics fusion data and the patient similarity network following AE dimensionality reduction. The graph nodes are then embedded through two residual networks with skip connections and a 4-layer GCN, and the classification results are then output using a softmax layer. We compared and assessed RRGCN's performance with several traditional machine learning models and deep learning methods in the third chapter of the paper. Figure 1 shows the workflow of RRGCN.

Datasets and data preprocessing

To train the model, we used information on GC from the TCGA (https://tcga-data.nci.nih.gov/tcga/). Transcriptomic data, copy number variations (CNV), and somatic mutation data are all included in our study. We got 272 labeled samples and four subtypes of data from the R tool "TCGAbiolinks" (Colaprico et al., 2016). We download the experimental data using the R package "TCGA-assembler 2" (Wei et al., 2018). The transcriptome data is from the Illumina HiSeq_RNASeqV2 sequencing platform, the CNV data is from the cna_cnv.hg19 sequencing platform and the somatic mutation data is from the somaticMutation_DNAseq sequencing platform. In addition, to make it easier for the model to categorize the input data, we define four GC subtypes as numbers, Epstein-Barr virus type (EBV) as 0, Microsatellite instability type (MSI) as 1, Genetically stable type (GS) as 2, and Chromosome instability type (CIN) as 3.

The dataset in TCGA has to be preprocessed because it contains a large amount of zero and missing value data. The preprocessing step helps to reduce the redundancy and inconsistency of the dataset, thus improving the accuracy and speed of the subsequent mining process. From the phenotypic data, sample information with labels for the various cancer subtypes was first retrieved, and the features that were absent from all samples or had a zero-expression level were subsequently eliminated. So we ended up with 272 samples. Second, among the genes that have been duplicated, we

TABLE 1 Overview of the STAD dataset.

Multi-omics	Number of features	Subtypes	Samples
mRNA	20,468	EBV	25
CNV	22,434	MSI	60
Somatic	19,600	GS	51
—	—	CIN	136
Total	62,502	Total	272

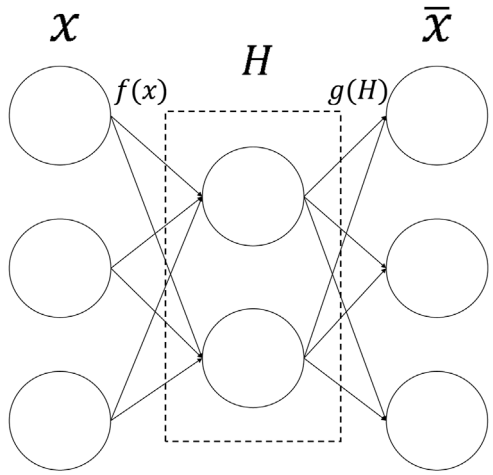

FIGURE 2
The structure of Autoencoder. The input data is first encoded by the encoder, then embedded in the hidden layer, and finally output by the decoder.

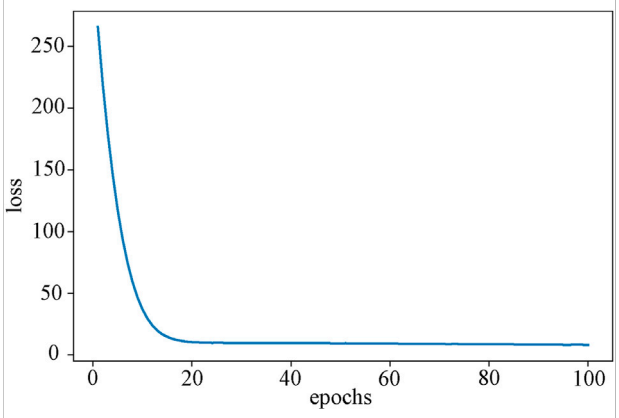

FIGURE 3
The loss curve of the AE training process. As the epoch grows, training tends to converge around epoch 20.

choose the one whose mean expression across all samples has the least absolute value. Finally, for the transcriptome data, we expressed expression levels in units of log2 (FPKM + .1), where FPKM stands for Fragments Per Kilobase of Exon Model per Million mapped Fragment. In this study, we removed the number of zero values and missing values in mRNA, CNV and somatic cells to be 62, 2,481 and 2 respectively, resulting in 20,468, 22,434 and 19,600 features for subsequent model construction. Table 1 shows the details of the dataset.

Autoencoder architecture

The autoencoder (AE) (Hinton and Salakhutdinov, 2006) is an unsupervised artificial neural network model that belongs to the deep learning category. AE can extract latent embedding representations from multi-omics datasets to reduce dimensionality and computational cost. It can first learn the hidden features of the input data through encoding, then output to the next hidden layer, and then decode and rebuild the original input data with the learned new features (Binbusayyis and Vaiyapuri, 2021). Figure 2 is the basic framework of Autoencoder. The formula is:

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{\delta}(\boldsymbol{\omega x} + \boldsymbol{b}) = \boldsymbol{H} \tag{1}$$

$$\boldsymbol{g}(\boldsymbol{H}) = \boldsymbol{\delta}(\boldsymbol{\omega}'\boldsymbol{H} + \boldsymbol{b}') = \bar{\boldsymbol{x}} \tag{2}$$

