SGFCCDA: Scale Graph Convolutional Networks and Feature Convolution for circRNA-Disease Association Prediction

Junliang Shang, *Member, IEEE*, Linqian Zhao, Xin He, Xianghan Meng, Limin Zhang, Daohui Ge, Feng Li, and Jin-Xing Liu, *Member, IEEE*

***Abstract*—Circular RNAs (circRNAs) have emerged as a novel class of non-coding RNAs with regulatory roles in disease pathogenesis. Computational models aimed at predicting circRNA-disease associations offer valuable insights into disease mechanisms, thereby enabling the development of innovative diagnostic and therapeutic approaches while reducing the reliance on costly wet experiments. In this study, SGFCCDA is proposed for predicting potential circRNA-disease associations based on scale graph convolutional networks and feature convolution. Specifically, SGFCCDA integrates multiple measures of circRNA and disease similarity and combines known association information to construct a heterogeneous network. This network is then explored by scale graph convolutional networks to capture both topological and attribute information. Additionally, convolutional neural networks are employed to further learn the features and obtain higher-order feature representations containing richer information about nodes. The Hadamard product is utilized to effectively combine circRNA features with disease features, and a multilayer perceptron is applied to predict the association between each pair of circRNA and disease. Five-fold cross validation experiments conducted on the CircR2Disease dataset demonstrate the accurate prediction capabilities of SGFCCDA in identifying potential circRNA-disease associations. Furthermore, case studies provide further confirmation of SGFCCDA's ability to identify disease-associated circRNAs.**

***Index Terms*—circRNA-disease association, association prediction, graph convolutional networks, multiscale features.**

Received 28 May 2024; revised 31 July 2024 and 30 August 2024; accepted 4 September 2024. Date of publication 9 September 2024; date of current version 7 November 2024. The work was supported by the National Natural Science Foundation of China under Grant 62472250, Grant 61972226, and Grant 62172254. *(Corresponding author: Junliang Shang.)*

Junliang Shang, Linqian Zhao, Xin He, Xianghan Meng, Limin Zhang, Daohui Ge, and Feng Li are with the School of Computer Science, Qufu Normal University, Rizhao 276826, China (e-mail: shangjunliang110@163.com; zlq20001012@163.com; hexinhhxx123@163.com; mengxianghan1115@163.com; zhanglimin_99@163.com; dhge@qfnu.edu.cn; lifeng_10_28@163.com).

Jin-Xing Liu is with the School of Health and Life Sciences, University of Health and Rehabilitation Sciences, Qingdao 266113, China (e-mail: sdcavell@126.com).

Digital Object Identifier 10.1109/JBHI.2024.3456478

I. Introduction

With the rapid development of RNA sequencing technology, an increasing number of circular RNAs (circRNAs) have been identified, with the diverse functions of circRNAs being successively discovered by biologists [1]. Compared with traditional linear RNAs, circRNAs have a more stable expression pattern through reverse splicing, forming a unique closed structure with the shape of a circle [2]. Through years of dedicated research, it has been demonstrated that aberrant expression of circRNAs has a crucial influence on various changes in cell proliferation and metabolism. In addition, circRNAs can also be used as biomarkers to provide highly valuable assistance in the diagnosis, treatment, and prognosis of diseases [3].

For a long time, the associations between circRNAs and diseases have been discovered by wet experiments carried out continuously by biologists. This traditional technique not only requires unpredictable costs in terms of manpower, money, and time to implement, but also the results of wet experiments are likely to be difficult to meet the researchers' expected standards [4]. In order to solve the above problems, researchers have developed a large number of computational models for the prediction of circRNA-disease pairs with a greater likelihood of associations [5].

For the existing computational models, we analyze them in terms of the algorithms employed by the models and classify them into two main categories. The first category is the prediction of potential circRNA-disease associations by traditional computational strategies. Wei et al. [6] proposed a computational predictor named iCircDA-MF, which corrects the false negative associations with neighbor interaction profiles. Shen et al. [7] designed the XGBCDA model, which integrates multiple networks, and the in-depth extraction of potential features is carried out by Extreme Gradient Boosting classifier. However, some conventional computational strategies usually cannot adequately capture the nonlinear features of the data, which is probably leading to insufficient model performance.

In recent years, deep learning has emerged as a research hotspot across various fields due to its powerful learning capabilities, leading to remarkable achievements. For instance, Lin et al. [8] proposed LENAS, a learning-based ensemble model that predicts 3D radiotherapy doses by integrating multiple search architectures. Guo et al. [9] introduced MCANet, a deep

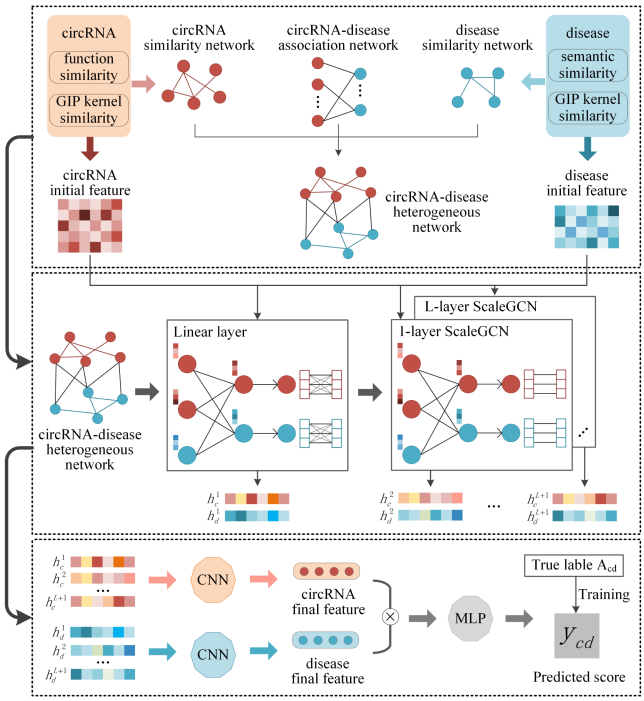


Fig. 1. The framework of SGFCCDA.

spatiotemporal neural network model that predicts Poly (A) signals by adaptively learning spatiotemporal contextual dependencies. Graph Neural Networks (GNNs) are a class of deep learning methods specifically designed for graph-structured data and are widely applied in circRNA-disease association prediction tasks. Chen et al. [10] designed a prediction model KGANCDA, based on the Knowledge Graph Attention Network, which fully considers the importance of node neighbor information to generate embedding representations. Lan et al. [11] developed an improved graph convolutional networks (GCNs) to learn node embeddings and employed a negative sampling method to reduce the impact of noisy data on model performance. Chen et al. [12] proposed a computational model LGCDA, combining local and global features to predict potential circRNA-disease associations. Wang et al. [13] proposed a GCNCDA model based on the GCNs algorithm, which uses fast learning with GCNs to extract the deep features embedded in the fusion descriptors of circRNAs and diseases. However, the aforementioned models have certain limitations: (i) Models based on traditional GCNs use excessive neural network layers in the information aggregation process, often leading to the degradation of initial feature representation. (ii) These models only use features extracted from GNN-based methods for prediction tasks, without considering interactions between features, thus lacking sufficient exploration of feature representation.

