IGNSCDA: Predicting CircRNA-Disease Associations Based on Improved Graph Convolutional Network and Negative Sampling

Wei Lan, Yi Dong, Qingfeng Chen*, Jin Liu, Jianxin Wang, Yi-Ping Phoebe Chen and Shirui Pan

Abstract—Accumulating evidences have shown that circRNA plays an important role in human diseases. It can be used as potential biomarker for diagnose and treatment of disease. Although some computational methods have been proposed to predict circRNA-disease associations, the performance still need to be improved. In this paper, we propose a new computational model based on Improved Graph convolutional network and Negative Sampling to predict CircRNA-Disease Associations. In our method, it constructs the heterogeneous network based on known circRNA-disease associations. Then, an improved graph convolutional network is designed to obtain the feature vectors of circRNA and disease. Further, the multi-layer perceptron is employed to predict circRNA-disease associations based on the feature vectors of circRNA and disease. In addition, the negative sampling method is employed to reduce the effect of the noise samples, which selects negative samples based on circRNAs expression profile similarity and Gaussian Interaction Profile kernel similarity. The 5-fold cross validation is utilized to evaluate the performance of the method. The results show that IGNSCDA outperforms than other state-of-the-art methods in the prediction performance. Moreover, the case study shows that IGNSCDA is an effective tool for predicting potential circRNA-disease associations.

Index Terms—circRNA-disease associations, graph convolutional network, circRNA similarity.

1 Introduction

CIRCRNA is a kind of non-coding RNA without 5 and 3 polyadenylated tails [1], [2]. It is different from linear RNA that the circRNA is relatively stable as it can escape the digestion of exonuclease [3]. As early as 1980s, it was found that the circRNA widely existed in plants and eukaryotes [4], [5]. According to the location in genome, circRNAs are classified as four categories: exonic circRNAs, intronic circRNAs, exonintron circRNAs and intergenic circRNAs [6], [7]. In the past, the circRNA was viewed as the transcription noise [8]. However, with the development of high-throughput sequencing technology, it has been discovered that circRNAs participate in plenty biological processes such as miRNA sponge [9], transcription of parent gene [10], RNA binding proteins regulators [11], etc. Furthermore, increasing studies have found that the circRNAs have a close relationship with human diseases [12], [13], [14], [15], [16]. For example, the deletion of cirs-7 in the brain tissue can lead to the up-regulation of miR-7 and the down regulation of multiple mRNA targets related to Alzheimers disease [17], [18]. In addition, it has been proved that the hsa_circ_002059 is down-regulated in plasma sample of gastric cancer [19]. The hsa_circ_001649 is prominently down-regulated, while hsa_circ_005075 is significantly up-regulated in hepatocellular carcinoma [20], [21]. Therefore, predicting the associations between circRNA and disease can help biologist to understand disease pathology, and further for diagnosis and treatment of disease.

Many circRNA-disease associations have been found by using biological experimental methods [22], [23]. However, it has some limitations such as time-consuming and expensive [24]. In order to address problems, some computational methods have been proposed to predict potential circRNA-disease associations. These computational approaches are based on the assumption that similar circRNAs tend to be related with similar diseases [25], [26], [27]. These methods can be divided into three categories. The first class of methods are based on the information propagation. Lei et al. [28] proposed a computational method (BRWSP) to predict circRNA-disease associations based on Biased Random Walk. Specifically, they constructed a large heterogeneous network by using associations between circRNAs, diseases and genes. Then, the partial random walk was utilized to search paths in the heterogeneous network. Finally, the final circRNA-disease association scores were computed

- *Wei Lan is School of Computer, Electronic and Information, Guangxi University, Nanning, Guangxi, 530004 and Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, Hunan, 410083, China. E-mail: lanwei@gxu.edu.cn*
- *Yi Dong is School of Computer, Electronic and Information, Guangxi University, Nanning, Guangxi, 530004, China. E-mail: dongyi@st.gxu.edu.cn*
- *Qingfeng Chen is School of Computer, Electronic and Information and State Key Laboratory for Conservation and Utilization of Subtropical Agro-bioresources, Guangxi University, Nanning, Guangxi, 530004, China. E-mail: qingfeng@gxu.edu.cn*
- *Jin Liu is Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, Hunan, 410083, China. E-mail:liujin06@mail.csu.edu.cn*
- *Jianxin Wang is Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, Hunan, 410083, China. E-mail: jxwang@mail.csu.edu.cn*
- *Yi-Ping Phoebe Chen is Department of Computer Science and Information Technology, La Trobe University, Melbourne Victoria 3086, Australia. E-mail:phoebe.chen@latrobe.edu.au*
- *Shirui Pan is the Department of Data Science and AI, Faculty of Information Technology, Monash University, Melbourne Victoria 3800, Australia. E-mail: shirui.pan@monash.edu*
- ** Corresponding author*

Manuscript received April 19, 2020; revised August 26, 2020.

based on the paths. Fan et al. [29] proposed a method (KATZHCDA) based on Katz to identify circRNA-disease associations. Similarly, Deng et al. [30] presented the method (KATZCPDA) to predict disease-related circRNA. Zhao et al. [31] combined Katz with bipartite network projection algorithm to infer potential circRNA-disease associations. Hseyin et al. [32] presented a computational method for circRNA-disease associations prediction by using random walk with restart. Ding et al. [33] proposed a computational model (RWLR) by combining random walk with restart with logistic regression to predict circRNA-disease associations. Similar work was developed by Lei et al. [34], they integrated random walk with restart and k-Nearest Neighbor to identify circRNA-disease associations. The second kind of methods are based on machine learning. Li et al. [35] proposed a computational method (SIMCCDA) to infer circRNA-disease associations based on matrix completion. It constructed a comprehensive matrix by integrating different kinds of biological data. Then, the speedup inductive matrix completion was employed to predict the affinity scores between circRNAs and diseases. Wei et al. [36] introduced a model (iCircDA-MF) based on matrix factorization to predict circRNA-disease associations, which corrected false negatives based on neighbors profile. Yan et al. [37] proposed a computational method (DWNN-RLS) to predict circRNA-disease associations based on the regularized least square method with Kronecker product kernel. In this method, the weighted nearest neighbor method was utilized to address the prediction problem of the new circRNA. The third type of methods are based on deep learning. Wang et al. [38] presented a computational framework to identify circRNA-disease associations based on deep learning. In this method, the convolutional neural network was used to extract hidden features and the extreme learning machine was utilized to predict circRNA-disease associations. Wang et al. [39] proposed a computational model (GCNCDA) to infer associations between circRNAs and diseases by using graph convolutional network. These models have achieved great successes in circRNA-disease prediction. However, there are some limitations in these methods. Some methods ignored the feature vectors of each layer and each node. Furthermore, the negative samples were chosen randomly, which may bring some noise as some negative samples may be unlabeled positive samples. In addition, the graph convolutional network (GCN) adjusted the embedding vectors of circRNAs and diseases based on their neighbor profile information. However, the relationship information extracted by GCN was limited as the sparsity of data. Therefore, the origin GCN need to be improved to obtain the feature vectors containing more information.