Where x is the input feature in the AE, which is encoded and decoded to $\bar{x}$. $f(x)$ represents the encoder function, H represents the hidden unit, $g(H)$ represents the decoder function, $\bar{x}$ represents the output, δ represents the activation function, ω represents the weight matrix, b represents the bias. We used the mean square error (MSE) (Sammut and Webb, 2010) as the loss function to calculate the loss between the predicted value and the true value, where the predicted value is $\bar{x}$ and the true value is x. The formula is:

$$\boldsymbol{mseloss}(\boldsymbol{x}, \bar{\boldsymbol{x}}) = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{n}} (\boldsymbol{x_i} - \bar{\boldsymbol{x}}_{\boldsymbol{i}})^{\boldsymbol{2}} \tag{3}$$

Since RRGCN uses three different forms of data, we gave each omics data a varied weight based on prior knowledge (Li et al., 2022) to emphasize their contributions to the model, and all weights sum up to 1. In light of this, the loss function is described as:

$$\boldsymbol{L_{AE}} = \boldsymbol{a}*\boldsymbol{mseloss}(\boldsymbol{x_1}, \bar{\boldsymbol{x}}_{\boldsymbol{1}}) + \boldsymbol{b}*\boldsymbol{mseloss}(\boldsymbol{x_2}, \bar{\boldsymbol{x}}_{\boldsymbol{2}}) + \boldsymbol{c}*\boldsymbol{mseloss}(\boldsymbol{x_3}, \bar{\boldsymbol{x}}_{\boldsymbol{3}}) \tag{4}$$

L_{AE} represents the MSE loss function, and a, b, and c represent the weights of the input data, respectively for 0.4, 0.3, and 0.3. As the input data are characterized by multi-omics data types and represented by multiple matrices x_1, x_2, and x_3, corresponding to the mRNA, CNV, and somatic matrices, $\bar{x}_1$, $\bar{x}_2$, and $\bar{x}_3$ correspond to the output of three types of data.

In this study, we took into account high-dimensional multi-omics data using an AE with three hidden layers. The three hidden layers were (500, 200, 500), and the training epoch was set to 100, which ultimately converged after 20 epochs (Figure 3). All layers employ the sigmoid function as their activation function. AE is trained by back-propagation through the Adam (Kingma and Ba, 2015) optimizer. Additionally, we used grid search to select the batch size from (32, 64, 128) and the learning rate (LR) from (0.01, 0.001, 0.0001). The final batch size is 32, and

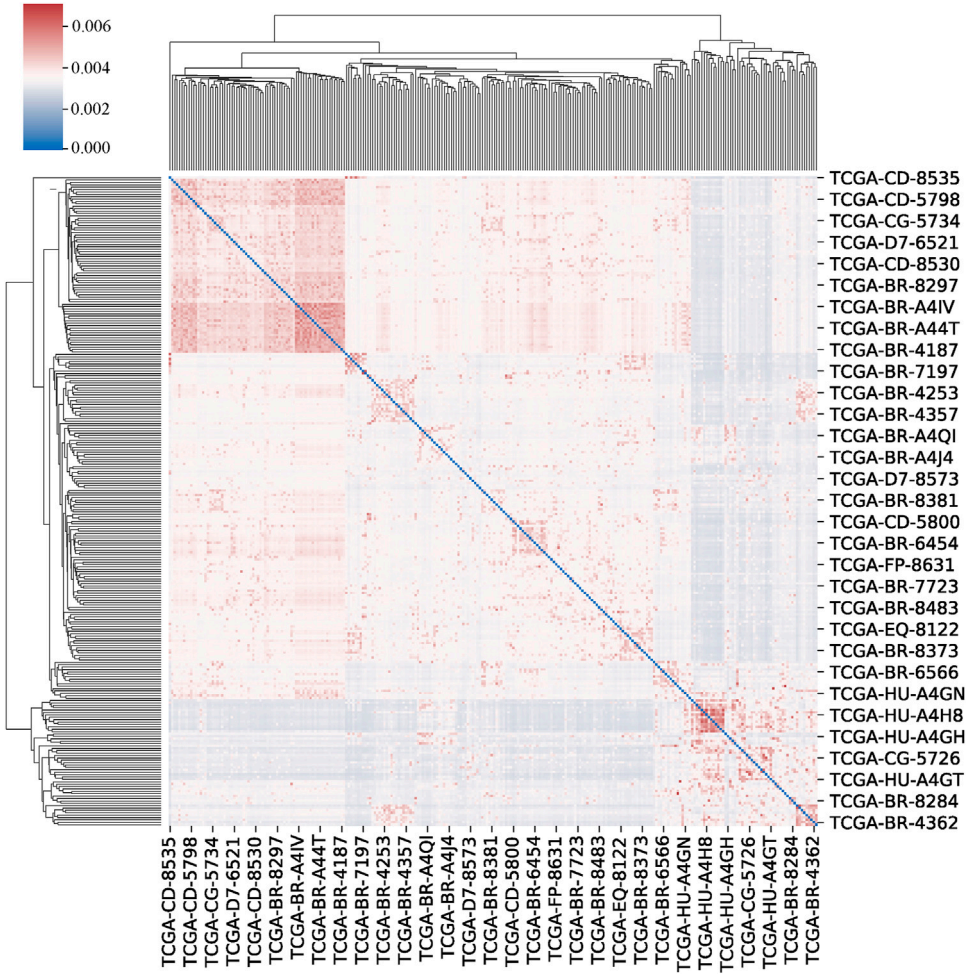

FIGURE 4
The clustermap of patients. The clustermap shows the clustering relationship among multiple samples, with red representing high correlation and blue representing low correlation.

LR is 0.001. Every model used in our study is built by PyTorch (v1.8.0) (Paszke et al., 2019). The feature matrix extracted by the AE hidden layer will be used as the input of the RRGCN.