To overcome the limitations existing in the aforementioned computational models, in this study, we propose a computational model, abbreviated as SGFCCDA, which realizes the sufficient mining of deep features of circRNAs and diseases and is further used to identify disease-associated circRNAs. First, SGFCCDA fuses multiple similarities of circRNAs and diseases respectively and generates their corresponding similarity networks. By combining the association network, the circRNA-disease heterogeneous network is constructed. Secondly, SGFCCDA introduces the Scale Graph Convolutional Network (ScaleGCN) to capture both the topological information of heterogeneous networks and the attribute information of nodes, thereby altering the linear layer structure of traditional GCNs. It applies channel-scale transformations in each convolutional layer to prevent feature mixing between channels. To better learn high-order feature representations of nodes, SGFCCDA applies Convolutional Neural Networks (CNNs) for feature-level convolutional learning, fully considering interactions between features to explore richer node information. Finally, the Hadamard product is employed to efffciently combine the features of circRNAs and diseases, and the association prediction of each pair of circRNAdisease is further achieved by multilayer perceptron (MLP). Experimental results from the CircR2Disease dataset and the CircFunBase dataset demonstrate that SGFCCDA can effectively predict the potential association between circRNAs and diseases, and the prediction ability of SGFCCDA is proven to be more signiffcant by comparing with the existing models.

II. Methods

A. Overview of SGFCCDA

This section details the implementation of the SGFCCDA method, as illustrated in Fig. 1. First, we integrated various similarities of circRNA and diseases, generating their corresponding similarity networks. By combining these association networks, we constructed a circRNA-disease heterogeneous network. Next, we used ScaleGCN to learn the topological information and node attribute information on the heterogeneous network. Subsequently, SGFCCDA applied a CNN to the stacked feature matrix to further conduct convolutional learning of the features, generating the final embedding representations of the nodes. Finally, we combined the features of circRNA and diseases using the Hadamard product and input them into the MLP to predict circRNA-disease associations.

B. CircRNA-Disease Associations

Extensive biological experiments have been done by researchers so far, confirming some of the associations that clearly exist between circRNAs and diseases. We utilized CircR2Disease [14] and CircFunBase [15] databases to evaluate the performance of SGFCCDA. CircR2Disease comprises 739 confirmed human disease associations, involving 661 circRNAs and 100 diseases. Following data processing, we identified 650 human disease associations, encompassing 585 circRNAs and 88 diseases. In CircFunBase, we collected 2984 confirmed human disease associations, comprising 2597 circRNAs and 67 diseases. Then, based on known information, an association network can be constructed, with the corresponding adjacency matrix defined as A, having dimensions of $R \times D$, where R and D represent the total numbers of RNA and disease nodes, respectively. The elements of this matrix indicate the presence or absence of an association between circRNA and disease. If

a confirmed association exists, the corresponding element value is 1, indicating an edge between the associated circRNA and disease in the network; if no association exists, the element value is 0, indicating no edge.

C. Construction of Disease Similarity

In SGFCCDA, the comprehensive result of semantic similarity and Gaussian Interaction Profile (GIP) kernel [16] similarity is used as the similarity measure for diseases. The following is the computational procedure for the similarity of diseases.

1) Disease Semantic Similarity: The MeSH [17] database contains unique classification rules for diseases, which can be used to represent the attribute information of diseases. The semantic relationships provided by MeSH enable the construction of the directed acyclic graph (DAG) of diseases. Given a disease d, its DAG can be described as $DAG(d) = (d, N(d), R(d))$, where $N(d)$ represents the set of nodes that have a relationship with disease d, and $R(d)$ represents all the relationships between nodes in the DAG. In the $DAG(d)$, the semantic contribution value $SC_{d \leftarrow p}$ of a certain disease p to d can be calculated as follows:

$$SC_{d \leftarrow p} = \begin{cases} 1, & p = d \\ max\left\{\Delta * SC_{d \leftarrow p'} | p' \in \text{children of } p\right\}, & p \neq d \end{cases} \tag{1}$$

where Δ is the semantic contribution factor. Referring to previous studies [18], Δ is set to 0.5. Then, by aggregating all semantic contributions, the semantic value of disease d can be obtained as follows:

$$S_d = \sum_{p \in N(d)} SC_{d \leftarrow p} \tag{2}$$

Based on the assumption that the more identical nodes exist in the DAG of two diseases, the more similar the two diseases will be. The semantic similarity between disease d_i and disease d_j is calculated as follows:

$$DS_{sim1}(d_i, d_j) = \frac{\sum_{p \in N(d_i) \cap N(d_j)} \left(SC_{d_i \leftarrow p} + SC_{d_j \leftarrow p}\right)}{S_{d_i} + S_{d_j}} \tag{3}$$

Based on the other hypothesis that the fewer times a disease appears in a DAG, the more important it may be. From this, another semantic contribution and semantic value is calculated as follows:

$$SC'_{d \leftarrow p} = -log\left(\frac{num(DAGs(p))}{num(disease)}\right) \tag{4}$$

By aggregating all semantic contributions, another semantic value for disease d can be obtained as follows:

$$S'_d = \sum_{p \in N(d)} SC'_{d \leftarrow p} \tag{5}$$

The semantic similarity corresponding to the above theory is calculated as follows:

$$DS_{sim2}(d_i, d_j) = \frac{\sum_{p \in N(d_i) \cap N(d_j)} \left(SC'_{d_i \leftarrow p} + SC'_{d_j \leftarrow p}\right)}{S'_{d_i} + S'_{d_j}} \tag{6}$$

The above two similarities are averaged to obtain the final disease semantic similarity.

$$DS_{sim}(d_i, d_j) = \frac{DS_{sim1}(d_i, d_j) + DS_{sim2}(d_i, d_j)}{2} \tag{7}$$

2) Disease GIP Kernel Similarity: In order to more fully describe the disease similarity information, SGFCCDA calculates the disease GIP kernel similarity on the basis of the circRNA-disease association matrix. The calculation rules are as follows:

$$DG_{sim}(d_i, d_j) = \exp\left(-\gamma_d \left\|A\_d_i - A\_d_j\right\|^2\right) \tag{8}$$

where A_d_i corresponds to the i-th column of the adjacency matrix A, indicating the associations between disease d_i and all circRNAs. As a parameter for controlling width, γ_d is calculated as follows:

$$\gamma_d = \frac{1}{D_{num}} \sum_{i=1}^{D_{num}} \left\|A\_d_i\right\|^2 \tag{9}$$

where D_{num} denotes the quantity value of diseases.