In this paper, we propose a new computation framework (IGNSCDA) based on an Improved Graph convolutional network and a Negative Sampling to infer Associations between CircRNAs and Diseases. Specifically, it constructs heterogeneous network based on known circRNA-disease associations. Then, an improved graph convolutional network is designed to obtain feature vectors of circRNAs and diseases, respectively. Further, the multi-layer perceptron is employed to predict circRNA-disease associations based on the feature vectors of circRNA and disease. In order to reduce the influence of noisy samples, a negative sampling method is developed to select negative samples in our model. The experimental results demonstrate IGNSCDA outperforms than other state-of-the-art methods in terms of 5-fold validation. In addition, the case study shows that our model is an effective tool for predicting circRNA-disease associations.

2 Materials and Methods

2.1 Data Collection

The benchmark dataset of circRNA-disease association is downloaded from CircR2Disease [40]. 612 known circRNA-disease associations including 533 circRNAs and 89 diseases are obtained in final. The matrix $Y \in R^{m \times n}$ is constructed to represent associations between circRNAs and diseases, where m and n represent the number of circRNAs and diseases, respectively.

$$Y_{ij} = \begin{cases} 1, & \text{if circRNA } i \text{ is associated with disease } j \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

The expression profile of circRNAs is downloaded from exoRBase database [41]. The dimension of circRNAs expression data is 90, which represents different expression levels based on normalized RNA-seq data spanning both normal individuals and patients with different diseases. Then, the circRNA id in exoRBase is mapped into CircR2Disease dataset, and 192 circRNAs expression profile is obtained in final.

2.2 Vector Embedding

Each circRNA and disease in matrix Y is extracted and represented by one hot encoding [42]. While the origin vector after one-hot encoding is long and sparse [43], the vector embedding layer is utilized for dimension reduction and feature extraction. Then, the embedding vector of circRNA i $vc_i \in R^e$ is obtained, where e represents the dimension of circRNA feature vector after embedding. Similar, the embedding vector of disease j $vd_j \in R^e$ is obtained, where e represents the vector dimension of circRNA after embedding. In final, the embedding matrix of circRNA $VC \in R^{m \times e}$ and embedding matrix of disease $VD \in R^{n \times e}$ are obtained by combining the embedding vectors of circRNAs and diseases, respectively. The forms of two matrices are shown as follows:

$$VC = \begin{bmatrix} vc_1 \\ vc_2 \\ vc_3 \\ \cdots \\ vc_M \end{bmatrix} \quad VD = \begin{bmatrix} vd_1 \\ vd_2 \\ vd_3 \\ \cdots \\ vd_N \end{bmatrix} \quad (2)$$

2.3 Improved Graph Convolutional Network

Fig 1(a) and 1(b) show the heterogeneous network and the tree layers structure of neighbor nodes. The adjacent association can be divided into first-order neighbors and multiple-order neighbors. Taking circRNA C_2 as an example, the node C_2 directly links to three disease nodes (D_1, D_2 and D_3). These three disease nodes are first-order neighbors of C_2, which connect to four circRNA nodes (C_1, C_3, C_4,

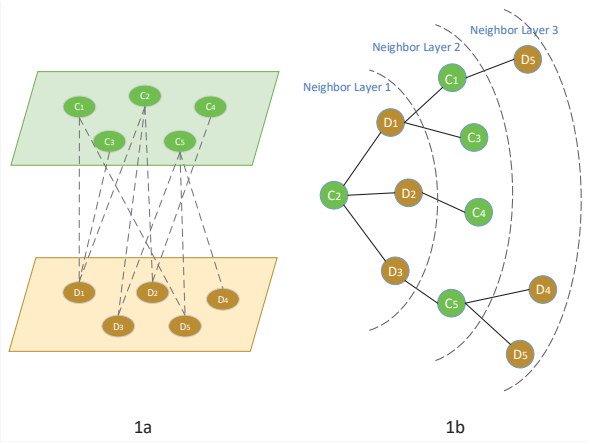

Fig. 1. (a) the heterogeneous network. (b) the tree layers structure of neighbor nodes.

C_5). These circRNA nodes are the second-order neighbors for circRNA C_2. Similarly, the D_4 and D_5 are third-order neighbors of C_2. The origin GCN could not effective extract the deep latent information as the sparse circRNA-disease associations. In order to extract more neighbor information, an improved GCN (IGCN) is designed. The detail of IGCN will be introduced as: *information aggregation from first-order neighbors* and *information aggregation from multiple-order neighbors*.

2.3.1 Information Aggregation from First-order Neighbors

In the heterogeneous network, the neighbor information contained in each connected path between circRNA and disease. The improved GCN aggregates neighbor information in these paths. For example, the path between disease j to circRNA i (E_{ji}) is defined as follows:

$$E_{ji} = \frac{1}{\gamma}\left(W_1 \cdot vd_j + W_2 \cdot (vd_j * vc_i)\right) \tag{3}$$

$$\gamma = \sqrt{|N_c|\,|N_d|} \tag{4}$$

where $W_1 \in R^{e' \times e}$ and $W_2 \in R^{e' \times e}$ denote two weight matrices which can be adjusted in IGCN during training process. e' denotes transformation size. N_c and N_d represent one-order neighbors of circRNAs and diseases, respectively. The symbols $\cdot$ and $*$ represent the matrix multiplication and elements-wise product, respectively. It is different from origin graph convolutional network that an extra term: $vd_j * vc_i$ is added in our model. The origin GCN only depends on the contribution of vd_j, which is not enough due to the sparse association dataset. Here, the added term considers more neighbor information between vd_j and vc_i. Therefore, the improved GCN can improve the prediction performance.

In each embedding propagation layer, in addition to the information transmitted from disease to circRNA, it also takes the influence of circRNAs embedding vector into account:

$$E_{ii} = W_1 \cdot vc_i \tag{5}$$

After adjusted by first layer of IGCN, the circRNAs embedding vectors aggregated all first-order neighbors of circRNA, which is defined as follow:

$$E_i^{(1)} = \text{LeakyReLU}\left(E_{ii} + \sum_{j \in N_c} E_{ji}\right) \tag{6}$$

where $E_i^{(1)}$ denotes refined embedding vector of circRNA through the first propagation layer of IGCN. Similarly, the refined embedding vector of disease $E_j^{(1)}$ is defined as follow:

$$E_j^{(1)} = \text{LeakyReLU}\left(E_{jj} + \sum_{i \in N_d} E_{ij}\right) \tag{7}$$

$$E_{ij} = \frac{1}{\gamma}\left(W_1 \cdot vc_i + W_2 \cdot (vc_i * vd_j)\right) \tag{8}$$

$$E_{jj} = W_1 \cdot vd_j \tag{9}$$

2.3.2 Information Aggregation from Multiple-order Neighbors

After adjusted by first-layer of IGCN, the first-order neighbor information of circRNAs and diseases are obtained. Then, multiple layers of IGCN are stacked to explore deep neighbor information in the heterogeneous network. This deep neighbor information is crucial to explore the latent associations between circRNAs and diseases.