Patient similarity network

The Similarity Network Fusion (SNF) (Wang et al., 2014) algorithm is a computational approach that creates a network of similarities across patients for each type of data to provide a holistic perspective of a certain disease or biological process. We used the SNF algorithm to compute and fuse patient similarity networks from each data type in the GC dataset to create an overall view of GC patients. The advantage of PSN is that it enables RRGCN to seek and obtain important information from the neighbour nodes of the patient, rather than relying solely on the level of gene expression. This improves the accuracy and applicability of the model. The SNF algorithm creates patient-patient similarity matrices for each data type and construct the patient adjacency matrix, then builds a network through the matrix, and lastly fuses various forms of patient-patient similarity networks to create a fusion network. SNF can fully exploit the complementarity of various source data (El-Manzalawy et al., 2018; Picard et al., 2021), which is far superior to the comprehensive analysis approach established by employing a single dataset and has significant advantages in the detection and classification of cancer subtypes (Wang T.-H. et al., 2021; Franco et al., 2021).

Assume there are n samples and m various categories of data (in this study, the data types include mRNA, CNV, and

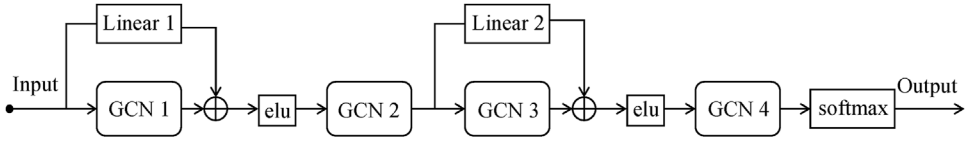

FIGURE 5
The structure of RRGCN.

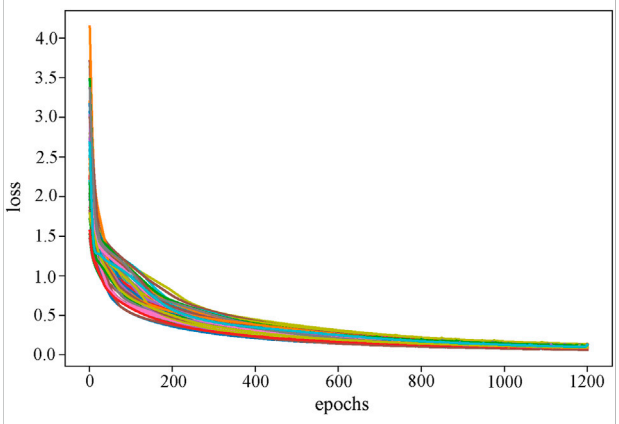

FIGURE 6
The loss curve of the RRGCN training process. Different colors reflect the loss curves of different cross-validation times.

TABLE 2 Confusion matrix.

Predicted	Actual	
	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

somatic data). We refer to a PSN as a graph $G = (V, E)$, where the vertex V is a collection of samples made up of $(x_1, x_2, \cdots x_n)$, and E makes up the edges of the graph. A similarity matrix defined by the scaled exponential similarity kernel was computed:

$$\boldsymbol{w(i,j) = \exp\left(-\frac{\rho^2(x_i, x_j)}{\mu \varepsilon_{i,j}}\right)} \tag{5}$$

Among them, w represents the similarity matrix between samples, $\rho(x_i, x_j)$ represents the Euclidean distance between the patient x_i and patient x_j, μ is a hyperparameter set by experience, and the commonly used range is (0.3, 0.8), and $\varepsilon_{i,j}$ is a parameter used to eliminate the scaling problem, which is defined as:

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

where N_i is the set of x_i's neighbors and $mean(\rho(x_i, N_i))$ is the mean distance from x_i to each neighbor. Thus, to compute fusion matrices from multiple data types, the similarity matrix is defined as:

$$\boldsymbol{P_{i,j} = \begin{cases} \frac{W_{i,j}}{2\sum_{k \neq i} W_{i,k}}, & j \neq i \\ \frac{1}{2}, & j = i \end{cases}} \tag{7}$$

Then, the affinity matrix S is calculated:

$$\boldsymbol{S_{i,j} = \begin{cases} \frac{W_{i,j}}{\sum_{k \in N_i} W_{i,k}}, & j \in N_i \\ 0, & otherwise \end{cases}} \tag{8}$$

In the case of various data types:

$$\boldsymbol{P^{(v)} = S^{(v)} \left(\frac{\sum_{k \neq v} P^{(k)}}{m-1}\right) \left(S^{(v)}\right)^T, v = 1, 2, \cdots, m} \tag{9}$$

where the $S^{(v)}$ represents the affinity matrix of vth type of data, the $P^{(v)}$ represents the similarity matrix of vth type data. The Pearson correlation coefficient is used to calculate the correlation (linear correlation) between two variables and has a value between −1 and 1. We determined how similar patients were to one another using the Pearson correlation coefficient, and if their similarity exceeded a predetermined threshold, we categorized this as a correlation between patients. The patient similarity network established by the merging of many types of data (multi-omics) is finally obtained by the continual update and iteration of the preceding algorithm.

We set the number of neighbours to consider when creating the affinity matrix to 20 and the scaling factor to 0.5. The clustermap of patients is shown in Figure 4.