3) Comprehensive Disease Similarity: The comprehensive similarity of diseases is generated by averaging the two aforementioned similarities, which adequately integrates the two attribute information of diseases. The fusion rule is shown as follows:

$$D_{sim}(d_i, d_j) = \frac{DS_{sim}(d_i, d_j) + DG_{sim}(d_i, d_j)}{2} \tag{10}$$

D. Construction of circRNA Similarity

1) CircRNA Function Similarity: In general, if two circRNAs are correlated with diseases that have the same semantic information, then we consider the two circRNAs are similar [19]. Therefore, the functional similarity of circRNAs is calculated as follows:

$$CF_{sim}(c_i, c_j) = \left[\sum_{1 \le u \le |D(c_i)|} S(d_u, D(c_j)) + \sum_{1 \le v \le |D(c_j)|} S(d_v, D(c_i))\right] / \left[|D(c_i)| + |D(c_j)|\right] \tag{11}$$

where $D(c_i)$ and $D(c_j)$ denote the set of diseases that have interactions with circRNA c_i and circRNA c_j, respectively. $S(d_u, D(c_j))$ stands the semantic similarity of disease d_u to $D(c_j)$, which is calculated as follows:

$$S(d_u, D(c_j)) = \max_{1 \le v \le |D(c_j)|} D_{sim}(d_u, d_v) \tag{12}$$

2) CircRNA GIP Kernel Similarity: Similar to the disease GIP kernel similarity, the circRNA GIP kernel similarity is also obtained by performing a computation on the adjacency matrix A as follows:

$$CG_{sim}(c_i, c_j) = \exp\left(-\gamma_c \left\|A\_c_i - A\_c_j\right\|^2\right) \tag{13}$$

$$\gamma_c = \frac{1}{C_{num}} \sum_{i=1}^{C_{num}} \left\|A\_c_i\right\|^2 \tag{14}$$

TABLE I
NOTATIONS AND THEIR ILLUSTRATION

Notation	Illustration
N	the set of nodes
A	the adjacency matrix
W	the weight transformation matrix
S	the similarity matrix
h	the embedding representation of nodes
l	the number of network layers
σ	the activation function

where A_c_i corresponds to the i-th row of the adjacency matrix A, signifying the associations between circRNA c_i and all diseases. γ_c is the width parameter and C_{num} is the amount of circRNA.

3) Comprehensive circRNA Similarity: By performing the averaging operation on the two aforementioned similarities of circRNAs, the comprehensive similarity of circRNAs can be gained. The rule is shown as follows:

$$C_{sim}\left(c_i, c_j\right) = \frac{CF_{sim}\left(c_i, c_j\right) + CG_{sim}\left(c_i, c_j\right)}{2} \tag{15}$$

E. Construction of Heterogeneous Network

Heterogeneous networks effectively represent data by organizing all objects at a systemic level and modeling associations between different biomolecules as graphs. Therefore, based on the circRNA-disease association network and their respective subnetworks, we constructed a heterogeneous network encompassing all circRNA and disease nodes. Edges between circRNAs or diseases denote similarity, whereas edges between circRNAs and diseases denote associations. Specifically, the sub-networks of circRNAs and diseases are generated via their respective comprehensive similarities. By setting a threshold on the similarity between nodes in the comprehensive similarity network, some connections with less similarity are removed to refine the topology of the heterogeneous network more adequately. The main notations utilized in SGFCCDA are shown in Table I The removal rules for connections in the comprehensive similarity networks of circRNAs and diseases are as follows:

$$N_C(c_i, c_j) = \begin{cases} 1, & C_{sim}(c_i, c_j) \geq T \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

$$N_D(d_i, d_j) = \begin{cases} 1, & D_{sim}(d_i, d_j) \geq T \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

where N_C and N_D denote the sub-networks of circRNAs and diseases, respectively. Combined with the circRNA-disease association network, the heterogeneous network can be constructed, while the corresponding adjacency matrix is defined as N_{CD}. The formula is as follows:

$$N_{CD} = \begin{bmatrix} N_C & A \\ A^T & N_D \end{bmatrix} \tag{18}$$

F. Feature Extraction by ScaleGCN With Feature Convolution

In recent years, the applications of GCNs have become more widespread and have shown superior performance in a number of tasks. The rule for extracting node features by traditional GCNs is shown as follows:

$$h^{l+1} = \sigma\left(\tilde{K}^{-\frac{1}{2}}\tilde{N}_{CD}\tilde{K}^{-\frac{1}{2}}h^l W^l\right) \tag{19}$$

where $\tilde{N}_{CD} = N_{CD} + I_{CD}$, which means to add a connection pointing to itself for each node in the heterogeneous network, and I_{CD} is identity matrix in the same dimension. $\tilde{K}$ is computed by $\tilde{K}(i,i) = \sum_j \tilde{N}_{CD}(i,j)$, W^l is the weight transformation matrix in the l-th layer, and $\sigma(\cdot)$ indicates the activation function.

The literature [20] proposes an improved graph convolution strategy. With a channel-wise scale transformation strategy, an efficient ScaleGCN is implemented on the graph. Compared with traditional GCNs, ScaleGCN changes the traditional linear layer structure by applying channel-wise scale transformation in each convolutional layer to avoid feature mixing between individual channels. Furthermore, ScaleGCN draws on two strategies, which are initial residual connection and identity mapping. The core idea of initial residual connection is to assign a connection from the input layer to each layer of the network in the model, transmitting a portion of the initial features of all nodes to each layer of the model. This operation ensures that a part of the initial feature information is covered in the final representation of the nodes, which can alleviate the over-smoothing problem of the model to some extent. For identity mapping, it centers on the idea of adding the identity matrix I_{CD} to the weight matrix W. This operation makes sure that the performance of the model does not fall below that of the shallow network model as the number of network layers deepens.

