GCNCDA: A New Method for Predicting CircRNA-Disease Associations Based on Graph Convolutional Network Algorithm

Lei Wang¹,*, Zhu-Hong You¹,*, Yang-Ming Li², Kai Zheng³, Yu-An Huang⁴

¹Xinjiang Technical Institutes of Physics and Chemistry, Chinese Academy of Sciences, Urumqi, 830011, China
²Department of Electrical Computer and Telecommunications Engineering Technology, Rochester Institute of Technology, Rochester, 14623, USA
³School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, 221116, China
⁴Department of Computing, Hong Kong Polytechnic University, Hong Kong 999077, China

*Corresponding authors
#The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

Email: zhuhongyou@ms.xjb.ac.cn; leiwang@ms.xjb.ac.cn

Abstract

Numerous evidences indicate that Circular RNAs (circRNAs) are widely involved in the occurrence and development of diseases. Identifying the association between circRNAs and diseases plays a crucial role in exploring the pathogenesis of complex diseases and improving the diagnosis and treatment of diseases. However, due to the complex mechanisms between circRNAs and diseases, it is expensive and time-consuming to discover the new circRNA-disease associations by biological experiment. Therefore, there is increasingly urgent need for utilizing the computational methods to predict novel circRNA-disease associations. In this study, we propose a computational method called GCNCDA based on the deep learning Fast learning with Graph Convolutional Networks (FastGCN) algorithm to predict the potential disease-associated circRNAs. Specifically, the method first forms the unified descriptor by fusing disease semantic similarity information, disease and circRNA Gaussian Interaction Profile (GIP) kernel similarity information based on known circRNA-disease associations. The FastGCN algorithm is then used to objectively extract the high-level features contained in the fusion descriptor. Finally, the new circRNA-disease associations are accurately predicted by the Forest by Penalizing Attributes (Forest PA) classifier. The 5-fold cross-validation experiment of GCNCDA achieved 91.2% accuracy with 92.78% sensitivity at the AUC of 90.90% on circR2Disease benchmark dataset. In comparison with different classifier models, feature extraction models and other state-of-the-art methods, GCNCDA demonstrates superior performance in predicting novel circRNA-disease associations.
methods, GCNCDA shows strong competitiveness. Furthermore, 10 of the top 15 circRNA-disease association candidates with the highest prediction scores were confirmed by recently published literature. These results suggest that GCNCDA can effectively predict potential circRNA-disease associations and provide highly credible candidates for biological experiments.

Author Summary

The recognition of circRNA-disease association is the key of disease diagnosis and treatment, and it is of great significance for exploring the pathogenesis of complex diseases. Computational methods can predict the potential disease-related circRNAs quickly and accurately. Based on the hypothesis that circRNA with similar function tends to associate with similar disease, GCNCDA model is proposed to effectively predict the potential association between circRNAs and diseases by combining FastGCN algorithm. The performance of the model was verified by cross-validation experiments, different feature extraction algorithm and classifier models comparison experiments. Furthermore, 10 of the top 15 disease-associated circRNAs with the highest prediction scores were confirmed by recently published literature. It is anticipated that GCNCDA model can give priority to the most promising circRNA-disease associations on a large scale to provide reliable candidates for further biological experiment.

1. Introduction

As a new type of endogenous non-coding RNA, circular RNA (circRNA) has a closed-loop structure without a 5’and 3’polyadenylated tails [1-3]. As early as 1971, researchers discovered the viroids genome composed of single-stranded closed RNA molecules in potatoes [4]. In 1979, Hsu et al. [5] observed the presence of circRNA in the cytoplasm of eukaryotic cells by electron microscopy. In 1995, the researchers [6] found that the mouse sperm determinant gene Sry has circular transcription during transcription. But these findings did not attract much attention of researchers at the time. Until 2012, Salzman et al. [7] reported about 80 circRNAs for the first time with the help of high-throughput sequencing technology. Since then, a large number of circRNA molecules have been identified.

