Systems biology

An Efficient Approach based on Multi-sources Information to Predict CircRNA-disease Associations Using Deep Convoltional Neural Network

Lei Wang[1,†], Zhu-Hong You[1,†,*], Yu-An Huang[2], De-Shuang Huang[3] and Keith C.C. Chan[2]

[1] Xinjiang Technical Institutes of Physics and Chemistry, Chinese Academy of Science, Urumqi 830011, China;

[2] Department of Computing, Hong Kong Polytechnic University, Hong Kong

[3] Institute of Machine Learning and Systems Biology, School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China

ABSTRACT

Motivation: Emerging evidence indicates that circular RNA (circRNA) plays a crucial role in human disease. Using circRNA as biomarker gives rise to a new perspective regarding our diagnosing of diseases and understanding of disease pathogenesis. However, detection of circRNA-disease associations by biological experiments alone is often blind, limited to small-scale, high-cost and time-consuming. Therefore, there is an urgent need for reliable computational methods to rapidly infer the potential circRNA-disease associations on a large scale and to provide the most promising candidates for biological experiments.

Results: In this paper, we propose an efficient computational method based on multi-source information combined with deep convolutional neural network to predict circRNA-disease associations. The method first fuses multi-source information including disease semantic similarity, disease Gaussian interaction profile kernel similarity, and circRNA Gaussian interaction profile kernel similarity, and then extracts its hidden deep feature through the convolutional neural network, and finally sends them to the extreme learning machine classifier for prediction. The five-fold cross-validation results show that the proposed method achieves 87.21% prediction accuracy with 88.50% sensitivity at the AUC of 86.67% on the CIRCR2Disease dataset. In comparison with the state-of-the-art SVM classifier and other feature extraction methods on the same dataset, the proposed model achieves the best results. In addition, we also obtained experimental support for prediction results by searching published literature. As a result, 7 of the top 15 circRNA-disease pairs with the highest scores were confirmed by literature. These results demonstrate that the proposed model is a suitable method for predicting circRNA-disease associations and can provide reliable candidates for biological experiments. The source code and datasets explored in this work are available at https://github.com/look0012/circRNA-Disease-association.

Contact: zhuhongyou@ms.xjb.ac.cn

Supplementary information: Supplementary data are available at Bioinformatics online.

1. INTRODUCTION

Circular RNA (circRNA) is a kind of special single-stranded circular endogenous non-coding RNA formed by back splicing (Danan, et al., 2012; Nigro, et al., 1991; Salzman, et al., 2013). It is widely distributed in organisms with certain biological stability, spatiotemporal specificity and evolutionary conservation. As early as 1976, circRNA was first found in plant viruses through electron microscopy (Sanger, et al., 1976). However, it was considered as a by-product of "shear noise" or abnormal shear at that time, which did not have important functions and had not been received corresponding attention. Until 2012, Salzman *et al.* detected a large number of highly expressed circular transcribed RNA molecules in bone marrow of children with acute lymphoblastic leukemia, cervical cancer Hela cell line and human embryonic stem cells (H9) using second-generation sequencing technology. It was first confirmed that circRNAs are widely present in human embryonic stem cells and malignant tissue cells, which really caught the attention of the scientific community. Since then, more and more circRNAs have been widely recognized and discovered in many eukaryotes such as humans, mice and nematodes (Chen, et al., 2017; Qin, et al., 2016; Zhou and Xu, 2017; Zhu, et al., 2016). Memczak *et al.* identified 1950 human circRNAs, 1903 mouse circRNAs (of which 81 were identical to human circRNAs) and 724 nematode circRNAs by RNA-seq data binding to the human leukocyte database (Memczak, et al., 2013). Jeck *et al.* detected more than 25000 kinds of circRNA in human fibroblasts (Jeck, et al., 2013). Bahn *et al.* detected more than 400 circRNAs in saliva using second-generation sequencing technology (Bahn, et al., 2015).

With the development of research, more and more circRNA have been proved to have specific biological functions, such as gene expression regulation, cell communication and protein translation by adsorbing microRNAs or interacting with proteins and other molecules, and participate in the occurrence and development of many diseases. Emerging experiment results indicate that circRNA molecule is rich in miRNA binding sites and acts as a miRNA sponge in the cell, thereby releasing the inhibitory effect of the miRNA on its target gene and increasing the expression level of the target gene (Hansen, et al., 2013; Rong, et al., 2017; Wang, et al., 2019). Nan *et al.* found that circRNA can up-regulate caspase-8 and p38 mitogen-activated protein kinase by competitive binding to miRNA-671, and indirectly affect the apoptosis pathway in lead-induced neuronal cell apoptosis (Nan, et al., 2017). Rybak-Wolf *et al.* found that there are a large number of circRNA molecules in mammalian brain tissue, and as the brain develops, its expression levels also increase, especially in synaptic structures (RybakWolf, et al., 2015). Chen *et al.* found five circRNAs: IQCK, MAP4K3, EFCAB11, DTNA, and MCTP1 in a study of patients with multiple system atrophy. These five circRNAs are specifically expressed in the frontal cortex white matter of patients with multiple system atrophy, but their corresponding linear RNA expression did not change significantly, suggesting that these circRNAs may be involved in the development of multiple system atrophy (Chen, et al., 2016). Besides, during the analysis of 13617 circRNAs expression profiles in multiple sclerosis patients and healthy controls, Leire *et al.* screened out disease-related circRNA and found 406 differentially expressed circRNAs, and finally found circRNAs from ANXA2 (circ-ANXA2) are low expression in patients with multiple sclerosis and can be used as new disease biomarkers (Leire, et al., 2017). From the above research we can see that the study of circRNA-disease association can provide theoretical basis and new ideas for the treatment of complex diseases. However, research on circRNA-disease association by experimental methods is often limited to small scales and requires a lot of time and labor. Therefore, there is an urgent need for reliable computational methods to study it on a large scale and quickly.

In this study, we propose a novel efficient computational method to predict potential circRNA-disease associations based on the hypothetical that functionally similar circRNAs are usually associated with phenotypically similar diseases, and vice versa. Firstly, we construct a numerical descriptor founded on circRNA similarity network and disease semantic similarity network. To take full advantage of the biological information of circRNA and disease, we also incorporate their Gaussian interaction profile kernel similarity network into the descriptor. Secondly, we use the deep learning Convolutional Neural Network (CNN) algorithm to automatically and objectively extract the deep features of the circRNA-disease descriptor. Finally, the Extreme Learning Machine (ELM) classifier is used to predict the potential circRNA-disease associations quickly and accurately. In order to fairly evaluate the performance of the proposed model, the five-fold cross-validation method was used in the experiment. As a result, 87.21% prediction accuracy with 88.50% sensitivity at the AUC of 86.67% was obtained on the CircR2Disease dataset. In comparison with the state-of-the-art SVM classifier model and different feature extraction models, the proposed model has achieved good results. Additionally, we also validated the predicted circRNA-disease pairs in published literature. Ultimately, 7 of the 15 pairs with the highest predicted scores were confirmed. These results demonstrated that our proposed model is very suitable for predicting circRNA-disease association. The flowchart of our proposed model is shown in figure 1.