The rules for updating the node features based on the channel-wise scale transformed graph convolution are as follows:

$$h' = (1-\alpha_l)\tilde{K}^{-\frac{1}{2}}\tilde{N}_{CD}\tilde{K}^{-\frac{1}{2}}h^l + \alpha_l h^0 \tag{20}$$

$$h^{l+1} = \sigma(h'((1-\beta_l)I_{CD} + \beta_l diag(\theta^l))) \tag{21}$$

where α_l and β_l denote two hyperparameters that control the proportion of the initial feature h^0 and the hidden feature h^l in the l-th layer, respectively. θ^l indicates the parameter vector used in the l-th layer. For the initial feature representation of nodes, the initial feature matrix h^0 is obtained by projecting the comprehensive similarity matrices corresponding to circRNAs and diseases respectively into the same feature space, and concatenating all the mapped node features together. The formula is as follows:

$$h^0 = \begin{bmatrix} C_{sim} \times W_c \\ D_{sim} \times W_d \end{bmatrix} \tag{22}$$

where W_c and W_d are two different types of transformation matrices, respectively.

After L layers of ScaleGCN, the topological information of the heterogeneous network as well as the attribute information of the nodes can be fully captured. In order to further enhance

the exploration of high-order features, SGFCCDA stacks the feature representations of circRNAs and diseases obtained from each layer to prepare for the realization of high-order feature extraction. Taking circRNA as example, the node features after stacking are represented as:

$$H_c = \begin{bmatrix} h_c^1 \\ h_c^2 \\ \vdots \\ h_c^{L+1} \end{bmatrix} \tag{23}$$

To adequately learn the different feature representations of nodes, SGFCCDA applies convolutional neural network on the stacked feature matrix to further perform convolutional learning on features, generating representations containing richer information [21]. The formula is as follows:

$$X_c = \delta(\phi * H_c + \gamma) \tag{24}$$

where $\delta(\cdot)$ represents the nonlinear activation function, ϕ and γ refer to the weight matrix and bias, respectively. Through similar steps, the feature representations X_c of diseases can be obtained.

G. Prediction of Potential Associations

There are still numerous circRNA-disease associations that have not been discovered so far. To explore these unknown associations, SGFCCDA performs pairwise combination of circRNAs and diseases, and computes its interaction features for each circRNA-disease pair through the Hadamard product. Given a pair of circRNA c and disease d, their corresponding features are calculated as follows:

$$F(c, d) = x_c \odot x_d \tag{25}$$

where $x_c \in X_c$ and $x_d \in X_d$ are the feature vectors corresponding to circRNA c and disease d, respectively. $\odot$ represents the Hadamard product, which operates by multiplying the elements at the same positions of x_c and x_c, respectively.

Eventually, SGFCCDA utilizes MLP to learn the potential interactions between each pair of circRNAs and diseases. The predicted association score is calculated as follows:

$$y_{cd} = \tau(\omega^l(\cdots(\tau(\omega^1 F(c, d) + b^1)) + \cdots) + b^l) \tag{26}$$

where $\tau(\cdot)$ denotes the activation function, ω^l and b^l are the weight matrix and bias vector in the l-th layer, respectively. y_{cd} is the association score of circRNA c and disease d predicted by the model, which is finally trained by minimizing the binary cross-entropy loss. The loss function is defined as follows:

$$Loss = -\left(A_{cd} \log y_{cd} + (1 - A_{cd}) \log\left(1 - y_{cd}\right)\right) \tag{27}$$

where A_{cd} is the element in the circRNA-disease association adjacency matrix A.

H. Complexity Analysis

The time complexity of SGFCCDA is composed of the following parts. Firstly, features of circRNAs and diseases are extracted using two layers of scale graph convolutional networks, with C_m and D_n representing the numbers of circRNAs and diseases, respectively. For circRNAs, let a_1 and a_2 be the numbers of nodes in the first and second layers of the scale graph convolutional networks, respectively. The time complexity for circRNAs is $O(C_m^2 a_1 + a_1^2 a_2)$. Similarly, for diseases, let b_1 and b_2 be the numbers of nodes in the first and second layers of the scale graph convolutional networks, respectively. The time complexity for diseases is $O(D_n^2 b_1 + b_1^2 b_2)$. Next, we use feature convolution to comprehensively learn different feature representations of nodes, with a time complexity of $O(C_m^2 a_3 + D_n^2 b_3)$. Finally, SGFCCDA learns the potential interactions between each pair of circRNA and disease, with a time complexity of $O(C_m D_n l_1 + l_1^2 l_2)$. Therefore, the time complexity of SGFCCDA is $O(C_m^2(a_1 + a_3) + a_1^2 a_2 + D_n^2(b_1 + b_3) + b_1^2 b_2 + C_m D_n l_1 + l_1^2 l_2)$. The source data and code can be downloaded at https://github.com/CDMBlab/SGFCCDA.

III. Experiments and Results

A. Evaluation Metrics

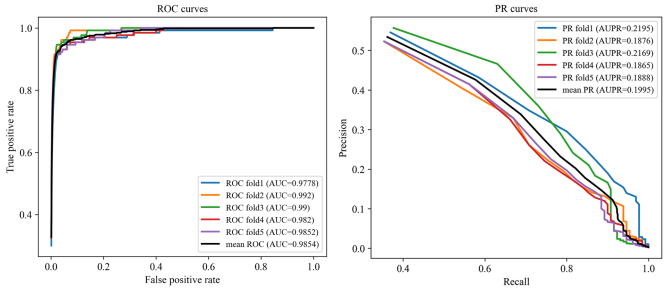
To evaluate the predictive performance of SGFCCDA, the five-fold cross validation (FF-CV) experiment is conducted on the CircR2Disease dataset. In FF-CV, the dataset is equally divided into five sub-parts, four of which are used in the training process of the model and the remaining one is used as the testing set. After repeating the process for five times, each of the five sub-datasets has experienced an opportunity to be used as testing set. In this study, we selected known circRNA-disease pairs from the dataset as positive samples, and randomly selected an equal number of circRNA-disease pairs from unknown associations as negative samples, ensuring a balanced dataset. The average values of the five results are finally calculated to obtain the final results. In addition, SGFCCDA adopts several evaluation metrics that are commonly used in the field of circRNA-disease associations prediction research, including the area under the receiver operating characteristic curve (AUC), the area under the precision-recall curve (AUPR), Accuracy, Recall, Specificity, and F1-value.