The embedding vectors of circRNAs and diseases obtain the neighbor information by stacking l-th layers of IGCN. As showed by Fig 1b, the multiple-order embedding vector is defined as follows:

$$E_i^{(l)} = \text{LeakyReLU}\left(E_{ii}^{(l)} + \sum_{j \in N_c} E_{ji}^{(l)}\right) \tag{10}$$

where $E_{ii}^{(l)}$ and $E_{ji}^{(l)}$ represent neighbor information propagated among l-th layer in IGCN, which are defined as follows:

$$E_{ji}^{(l)} = \frac{1}{\gamma}\left(W_1^{(l)} \cdot E_j^{(l-1)} + W_2^{(l)} \cdot \left(E_j^{(l-1)} * E_i^{(l-1)}\right)\right) \tag{11}$$

$$E_{ii}^{(l)} = W_1^{(l)} E_i^{(l-1)} \tag{12}$$

where $W_1^{(l)}$ and $W_2^{(l)} \in R^{d_l \times d_{l-1}}$ are the weighted matrices that can transform the embedding vector size in the training process. d_l denotes the transformation size in l-th layer.

In final, the l-th layer of embedding vector is defined as follow:

$$\begin{aligned} E^{(l)} = \text{LeakyReLU}\,&\left((L+I) \cdot E^{(l-1)} \cdot W_1^{(l)}\right. \\ &\left. + L \cdot E^{(l-1)} * E^{(l-1)} \cdot W_2^{(l)}\right) \end{aligned} \tag{13}$$

$$L = D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \tag{14}$$

$$A = \begin{bmatrix} \mathbf{0} & Y \\ Y^T & \mathbf{0} \end{bmatrix} \tag{15}$$

where $\mathbf{0}$ represents all-zeros matrix. D denotes a diagonal matrix that adds up the values of each row in Y. $E^{(l)}$ denotes circRNAs and diseases embedding vectors after passing through l-th layer of IGCN. I denotes identity matrix. $E^{(0)}$ denotes the matrix consisting of the initial embedding matrices of circRNA and disease which is defined as follows:

$$E^{(0)} = \begin{bmatrix} VC \\ VD \end{bmatrix} \tag{16}$$

2.4 Prediction and Optimization

2.4.1 Multi-layer Perceptron

After adjusted by the propagation layers of IGCN, the feature vectors of circRNA and disease are connected. Then, the multi-layer perceptron (MLP) is utilized to predict circRNA-disease associations. The structure of MLP is designed by optimizing the best one among repeated experiments, which is defined as:

$$\hat{y}_{ij} = W^l \sigma^l \Big(W^{l-1} \sigma^{l-1} \left(\cdots \left(W^2 \sigma \left(W^1 E + b^1 \right) + b^2 \right) \cdots \right) + b^{l-1} \Big) + b^l \tag{17}$$

where W^l and b^l represent the weight matrix and bias vector in l-th layer, respectively. σ^l represents the activation function of l-th layer. E denotes the concatenated embedding vector of circRNA i and disease j in l-th layer. In here, the ReLU function is selected to avoid the oversaturation [44]. In order to limit the output of the model in range of 0 to 1, the sigmoid function is used as activation function of the final layer.

2.4.2 Loss Function

In order to make the prediction scores of positive samples higher than that of negative samples [45], the loss function is defined as follows:

$$Loss = \sum_{\substack{(i,j)\in P \\ (i',j')\in N}} -\ln \operatorname{sigmoid}\left(\hat{y}_{ij} - \hat{y}_{i'j'}\right) + \lambda \|\alpha\|_2^2 \tag{18}$$

where P and N denote positive samples and negative samples, respectively. α denotes all trainable parameters inlucding W_1^l, W_2^l and W^l. λ is used to control the regularization strength to avoid overfitting. The Adam optimizer is utilized to update the parameter of the model.

2.5 Negative Sampling Method

The selection of negative samples may influence the prediction performance of the model as negative samples contain a lot of unlabeled samples. Based on the hypothesis that the similar circRNAs tend to associate with similar diseases, a negative sampling method is employed to overcome this limitation. First, two kinds of circRNAs similarities are calculated based on the known circRNA-disease associations and circRNA expression profile. Then, two kinds of similarities are integrated to enhance the reliability of final similarity. Further, the negative samples are selected based on the final similarity.

2.5.1 CircRNA Gaussian Interaction Profile Kernel Similarity

The Gaussian interaction profile (GIP) kernel function is used to compute circRNA similarity based on the known association matrix Y. The binary vector $c(i)$ represents the interaction profiles of circRNA i, which is the i-th row vector of matrix Y. Thus, the Gaussian interaction profile kernel similarity for circRNAs can be calculated as follows:

$$GIP\_circ\,(i,j) = \exp\left(-\theta_c \|c(i) - c(j)\|^2\right) \tag{19}$$

$$\theta_c = 1/\left(\frac{1}{m}\sum_{i=1}^{m} \|c(i)\|^2\right) \tag{20}$$

where θ_c is the width parameter of the kernel.

2.5.2 CircRNA Expression Profile Similarity

Considering that some circRNAs are only related with one disease, the GIP similarity of circRNA is sensitive to known circRNA-disease associations. The circRNAs expression profile similarity is calculated by using Pearson correlation coefficient. Given two circRNAs i and j, the expression profiles of two circRNAs are represented by $\mathrm{circexp}(i) = (\alpha_1, \alpha_2, \cdots, \alpha_N)$ and $\mathrm{circexp}(j) = (\beta_1, \beta_2, \cdots, \beta_N)$, respectively. The expression profile similarity between circRNAs i and j is defined as follows:

$$EX\_circ(i,j) = \frac{\sum_{p=1}^{N} (\alpha_p - \bar{\alpha})(\beta_p - \bar{\beta})}{\sqrt{\sum_{p=1}^{N} (\alpha_p - \bar{\alpha})^2 \sum_{p=1}^{N} (\beta_p - \bar{\beta})^2}} \tag{21}$$

where α_p and β_p denote the expression value of circRNA i and j, respectively. N represents the total number of dimensions in expression profile.

2.5.3 Similarity Fusion for CircRNAs

In order to enhance the reliability of circRNA similrity, the GIP similarity and expression similarity are integrated to obtain the final circRNA similarity. It is defined as follows:

$$CircSim(i,j) = \begin{cases} GIP\_circ(i,j), & \text{if c(i) and c(j)} \neq 0 \\ EX\_circ\,(i,j)\,, & \text{if c(i) or c(j)=0} \end{cases} \tag{22}$$

It means that if circRNA i or circRNA j are not related with any diseases, the circRNA expression profile similarity will be used instead of GIP similarity.