With the rapid development of bioinformatics and the continuous innovation of high-throughput sequencing technology, a large number of endogenous circRNA have been found in eukaryotic cells. CircRNA has the characteristics of universality, conservativeness, tissue-specificity and stability. Its unique sequence structure makes it have the functions of microRNA sponge [8], regulators of RNA binding proteins[9] and transcription of parental genes [10]. In addition, it is involved in the development and progression of diseases such as cancer [11,12], diabetes [13], nervous system diseases [14] and atherosclerosis [15]. For example, Burd et al. [16] found that the expression of cANRIL (circular antisense non-coding RNA in the INK4 locus) is an antisense transcript of INK4/ARF gene, which can inhibit the expression of INK4/ARF through specific multi comb family complex, thereby affecting the risk of atherosclerosis. Du et al. [17] found that circ-Foxo3, a member of the transcription factor foxo3, is highly expressed in myocardial samples from elderly patients and rats. It can prevent and reposition ID-1, E2F1, FAK and H1F1a in the cytoplasm and prevent their anti-aging function. By establishing the HT22 cell model of oxygen-glucose deprivation/reoxygenation (OGD/R), Lin et al. [18] found that the expression of mmu-circRNA-015947 was higher than that of normal cells,
indicating that the expression of circRNA was involved in OGD/R-induced neuron injury. Lukiw [19] found that in the hippocampal CA1 region of Alzheimer's disease (AD), there is a dysregulation of the miRNA-circRNA system. When the expression of CDRlas (CiRS-7) decreased or the ability to adsorb microRNA-7 weakened, the expression of miR-7 is increased and directly leads to down-regulation of ubiquitin ligase an expression in the human central nervous system, thereby affecting the normal function of the central nervous system and causing serious damage to brain tissue. Numerous studies have shown that circRNA can be a new clinical diagnostic marker or a potential target for human disease treatment. Therefore, the identification of disease-related circRNA may help to reveal the mechanism of disease occurrence and development, and further promote the understanding of complex human diseases.

As the number of detected circRNAs increases, multiple databases have been created to store information on circRNAs, such as Circ2Traits [20], circBase [21], deepBase [22] and CircNet [23]. Furthermore, researchers have gradually collected circRNA-disease associations supported by experiments and established databases, such as circR2Disease [24], circRNADb [25], circRNADisease [26] and Circ2Disease [27]. The accumulation of these data provides an opportunity for computational methods to predict potential circRNA-disease associations. For example, Xiao et al. [28] proposed an integrated computational framework called MRLDC to identify disease-associated circRNAs based on the hypothesis that circRNAs with similar functions are usually associated with similar diseases, and vice versa. Yan et al. [29] developed the DWNN-RLS method using Regularized Least Squares of Kronecker product kernel to predict circRNA-disease associations. In the experiment, this method achieved AUC of 0.8854, 0.9205 and 0.9701 in 5-fold CV, 10-fold CV and LOOCV, respectively. Fan et al. [30] proposed the KATZHCDA model for predicting circRNA-disease associations based on a heterogeneous network constructed by disease phenotype similarity, circRNA expression profiles and Gaussian interaction profile kernel similarity. As a result, KATZHCDA reached the AUC values of 0.7936 and 0.8469 in 5-fold cross-validation and LOOCV, respectively. Although the above models play important roles in the development of circRNA-disease association prediction computational methods and have achieved fruitful results, they are limited by certain problems: (1) the existing data are derived from incompletely related biological information, which cannot fully describe the complex association between circRNA and disease. (2) The experimentally verified circRNA-disease associations are limited in number and have some noise information, which easily leads to many false negative associations predicted by the model.

The purpose of this study is to propose a new computational model to predict the potential circRNA-disease associations in an attempt to overcome these problems. The proposed model GCNCDA has the following advantages: (1) Comprehensive use of disease semantic similarity information, disease GIP kernel similarity information, circRNA GIP kernel similarity information and known circRNA-disease association information to accurately predict potential circRNA-disease associations. (2) The advanced features of circRNA-disease associations are extracted by the deep learning FastGCN algorithm to reduce false negative associations and improve model performance. In the 5-fold cross-validation experiment on the benchmark dataset, GCNCDA achieved an AUC value of 90.90%. The results of comparative experiments show that GCNCDA is superior to other competing models and can effectively predict potential circRNA-disease associations. Furthermore, case studies show that GCNCDA can identify new circRNA-disease associations, which are validated by the latest literature. It is worth noting that
the performance of GCNCDA is underestimated due to experimentally verified limitations on the number of circRNA-disease associations.

2. Materials and Methods

2.1 Method overview

In this study, we propose a computational method called GCNCDA to predict potential circRNA-disease associations. The execution process of GCNCDA is divided into three steps, and its framework is shown in Figure 1. Firstly, we construct descriptor by synthesizing disease semantic similarity information, disease GIP kernel similarity information, circRNA GIP kernel similarity information and known circRNA-disease association information. Secondly, the FastGCN algorithm of deep learning is used to extract high-level features of descriptors and reduce noise information. Finally, the extracted high-level features are fed into Forest PA classifier to accurately predict the potential circRNA-disease associations.