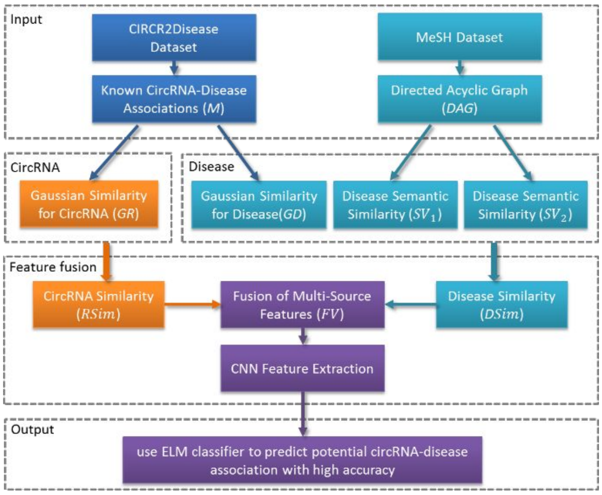
Input
CIRCR2Disease Dataset
MeSH Dataset
Known CircRNA-Disease Associations (M)
Directed Acyclic Graph (DAG)
CircRNA
Disease
Gaussian Similarity for CircRNA (GR)
Gaussian Similarity for Disease(GD)
Disease Semantic Similarity (SV_1)
Disease Semantic Similarity (SV_2)
Feature fusion
CircRNA Similarity ($RSim$)
Fusion of Multi-Source Features (FV)
Disease Similarity ($DSim$)
CNN Feature Extraction
Output
use ELM classifier to predict potential circRNA-disease association with high accuracy

Figure 1. Flowchart of the proposed model to predict potential circRNA-disease associations

2. MATERIALS AND METHODS

2.1 CircRNA-Disease Association Dataset

In the study, we validate our model using the CircR2Disease dataset compiled by Fan *et al.* (Fan, et al., 2018). Currently, the latest CIRCR2Disease dataset collects 739 entities supported by experiments from 246 published literature up to Mar 31, 2018, including 661 circRNAs and 100 diseases, with detailed and comprehensive annotations. This information provides valuable resources for us to further study circRNA-disease associations. The CIRCR2Disease dataset can be downloaded from http://bioinfo.snnu.edu.cn/CircR2Disease/. For easy viewing, we attach the CIRCR2Disease dataset as supplementary material to this study.

In experiments, we use the known circRNA-disease associations provided in the CIRCR2Disease dataset as the positive sample set. To construct the balanced dataset, we randomly select the same number of associations from other unknown circRNA-disease associations as the negative sample set. Although it is possible to use unconfirmed circRNA-disease pairs with association as negative samples, from the perspective of probability, the circRNA-disease pairs we selected as negative samples account for only $739 \div (661 \times 100) \approx 1.12\%$ of all circRNA-disease pairs, which is negligible. Ultimately, the dataset we constructed contains 1478 samples, of which positive and negative samples account for half.

On the basis of CIRCR2Disease dataset, we constructed circRNA and disease adjacent matrix M, which contains 661 rows and 100 columns, corresponding to 661 circRNAs and 100 diseases, respectively. When circRNA $c(i)$ and disease $d(j)$ are marked as associated by the CIRCR2Disease dataset, the element $M(c(i),d(j))$ of the matrix M is assigned to value of 1, otherwise it is assigned to value of 0.

2.2 Construction of Disease Semantic Similarity Model 1

The information we used to construct the disease semantic similarity model was based on the MeSH database (Folador, et al., 2014; Macintyre, et al., 2014; Xiang, et al., 2013; Zheng, et al., 2019) from the National Library of Medicine (NLM), the world's largest medical library operated by the United States federal government. The MeSH database provides a rigorous system for disease classification, which can be downloaded from https://www.nlm.nih.gov/. In MeSH database, the relationship between diseases is described as the form of Directed Acyclic Graph (DAG), in which nodes represent diseases and edges represent relationships between diseases. Given a disease d, it can be described as $DAG_d = (d,N_d,E_d)$, where N_d is an ancestral node set of d including itself, and E_d is the corresponding edge set connecting these diseases. If the disease e is in DAG_d, its contribution to disease d can be calculated as follows:

$$\begin{cases} D_d(e) = 1 & if\ \ e = d \\ D_d(e) = \max\left\{\varepsilon \cdot D_d(e') \,\middle|\, e' \in children\ of\ e\right\} & if\ e \neq d \end{cases} \quad (1)$$

where ε is the semantic contribution factor linking disease e and its child disease e'. In the DAG of d, the contribution of disease d to its own semantic value is defined as 1. Thus, we can calculate the semantic value $DV(d)$ of disease d by the following formula

$$DV(d) = \sum\nolimits_{e \in N_d} D_d(e) \quad (2)$$

Here, we assume that the more parts shared in the DAG of the two diseases, the more semantic similarity the two diseases are. Thus, we can calculate the first semantic similarity value $SV_1(d(i),d(j))$ between disease $d(i)$ and disease $d(j)$ according to their relative positions in MeSH disease DAG, and the formula is as follows:

$$SV_1(d(i),d(j)) = \frac{\sum_{e \in N_{d(i)} \cap N_{d(j)}} (D_{d(i)}(e) + D_{d(j)}(e))}{DV(d(i)) + DV(d(j))} \quad (3)$$

2.3 Construction of Disease Semantic Similarity Model 2

In disease semantic similarity model 1, we mainly consider the layer relationship of the ancestral nodes of the disease in the DAG, that is, the contribution of different ancestral nodes of the disease in the same layer to the disease is the same. However, this model does not consider the number of diseases occurring in DAGs. In general, diseases that are less common in DAGs should be given a higher contribution value. Therefore, in order to reflect this situation, we construct the second model to calculate the contribution of disease e, the formula is as follows:

$$D'_d(e) = -\log\left(\frac{num(DAGs(e))}{num(diseases)}\right) \quad (4)$$

where $num(DAGs(e))$ represents the number of DAGs including disease e, and $num(diseases)$ represents the number of all diseases. Thus, we can get the second semantic similarity model of disease. For disease $d(i)$ and disease $d(j)$, the semantic similarity value $SV_2(d(i),d(j))$ between them can be calculated as follows:

$$SV_2(d(i),d(j)) = \frac{\sum_{e \in N_{d(i)} \cap N_{d(j)}} (D'_{d(i)}(e) + D'_{d(j)}(e))}{DV(d(i)) + DV(d(j))} \quad (5)$$

where $DV(d(i))$ and $DV(d(j))$ has the same meaning as disease semantic similarity model 1, which can be calculated by equation 2.