2.5.4 Negative Sampling Method

Since the value 0 in matrix Y represents that the associations remain to be validated by experiment, the association matrix Y contains lots of unknown associations between circRNAs and diseases. Therefore, it is necessary to select some negative samples for training. While it may bring noise data if we choose negative samples randomly. Then, based on the hypothesis that the similar circRNA tend to be associated with the similar diseases, a new negative sampling method is developed based on circRNA similarity. Specifically, the similarity matrix is constructed based on the fusion similarity of circRNAs. Then, the top-k most similar circRNAs are selected based on similarity matrix. The similar circRNAs related diseases will not be chosen

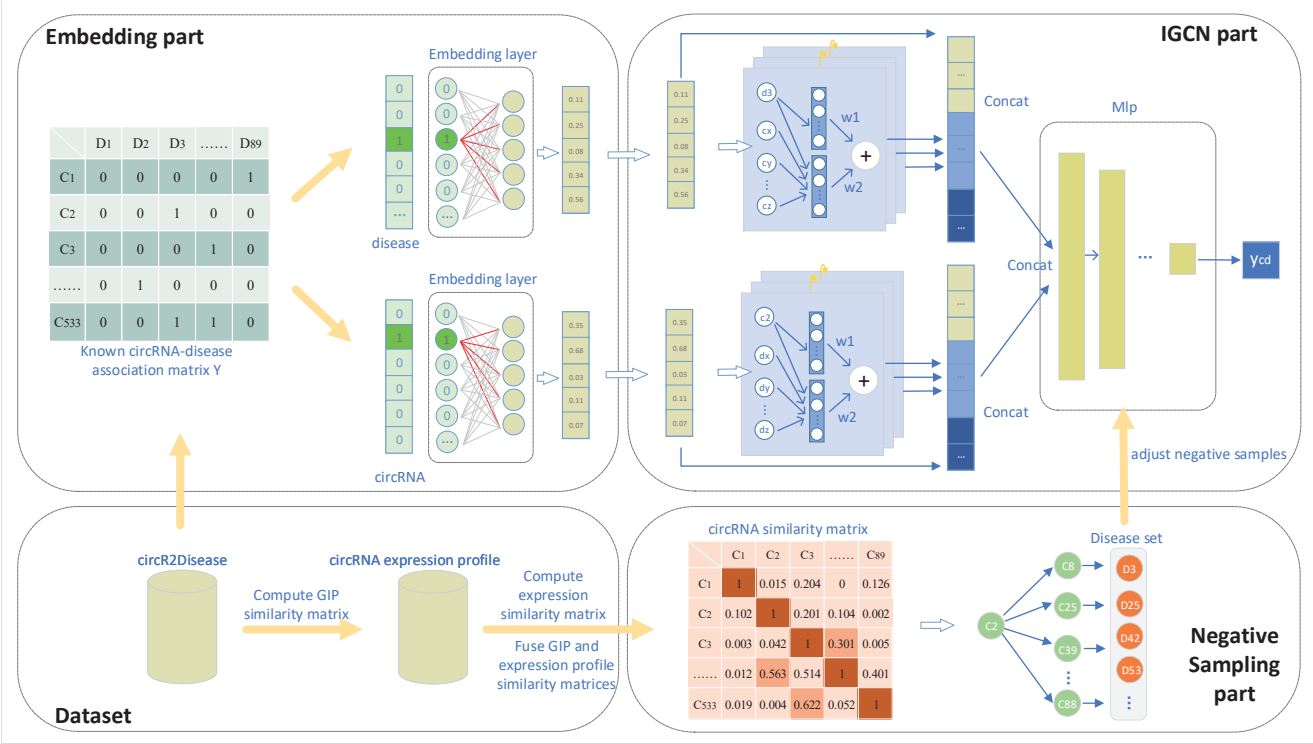

Fig. 2. The architecture of IGNSCDA

as the negative samples. For example, given circRNA i is related with disease j. Then, the top-k circRNAs of circRNA i are selected as a set: $Sc = \{c_1, c_2, \cdots, c_k\}$. Further, the related diseases of all circRNAs in Sc are chosen as a set: $Sd = \{d_1, d_2, \cdots, d_h\}$, where h denotes the number of related diseases. When choosing the negative samples for training, the disease in S_d will not be chosen.

The whole architecture of IGNSCDA is shown in Fig 2, which includes three main part: embedding part, IGCN part and negative sampling part. First, all known association data and circRNA expression profile are downloaded in circR2Disease dataset [40] and exoRBase [41] database, respectively. Then, after processed in embedding part, all vectors of circRNAs and diseases are inputted into IGCN part. Adjusted by multiple IGCN layers, the embedding vectors of circRNAs and diseases contain neighbor information. Further, the multi-layer perceptron is used to predict circRNA-disease associations based on the feature vectors. In addition, after adjusted by negative sampling part, the effect brought by false negative samples is reduced.

3 Experiments and Results

3.1 Evaluation Metrics

In order to evaluate the performance, the 5-fold cross validation is conducted in the experiment [46]. In the 5-fold cross validation, all samples are divided into five subsets and each subset is removed in turn as test samples, while the rest sets are regarded as training samples for model learning. Then, the scores of unknown samples and test samples are predicted and sorted in descending order. The true positive rate (TPR) and false positive rate (FPR) are calculated and the ROC curve is plotted by varying rank thresholds. The area under the curve (AUC) is computed to evaluate the performance of model. The higher AUC value is, the better this model is. In addition, we also compute accuracy (Acc), recall (Rec), precision (Pre), F1-score (F1) as the evaluation criterion. The definitions of evaluation criteria are listed as follows:

$$FPR = \frac{FP}{FP + TN} \tag{23}$$

$$TPR = \frac{TP}{TP + FN} \tag{24}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{25}$$

$$Rec = \frac{TP}{TP + FN} \tag{26}$$

$$Pre = \frac{TP}{TP + FP} \tag{27}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{28}$$

where TP indicates the number of true positive samples. FP indicates the number of false positive samples. TN indicates the number of true negative samples. FN indicates the number of false negative samples.