B. Performance Analysis

The performance of SGFCCDA in predicting circRNA-disease associations is validated by performing FF-CV on the CircR2Disease dataset. The ROC curves and PR curves achieved by the model are shown in Fig. 2(a) and (b). The detailed results of the remaining metrics are listed in Table II. As can be seen from the table, SGFCCDA has achieved results of 0.9854, 0.1995, 0.9955, 0.9853, 0.9985 and 0.8929 for AUC, AUPR, Accuracy, Recall, Specificity and F1-value, with standard deviations of 0.0052, 0.0150, 0.0021, 0.0051, 0.0005 and 0.0293, respectively. It can be observed that on the AUC, which responds to the comprehensive performance of the model, the FF-CV of SGFCCDA achieves an excellent result of more than 0.9500 for each fold. In addition, Accuracy reflects the degree of correctness of circRNA-disease associations predicted by the model, and SGFCCDA achieves more than 0.99 for each fold on this metric. The experimental results on each of these evaluation metrics demonstrate that SGFCCDA has the ability to accurately

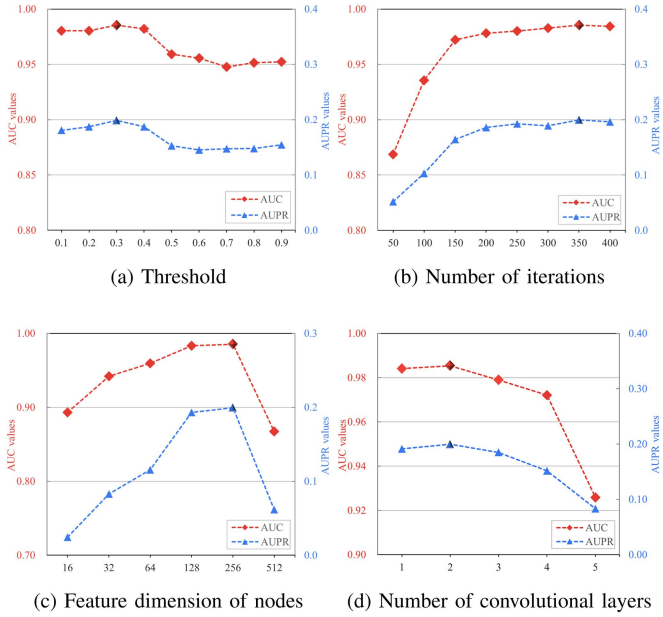
TABLE II
RESULTS OF FF-CV ACHIEVED BY SGFCCDA

Fold	AUC	AUPR	Accuracy	Recall	Specificity	F1-score
1	0.9778	0.2195	0.9922	0.9778	0.9993	0.8437
2	0.9920	0.1876	0.9978	0.9918	0.9977	0.9109
3	0.9900	0.2169	0.9975	0.9899	0.9983	0.9311
4	0.9820	0.1865	0.9948	0.9819	0.9987	0.8844
5	0.9852	0.1888	0.9953	0.9851	0.9984	0.8943
Average	0.9854±0.0052	0.1995±0.0150	0.9955±0.0021	0.9853±0.0051	0.9985±0.0005	0.8929±0.0293



(a) ROC curve (b) PR curve

Fig. 2. FF-CV achieved by SGFCCDA.



(a) Threshold (b) Number of iterations

(c) Feature dimension of nodes (d) Number of convolutional layers

Fig. 3. Parametric analysis of SGFCCDA.

predict circRNA-disease associations with superior predictive performance.

C. Parameter Analysis

In SGFCCDA, there are some parameters that affect the performance of the model in different degrees. In order to improve the predictive ability, different parameters are adjusted to analyze SGFCCDA. Several experiments are implemented on the CircR2Disease dataset, and when adjusting one parameter, it is ensured that each of the other parameters are kept at a consistent setting. During the construction of the circRNA-disease heterogeneous network, SGFCCDA sets a threshold for the comprehensive similarity of circRNAs and diseases to generate their respective sub-networks. The threshold is set to range from 0.1 to 0.9. The larger the selected threshold, the more edges will be deleted in generating the sub-network, and the sparser the constructed heterogeneous network will be. It can be seen from Fig. 3(a) that SGFCCDA achieves the best performance when the threshold is set to 0.3. Fig. 3(b) displays the effect of the number of iterations in the training process on the model performance. Too few iterations will make it difficult for the model to reach convergence, and too many iterations may consume a lot of unnecessary runtime. After a series of experiments, it is noticed that when the number of iterations is set to 350, SGFCCDA is able to achieve relatively good prediction performance.

SGFCCDA utilizes ScaleGCN for information aggregation and embedding update of network nodes, where the feature dimension of the nodes and the number of convolutional layers both affect the performance of SGFCCDA. The larger the feature dimension of a node, the more accurate the information it represents. However, if it carries excessive information, the model performance may decrease due to the presence of redundant features. We set the values of the feature dimension of nodes to 16, 32, 64, 128, 256, 512. From Fig. 3(c), it can be seen that the performance of SGFCCDA reaches the optimal result when the feature dimension is set to 256. The number of convolutional layers is determined from 1, 2, 3, 4, 5. Excessive number of convolutional layers may lead to the risk of over-fitting. Fig. 3(d) shows the performance of SGFCCDA with different number of convolutional layers, from which it can be inferred that SGFCCDA performs best when the number of layers is set to 2.

D. Ablation Studies

To verify the impact of ScaleGCN and feature convolution on the performance of SGFCCDA, ablation experiments are carried out. In SGFCCDA, ScaleGCN is applied for capturing the topological information of heterogeneous networks as well as the attribute information of nodes. In addition, in order to explore richer node information, SGFCCDA utilizes a CNN approach to implement feature-level convolutional learning to further learn higher-order feature representations of nodes. For the former, we replace the ScaleGCN in SGFCCDA with the basic GCN algorithm, with all other setups the same, which is named SGFCCDA_noSG. For the latter, we remove the CNN portion of SGFCCDA, i.e., the features extracted by the ScaleGCN are directly used for the prediction of associations, which is

TABLE III
RESULTS OF ABLATION STUDIES REALIZED BY SGFCCDA

Model	AUC	AUPR	Accuracy	Recall	Specificity	F1-score
SGFCCDA	0.9854	0.1995	0.9955	0.9853	0.9985	0.8929
SGFCCDA_noFC	0.9655	0.1941	0.9754	0.9656	0.9784	0.8698
SGFCCDA_noSG	0.8565	0.0379	0.8701	0.8580	0.8625	0.7531