In the experiment, the disease semantic similarity model 1 and model 2 we constructed are based on the MeSH database. But for the CIRCR2Disease dataset, only DAG of partial diseases can be found in the MeSH database. Therefore, in order to make the disease information more comprehensive, we introduced Gaussian interaction profile kernel similarity to calculate the similarity of other diseases. In describing the disease information, we fuse the disease semantic similarity model 1, model 2 and Gaussian interaction profile kernel similarity for disease to obtain the final disease similarity descriptor.

2.4 Gaussian Interaction Profile Kernel Similarity for Disease

Under the hypothesis that similar diseases tend to associate functionally similar circRNAs, and vice versa, Gaussian interaction profile kernel similarity for diseases can be calculated (Wang, et al., 2010; Xuan, et al., 2013). Gaussian interaction profile kernel function is a scalar function that is symmetric along the radial direction. It can better measure the similarity between samples, so that similar samples can be better clustered in a space describing similarity, and then can be linearly separable. We define binary vector $V(d(i))$ to represent the interaction profiles of disease $d(i)$, whose value is derived from the relationship between disease $d(i)$ and each of the 661 circRNAs in the CIRCR2Disease dataset. When the disease $d(i)$ is associated with a certain circRNA, the corresponding position of the circRNA in the binary vector $V(d(i))$ is assigned to 1, otherwise it is assigned to 0. In other words, the binary vector $V(d(i))$ is the row vector of the disease $d(i)$ in the adjacency matrix M of the CIRCR2Disease dataset. Thus, the Gaussian interaction profile kernel similarity for diseases $GD(d(i),d(j))$ of disease $d(i)$ and disease $d(j)$ can be calculated using the following formula:

$$GD(d(i),d(j)) = \exp\left(-\theta_d \|V(d(i)) - V(d(j))\|^2\right) \quad (6)$$

where θ_d is the width parameter of the function, which can be calculated by normalizing the original parameters. The formula is as follows:

$$\theta_d = \frac{1}{m}\sum\nolimits_{i=1}^{m} \|V(d(i))\|^2 \quad (7)$$

where m is the number of rows of the adjacency matrix M.

2.5 Gaussian Interaction Profile Kernel Similarity for CircRNA

For circRNA, we calculated its Gaussian interaction profile kernel similarity in the experiment. Similar to disease, we use binary vector $V(c(i))$ to represent interaction profiles of circRNA $c(i)$, which is the column vector of circRNA $c(i)$ in the adjacency matrix M of the CIRCR2Disease dataset. Thus, the Gaussian interaction profile kernel similarity for circRNA $GR(c(i),c(j))$ of circRNA $c(i)$ and circRNA $c(j)$ can be calculated using the following formula:

$$GR(c(i),c(j)) = \exp\left(-\theta_c \|V(c(i)) - V(c(j))\|^2\right) \quad (8)$$

$$\theta_c = \frac{1}{n}\sum\nolimits_{i=1}^{n} \|V(c(i))\|^2 \quad (9)$$

where θ_c is the width parameter of the function, n is the number of columns of the adjacency matrix M.

2.6 Multi-source Data Fusion

In this study, the descriptor we finally use combines disease semantic similarity, disease Gaussian interaction profile kernel similarity and circRNA Gaussian interaction profile kernel similarity. This descriptor can contain more abundant information and help to dig deeper potential circRNA-disease associations.

For disease, we constructed disease semantic similarity model SV_1, disease semantic similarity model SV_2, and disease Gaussian interaction profile kernel similarity GD. In order to make full use of the information in the MeSH database, if there is semantic similarity between disease $d(i)$ and $d(j)$, the disease semantic similarity models are used to construct the descriptor $DSim(d(i),d(j))$, otherwise disease Gaussian interaction profile kernel similarity is used. The formula can be expressed as follows.

$$DSim(d(i),d(j)) = \begin{cases} \frac{SV_1(d(i),d(j)) + SV_2(d(i),d(j))}{2} & if\ d(i)\ and\ d(j)\ has\ semantic\ similarity \\ GD(d(i),d(j)) & otherwise \end{cases} \tag{10}$$

For circRNA, we directly use Gaussian interaction profile kernel similarity $GR(c(i),c(j))$ between circRNA $c(i)$ and $c(j)$ to represent circRNA similarity $RSim(c(i),c(j))$. Thus, we established the circRNA-disease fusion descriptor based on the association among circRNAs and diseases provided by CIRCR2Disease dataset. The fusion descriptor $FV(c(i),d(j))$ composed of circRNA $c(i)$ and diseases $d(j)$ can be expressed as follows:

$$FV(c(i),d(j)) = [RSim(c(i)),DSim(d(j))] \tag{11}$$

where $\mathrm{RSim(c(i))}$ represents Gaussian interaction profile kernel similarity of circRNA $\mathrm{c(i)}$, $\mathrm{DSim(d(j))}$ represents the j row vector of disease $\mathrm{d(j)}$ in the disease similarity matrix Dsim. In the fusion of multi-source data, we used normalization operation, and the dimension of the fused descriptor is 761.

2.7 Convolutional Neural Network

Deep learning can form more abstract high-level representation attributes or features by combining low-level features, thus discovering the distributed feature representation of data, which brings hope to solve the optimization problems related to deep structure(Pan and Shen, 2018; Sss, et al., 2017; Wang, et al., 2019). Among several deep learning architectures, Convolutional Neural Network (CNN) can use spatial relationships to reduce the number of learning parameters (Krizhevsky, et al., 2012; Yu, et al., 2017), thereby improving training performance and efficiently extracting data features. Therefore, we introduce the deep learning CNN algorithm into the circRNA-disease association prediction model to extract the hidden deep features of the fused multi-source descriptors.