3.2 Five-fold Cross Validation

In order to validate the effectiveness of our model, we compared IGNSCDA with other four state-of-the-art methods (DWNN-RLS [37], KATZHCDA [29], RWR [32], GCNCDA [39]). As shown in Fig 3, IGNSCDA achieves AUC of 0.8291

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2021.3111607, IEEE/ACM Transactions on Computational Biology and Bioinformatics

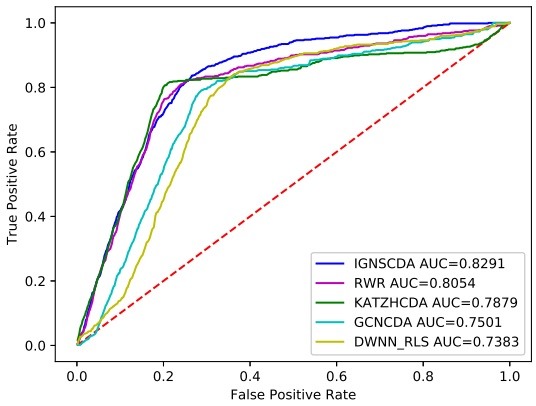

Fig. 3. The performance comparison of IGNSCDA, RWR, KATZHCDA, GCNCDA and DWNN-RLS in term of AUC value based on 5-fold cross validation

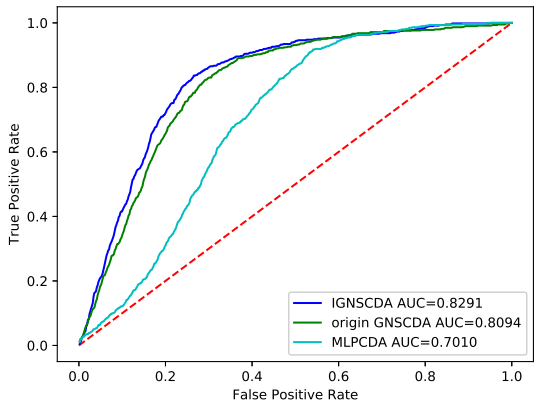

Fig. 4. The performance comparison of IGNSCDA, origin-GNSCDA and MLPCDA in term of AUC value based on 5-fold cross validation

TABLE 1
The performance of five models based on 5-fold cross validation

Methods	Acc	Rec	Pre	F1
IGNSCDA	**0.4958**	**0.8310**	**0.0057**	**0.0111**
DWNN-RLS	0.4953	0.7412	0.0042	0.0083
KATZHCDA	0.4956	0.7902	0.0057	0.0111
GCNCDA	0.4954	0.7530	0.0044	0.0086
RWR	0.4956	0.8075	0.0055	0.0103

TABLE 2
The effect of parameters (epoch and negative samples) on model performance

Epoch/Neg	10	15	20	25	30
100	0.6881	0.6997	0.7028	0.7162	0.7359
150	0.7246	0.7532	0.7649	0.7884	0.7745
200	0.7857	0.8029	0.8097	0.8085	0.7988
250	0.8092	0.8127	**0.8291**	0.8128	0.8104
300	0.8031	0.8010	0.8223	0.8115	0.8009

in the 5-fold cross validation, which is better than other four methods (DWNN-RLS: 0.7383, GCNCDA: 0.7501, KATZHCDA: 0.7879, RWR: 0.8054). In addition, other experiments results (Accuracy, Recall, Precision, F1) are listed in Table 1. The IGNSCDA achieves 0.4958 in accuracy, 0.8310 in Recall, 0.0057 in Precision, 0.0111 in F1-value. The result shows our model outperforms other four state-of-the-art approaches.

3.3 The effect of Propagation method in IGCN

In order to verify how the propagating method of IGCN effect the prediction performance, we remove the element wise product representation of Equation 3 and set it as origin GCN model named origin-GNSCDA. In addition, we remove IGCN part only leaving the multi-layer perceptron part and set it as the variant model named MLPCDA. Then, we compare IGNSCDA with origin-GNSCDA and MLPCDA. The results are shown in Fig 4, we have following findings:

• The origin-GNSCDA outperforms than MLPCDA, which illustrates that GCN could extract the latent features of circRNAs and diseases in the heterogeneous network and achieve the base promising performance.

• IGNSCDA is superior to origin-GNSCDA and MLPCDA, which demonstrates the element-wise product $(vd_j * vc_i)$ could obtain more neighbor information of circRNAs and diseases into the model than other variants .

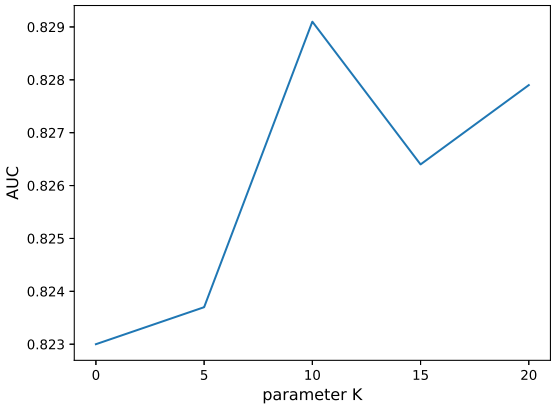

Fig. 5. The AUC value of the parameter k

3.4 Parameters Setting

For IGCN part, the Xavier initializer [47] is used to initialize the model parameters and the Adam [48] is used to train model for 250 epochs. The learning rate is set as 0.001 and L2 normalization is set as 0.0001. The ratio of node dropout is set as 0.1. In addition, it is shown in Fig 5 that the whole performance will be best when the parameter top k is set as 10, which means top-10 most similar circRNAs related disease will not be selected as the negative samples. Further, the most important two parameters in GCN are the number of embedding propagation layers and the embedding dimension of GCN. In order to analyze the effect of the number of embedding propagation layers, we test different propagation layers from 1 to 3. The result is shown in Fig 6. It can be concluded that one layer is enough for this dataset and too many layers may bring new noise. In addition, we

also test the effect of different dimensions of embedding vector. It can be found from Fig 7 that the result is the best when the dimension of embedding vector is set to 32.

In MLP, we design a four-layers structure (256-128-64-1). In order to test the effect of the number of negative samples and training epochs, we change the number of negative samples and training epochs from 10 to 30 and 100 to 300, respectively. The result is shown in Fig 8 and the detail result is listed in table 2. It can be observed that the performance is the best when the number of negative samples and training epochs are set to 20 and 250, respectively.

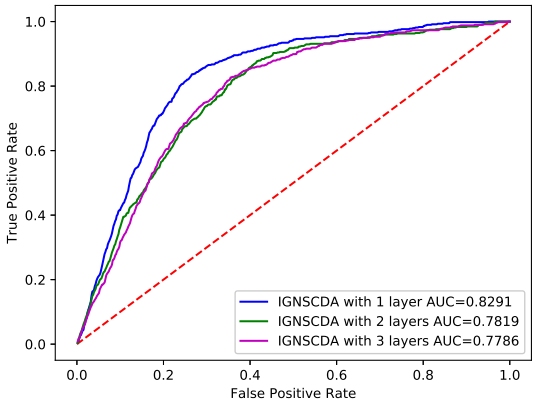

Fig. 6. Average AUC value on different number of propagation layers in IGNSCDA

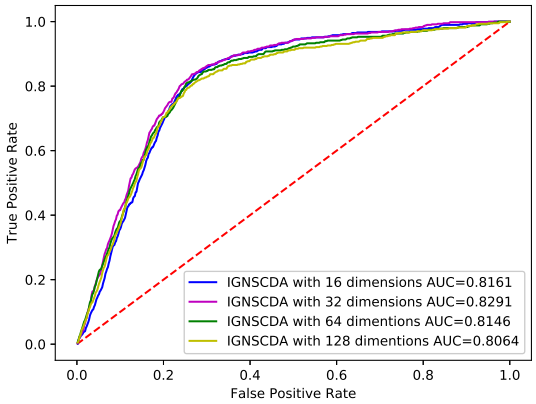

Fig. 7. Average AUC value on different dimensions of feature vectors in IGNSCDA

3.5 Case Study

To further verify the predictive ability of IGNSCDA, we implement the case study on bladder cancer. Bladder cancer is a very common disease and associated with substantial morbidity and mortality [49]. Therefore, identifying circRNAs related with bladder cancer may contribute to diagnose and treatment. Specifically, we first train the proposed model using all known associations. Then, the unknown interactions are predicted by trained model. The top-10 predicted circRNA of bladder are listed in Table 3.