Construction of RRGCN

We use GCN to process non-Euclidean data computed using the SNF algorithm. The purpose of GCN is to learn latent

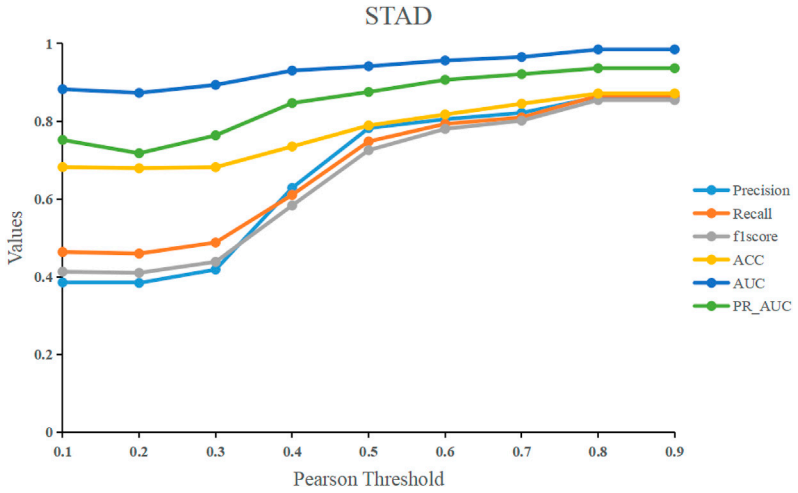

FIGURE 7
The effect of different Pearson correlation thresholds on model performance.

representations based on the node feature matrix X (input, graph nodes) and the similarity matrix A (similarities between nodes). Mathematically, the propagation formula between GCN layers is:

$$H^L = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{L-1}W^{L-1}\right) \quad (10)$$

H^L represents the output of the Lth layer, that is, the node features learned by the Lth layer, W^{L-1} represents the weight matrix of the L-1-th layer, and σ represents the non-linear activation function in the GCN. D is the degree matrix of A, $\tilde{A} = A + E$, E represents the identity matrix.

We use the ResNet (He et al., 2016) concept and add skip connections between GCN layers to overcome the problem of model degradation in deep neural network training. The insertion of skip connections can compensate for the loss of features between the data of the previous layer and the data of the following layer, reducing information loss and improving model performance (Yamanaka et al., 2017). At the same time, to avoid the inconsistency between the output of GCN and the dimension of the input data, we add an independent linear layer to the skip connection. The formula for skip connection can be defined as:

$$H^{L+1} = elu\left(H^L + linear\left(X\right)\right) \quad (11)$$

H^L represents the output of the previous GCN layer. The input feature matrix is sent to the linear layer, and the result is added to the GCN layer and then passed to the non-linear activation function Exponential Linear Units (ELU) (Clevert et al., 2016) to generate the output H^{L+1}, which is utilized as the input of the next skip connection.

RRGCN, which has more skip connections than ERGCN, which only has one, improves model performance by increasing the information flow between layers, making up for information loss, increasing the connectivity between the upper and lower information, and improving the flow of information between layers. The RRGCN as a whole consists of 2 residual networks with skip connections, 4 GCN layers, and 1 softmax layer for generating classification results. To compute the difference between the classification results and the true labels, we utilize the cross-entropy loss function:

$$L = -\left[y \cdot log\left(\bar{y}\right) + \left(1 - y\right) \cdot \log\left(1 - \bar{y}\right)\right] \quad (12)$$

where y represents the true label corresponding to the sample, and $\bar{y}$ is the probability value output by the softmax layer. The structure of RRGCN is shown in Figure 5.

We set the dimensions of the four graph convolutional layers to 64, 32, 16, and the number of subtypes, respectively. By performing grid search on the LR and weight decay in (0.1, 0.01, 0.001, 0.0001) and (0.1, 0.01, 0.001), respectively, the optimal LR and weight decay are determined to be 0.0001 and 0.01. We use the Adam optimizer function and set the epoch to 1,200, the training process finally converges at 600 (Figure 6). RRGCN employs ELU as the non-linear activation function, and the

TABLE 3 Results of multi-omics data compared with single-dimensional data.

Omics	Accuracy	AUC
mRNA	0.7384	0.9339
CNV	0.4766	0.7809
Somatic	0.5190	0.8272
Multi-omics	**0.8713**	**0.9848**

Bold values emphasize that the experimental results of multi-omics are better than other groupings.

TABLE 4 Results in comparison to other methods.

Model	Accuracy	F1 score	Precision	Recall
RF	0.8363	0.7665	0.8172	0.7471
SVM	0.7455	0.7340	0.7792	0.7460
MoGCN	0.7944	0.8078	0.8034	0.7407
ERGCN	0.7901	0.6826	0.7160	0.6120
RRGCN	**0.8713**	**0.8544**	**0.8621**	**0.8654**

Bold values are to highlight the performance of our model over other classical models.

classification results are finally output *via* the softmax layer. We used 80% of the multi-omics fusion data as the training set and reserved 20% for validation. Model performance was evaluated using 5-fold cross-validation on the training set. Furthermore, to eliminate the bias introduced by a single trial, we took the average outcome of ten iterations of the 5-fold cross-validation test set as the evaluation metric.

Model evaluation metrics

In the classification task, the model produces four main prediction results: True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). The confusion matrix in Table 2 can be constructed based on the four different prediction outcomes.