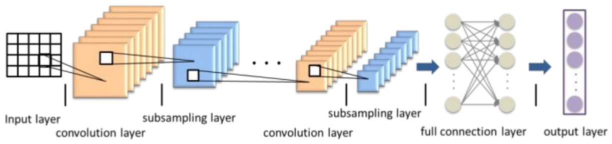

Figure 2. Schematic diagram of CNN structure

CNN is a multi-layer neural network structure consisting of input layer, convolution layer, subsampling layer, full connection layer and the output layer. Its structure is shown in figure 2. Suppose L_i is the feature map of the ith layer, which is expressed by the following formula:

$$L_i = f(L_{i-1} \otimes W_i + b_i) \tag{12}$$

where $f(x)$ denotes the activation function, L_{i-1} is the feature map of the $(i-1)th$ layer, W_i represents the weight matrix of the convolution kernel of ith layer, b_i represents the offset vector and operator $\otimes$ denotes convolution operations. In CNN, the subsampling layer usually behind the convolution layer and samples the feature graph according to certain rules. When L_i is the subsampling layer, its sampling formula can be expressed as:

$$L_i = subsampling(L_{i-1}) \tag{13}$$

After multiple convolution and sampling, CNN classifies the features extracted by the full connection layer and obtains the input-based probability distribution S. CNN is essentially a mathematical model that allows the original input matrix L_0 to be mapped to the new feature expression S through multi-level data transformation or dimensionalization.

$$S(i) = Map(C = c_i|L_0;\ (W,b)) \tag{14}$$

where c_i denotes the ith label class, L_0 represents the raw input matrix and S represents the feature expression.

The training goal of CNN is to minimize the loss function of the network. Meanwhile, in order to alleviate the over-fitting problem, the ultimate loss function $E(W,b)$ is usually controlled by norm, and through the parameters δ to control the intensity of the over-fitting.

$$E(W,b) = F(W,b) + \frac{\delta}{2}W^T W \tag{15}$$

In the CNN training process, the gradient descent method is usually used to optimize the network, update the parameters $(W,b)$of the network layer by layer, and control the back-propagation intensity by using the learning rate ϑ.

$$W_i = W_i - \vartheta\frac{\partial E(W,b)}{\partial W_i} \tag{16}$$

$$b_i = b_i - \vartheta\frac{\partial E(W,b)}{\partial b_i} \tag{17}$$

2.8 Extreme Learning Machine

Extreme Learning Machine (ELM) is a learning algorithm based on single hidden layer feedforward neural network model proposed by Huang *et al.* (Huang, et al., 2011; Huang, et al., 2006). It has the advantages of fast training and good normalization performance (Al-Yaseen, et al., 2017; Iosifidis, et al., 2017). Therefore, we use ELM as the classifier for predicting circRNA-disease associations. The structure of the ELM classifier is shown in figure 3.

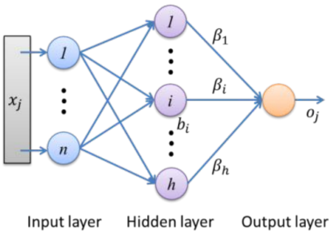

Figure 3. Structure diagram of extreme learning machine

Suppose there are n arbitrary samples (X_i,l_i), where $X_i = [x_{i1},x_{i2,...,}x_{in}]^T \in \mathbb{R}^n$ represents the sample attribute, and $l_i = [l_{i1},l_{i2},...,l_{im}]^T \in \mathbb{R}^m$ represents the sample label. The ELM classifier with h hidden layer nodes can be described as follows.

$$\sum_{i=1}^{h}\beta_i f(W_i \cdot X_j + b_i) = o_j,\ j = 1,\ldots,n \tag{18}$$

where $W_i = [w_{i1}, w_{i2}, \ldots, w_{in}]^T$ represents the input weight, β_i represents the output weight, $W_i \cdot X_j$ indicates the inner product of W_i and X_j, b_i represents the offset of the ith hidden layer, $f(x)$ indicates the activation function and o_j represents the output.

The ELM's learning goal is to minimize output errors, that is:

$$\sum_{j=1}^{n}\|o_j - l_j\| = 0 \tag{19}$$

In order to achieve this goal, ELM needs to adjust the parameters W_i, b_i and β_i to make

$$\sum_{i=1}^{h}\beta_i f(W_i \cdot X_j + b_i) = l_j,\ j = 1,\ldots,n \tag{20}$$

It can be represented by matrix:

$$F\beta = L \tag{21}$$

$$F = \begin{bmatrix} f(W_1 \cdot X_1 + b_1) & \cdots & f(W_h \cdot X_1 + b_h) \\ \vdots & \vdots & \vdots \\ f(W_1 \cdot X_n + b_1) & \cdots & f(W_h \cdot X_n + b_h) \end{bmatrix}_{n \times h}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_h^T \end{bmatrix}_{h \times m} \quad L = \begin{bmatrix} L_1^T \\ \vdots \\ L_n^T \end{bmatrix}_{n \times m} \tag{22}$$

where L is the expected output, F is the output of the hidden layer node, and β is the output weight. Therefore, we hope to get W_i, b_i and β_i by training a single hidden layer neural network, making

$$\left\|F(\hat{W}_i, \hat{b}_i)\hat{\beta}_i - L\right\| = \min_{W,b,\beta}\|F(W_i, b_i)\beta_i - L\| \quad i = 1,2,\cdots,h \tag{23}$$

This is equivalent to minimizing the loss function

$$E = \sum_{j=1}^{n}\left(\sum_{i=1}^{h}\beta_i f(W_i \cdot X_j + b_i) - l_j\right)^2 \tag{24}$$

In ELM algorithm, once the input weight W_i and the offset b_i are randomly determined, the output matrix F of the hidden layer can be uniquely determined. Thus, the training single hidden layer neural network is transformed into solving a linear system $F\beta = L$. The output weight β can be determined and the norm of the solution can be proved to be minimal and unique. In the experiment, we set ELM's parameter 'Elm_Type' to 1, 'NumberofHiddenNeurons' to 5, 'ActivationFunction' to Sigmoidal function, and other parameters to default values.

3. RESULTS AND DISCUSSION

3.1 Evaluation Criteria

To obtain a reliable and stable model, and to facilitate comparison with other models, we use the five-fold cross-validation method to verify the proposed model. The method first randomly divides all samples into five approximately identical parts, and then takes four of them to train the model, and the remaining one to test the model. This process continues until all five parts been tested once and only tested once(Wang, et al., 2018). In the specific implementation process, we performed 5 times five-fold cross-validation to reduce the variation resulting from samples partition. The final experimental results are the average values of the five experiments. In addition, in order to prevent the leakage of association information, we recomputed the Gaussian interaction profile kernel similarity for each fold according to the method of Laarhoven *et al.* (van Laarhoven, et al., 2011). In the experiment, we use accuracy (Accu.), sensitivity (Sen.), F1-Score (F1) and Matthews Correlation Coefficient (MCC) as the evaluation criteria to evaluate the proposed model. They are defined as follows:

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

$$Sen. = \frac{TP}{TP + FN} \tag{26}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{27}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{28}$$

where TP indicates the number of true positives; TN indicates the number of true negative; FP indicates the number of false positive; FN indicates the number of false negative. In addition, we also plot the Receiver Operating Characteristic (ROC) (Swets, 1988; Zweig and Campbell, 1993) generated by the proposed model and calculated the area under the ROC curve (AUC) (Bradley, 1997).