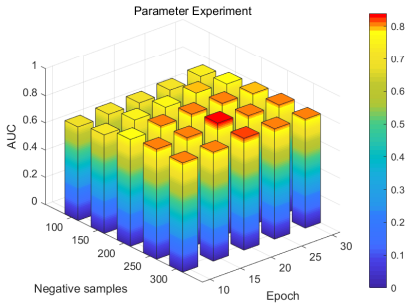

Fig. 8. The influence of parameters (epoch and negative samples) on IGNSCDA

TABLE 3
The top 10 predicted results of bladder-related circRNAs based on IGNSCDA

Rank	CircRNA	Evidence
1	Mmu_circ_0000254	Unknown
2	CircSLC8A1-1	PMID:31228937
3	Hsa_circ_0037911	Unknown
4	Hsa_circ_0002161	Unknown
5	Circ-ZKSCAN1	PMID:33046073
6	Hsa_circ_0067531	Unknown
7	CircZNF139	PMID:32454461
8	Hsa_circ_0001073	PMID:31101108
9	Hsa_circ_0000144	PMID:30098434
10	Hsa_circ_0051239	Unknown

There are 5 of top 10 predicted candidate circRNAs (CircSLC8A1-1, Circ-ZKSCAN1, CircZNF139, Hsa_circ_0001073, Hsa_circ_0000144) that have been confirmed to be associated with bladder cancer by recent studies. Recent study has shown that CircSLC8A1-1 ranked at top 2 is associated with pathological stage and histological grade of bladder cancer [50]. The evidence proves that Circ-ZKSCAN1 ranked at top 5 can suppress the progression of bladder cancer through miR-1178-3p/p21 axis [51]. It has been demonstrated that the CircZNF139 ranked at top 7 promotes cell proliferation, migration and invasion of bladder cancer [52]. The hsa_circ_0001073 ranked at top 8 has been proved that it is a tumor suppressor to inhabit proliferation and metastasis of bladder cancer cell [53]. It has been proved that hsa_circ_0000144 ranked at top 9 exerts oncogenic roles in bladder cancer through suppressing miR-217 to facilitate RUNX2 expression [54]. Although, some unknown associations are not supported by the literature, it deserves biologists to further validate it by using experimental methods.

4 Conclusion

Increasing evidences have been shown that circRNAs have close associations with many diseases [55]. In this paper, we propose a computational model (IGNSCDA) to predict associations between circRNAs and diseases. Firstly, the heterogeneous network is constructed based on the association data. Then, the improved graph convolutional network is

designed to obtain the feature vectors of circRNAs and diseases. Further, the multi-layer perceptron is used to predict latent associations based on the feature vectors of circRNAs and diseases. In addition, a negative sampling method is designed to reduce the effect of unlabeled positive samples based on the similarity of circRNA. In order to verify the effectiveness of our model, we compare IGNSCDA with other four state-of-the-art methods (DWNN-RLS, RWR, KATZHCDA, GCNCDA). The results demonstrate that our model is effective to predict circRNA-disease associations. In future research, we will integrate more different kinds of biological data to improve the prediction performance and explore more methods that can extract latent features containing more neighbor information in the field of graph neural network [56], [57], [58].

FUNDING

This work was partially supported by the National Natural Science Foundation of China (Nos. 62072124, 61963004 and 61972185), the Natural Science Foundation of Guangxi (Nos. 2021GXNSFAA075041 and 2018GXNSFBA281193), the Science and Technology Base and Talent Special Project of Guangxi (No. AD20159044), the Natural Science Foundation of Yunnan Province of China (No. 2019FA024), the Hunan Provincial Science and Technology Program (No. 2018WK4001).

REFERENCES

[1] J. Salzman, R. E. Chen, M. N. Olsen, P. L. Wang, and P. O. Brown, "Cell-type specific features of circular rna expression," *PLoS Genet*, vol. 9, no. 9, p. e1003777, 2013.

[2] W. R. Jeck and N. E. Sharpless, "Detecting and characterizing circular rnas," *Nature biotechnology*, vol. 32, no. 5, pp. 453–461, 2014.

[3] M. Danan, S. Schwartz, S. Edelheit, and R. Sorek, "Transcriptome-wide discovery of circular rnas in archaea," *Nucleic acids research*, vol. 40, no. 7, pp. 3131–3142, 2012.

[4] T. Diener, "Potato spindle tuber virus: Iv. a replicating, low molecular weight rna," *Virology*, vol. 45, no. 2, pp. 411–428, 1971.

[5] M.-T. Hsu and M. Coca-Prados, "Electron microscopic evidence for the circular form of rna in the cytoplasm of eukaryotic cells," *Nature*, vol. 280, no. 5720, pp. 339–340, 1979.

[6] F. Wang, A. J. Nazarali, and S. Ji, "Circular rnas as potential biomarkers for cancer diagnosis and therapy," *American journal of cancer research*, vol. 6, no. 6, p. 1167, 2016.

[7] M. Zeng, F. Zhang, W. FangXiang, M. Li, J. Wang *et al.*, "Deep matrix factorization improves prediction of human circrna-disease associations," *IEEE Journal of Biomedical and Health Informatics*, 2020.

[8] L.-L. Chen, "The biogenesis and emerging roles of circular rnas," *Nature reviews Molecular cell biology*, vol. 17, no. 4, pp. 205–211, 2016.

[9] T. B. Hansen, T. I. Jensen, B. H. Clausen, J. B. Bramsen, B. Finsen, C. K. Damgaard, and J. Kjems, "Natural rna circles function as efficient microrna sponges," *Nature*, vol. 495, no. 7441, pp. 384–388, 2013.

[10] J. T. Granados-Riveron and G. Aquino-Jarquin, "The complexity of the translation ability of circrnas," *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, vol. 1859, no. 10, pp. 1245–1251, 2016.

[11] Z. Li, C. Huang, C. Bao, L. Chen, M. Lin, X. Wang, G. Zhong, B. Yu, W. Hu, L. Dai *et al.*, "Exon-intron circular rnas regulate transcription in the nucleus," *Nature structural & molecular biology*, vol. 22, no. 3, p. 256, 2015.