Precision refers to the probability that the prediction is correct in the sample that is predicted to be true. It is defined as:

$$\boldsymbol{precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \tag{13}$$

Recall, also known as sensitivity, is the measure of how many samples are selected as being true. It is defined as:

$$\boldsymbol{recall} = \boldsymbol{sensitivity} = \frac{\boldsymbol{TP}}{\boldsymbol{P}} \tag{14}$$

The F1 score is a weighted harmonic average of precision and recall that is unaffected by imbalanced samples. The F1 score has a maximum value of one and a minimum value of zero. The higher the value, the higher the model quality. In most circumstances, the f1 score can be used directly to evaluate and pick the model, and some well-known machine learning competitions do as well. It is defined as:

$$\boldsymbol{F1\ score} = \frac{\boldsymbol{2{*}precision{*}recall}}{\boldsymbol{precision} + \boldsymbol{recall}} \tag{15}$$

Accuracy is defined as the ratio of accurately predicted samples to total samples. It is defined as:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP} + \boldsymbol{TN}}{\boldsymbol{TP} + \boldsymbol{TN} + \boldsymbol{FP} + \boldsymbol{FN}} \tag{16}$$

The area contained by the curve with the false positive rate (FPR) on the abscissa and the true positive rate (TPR) on the ordinate is known as the area under the receiver operating characteristic curve (ROC) curve (AUC). The categorization skill given by the ROC curve is intuitively reflected by AUC. The AUC value ranges between 0 and 1, and the higher the value, the better the classifier's performance. FPR is the likelihood that the prediction is a positive sample but the prediction is incorrect. It is defined as:

$$\boldsymbol{FPR} = \frac{\boldsymbol{FP}}{\boldsymbol{TN} + \boldsymbol{FP}} \tag{17}$$

TPR reflects the likelihood that the forecast is a positive sample and that the prediction is right. It is defined as:

$$\boldsymbol{TPR} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \tag{18}$$

The area contained by the curve with recall on the abscissa and precision on the ordinate is known as PR-AUC, and it is the mean value of precision calculated for each recall threshold (Géron, 2017). All model evaluation metrics are based on Scikit-learn (Pedregosa et al., 2011).

3 Results

Determination of pearson correlation threshold

We used the Pearson correlation threshold to see if there was a link between samples. If the Pearson correlation coefficient between samples is larger than the threshold, we connect the two samples with an edge and set the corresponding value in the adjacency matrix to 1. In contrast, there is no edge connecting the two samples, and the corresponding values in the adjacency matrix are 0. To examine the performance of the models, we fixed the threshold to a value ranging from 0.1 to 0.9. Figure 7 shows that before 0.5, the model's performance improves significantly as the threshold is raised. After 0.5, it tends to be flat, and the model's performance peaks at the final threshold of 0.8. Our model RRGCN performed best when Pearson correlation threshold was 0.8, where the Precision, Recall, F1score, ACC, AUC and PR_AUC reached 0.862, 0.865, 0.854, 0.871, 0.984, and 0.936, respectively.

Performance of RRGCN in multi-omics

To verify the superiority of multi-omics data, as well as the validity and contribution of each type of data to the model, we conduct experiments on different types of data separately. From the experimental results (Table 3), it can be seen that using a single omics data training model, the highest performance is the

mRNA group with an accuracy of 0.7384, followed by the somatic group with an accuracy of 0.5190. The CNV group has the lowest accuracy, only 0.4766. It can be seen that although RNA-seq data has good performance and is indeed used in most studies, its effect is still inferior to multi-omics data. Of course, this also reflects from a certain level that RNA-seq data contains extremely important biological information, has good performance in the classification of cancer subtypes, and is extremely important for cancer diagnosis and treatment (Yang et al., 2021). It can be verified that there is complementary information between different omics, which can explain the nature of cancer from different perspectives and improve the diagnostic efficiency of cancer.

Comparison with other classical methods

To validate RRGCN's classification performance, we compare it to two other classical machine learning methods and two graph convolution-based classification approaches and evaluate it using four standard external evaluation measures. We employ four classification methods: Random Forest (Breiman, 2001), Support Vector Machine (Cortes and Vapnik, 1995), MoGCN (Li X. et al., 2022), and ERGCN (Dai et al., 2021).

- Random Forest (RF) is essentially a bagging algorithm, which randomly selects a feature from the most important features for branching, creates multiple decision trees, and finally votes on which category the data finally belongs to.
- Support Vector Machine (SVM), a binary classification model whose basic model is defined as a linear classifier with the biggest margin on the feature space. The goal is to build an objective function based on the structural risk reduction principle that distinguishes between the two types as much as possible.
- MoGCN is a multi-omics integration model based on GCN. The model utilizes feature extraction and network visualization for further biological knowledge discovery and subtype classification.
- ERGCN is a cancer subtype classification method based on residual graph convolutional networks and sample similarity networks for gene co-expression patterns.

To begin, we unified the AE latent layer feature matrix as input data for each model to ensure the rigor of the compared tests. Then, we utilized scikit-learn to construct these algorithms and grid search to optimize the RF and SVM parameters. The best number of sub-decision trees (n_esimators) for RF is between 1 and 101, with a step size of 10. Finally, 5-fold cross-validation yielded an optimal n_esimators of 26. The maximum number of features (max_features) should ideally be between 1 and 21, with a stride of 1. Finally, the optimal max_features is selected as 20 through 5-fold cross-validation. We also use grid search for SVM, choosing the penalty coefficient (C) from (0.1, 1, 100, 1,000) and the kernel function coefficient (gamma) from (0.0001, 0.001, 0.005, 0.1, 1, 3, 5), as well as the kernel function (kernel) from ("linear," "rbf"). The final optimized C is 1,000, the gamma is 0.001, and the kernel is "rbf." For MoGCN and ERGCN, we use the optimal parameters already set by their authors. The model comparison results are shown in Table 4.