3.2 Assessment of Prediction Ability

To evaluate the performance of the proposed model for predicting circRNA-disease associations comprehensively, we implemented it on CIRCR2Disease dataset using the five-fold cross-validation method. The detailed experimental results are summarized in table 1. From the table we can see that our model achieves an accuracy of 87.21% on CIRCR2Disease dataset. The accuracy of the five verifications is 88.47%, 87.46%, 82.37%, 90.85% and 86.91%, respectively, and their standard deviation is 3.10%. In the sensitivity indicating the accuracy for the classification of true positive samples, our model achieved 88.50% of the results with the standard deviation of 3.26%. On the evaluation criteria F1-score and MCC reflecting the comprehensive performance of the model, the proposed model obtains 87.39% and 74.46 results, and their standard deviations are 2.84% and 6.18%, respectively. Besides, we plot the ROC curves generated by the proposed model on the CIRCR2Disease dataset and calculate their AUC respectively. As can be seen from figure 4, the ROC curves of the five-fold model can reach the upper left of the graph. Their AUCs reached 90.61%, 86.37%, 82.63%, 88.68% and 85.05% respectively, with the average of 86.67% and the standard deviation of 3.11%. From the above experimental results, we can see that the proposed model performs well on the CIRCR2Disease dataset and can effectively predict the potential circRNA-disease associations.

Table 1. Five-fold cross-validation results performed by proposed model on CIRCR2Disease dataset

Validation set	Accu.(%)	Sen.(%)	F1(%)	MCC(%)	AUC(%)
1	88.47	91.50	89.17	76.98	90.61
2	87.46	87.23	86.93	74.88	86.37
3	82.37	86.09	83.33	64.82	82.63
4	90.85	92.41	90.85	81.75	88.68
5	86.91	85.23	86.69	73.87	85.05
Average	**87.21±3.10**	**88.50±3.26**	**87.39±2.84**	**74.46±6.18**	**86.67±3.11**

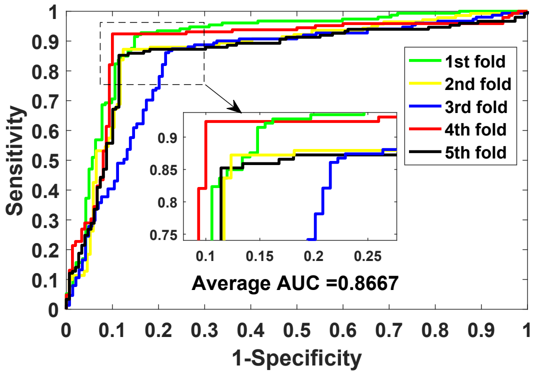

Figure 4. ROC curves performed by proposed method on CIRCR2Disease dataset

To further evaluate the performance of the model, we implemented the experiment on the independent test set. Specifically, we first randomly select 20% of the samples as independent test set, which are $1478 \times 20\% \approx 295$ samples. The remaining $1478 - 295 = 1183$ samples are divided into 5 approximately equal parts, which are used as training sets and verification sets to cross-validate the model. The number of these five subsets is 236, 236, 236, 236 and 239, respectively. In this way, we divide the data set into completely independent test set, validation set and training set, and ensure that there is no overlap between them. The divided data set is given in supplementary material 2. In this data set, the first column is the dataset label, where 'independent test set' and 'subset 1-5' represent independent test set and five subsets respectively; the second column is the flag, where '1' indicates an association between circRNA and disease, '-1' indicates no association; the third column is the name of the circRNAs; and the fourth column is the name of the diseases. Table 2 summarizes the experimental results of the proposed model on an independent test set. From the table we can see that 90.85% accuracy and 90.32% AUC were achieved on the independent test set, which is significantly better than the cross-validation results. The reason for this is mainly due to the larger number of training samples used in the independent test experiments, which enables more adequate training of the model.

Table 2. The performance on the independent test set by proposed model on CIRCR2Disease dataset

Validation set	Accu.(%)	Sen.(%)	F1(%)	MCC(%)	AUC(%)
1	83.47	85.19	82.51	67.00	82.41
2	80.93	82.14	80.35	61.90	83.75
3	88.14	85.37	88.24	76.47	89.43
4	91.53	93.60	92.13	83.00	92.54
5	85.36	85.59	85.23	70.71	83.77
Average	85.88±4.11	86.38±4.28	85.69±4.66	71.82±8.21	86.38±4.38
Independent test set	**90.85**	**86.27**	**90.72**	**82.16**	**90.32**

3.3 Comparison on Different DataSet

To further verify the performance of the proposed model, we performed experiments on another circRNADisease dataset. In circRNADisease dataset, more than 800 published literature were systematically reviewed, and manually curated and collected 330 circRNAs and 48 diseases in 354 associations. It can be downloaded at http://cgga.org.cn:9091/circRNADisease/.

The experimental results on circRNADisease dataset are shown in table 3. From the table we can see that the proposed model achieves the prediction accuracy of 89.36%, sensitivity of 86.67%, F1-score of 89.66%, MCC of 78.93% and AUC of 89.64% on circRNADisease dataset. Although the results of the proposed model on circRNADisease dataset are slightly inferior to those on CIRCR2Disease, they are also quite good. This is mainly because the circRNADisease dataset contains less data than the CIRCR2Disease dataset, so the training of the model is not as good as using the CIRCR2Disease dataset. Figure 5 plots the ROC curves generated by the proposed model on circRNADisease dataset. These results show that the proposed model can be applied to different datasets and is robust under various conditions.

Table 3. The performance on the independent test set by proposed model on circRNADisease dataset

Validation set	Accu.(%)	Sen.(%)	F1(%)	MCC(%)	AUC(%)
1	69.91	67.80	70.18	39.99	70.97
2	90.27	87.50	89.91	80.63	90.91
3	91.15	92.59	90.91	82.35	88.20
4	85.84	73.21	83.67	73.93	86.28
5	78.26	81.48	77.88	56.81	79.20
Average	83.09±8.96	80.52±10.13	82.51±8.66	66.74±18.04	83.11±8.06
Independent test set	**89.36**	**86.67**	**89.66**	**78.93**	**89.64**

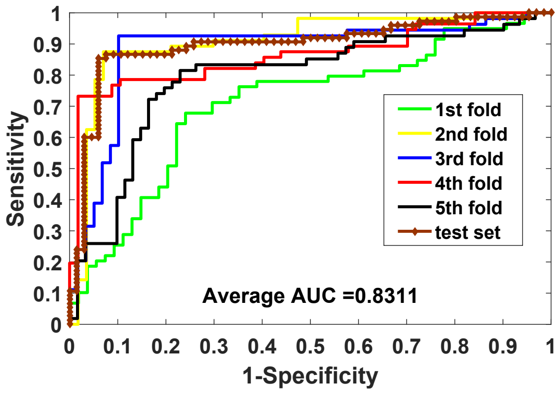

Figure 5. ROC curves performed by proposed method on circRNADisease dataset

3.4 Comparison with Different Classifier

In order to have a comprehensive assessment of the performance of the proposed model, we compare it to the state-of-the-art Support Vector Machine (SVM) classifier model. SVM is a classification algorithm based on the VC dimension theory and structural risk minimization principle. It has been widely used in the field of biological science because of its outstanding performance in solving small sample, nonlinear and high-dimensional pattern recognition. To ensure the fairness of the experiment, the descriptors fed into the SVM model are consistent with the proposed model.