Figure 1. The framework of GCNCDA to predict potential circRNA-disease associations

2.2 Benchmark Dataset

In this study, we used the recently established experimentally verified circRNA-disease association dataset circR2Disease [24] as the benchmark dataset to evaluate the performance of various models. CircR2Disease is a dedicated database and comprehensive platform that collects disease-related circRNAs from experimental support. The database currently hosts 739 entries from published literature, including 661 circRNAs, and 100 diseases. The benchmark dataset can be expressed as:

\[ \mathbb{R} = \mathbb{R}^+ \cup \mathbb{R}^- \]

where \( \cup \) denotes the union symbol in set theory, \( \mathbb{R}^+ \) represents the positive dataset, which contains 739 circRNA-disease associations with experimentally verified, \( \mathbb{R}^- \) represents the negative dataset, which contains 739 circRNA-disease associations without experimentally verified. The circR2Disease dataset can be available on the website http://bioinfo.snnu.edu.cn/CircR2Disease/.

In the circR2Disease dataset, there were a total of \( 661 \times 100 - 739 = 65361 \) circRNA-disease associations without experimental verified. If they are all treated as negative samples, they will form an unbalanced dataset. In order to avoid bias in the prediction results caused by unbalanced data, we solve this problem by reducing the number of negative samples by the down-sampling method. Specifically, we select 739 negative samples from all negative samples using random sampling without replacement, and then combine the positive samples to form a distributed equilibrium dataset. In theory, there may be unconfirmed circRNA-disease associations in these 65361 negative samples. But in the 739 negative samples we selected, this probability is much less than \( 739 \div (661 \times 100 - 739) \approx 1.13\% \). Thus, we constructed the dataset containing 1478 samples in this way, in which the number of positive samples is the same as that of negative samples. Known circRNA-disease associations and their names obtained from circR2Disease database can be seen in Supplementary Tables 1-3.

Based on the circR2Disease dataset, we constructed \( 661 \times 100 \) dimensional adjacency matrix \( AM \), where 661 represents the number of circRNAs, and 100 represents the number of diseases. When circRNA \( c(i) \) is associated with disease \( d(j) \), element \( AM(i,j) \) of matrix \( AM \) is assigned a value of 1. Otherwise, it is assigned a value of 0.
2.3 Construction of CircRNA Similarity Model

In this study, we used the Gaussian interaction profile (GIP) kernel similarity to construct the similarity model of circRNA. Based on the hypothesis that circRNAs with similar function are often associated with similar diseases, and vice versa, we established the GIP kernel similarity model of circRNA according to the known circRNA-disease association network. Specifically, we define the binary vector $V(c(i))$ to represent the interaction profiles of circRNA $c(i)$. The dimension of the vector $V(c(i))$ is 100, which corresponds to 100 diseases in adjacent matrix $AM$. When disease $c(i)$ is associated with one of 100 diseases, the corresponding bit in vector $V(c(i))$ is set to 1. Otherwise, it is set to 0. That is to say, the interaction profiles binary vector $V(c(i))$ is the row vector of the row corresponding to disease $c(i)$ in the adjacency matrix $AM$. Thus, we can get the circRNA GIP kernel similarity $GC(c(i),c(j))$ of circRNA $c(i)$ and circRNA $c(j)$:

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

where $\theta_c$ is the width parameter, which can be calculated using the normalized original parameters of the following formula:

$$\theta_c = \frac{1}{n} \sum_{i=1}^{n} \| V(c(i)) \|^2$$ (3)

where $n$ is the column number of adjacent matrix $AM$.

2.4 Construction of Disease Similarity Model

The disease similarity model consists of two parts: the disease GIP kernel similarity and the disease semantic similarity. For the disease GIP kernel similarity, our construction method is similar to the GIP kernel similarity of circRNA. More concretely, we define a binary vector $V(d(i))$ to represent the interaction profiles of disease $d(i)$ according to the adjacent matrix $AM$ provided by circR2Disease dataset. The dimension of the vector $V(d(i))$ is 661, which corresponds to 661 circRNAs in adjacent matrix $AM$. When disease $d(i)$ is associated with one of 661 circRNAs, the corresponding bit in vector $V(d(i))$ is set to 1. Otherwise, it is set to 0. That is to say, the interaction profiles binary vector $V(d(i))$ is the column vector of the column corresponding to disease $d(i)$ in the adjacency matrix $AM$. Through the above definition, we can calculate the disease GIP kernel similarity $GD(d(i),d(j))$ of disease $d(i)$ and disease $d(j)$:

$$GD(d(i),d(j)) = \exp \left( -\theta_d \| V(d(i)) - V(d(j)) \|^2 \right)$$ (4)

$$\theta_d = \frac{1}{m} \sum_{i=1}^{m} \| V(d(i)) \|^2$$ (5)

where $\theta_d$ is the width parameter and $m$ is the row number of adjacent matrix $AM$.