[12] S. Xia, J. Feng, K. Chen, Y. Ma, J. Gong, F. Cai, Y. Jin, Y. Gao, L. Xia, H. Chang *et al.*, "Cscd: a database for cancer-specific circular rnas," *Nucleic acids research*, vol. 46, no. D1, pp. D925–D929, 2018.

[13] J. N. Vo, M. Cieslik, Y. Zhang, S. Shukla, L. Xiao, Y. Zhang, Y.-M. Wu, S. M. Dhanasekaran, C. G. Engelke, X. Cao *et al.*, "The landscape of circular rna in cancer," *Cell*, vol. 176, no. 4, pp. 869–881, 2019.

[14] Z. Zhang, T. Yang, and J. Xiao, "Circular rnas: promising biomarkers for human diseases," *EBioMedicine*, vol. 34, pp. 267–274, 2018.

[15] S. Qin, Y. Zhao, G. Lim, H. Lin, X. Zhang, and X. Zhang, "Circular rna pvt1 acts as a competing endogenous rna for mir-497 in promoting non-small cell lung cancer progression," *Biomedicine & pharmacotherapy*, vol. 111, pp. 244–250, 2019.

[16] W. Lan, J. Wang, M. Li, W. Peng, and F.-X. Wu, "Computational approaches for prioritizing candidate disease genes based on ppi networks," *Tsinghua Science and Technology*, vol. 20, no. 5, pp. 500–512, 2015.

[17] J. Greene, A.-M. Baird, L. Brady, M. Lim, S. G. Gray, R. McDermott, and S. P. Finn, "Circular rnas: biogenesis, function and role in human diseases," *Frontiers in molecular biosciences*, vol. 4, p. 38, 2017.

[18] W. Lukiw, "Circular rna (circrna) in alzheimer's disease (ad)," *Frontiers in genetics*, vol. 4, p. 307, 2013.

[19] P. Li, S. Chen, H. Chen, X. Mo, T. Li, Y. Shao, B. Xiao, and J. Guo, "Using circular rna as a novel type of biomarker in the screening of gastric cancer," *Clinica chimica acta*, vol. 444, pp. 132–136, 2015.

[20] M. Qin, G. Liu, X. Huo, X. Tao, X. Sun, Z. Ge, J. Yang, J. Fan, L. Liu, and W. Qin, "Hsa_circ_0001649: a circular rna and potential novel biomarker for hepatocellular carcinoma," *Cancer Biomarkers*, vol. 16, no. 1, pp. 161–169, 2016.

[21] L. Yu, X. Gong, L. Sun, Q. Zhou, B. Lu, and L. Zhu, "The circular rna cdr1as act as an oncogene in hepatocellular carcinoma through targeting mir-7 expression," *PloS one*, vol. 11, no. 7, p. e0158347, 2016.

[22] W. Lan, M. Zhu, Q. Chen, B. Chen, J. Liu, M. Li, and Y.-p. P. Chen, "Circr2cancer: a manually curated database of associations between circrnas and cancers," *Database The Journal of Biological Databases and Curation*, vol. 2020, no. 2020, p. baaa085, 2020.

[23] W. Lan, L. Huang, D. Lai, and Q. Chen, "Identifying interactions between long noncoding rnas and diseases based on computational methods," *Computational Systems Biology*, pp. 205–221, 2018.

[24] W. Lan, J. Wang, M. Lin, J. Liu, F.-X. Wu, and P. Yi, "Predicting microrna-disease associations based on improved microrna and disease similarities," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 6, pp. 1774–1782, 2018.

[25] C. Lu, M. Zeng, F.-X. Wu, M. Li, and J. Wang, "Improving circrna-disease association prediction by sequence and ontology representations with convolutional and recurrent neural networks," *Bioinformatics*, 2020.

[26] E. Ge, Y. Yang, M. Gang, C. Fan, and Q. Zhao, "Predicting human disease-associated circrnas based on locality-constrained linear coding," *Genomics*, vol. 112, no. 2, pp. 1335–1342, 2020.

[27] Q. Xiao, J. Luo, and J. Dai, "Computational prediction of human disease-associated circrnas based on manifold regularization learning framework," *IEEE journal of biomedical and health informatics*, vol. 23, no. 6, pp. 2661–2669, 2019.

[28] X. Lei and W. Zhang, "Brwsp: predicting circrna-disease associations based on biased random walk to search paths on a multiple heterogeneous network," *Complexity*, vol. 2019, 2019.

[29] C. Fan, X. Lei, and F.-X. Wu, "Prediction of circrna-disease associations using katz model based on heterogeneous networks," *International journal of biological sciences*, vol. 14, no. 14, p. 1950, 2018.

[30] L. Deng, W. Zhang, Y. Shi, and Y. Tang, "Fusion of multiple heterogeneous networks for predicting circrna-disease associations," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.

[31] Q. Zhao, Y. Yang, G. Ren, E. Ge, and C. Fan, "Integrating bipartite network projection and katz measure to identify novel circrna-disease associations," *IEEE transactions on nanobioscience*, vol. 18, no. 4, pp. 578–584, 2019.

[32] H. Vural, M. Kaya, and R. Alhajj, "A model based on random walk with restart to predict circrna-disease associations on heterogeneous network," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2019, pp. 929–932.

[33] Y. Ding, B. Chen, X. Lei, B. Liao, and F.-X. Wu, "Predicting novel circrna-disease associations based on random walk and logistic regression model," *Computational Biology and Chemistry*, vol. 87, p. 107287, 2020.