From the results, we can see that RRGCN has an excellent performance in the classification of GC subtypes. The classification accuracy of RRGCN is as high as 0.8713, which is 5.49% higher than the best RF among the other four methods, and 11.47%, 8.83%, and 9.32% higher than the other three methods, respectively. The F1 score, Precision, and Recall of RRGCN are 0.8544, 0.8621, and 0.8654 respectively, and the values are also much higher than other methods. Most crucially, as compared to ERGCN, RRGCN performs better on each of the four evaluation metrics by 10.28%, 25.17%, 20.41%, and 41.41%, respectively. In defining the various subtypes of GC, RRGCN has more advantages. In the future, it might be used to treat more diseases, offering novel perspectives on how to diagnose and treat clinical illnesses.

4 Discussion

Heterogeneity causes cancer to differentiate into different subtypes, and subtypes with different degrees of differentiation and malignancy have different sensitivities to clinical therapeutic drugs, which brings great challenges to the diagnosis and treatment of the disease (Lin et al., 2021; Yuan et al., 2022). GC is a highly heterogeneous tumor, and its average somatic gene copy number changes are much higher than those of other tumor types (Joshi and Badgwell, 2021). Therefore, in clinical studies, the progression of GC is the slowest (Li et al., 2021).

Therefore, by integrating multi-omics data, we propose a graph convolutional network based on residual networks to realize the subtype classification of GC. Multi-omics datasets are dimensionally reduced by AE to extract representative latent layer features. The SNF algorithm is used to find the associations existing between patients. Finally, PSN combined with the feature matrix was input into RRGCN, and the classification results were output through the softmax layer. The results show that the accuracy of RRGCN reaches 0.8713.

The improvement of RRGCN over previous models is that multi-omics data is used as the basis of research, and the neglected similarity between patients is combined as the input of the model. For model selection, we introduce two skip connections to alleviate the loss of information during training and solve the model degradation problem.

To explain the advantage of multi-omics data, we retrain three different types of data separately and compare the results with multi-omics data. The results show that the performance of the model trained with multi-omics data is much higher than that of the single-omics data, and the accuracy is improved by about 18.00%. To prove the superiority of RRGCN, we compare RRGCN with classical machine learning methods and well-performing deep learning models, respectively. The results show that the performance of RRGCN is higher than other methods in all aspects. Most importantly, the accuracy of RRGCN is 10.28% higher than that of ERGCN.

The model is sensitive to the selection of the Pearson threshold, and the supervised learning method also brings inconvenience to the selection of data. In the future, we will focus on studying the application of graph convolution combined with other classical convolutional neural networks, considering the development of new unsupervised learning methods for cancer subtype recognition and classification.

5 Conclusion

In summary, we proposed a new classification method for gastric cancer subtypes called RRGCN by borrowing skip connections in residual networks. Through the deep mining of GC multi-omics data and the consideration of the relationship between patients, and comparing RRGCN with other classical machine learning methods and deep learning models, we verify the excellent performance of RRGCN in various aspects and improve the cancer subtype classification method to a higher level. The development of new models opens up new avenues for precise treatment. Li J. et al., (2022), Yang et al., (2022), and Hu et al., (2021). have tried to combine GCN with spatial transcriptomics for cell clustering and the identification of cancer subtypes. In the future, we will look into the spatial coordinate information of gastric cancer cells and employ unsupervised learning algorithms to provide more robust support for clinical diagnosis and treatment of gastric cancer.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

CL designed the study and wrote the manuscript. YW collected data. YG and HK validated the findings of the experiment. QZ and YD analyzed the data. JH supervised the study, revised the manuscript and gave the final approval of the version to be published. All authors reviewed and approved this paper.

Funding

This work was supported by the University Excellent Talent Funding Project of Anhui Province (Grant no. gxgnfx2020088); the Natural Science Project of Anhui University of Chinese Medicine (Grant no. 2020wtzx02); and the Industry-University Cooperation Collaborative Education Project of the Ministry of Education of the People's Republic of China (Grant no. 202101123001).

Acknowledgments

I'd like to thank all of the participants for their contributions to the study, especially my mentor JH instruction and assistance, as well as the fund's assistance.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Baul, S., Ahmed, K. T., Filipek, J., and Zhang, W. (2022). omicsGAT: Graph attention network for cancer subtype analyses. *IJMS* 23, 10220. doi:10.3390/ijms231810220

Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). "Pearson correlation coefficient," in *Noise reduction in speech processing springer topics in signal processing* (Berlin, Heidelberg: Springer Berlin Heidelberg), 1–4. doi:10.1007/978-3-642-00296-0_5

Binbusayyis, A., and Vaiyapuri, T. (2021). Unsupervised deep learning approach for network intrusion detection combining convolutional autoencoder and one-class SVM. *Appl. Intell.* 51, 7094–7108. doi:10.1007/s10489-021-02205-9

Breiman, L. (2001). Random forests. *Mach. Learn.* 45 (1), 5–32. doi:10.1023/a:1010933404324

Clevert, D. A., Unterthiner, T., and Hochreiter, S. (2016). *Fast and accurate deep network learning by exponential linear units (elus)*. International Conference on Learning Representations (ICLR).