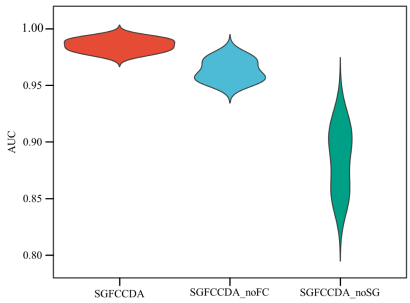
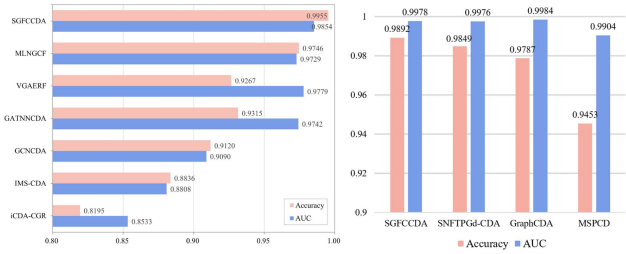


Fig. 4. AUC analysis of ablation studies.



(a) CircR2Disease dataset (b) CircFunBase dataset

Fig. 5. The comparison with other models on the CircR2Disease dataset and the CircFunBase dataset.

named SGFCCDA_noFC. The results are shown in Table III. From the data in the table, the AUC of SGFCCDA_noSG and SGFCCDA_noFC are 0.8565 and 0.9655, respectively, which are both lower than the AUC results of SGFCCDA. Moreover, the results of SGFCCDA are better than those of the other two models in every evaluation metrics. In order to further compare the prediction performance, we analyze the distribution of AUC values obtained by the above three models in the FF-CV, and the results are shown in Fig. 4. It can be seen that SGFCCDA not only has the highest AUC index, but also its prediction results are more stable compared to the other two models. The aforementioned results indicate that both the ScaleGCN introduced in SGFCCDA and the CNN used for extracting higher-order features play a vital role in improving the model's ability towards predicting circRNA-disease associations.

E. Comparison With Other Models

To demonstrate the superior predictive performance of SGFCCDA, we compare it with six representative models on the CircR2Disease dataset. The compared models include:

iCDA-CGR [22]: The model introduces the position information of circRNA sequences into the circRNA-disease association prediction model for the first time.

IMS-CDA [23]: The model integrates diverse similarity information of circRNAs and diseases, and then uses a deep learning stacked autoencoder to extract hidden features.

GCNCDA [15]: The model also employs deep learning algorithms to learn high-level features of circRNAs and diseases through the fast learning with GCNs.

GATNNCDA [24]: The model has designed an end-to-end prediction model in which graph attention networks are employed to extract low-dimensional feature representations.

VGAERF [25]: The model proposes a predictor that combines the variational graph auto-encoder with random forest.

MLNGCF [26]: The model develops a collaborative filtering method based on a multi-layer attentional neural graph and utilizes it for the prediction of circRNA-disease associations.

To further validate the superiority of SGFCCDA, we compared it with three advanced methods using the CircFunBase dataset. The models compared include:

SNFTPGd-CDA [27]: The model uses similarity network fusion to integrate multiple networks and applies tensor product graph diffusion to propagate node similarities.

GraphCDA [28]: The model combines GCNs and GATs to simultaneously learn feature representations of nodes.

MSPCD [29]: This model integrates diverse biological information and utilizes neural networks to extract high-order feature representations of nodes.

We select two of these evaluation metrics to compare the seven models, that is AUC, which can reflect the comprehensive performance, and Accuracy, which can reflect the overall prediction accuracy. The experimental results on the CircR2Disease dataset are shown in Fig. 5(a). As can be seen from the figure, the SGFCCDA proposed in this paper achieves the best results on both evaluation metrics. Although the other six computational models are also capable of handling the circRNA-disease association prediction task, it is obvious that SGFCCDA is more competitive than them. On the CircFunBase dataset, SGFCCDA achieved the highest Accuracy and the second-best area AUC. These results are attributed to SGFCCDA's use of ScaleGCN for feature extraction from heterogeneous networks, which avoids feature mixing through channel scaling and thereby mitigates the over-smoothing problem. Additionally, SGFCCDA employs CNNs for convolutional learning of features, thoroughly considering the interactions between features, which leads to comprehensive exploration and provides rich feature representations for subsequent prediction tasks.

To validate the stability of SGFCCDA, we performed a 5-fold cross-validation comparison with six other methods on the CircR2Disease dataset, as shown in the AUC box plot in Fig. 6. In the box plot, a larger box indicates greater result dispersion. Based on the data positions and box sizes, it is evident that the proposed model exhibits good predictive performance and stability, further validating SGFCCDA's capability to predict potential circRNA-disease associations.

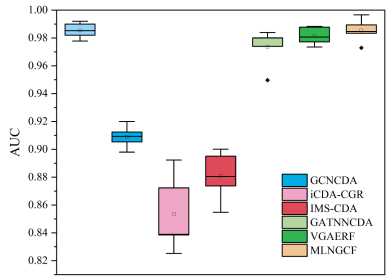


Fig. 6. FF-CV on the CircR2Disease dataset with other models.

TABLE IV
TOP 20 ASSOCIATED WITH COLORECTAL CANCER

Rank	circRNA	Evidence
1	circPVT1	36052265
2	circFAT1	32379550
3	circFUT8	34948079
4	hsa_circ_0001666	34841662
5	circ-LDLRAD3	37587156
6	circRHOBTB3	34158864
7	circARHGAP5	unconfirmed
8	hsa_circ_101555	33824281
9	circVRK1	unconfirmed
10	circZFR	36092343
11	hsa_circRNA_102051	37794386
12	hsa_circRNA_101717	35421275
13	circ-Foxo3	34549306
14	hsa_circ_006054	30585259
15	circDENND4C	32196590
16	circSMARCA5	34948079
17	hsa_circ_0005927	33732022
18	circ-FBXW7	33147048
19	hsa_circRNA_104270	unconfirmed
20	circPTK2	31973707

F. Case Studies

To estimate the ability of SGFCCDA in predicting novel circRNA-disease associations, case studies are conducted. In particular, all the validated associations between circRNAs and diseases are used for the training of the model, and then all the pairs of unknown associations in the dataset are predicted. In the prediction results of SGFCCDA, we choose two diseases, colorectal cancer, and hepatocellular carcinoma.