Table 4 lists the experimental results of the SVM model with the proposed feature on the CIRCR2Disease dataset. From the table it can be seen that the SVM model achieves the prediction accuracy of 87.12%, sensitivity of 75.16%, F1-score of 85.82%, MCC of 77.00% and AUC of 87.43%, respectively. It can be seen from the comparison of these evaluation criteria that the accuracy, sensitivity, F1-score, MCC and AUC of the proposed model are 3.73%, 11.11%, 4.90%, 5.16% and 2.89% higher than those of the SVM model, respectively. Figure 6 plots the ROC curves generated by the proposed model and the SVM model on CIRCR2Disease dataset, where the solid line annotates the ROC curves generated for the proposed model and the dashed line annotates the ROC curves generated for the SVM model. It is obvious from the figure that the area under the ROC curve generated by the proposed model is larger than that generated by the SVM model. This indicates that the performance of the proposed model is better than that of the SVM model, and is more suitable for predicting the circRNA-disease associations.

Table 4. The performance on the independent test set by SVM model combined with the proposed feature descriptor

Validation set	Accu.(%)	Sen.(%)	F1(%)	MCC(%)	AUC(%)
1	90.68	79.63	88.66	82.43	90.05
2	88.98	76.79	86.87	79.67	89.54
3	72.03	46.34	63.33	54.09	72.98
4	76.27	55.20	71.13	60.57	79.03
5	77.41	54.24	70.33	61.24	74.63
Average	81.07±8.26	62.44±14.83	76.07±11.12	67.60±12.63	81.25±8.11
Independent test set	**87.12**	**75.16**	**85.82**	**77.00**	**87.43**

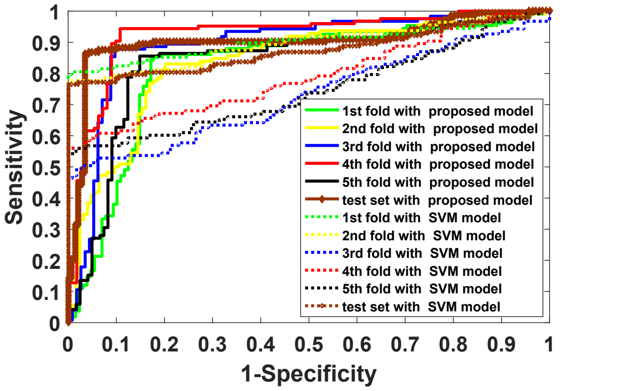

Figure 6. Comparison of ROC curves of different classifier models on CIRCR2Disease dataset

3.5 Comparison among Different Feature Extraction Algorithm

To verify the ability of the CNN algorithm we used to extract circRNA-disease association information, we compare it with Auto Covariance (AC) and Discrete Cosine Transform (DCT) feature extraction algorithms. Both AC and DCT algorithms have powerful feature extraction capabilities and have been applied to biological data processing by many researchers(Gao, et al., 2016; Guo, et al., 2006; Guo, et al., 2008; Wang, et al., 2017; Wang, et al., 2018). Similarly, in order to ensure the fairness of the comparison, we only change the feature extraction algorithm of the model in the experiment. Table 5 summarizes the experimental results generated on CIRCR2Disease dataset using the AC feature extraction algorithm. As can be seen from the table, the accuracy, sensitivity, F1-score, MCC and AUC of the AC model are 73.90%, 83.57%, 75.24%, 49.30% and 76.58% respectively. Table 6 shows the experimental results generated on CIRCR2Disease dataset using the DCT feature extraction algorithm. We can see that the accuracy, sensitivity, F1-score, MCC and AUC of DCT model are 74.92%, 74.23%, 76.58%, 49.75% and 75.81% respectively. For the sake of facilitating comparison, we present the results generated by these three models in the form of the histogram. It is obvious from figure 7 that the proposed model achieves the best results in the above evaluation criteria. From the comparison of feature extraction methods, it can be concluded that the CNN algorithm used in the proposed model can effectively mine the deep information contained in the circRNA-disease association data, and help to improve the performance of the model.

Table 5. The performance on the independent test set by AC model combined with the proposed classifier

Validation set	Accu.(%)	Sen.(%)	F1(%)	MCC(%)	AUC(%)
1	73.73	80.80	76.52	47.24	73.98
2	71.19	73.50	71.67	42.45	72.56
3	71.19	74.79	72.36	42.43	72.60
4	69.07	80.36	71.15	39.94	71.61
5	70.71	78.57	73.88	41.20	69.37
Average	71.18±1.67	77.60±3.30	73.11±2.16	42.65±2.76	72.02±1.71
Independent test set	**73.90**	**83.57**	**75.24**	**49.30**	**76.58**

Table 6. The performance on the independent test set by DCT model combined with the proposed classifier

Validation set	Accu.(%)	Sen.(%)	F1(%)	MCC(%)	AUC(%)
1	71.19	69.23	70.43	42.38	70.90
2	69.07	75.61	71.81	37.97	69.65
3	70.34	70.27	69.03	40.61	73.49
4	69.07	76.58	69.96	39.18	68.95
5	67.36	59.65	63.55	34.48	67.47
Average	69.41±1.45	70.27±6.75	68.96±3.19	38.92±2.98	70.09±2.27
Independent test set	**74.92**	**74.23**	**76.58**	**49.75**	**75.81**

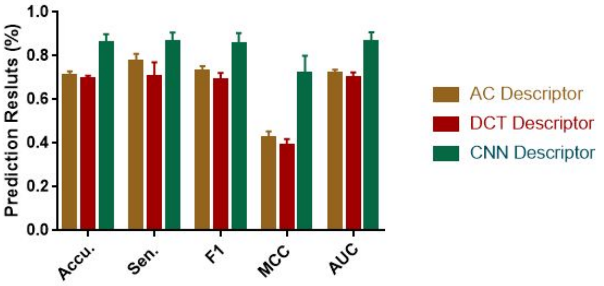

Figure 7. Comparison of results of different descriptor models on CIRCR2Disease dataset

3.6 Comparison with previous studies

To further evaluate the performance of the proposed method, we compared it with previous studies. Here we collect methods for predicting circRNA-disease associations on CircR2Disease dataset using five-fold cross-validation, including DWNN-RLS (Yan, et al., 2018), PWCDA (Lei, et al., 2018) and Fan’s method (Fan, et al., 2018). Since these methods only reported the value of AUC in the paper, we compared the proposed method with them using AUC. The experimental results are summarized in table 7. It can be seen from the table that although the proposed model does not achieve the optimal result, it is only about 2% lower than the highest score.