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., et al. (2016). TCGAbiolinks: An R/bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* 44, e71. doi:10.1093/nar/gkv1507

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20 (3), 273–297. doi:10.1007/bf00994018

Dai, W., Yue, W., Peng, W., Fu, X., Liu, L., and Liu, L. (2021). Identifying cancer subtypes using a residual graph convolution model on a sample similarity network. *Genes* 13, 65. doi:10.3390/genes13010065

Dong, X., Chen, C., Deng, X., Liu, Y., Duan, Q., Peng, Z., et al. (2021). A novel mechanism for C1GALT1 in the regulation of gastric cancer progression. *Cell Biosci.* 11, 166. doi:10.1186/s13578-021-00678-2

Duan, R., Gao, L., Gao, Y., Hu, Y., Xu, H., Huang, M., et al. (2021). Evaluation and comparison of multi-omics data integration methods for cancer subtyping. *PLoS Comput. Biol.* 17, e1009224. doi:10.1371/journal.pcbi.1009224

El-Manzalawy, Y., Hsieh, T.-Y., Shivakumar, M., Kim, D., and Honavar, V. (2018). Min-redundancy and max-relevance multi-view feature selection for predicting ovarian cancer survival using multi-omics data. *BMC Med. Genomics* 11, 71. doi:10.1186/s12920-018-0388-0

Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., et al. (2021). Cancer statistics for the year 2020: An overview. *Int. J. Cancer* 149, 778–789. doi:10.1002/ijc.33588

Franco, E. F., Rana, P., Cruz, A., Calderón, V. V., Azevedo, V., Ramos, R. T. J., et al. (2021). Performance comparison of deep learning autoencoders for cancer subtype detection using multi-omics data. *Cancers* 13, 2013. doi:10.3390/cancers13092013

Géron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. Sebastopol, CA, United States: O'Reilly Media, Inc.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27-30 June 2016 (IEEE), 770–778. doi:10.1109/CVPR.2016.90

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi:10.1126/science.1127647

Hu, J., Li, X., Coleman, K., Schroeder, A., Ma, N., Irwin, D. J., et al. (2021). SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods* 18, 1342–1351. doi:10.1038/s41592-021-01255-8

Jia, Q., Chu, H., Jin, Z., Long, H., and Zhu, B. (2022). High-throughput single-cell sequencing in cancer research. *Sig Transduct. Target Ther.* 7, 145. doi:10.1038/s41392-022-00990-4

Joshi, S. S., and Badgwell, B. D. (2021). Current treatment and recent progress in gastric cancer. *CA A Cancer J. Clin.* 71, 264–279. doi:10.3322/caac.21657

Kim, J., Park, C., Kim, K. H., Kim, E. H., Kim, H., Woo, J. K., et al. (2022). Single-cell analysis of gastric pre-cancerous and cancer lesions reveals cell lineage diversity and intratumoral heterogeneity. *npj Precis. Onc.* 6, 9. doi:10.1038/s41698-022-00251-1

Kim, S., Bae, S., Piao, Y., and Jo, K. (2021). Graph convolutional network for drug response prediction using gene expression data. *Mathematics* 9, 772. doi:10.3390/math9070772

Kingma, D. P., and Ba, J. A. (2015). *A method for stochastic optimization*. ICLR. arXiv preprint arXiv:1412.6980. doi:10.48550/arXiv.1412.6980

Kipf, T. N., and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. Available at: http://arxiv.org/abs/1609.02907 (Accessed November 3, 2022).

Li, J., Chen, S., Pan, X., Yuan, Y., and Shen, H.-B. (2022). Cell clustering for spatial transcriptomics data with graph neural networks. *Nat. Comput. Sci.* 2, 399–408. doi:10.1038/s43588-022-00266-5

Li T., T., Liu, Y., Liu, Q., Xu, W., Xiao, Y., and Liu, H. (2022). A malware propagation prediction model based on representation learning and graph convolutional networks. *Digital Commun. Netw.*, S2352-8648(22)00106-7. doi:10.1016/j.dcan.2022.05.015

Li, X., Ma, J., Leng, L., Han, M., Li, M., He, F., et al. (2022). MoGCN: A multi-omics integration method based on graph convolutional network for cancer subtype analysis. *Front. Genet.* 13, 806842. doi:10.3389/fgene.2022.806842

Li, Y.-K., Zhu, X.-R., Zhan, Y., Yuan, W.-Z., and Jin, W.-L. (2021). NEK7 promotes gastric cancer progression as a cell proliferation regulator. *Cancer Cell Int.* 21, 438. doi:10.1186/s12935-021-02148-8

Liang, C., Shang, M., and Luo, J. (2021). Cancer subtype identification by consensus guided graph autoencoders. *Bioinformatics* 37, 4779–4786. doi:10.1093/bioinformatics/btab535

Lin, Y., Pan, X., Zhao, L., Yang, C., Zhang, Z., Wang, B., et al. (2021). Immune cell infiltration signatures identified molecular subtypes and underlying mechanisms in gastric cancer. *npj Genom. Med.* 6, 83. doi:10.1038/s41525-021-00249-x

Lindskrog, S. V., Prip, F., Lamy, P., Taber, A., Groeneveld, C. S., Birkenkamp-Demtröder, K., et al. (2021). An integrated multi-omics analysis identifies prognostic molecular subtypes of non-muscle-invasive bladder cancer. *Nat. Commun.* 12, 2301. doi:10.1038/s41467-021-22465-w

Menyhárt, O., and Győrffy, B. (2021). Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. *Comput. Struct. Biotechnol. J.* 19, 949–960. doi:10.1016/j.csbj.2021.01.009

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Adv. neural Inf. Process. Syst.* 32.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Picard, M., Scott-Boyer, M.-P., Bodein, A., Périn, O., and Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. *Comput. Struct. Biotechnol. J.* 19, 3735–3746. doi:10.1016/j.csbj.2021.06.030

Ramirez, R., Chiu, Y.-C., Hererra, A., Mostavi, M., Ramirez, J., Chen, Y., et al. (2020). Classification of cancer types using graph convolutional neural networks. *Front. Phys.* 8, 203. doi:10.3389/fphy.2020.00203

Sammut, C., and Webb, G. I. (2010). "Mean squared error," in *Encyclopedia of machine learning* (Boston: Springer), 653.