Colorectal cancer is a malignant tumor that occurs in the colon or rectum, with an increasing incidence with age, especially common in people over 50 years old [30]. As can be seen in Table IV, 17 of the top 20 circRNAs predicted by SGFCCDA to be associated with colorectal cancer can be confirmed through the literature. For example, the expression level of CircPVT1 is significantly elevated in colorectal cancer, which promotes the proliferation and metastasis of colorectal cancer [31]. Hepatocellular carcinoma is the most common type of primary liver cancer worldwide [32]. The top 20 circRNAs predicted by SGFCCDA to be associated with hepatocellular carcinoma are listed in Table V.

Additionally, to further test the ability of SGFCCDA to predict unknown circRNA-disease associations, we trained it using all known associations in the CircR2Disease dataset and then predicted all unknown associations in the same dataset. We ranked the scores of all predicted circRNA-disease pairs with unknown associations and listed the top 20 predicted pairs in Table VI. As can be seen from the table, 17 of the top 20 association pairs can be verified by the relevant literature. For instance, Chen et al. proposed that circRNA_100290 is a potential biomarker for most oral squamous cell carcinomas [33]. CircGFRA can inhibit the progression of breast cancer by suppressing the proliferation of HER-2 positive breast cancer cells [34]. CircPVT1 has been shown through gene expression analysis to be associated with certain clinicopathological features of colorectal cancer [31]. The results above indicate that SGFCCDA possesses the ability to predict potential circRNA-disease associations. The results of the case studies can facilitate the conduct of wet experiments to a certain extent by providing biologists with disease-associated candidate circRNAs.

TABLE V
TOP 20 ASSOCIATED WITH HEPATOCELLULAR CARCINOMA

Rank	circRNA	Evidence
1	circFUT8	36474147
2	hsa_circRNA_104348	33311442
3	circZFR	33433801
4	circ-BANP	34985013
5	hsa_circ_0000677	30086881
6	hsa_circRNA_404833	unconfirmed
7	hsa_circ_0000567	30795787
8	hsa_circ_0000517	32774154
9	hsa_circ_0012673	unconfirmed
10	hsa_circRNA_401977	unconfirmed
11	hsa_circRNA_103809	32683589
12	hsa_circ_100226	32801909
13	hsa_circ_0014717	33598424
14	circMylk	32904604
15	hsa_circ_0000096	31108351
16	hsa_circ_0015756	31522588
17	hsa_circRNA_104700	unconfirmed
18	hsa_circ_0085616	31496798
19	hsa_circ_0000520	35322101
20	hsa_circRNA_104268	31222831

TABLE VI
TOP 20 CIRCRNA-DISEASE ASSOCIATIONS PREDICTED

Rank	circRNA	disease	Evidence
1	circRNA_100290	Oral squamous cell carcinoma	31187488
2	circGFRA	Breast cancer	34668628
3	circPVT1	Colorectal cancer	36052265
4	circRTN4	Breast cancer	30810051
5	hsa_circ_0088452	Pulmonary tuberculosis	unconfirmed
6	CircDOCK1	Oral squamous cell carcinoma	29286141
7	circFAT1	Colorectal cancer	32379550
8	circCCDC66	Colon cancer	30630646
9	circRNA.22054	Diabetes mellitus	29526755
10	hsa_circ_0011021	Pancreatic cancer	unconfirmed
11	circHIPK3	Type 2 diabetes mellitus	36209373
12	hsa_circRNA_404833	Lung cancer	29241190
13	circSMARCA5	Glioma	29415469
14	CDR1as	Pancreatic cancer	33593338
15	hsa_circRNA_003251	Diabetes mellitus	28779132
16	CDR1as	Breast cancer	30884120
17	hsa_circ_0014717	Colon cancer	unconfirmed
18	hsa_circRNA_104075	Glioma	31112718
19	circRNA_0001073	Lung cancer	31484581
20	hsa_circ_0015278	Liver cancer	31421050

IV. Conclusion

In this study, an effective computational model, SGFCCDA, is proposed to realize the task of predicting unknown associations by fully exploiting the nonlinear features of circRNAs and diseases, the highlights of which involve two parts. On one hand, to address the issue of over-smoothing caused by excessive neural network layers aggregating node information in traditional GCN methods, SGFCCDA implements ScaleGCN to extract features in heterogeneous networks. It avoids feature mixing through channel-scale transformations, thereby alleviating the over-smoothing issue. On the other hand, to fully consider interactions between features, SGFCCDA applies CNNs for convolutional learning on features. This enables a more comprehensive exploration, resulting in richer feature representations for subsequent prediction tasks. Experimental results on the CircR2Disease dataset and the CircFunBase dataset demonstrate that SGFCCDA achieves superior predictive performance and is capable of accurately predicting potential associations between circRNAs and diseases.

However, there are still several limitations to be addressed for better prediction performance. The similarity calculation methods used by SGFCCDA in constructing heterogeneous networks may lack diversity. We aim to apply more advanced algorithms for calculating prediction scores in the future.

References

[1] C.-X. Liu and L.-L. Chen, "Circular RNAs: Characterization, cellular roles, and applications," *Cell*, vol. 185, no. 12, pp. 2016–2034, 2022.

[2] H. L. Sanger, G. Klotz, D. Riesner, H. J. Gross, and A. K. Kleinschmidt, "Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures," *Proc. Nat. Acad. Sci.*, vol. 73, no. 11, pp. 3852–3856, 1976.

[3] J. Dong, Z. Zeng, Y. Huang, C. Chen, Z. Cheng, and Q. Zhu, "Challenges and opportunities for circRNA identification and delivery," *Crit. Rev. Biochem. Mol. Biol.*, vol. 58, no. 1, pp. 19–35, 2023.

[4] Q. Xiao, J. Dai, and J. Luo, "A survey of circular RNAs in complex diseases: Databases, tools and computational methods," *Brief. Bioinf.*, vol. 23, no. 1, 2022, Art. no. bbab444.

[5] W. Lan et al., "Benchmarking of computational methods for predicting circRNA-disease associations," *Brief. Bioinf.*, vol. 24, no. 1, 2023, Art. no. bbac613.

[6] H. Wei and B. Liu, "iCircDA-MF: Identification of circRNA-disease associations based on matrix factorization," *Brief. Bioinf.*, vol. 21, no. 4, pp. 1356–1367, 2020.

[7] S. Shen, J. Liu, C. Zhou, Y. Qian, and L. Deng, "XGBCDA: A multiple heterogeneous networks-based method for predicting circRNA-disease associations," *BMC Med. Genomic.*, vol. 13, no. 1, pp. 1–10, 2022.