[34] X. Lei and C. Bian, "Integrating random walk with restart and k-nearest neighbor to identify novel circrna-disease association," *Scientific reports*, vol. 10, no. 1, pp. 1–9, 2020.
[35] M. Li, M. Liu, Y. Bin, and J. Xia, "Prediction of circrna-disease associations based on inductive matrix completion," *BMC medical genomics*, vol. 13, pp. 1–13, 2020.
[36] H. Wei and B. Liu, "icircda-mf: identification of circrna-disease associations based on matrix factorization," *Briefings in bioinformatics*, vol. 21, no. 4, pp. 1356–1367, 2020.
[37] C. Yan, J. Wang, and F.-X. Wu, "Dwnn-rls: regularized least squares method for predicting circrna-disease associations," *BMC bioinformatics*, vol. 19, no. 19, pp. 73–81, 2018.
[38] L. Wang, Z.-H. You, Y.-A. Huang, D.-S. Huang, and K. C. Chan, "An efficient approach based on multi-sources information to predict circrna–disease associations using deep convolutional neural network," *Bioinformatics*, vol. 36, no. 13, pp. 4038–4046, 2020.
[39] 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 Computational Biology*, vol. 16, no. 5, p. e1007568, 2020.
[40] 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.
[41] S. Li, Y. Li, B. Chen, J. Zhao, S. Yu, Y. Tang, Q. Zheng, Y. Li, P. Wang, X. He *et al.*, "exorbase: a database of circrna, lncrna and mrna in human blood exosomes," *Nucleic acids research*, vol. 46, no. D1, pp. D106–D112, 2018.
[42] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 173–182.
[43] H. Zhao, Z. Lu, and P. Poupart, "Self-adaptive hierarchical sentence model," *arXiv preprint arXiv:1504.05070*, 2015.
[44] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 315–323.
[45] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua, "Neural graph collaborative filtering," in *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*, 2019, pp. 165–174.
[46] Y. Jiao and P. Du, "Performance measures in evaluating machine learning based bioinformatics predictors for classifications," *Quantitative Biology*, vol. 4, no. 4, pp. 320–330, 2016.
[47] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.
[48] K. Da, "A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[49] O. Sanli, J. Dobruch, M. A. Knowles, M. Burger, M. Alemozaffar, M. E. Nielsen, and Y. Lotan, "Bladder cancer," *Nature reviews Disease primers*, vol. 3, no. 1, pp. 1–19, 2017.
[50] Q. Lu, T. Liu, H. Feng, R. Yang, X. Zhao, W. Chen, B. Jiang, H. Qin, X. Guo, M. Liu *et al.*, "Circular rna circslc8a1 acts as a sponge of mir-130b/mir-494 in suppressing bladder cancer progression via regulating pten," *Molecular cancer*, vol. 18, no. 1, pp. 1–13, 2019.
[51] J. Bi, H. Liu, W. Dong, W. Xie, Q. He, Z. Cai, J. Huang, and T. Lin, "Circular rna circ-zkscan1 inhibits bladder cancer progression through mir-1178-3p/p21 axis and acts as a prognostic factor of recurrence," *Molecular cancer*, vol. 18, no. 1, pp. 1–14, 2019.
[52] J. Yao, K. Qian, C. Chen, X. Liu, D. Yu, X. Yan, T. Liu, and S. Li, "Znf139/circznf139 promotes cell proliferation, migration and invasion via activation of pi3k/akt pathway in bladder cancer," *Aging (Albany NY)*, vol. 12, no. 10, p. 9915, 2020.
[53] W. Dong, J. Bi, H. Liu, D. Yan, Q. He, Q. Zhou, Q. Wang, R. Xie, Y. Su, M. Yang *et al.*, "Circular rna acvr2a suppresses bladder cancer cells proliferation and metastasis through mir-626/eya4 axis," *Molecular cancer*, vol. 18, no. 1, pp. 1–16, 2019.
[54] W. Huang, Y. Lu, F. Wang, X. Huang, and Z. Yu, "Downregulation of circular rna hsa_circ_0000144 inhibits bladder cancer progression via stimulating mir-217 and suppressing runx2 expression," *Gene*, vol. 678, pp. 337–342, 2018.
[55] W. Lan, M. Li, Z. Kaijie, L. Jin, F.-X. Wu, P. Yi, and W. Jianxin, "Ldap: a web server for lncrna-disease association prediction," *Bioinformatics*, no. 3, pp. 458–460, 2017.
[56] Q. Chen, D. Lai, W. Lan, X. Wu, B. Chen, J. Liu, Y.-P. P. Chen, and J. Wang, "Ildmsf: Inferring associations between long non-coding rna and disease based on multi-similarity fusion," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 18, no. 3, pp. 1106–1112, 2021.
[57] W. Lan, J. Wang, M. Li, J. Liu, Y. Li, F.-X. Wu, and Y. Pan, "Predicting drugtarget interaction using positive-unlabeled learning," *Neurocomputing*, vol. 206, no. sep.19, pp. 50–57, 2016.
[58] W. Lan, D. Lai, Q. Chen, X. Wu, B. Chen, J. Liu, Y.-P. P. Chen, and J. Wang, "Ldicdl: Lncrna-disease association identification based on collaborative deep learning," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 99, no. 99, pp. 1–1, 2020.

Wei Lan received the PhD degree in computer science from the Central South University in 2016. He is now a lecturer at Guangxi University, China. His research interests include bioinformatics and Machine learning especially in non-coding RNA and disease gene.

Yi Dong pursues the master degree in School of Computer, Electronic and Information, Guangxi University, Nanning, China. His current research interests include data mining, computational biology and machine learning.

Qingfeng Chen received the BSc and MSc degrees in mathematics from Guangxi Normal University, China, in 1995 and 1998, respectively, and the PhD degree in computer science from the University of Technology Sydney in September 2004. He is now a Professor at Guangxi University, China and a hundred talent program of Guangxi. His research interests include bioinformatics, data mining, and artificial intelligence. He has published 40 refereed papers and two monographs by Springer, including IEEE Transaction on Knowledge and Data Engineering, Data Mining and Knowledge Discovery. He has been serving as an associate editor for Engineering Letters, and is invited to be a guest editor of two special issues for Current Protein & Peptide Science, and co-chairs for several international conferences.

Jin Liu is a Association Professor in School of Computer Science, Central South University, Changsha, Hunan, P.R. China. His current research interests include medical image analysis, machine learning and pattern recognition.

Jianxin Wang received the BEng and MEng degrees in computer engineering from Central South University, China, in 1992 and 1996, respectively, and the PhD degree in computer science from Central South University, China, in 2001. He is the dean of and a professor in School of Computer Science, Central South University, Changsha, Hunan, P.R. China. His current research interests include algorithm analysis and optimization, parameterized algorithm, Bioinformatics and computer network. He is a senior member of the IEEE.

Yi-Ping Phoebe Chen received the BInfTech degree and the PhD degree in computer science from the University of Queensland. Prof Chen is currently Professor and Chair of Department of Computer Science and Computer Engineering, La Trobe University. She is the steering committee chair of Asia Pacific Bioinformatics Conference (founder) and Multimedia Modelling. She has co-authored over 150 research papers with many published in top journals and conferences such as IEEE Transactions on Biomedical Engineering, Nucleic Acids Research, and ACM Transactions. She has been on the program committees of over 100 international conferences, including top ranking conferences such as ICDE, ICPR, ISMB, CIKM etc. Phoebe is a senior member of IEEE, member of ACM.

Shirui Pan received the Ph.D. degree in computer science from the University of Technology Sydney (UTS), Ultimo, NSW, Australia. He was a Lecturer with the School of Software, UTS. He is currently a Senior Lecturer with the Faculty of Information Technology, Monash University, Clayton, VIC, Australia. He has published over 60 research articles in top-tier journals and conferences, including the IEEE Transactions on Neural Networks and Learning Systems (TNNLS), the IEEE Transactions on Knowledge and Data Engineering (TKDE), the IEEE Transactions on Cybernetics (TCYB), the IEEE International Conference on Data Engineering (ICDE), the AAAI Conference on Artificial Intelligence (AAAI), the International Joint Conferences on Artificial Intelligence (IJCAI), and the IEEE International Conference on Data Mining (ICDM). His research interests include data mining and machine learning.