Shao, W., Yang, Z., Fu, Y., Zheng, L., Liu, F., Chai, L., and Jia, J. (2021). The pyroptosis-related signature predicts prognosis and indicates immune microenvironment infiltration in gastric cancer. *Front. Cell Dev. Biol.* 9, 676485. doi:10.3389/fcell.2021.676485

Shin, J., Shin, D. W., Lee, J., Hwang, J., Lee, J. E., Cho, B., et al. (2022). Exploring socio-demographic, physical, psychological, and quality of life-related factors related with fear of cancer recurrence in stomach cancer survivors: A cross-sectional study. *BMC Cancer* 22, 414. doi:10.1186/s12885-022-09507-2

Sivadas, A., Kok, V. C., and Ng, K.-L. (2022). Multi-omics analyses provide novel biological insights to distinguish lobular ductal types of invasive breast cancers. *Breast Cancer Res. Treat.* 193, 361–379. doi:10.1007/s10549-022-06567-7

Sun, J., Zhou, Q., and Hu, X. (2019). Integrating multi-omics and regular analyses identifies the molecular responses of zebrafish brains to graphene oxide: Perspectives in environmental criteria. *Ecotoxicol. Environ. Saf.* 180, 269–279. doi:10.1016/j.ecoenv.2019.05.011

Sun, Y. V., and Hu, Y.-J. (2016). Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. *Adv. Genet.* 93, 147–190. doi:10.1016/bs.adgen.2015.11.004

Tao, G.-Y., Ramakrishnan, M., Vinod, K. K., Yrjälä, K., Satheesh, V., Cho, J., et al. (2020). Multi-omics analysis of cellular pathways involved in different rapid growth stages of moso bamboo. *Tree Physiol.* 40, 1487–1508. doi:10.1093/treephys/tpaa090

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11, 333–337. doi:10.1038/nmeth.2810

Wang, B., Zhang, Y., Qing, T., Xing, K., Li, J., Zhen, T., et al. (2021). Comprehensive analysis of metastatic gastric cancer tumour cells using single-cell RNA-seq. *Sci. Rep.* 11, 1141. doi:10.1038/s41598-020-80881-2

Wang, T.-H., Lee, C.-Y., Lee, T.-Y., Huang, H.-D., Hsu, J. B.-K., and Chang, T.-H. (2021). Biomarker identification through multiomics data analysis of prostate cancer prognostication using a deep learning model and similarity network fusion. *Cancers* 13, 2528. doi:10.3390/cancers13112528

Wang, T., Shao, W., Huang, Z., Tang, H., Zhang, J., Ding, Z., et al. (2021). MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat. Commun.* 12, 3445. doi:10.1038/s41467-021-23774-w

Wei, L., Jin, Z., Yang, S., Xu, Y., Zhu, Y., and Ji, Y. (2018). TCGA-Assembler 2: Software pipeline for retrieval and processing of TCGA/CPTAC data. *Bioinformatics* 34, 1615–1617. doi:10.1093/bioinformatics/btx812

Xu, J., Wu, P., Chen, Y., Meng, Q., Dawood, H., and Dawood, H. (2019). A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data. *BMC Bioinforma.* 20, 527. doi:10.1186/s12859-019-3116-7

Yamanaka, J., Kuwashima, S., and Kurita, T. (2017). "Fast and accurate image super resolution by deep CNN with skip connection and network in network," in International Conference on Neural Information Processing. Bangkok, Thailand, November, 2017 (Cham: Springer), 217–225. doi:10.1007/978-3-319-70096-0_23

Yang, F., Wang, W., Wang, F., Fang, Y., Tang, D., Huang, J., et al. (2022). scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat. Mach. Intell.* 4, 852–866. doi:10.1038/s42256-022-00534-z

Yang, H., Chen, R., Li, D., and Wang, Z. (2021). Subtype-GAN: A deep learning approach for integrative cancer subtyping of multi-omics data. *Bioinformatics* 37, 2231–2237. doi:10.1093/bioinformatics/btab109

Yuan, Q., Deng, D., Pan, C., Ren, J., Wei, T., Wu, Z., et al. (2022). Integration of transcriptomics, proteomics, and metabolomics data to reveal HER2-associated metabolic heterogeneity in gastric cancer with response to immunotherapy and neoadjuvant chemotherapy. *Front. Immunol.* 13, 951137. doi:10.3389/fimmu.2022.951137

Zhang, G., Peng, Z., Yan, C., Wang, J., Luo, J., and Luo, H. (2022). A novel liver cancer diagnosis method based on patient similarity network and DenseGCN. *Sci. Rep.* 12, 6797. doi:10.1038/s41598-022-10441-3

Zhang, X.-M., Liang, L., Liu, L., and Tang, M.-J. (2021). Graph neural networks and their current applications in bioinformatics. *Front. Genet.* 12, 690049. doi:10.3389/fgene.2021.690049