[8] Y. Lin et al., "LENAS: Learning-based neural architecture search and ensemble for 3-D radiotherapy dose prediction," *IEEE Trans. Cybern.*, 2024, early access, May 10, 2024, doi: 10.1109/TCYB.2024.3390769.

[9] Y. Guo, D. Zhou, P. Li, C. Li, and J. Cao, "Context-aware poly (A) signal prediction model via deep spatial-temporal neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 8241–8253, Jun. 2024.

[10] W. Lan et al., "KGANCDA: Predicting circRNA-disease associations based on knowledge graph attention network," *Brief. Bioinf.*, vol. 23, no. 1, 2022, Art. no. bbab494.

[11] W. Lan et al., "IGNSCDA: Predicting circRNA-disease associations based on improved graph convolutional network and negative sampling," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 6, pp. 3530–3538, Nov./Dec. 2022.

[12] W. Lan et al., "LGCDA: Predicting CircRNA-Disease association based on fusion of local and global features," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2024, early access, Apr. 12, 2024, doi: 10.1109/TCBB.2024.3387913.

[13] L. Wang, Z.-H. You, Y.-M. Li, K. Zheng, and Y.-A. Huang, "GCNCDA: A new method for predicting circRNA-disease associations based on graph convolutional network algorithm," *PLoS Comput. Biol.*, vol. 16, no. 5, 2020, Art. no. e1007568.

[14] C. Fan, X. Lei, Z. Fang, Q. Jiang, and F.-X. Wu, "CircR2Disease: A manually curated database for experimentally supported circular RNAs associated with various diseases," *Database*, vol. 2018, 2018, Art. no. bay044.

[15] X. Meng et al., "CircFunBase: A database for functional circular RNAs," *Database*, vol. 2019, 2019, Art. no. baz003.

[16] T. Van Laarhoven, S. B. Nabuurs, and E. Marchiori, "Gaussian interaction profile kernels for predicting drug-target interaction," *Bioinformatics*, vol. 27, no. 21, pp. 3036–3043, 2011.

[17] C. E. Lipscomb, "Medical subject headings (MeSH)," *Bull. Med. Library Assoc.*, vol. 88, no. 3, 2000, Art. no. 265.

[18] D. Wang, J. Wang, M. Lu, F. Song, and Q. Cui, "Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases," *Bioinformatics*, vol. 26, no. 13, pp. 1644–1650, 2010.

[19] X. Chen, C. Clarence Yan, C. Luo, W. Ji, Y. Zhang, and Q. Dai, "Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity," *Sci. Rep.*, vol. 5, no. 1, 2015, Art. no. 11338.

[20] T. Zhang, Q. Wu, J. Yan, Y. Zhao, and B. Han, "ScaleGCN: Efficient and effective graph convolution via channel-wise scale transformation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 4, pp. 4478–4490, Apr. 2024.

[21] L. Han, H. Wu, N. Hu, and B. Qu, "Convolutional neural collaborative filtering with stacked embeddings," in *Proc. Asian Conf. Mach. Learn.*, 2019, pp. 726–741.

[22] K. Zheng, Z.-H. You, J.-Q. Li, L. Wang, Z.-H. Guo, and Y.-A. Huang, "iCDA-CGR: Identification of circRNA-disease associations based on chaos game representation," *PLoS Comput. Biol.*, vol. 16, no. 5, 2020, Art. no. e1007872.

[23] L. Wang, Z.-H. You, J.-Q. Li, and Y.-A. Huang, "IMS-CDA: Prediction of CircRNA-disease associations from the integration of multisource similarity information with deep stacked autoencoder model," *IEEE Trans. Cybern.*, vol. 51, no. 11, pp. 5522–5531, Nov. 2021.

[24] C. Ji, Z. Liu, Y. Wang, J. Ni, and C. Zheng, "GATNNCDA: A method based on graph attention network and multi-layer neural network for predicting circRNA-disease associations," *Int. J. Mol. Sci.*, vol. 22, no. 16, 2021, Art. no. 8505.

[25] Y. Fu, R. Yang, and L. Zhang, "Association prediction of CircRNAs and diseases using multi-homogeneous graphs and variational graph autoencoder," *Comput. Biol. Med.*, vol. 151, 2022, Art. no. 106289.

[26] Q. Wu et al., "MLNGCF: CircRNA-disease associations prediction with multilayer attention neural graph-based collaborative filtering," *Bioinformatics*, vol. 39, no. 8, 2023, Art. no. btad499.

[27] H. Liu et al., "Tensor product graph diffusion based on nonlinear fusion of multi-source information to predict circRNA-disease associations," *Appl. Soft Comput.*, vol. 152, 2024, Art. no. 111215.

[28] Q. Dai et al., "GraphCDA: A hybrid graph representation learning framework based on GCN and GAT for predicting disease-associated circRNAs," *Brief. Bioinf.*, vol. 23, no. 5, 2022, Art. no. bbac379.

[29] L. Deng et al., "MSPCD: Predicting circRNA-disease associations via integrating multi-source data and hierarchical neural network," *BMC Bioinf.*, vol. 23, no. 3, 2022, Art. no. 427.

[30] I. Mármol, C. Sánchez-de-Diego, A. Pradilla Dieste, E. Cerrada, and M. J. Rodriguez Yoldi, "Colorectal carcinoma: A general overview and future perspectives in colorectal cancer," *Int. J. Mol. Sci.*, vol. 18, no. 1, 2017, Art. no. 197.

[31] M. Zamani, A.-M. Foroughmand, M.-R. Hajjari, B. Bakhshinejad, R. Johnson, and H. Galehdari, "CASC11 and PVT1 spliced transcripts play an oncogenic role in colorectal carcinogenesis," *Front. Oncol.*, vol. 12, 2022, Art. no. 954634.

[32] J. Balogh et al., "Hepatocellular carcinoma: A review," *J. Hepatocellular Carcinoma*, vol. 3, pp. 41–53, 2016.

[33] X. Chen et al., "Circle RNA hsa_circRNA_100290 serves as a ceRNA for miR-378a to regulate oral squamous cell carcinoma cells growth via Glucose transporter-1 (GLUT1) and glycolysis," *J. Cellular Physiol.*, vol. 234, no. 11, pp. 19130–19140, 2019.

[34] M. Bazhabayi et al., "CircGFRA1 facilitates the malignant progression of HER-2-positive breast cancer via acting as a sponge of miR-1228 and enhancing AIFM2 expression," *J. Cellular Mol. Med.*, vol. 25, no. 21, pp. 10248–10256, 2021.