The main reason why our proposed method performs a litter bit low is mainly attributed to two reasons: (1) the data used by these four methods are not entirely consistent. We used all the data in the CircR2Disease dataset, while DWNN-RLS used data that removed redundancy, PWCDA used three main species of data including human, mouse and rat, and Fan's method used only human data. Comparatively speaking, they all have some optimizations on the raw data, while we use all the data in order to maintain the integrity of the dataset. However, the raw data without optimization may contain more noise information, which has a certain impact on the experimental results. (2) Our approach belongs to the machine learning based approach while other three methodsbelong to the network-based approach. The network-based approach has the characteristics of achieving higher accuracy in small-scale data, but it also has significant shortcomings: (a) Most of these methods can't be applied to new diseases without known associated circRNAs and/or new circRNAs without any known associated diseases or known circRNA interaction partners. (b) The network-based methods usually take a lot of time to build network models. This is because the network model generally has a high degree of complexity and with the increase of the number of samples, the modeling time increases exponentially. (c) It is not easy to add new nodes to the completed network model unless it is rebuilt. This greatly limits the flexibility of the model. However, we can easily avoid the above defects by using the machine learning based approach. For new diseases and circRNAs data, we only need to extract their features, and then put them into the training set to train the model, so that we can easily predict new circRNA-desease pairs of unknown association, Therefore, the proposed model has better adaptability. Another advantage of the proposed method is that the more adequate the model is trained, the better the performance of the model. With the development of circRNA-desease association research, more and more disease-related circRNAs will be found, and the prediction results of the proposed model will be further improved.

Table 7. Comparisons the proposed model with previous studies on CIRCR2Disease dataset

Methods	Our model	DWNN-RLS	PWCDA	Fan's method
AUC(%)	86.67	88.54	89.00	79.36

3.7 Case Studies

To verify the ability of the proposed model to predict potential circRNA-disease associations, we implemented the case studies on CIRCR2Disease dataset. Specifically, we first train the proposed model using all known circRNA-disease associations in the CIRCR2Disease dataset, then use the trained model to predict the circRNA-disease associations of the unknown association compiled from the CIRCR2Disease dataset. Finally, we sort the circRNA-disease pairs based on the predicted scores, and searched for the circRNA-disease pairs in the recently published literature. As a result, among the top 15 unknown circRNA-disease pairs with the highest scores, 7 of them are confirmed by the related literature. The detailed results are shown in table 8. It should be noted that although the other circRNA-disease pairs of unknown association are not supported by the literature, there is no denying the possibility of association between them.

Table 8. Top 15 circRNA-disease pairs were predicted by the proposed model based on known miRNA-disease associations in CIRCR2Disease database

circRNA	Disease	Evidence (PMID)
hsa_circ_0004214	Glioma	28622299
hsa_circ_0037911	Heart disease	unconfirmed
hsa_circ_0004214	Cervical cancer	28622299
hsa_circ_0013958	Lung cancer	28685964
hsa_circ_0002161	Intracranial aneurysms	unconfirmed
hsa_circ_0004458	Gastric cancer	28544609
hsa_circ_0013339	Heart disease	unconfirmed
hsa_circ_0054633	Diabetes mellitus	27878383
hsa_circ_0021001	Immunosenescence	unconfirmed
hsa_circ_0023404	Lung cancer	28343871
hsa_circ_0001955	Dilated cardiomyopathy	unconfirmed
hsa_circ_0021001	Oral squamous cell carcinomas	unconfirmed
hsa_circ_0001187	Acute myeloid leukemia	28282919
hsa_circ_0005836	Ebola virus disease	unconfirmed
hsa_circ_0013339	Prostate cancer	unconfirmed

4. CONCLUSION

In this study, we proposed an efficient computational method for predicting circRNA-disease associations by integrating multi-source biological data including disease semantic similarity, disease and circRNA Gaussian interaction profile kernel similarity, and convolutional neural network is utilized to mine its hidden deep features. The proposed model exhibits good performance on the CIRCR2Disease dataset. In order to fully verify the reliability and robustness of the proposed model, we also compare it with the state-of-the-art SVM model and different feature extraction algorithm models. The experimental results demonstrated that the proposed model achieved excellent prediction performance. In addition, the case study is designed to demonstrate the prediction performance, and 7 of the 15 circRNA-disease pairs with the highest predicted scores were supported by the published literature. These results indicate that the proposed model is an effective method for predicting circRNA-disease associations and can provide theoretical support for biological experiments. In future research, we will consider incorporating more biological information into the data to expect achieves better performance of the model.

COMPETING INTERESTS

The authors declare that they have no competing interests.

ACKNOWLEDGEMENTS

This work is supported is supported in part by Awardee of the NSFC Excellent Young Scholars Program, under Grant 61722212, in part by the National Natural Science Foundation of China, under Grants 61873212, 61702444, 61572506, in part by the Chinese Postdoctoral Science Foundation, under Grant 2019M653804, in part by the West Light Foundation of The Chinese Academy of Sciences, under Grant 2018-XBQNXZ-B-008. The authors would like to thank all anonymous reviewers for their constructive advices.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interests regarding the publication of this paper.

REFERENCE

Al-Yaseen, W.L., Othman, Z.A. and Nazri, M.Z.A. (2017) Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system, Expert Systems with Applications, 67, 296-303.

Bahn, J.H., et al. (2015) The Landscape of MicroRNA, Piwi-Interacting RNA, and Circular RNA in Human Saliva, Clinical Chemistry, 61, 221-230.

Bradley, A.P. (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern recognition, 30, 1145-1159.

Chen, B.J., et al. (2016) Characterization of circular RNAs landscape in multiple system atrophy brain, Journal of Neurochemistry, 139, 485-496.

Chen, L., et al. (2017) circRNA_100290 plays a role in oral cancer by functioning as a sponge of the miR-29 family, Oncogene, 36, 4551-4561.

Danan, M., et al. (2012) Transcriptome-wide discovery of circular RNAs in Archaea, Nucleic Acids Research, 40, 3131-3142.

Fan, C., et al. (2018) CircR2Disease: a manually curated database for experimentally supported circular RNAs associated with various diseases, Database, 1, 6.

Fan, C., Lei, X. and Wu, F.-X. (2018) Prediction of CircRNA-Disease Associations Using KATZ Model Based on Heterogeneous Networks, International journal of biological sciences, 14, 1950.

Folador, E.L., et al. (2014) An improved interolog mapping-based computational prediction of protein-protein interactions with increased network coverage, Integrative Biology, 6, 1080-1087.

Gao, Z.G., et al. (2016) Ens-PPI: A Novel Ensemble Classifier for Predicting the Interactions of Proteins Using Autocovariance Transformation from PSSM, Biomed Research International, 8.

Guo, Y., et al. (2006) Predicting G-protein coupled receptors-G-protein coupling specificity based on autocross-covariance transform, Proteins-Structure Function and Bioinformatics, 65, 55-60.

Guo, Y., et al. (2008) Using support vector machine combined with auto covariance to predict proteinprotein interactions from protein sequences, Nucleic Acids Research, 36, 3025-3030.

Hansen, T.B., et al. (2013) Natural RNA circles function as efficient microRNA sponges, Nature, 495, 384-388.

Huang, G.B., Wang, D.H. and Lan, Y. (2011) Extreme learning machines: a survey, Int. J. Mach. Learn. Cybern., 2, 107-122.

Huang, G.B., Zhu, Q.Y. and Siew, C.K. (2006) Extreme learning machine: Theory and applications, Neurocomputing, 70, 489-501.

Iosifidis, A., Tefas, A. and Pitas, I. (2017) Graph Embedded Extreme Learning Machine, IEEE Transactions on Cybernetics, 46, 311-324.

Jeck, W.R., et al. (2013) Circular RNAs are abundant, conserved, and associated with ALU repeats, Rna-a Publication of the Rna Society, 19, 141-157.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) ImageNet classification with deep convolutional neural networks. International Conference on Neural Information Processing Systems. pp. 1097-1105.

Lei, X., et al. (2018) PWCDA: Path Weighted Method for Predicting circRNA-Disease Associations, International journal of molecular sciences, 19, 3410.

Leire, I., et al. (2017) Circular RNA profiling reveals that circular RNAs from ANXA2 can be used as new biomarkers for multiple sclerosis, Human Molecular Genetics, 26.

Macintyre, G., et al. (2014) Associating disease-related genetic variants in intergenic regions to the genes they impact, Peerj, 2, e639.

Memczak, S., et al. (2013) Circular RNAs are a large class of animal RNAs with regulatory potency, Nature, 495, 333-338.

Nan, A., et al. (2017) A novel regulatory network among LncRpa, CircRar1, MiR-671 and apoptotic genes promotes lead-induced neuronal cell apoptosis, Archives of Toxicology, 91, 1-14.

Nigro, J.M., et al. (1991) Scrambled exons, Cell, 64, 607.

Pan, X. and Shen, H.-B. (2018) Learning distributed representations of RNA sequences and its application for predicting RNA-protein binding sites with a convolutional neural network, Neurocomputing, 305, 51-58.

Qin, M., et al. (2016) Hsa_circ_0001649: A circular RNA and potential novel biomarker for hepatocellular carcinoma, Cancer Biomarkers, 16, 161.

Rong, D., et al. (2017) An emerging function of circRNA-miRNAs-mRNA axis in human diseases, Oncotarget, 8, 73271-73281.

RybakWolf, et al. (2015) Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed, Molecular Cell, 58, 870.

Salzman, J., et al. (2013) Cell-type specific features of circular RNA expression, Plos Genetics, 9, e1003777.

Sanger, H.L., et al. (1976) Viroids are Single-Stranded Covalently Closed Circular RNA Molecules Existing as Highly Base-Paired Rod-Like Structures, Proceedings of the National Academy of Sciences of the United States of America, 73, 3852-3856.

Sss, K., Ayush, K. and Babu, R.V. (2017) DeepFix: A Fully Convolutional Neural Network for Predicting Human Eye Fixations, IEEE Transactions on Image Processing, PP, 1-1.

Swets, J.A. (1988) Measuring the accuracy of diagnostic systems, Science, 240, 1285.

van Laarhoven, T., Nabuurs, S.B. and Marchiori, E. (2011) Gaussian interaction profile kernels for predicting drug–target interaction, Bioinformatics, 27, 3036-3043.

Wang, D., et al. (2010) Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases, Bioinformatics, 26, 1644-1650.

Wang, L., et al. (2019) Predicting Protein-Protein Interactions from Matrix-Based Protein Sequence Using Convolution Neural Network and Feature-Selective Rotation Forest, Scientific reports, 9, 9848.

Wang, L., et al. (2019) LMTRDA: Using logistic model tree to predict MiRNA-disease associations by fusing multi-source information of sequences and similarities, PLoS computational biology, 15, e1006865.

Wang, L., et al. (2017) Advancing the prediction accuracy of protein-protein interactions by utilizing evolutionary information from position-specific scoring matrix and ensemble classifier, Journal Of Theoretical Biology, 418, 105-110.

Wang, L., et al. (2018) Using Two-dimensional Principal Component Analysis and Rotation Forest for Prediction of Protein-Protein Interactions, Scientific reports, 8, 12874.

Wang, L., et al. (2018) RFDT: A Rotation Forest-based Predictor for Predicting Drug-Target Interactions Using Drug Structure and Protein Sequence Information, Current Protein & Peptide Science, 19, 445-454.

Xiang, Z., et al. (2013) A genome-wide MeSH-based literature mining system predicts implicit gene-to-gene relationships and networks, BMC systems biology, 7, S9.

Xuan, P., et al. (2013) Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors, PloS one, 8, e70204.

Yan, C., Wang, J. and Wu, F.-X. (2018) DWNN-RLS: regularized least squares method for predicting circRNA-disease associations, BMC bioinformatics, 19, 520.

Yu, Y., et al. (2017) Unsupervised Representation Learning with Deep Convolutional Neural Network for Remote Sensing Images.

Zheng, K., et al. (2019) MLMDA: a machine learning approach to predict and validate MicroRNA–disease associations by integrating of heterogenous information sources, Journal of translational medicine, 17, 260.

Zhou, B. and Xu, H.Y. (2017) A novel identified circular RNA, circRNA_010567, promotes myocardial fibrosis via suppressing miR-141 by targeting TGF-β1, Biochemical & Biophysical Research Communications, 487, 769-775.

Zhu, W., et al. (2016) Gut Microbial Metabolite TMAO Enhances Platelet Hyperreactivity and Thrombosis Risk, Cell, 165, 111-124.

Zweig, M.H. and Campbell, G. (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine, Clinical chemistry, 39, 561-577.