For disease semantic similarity, we construct it through the MeSH database [31-33] from the National Library of Medicine (NLM). It can be downloaded at https://www.nlm.nih.gov/. The MeSH database gives a rigorous disease classification system that uses a Directed Acyclic Graph (DAG) to reflect relationships between different diseases. The MeSH dataset can be seen in Supplementary Table 4. In DAG, a node represents disease, and an edge represents the relationship between diseases. Given a disease $d$ whose structure can be expressed as $DAG_d = (d,N_d,E_d)$, where $N_d$ represents the set of diseases associated with $d$ including disease $d$ itself, and $E_d$ represents the relationship between these diseases. For a disease $s$ within $DAG_d$, its contribution value $D_d(s)$ can be calculated by the following formula:

$$D_d(s) = \begin{cases} 1 & \text{if } s = d \\ \max \left\{ \mu \cdot D_d(s) \mid s \in \text{children of } s \right\} & \text{if } s \neq d \end{cases}$$
where $\mu$ indicates the semantic contribution factor between disease $s$ and its child disease $s'$. According to the previous study by Wang et al. [34], we set the semantic contribution factor $\mu$ to the optimal value of 0.5. Thus, by accumulating the contribution values of all children with disease $d$, we can get their semantic values $DV(d)$:

$$DV(d) = \sum_{s \in N_d} D_d(s)$$ (7)

In general, the more nodes that are shared between DAGs of different diseases, the more similar they are. Based on this assumption, we construct the first disease semantic similarity model $SV_1(d(i), d(j))$ of disease $d(i)$ and disease $d(j)$ through the DAG hierarchical relationship of disease:

$$SV_1(d(i), d(j)) = \frac{\sum_{s \in N_d(i) \cap N_d(j)} (d_{ai}(s) + d_{aj}(s))}{DV(d(i)) + DV(d(j))}$$ (8)

In disease semantic similarity model $SV_1$, we mainly consider the hierarchical relationship of disease DAG, that is, the disease in the same layer in the DAG contributes the same value to the disease $d$. However, the number of different diseases in DAGs can also affect the semantic similarity of disease. The fewer diseases appear in DAGs, the more important they are. Therefore, we constructed the second method for calculating the disease contribution value based on this hypothesis:

$$DS_i(s) = -\log\left(\frac{\text{num}(\text{DAGs}(s))}{\text{num}(\text{diseases})}\right)$$ (9)

where $\text{num}(\text{DAGs}(s))$ denotes the number of DAGs that contain disease $s$, and $\text{num}(\text{diseases})$ denotes the number of all diseases. Thus, the second disease semantic similarity model $SV_2(d(i), d(j))$ of disease $d(i)$ and disease $d(j)$ can be calculated as follows:

$$SV_2(d(i), d(j)) = \frac{\sum_{s \in N_d(i) \cap N_d(j)} (d_{ai}(s) + d_{aj}(s))}{DV(d(i)) + DV(d(j))}$$ (10)

where $DV(d(i))$ and $DV(d(j))$ have the same meaning as disease semantic similarity model $SV_1$, which can be calculated from formula 7.

2.5 Multi-source Data Fusion

In order to make full use of information from different sources, we used the fusion method to fuse circRNA similarity information and disease similarity information with known circRNA-disease associations. The fused information can absorb the characteristics of different data sources, thus describing the complex relationship between circRNAs and diseases more comprehensively.

For the circRNA, we use the constructed circRNA GIP kernel similarity $GR$ directly to represent the circRNA descriptor $RSim$. For the disease, we need to fuse the disease semantic similarity model $SV_1$ and $SV_2$, and disease GIP kernel similarity $GD$. Since the MeSH database provides a strict disease association, we use it as much as possible. More specifically, if there is the semantic similarity between disease $d(i)$ and disease $d(j)$, then the disease semantic similarity is used to construct the descriptor $DSim$. Otherwise, it is constructed using disease GIP kernel similarity. This construction rule can be described by the following formula:

$$DSim(d(i), d(j)) = \begin{cases} 
\text{if } SV_1(d(i), d(j)) \neq 0, & \text{use } SV_1(d(i), d(j)) \\
\text{otherwise, } & \text{use } SV_2(d(i), d(j))
\end{cases}$$
Finally, we match circRNA similarity $RSim$ with disease similarity $DSim$ based on known circRNA-disease associations to form a complete fusion descriptor. The fusion descriptor $FV((c(i),d(j)))$ of circRNA $c(i)$ and disease $d(j)$ can be described as follows:

$$FV((c(i),d(j))) = [RSim(i),DSim(j)]$$

where $RSim(i)$ indicates the $i$ row vector of circRNA $c(i)$ in the circRNA similarity matrix $RSim$, and $DSim(j)$ indicates the $j$ column vector of disease $d(j)$ in the disease similarity matrix $DSim$.

### 2.6 Feature Extraction by Fast Learning with Graph Convolutional Networks

After getting the fusion descriptors, we used the Fast learning with Graph Convolutional Networks (FastGCN) [35] algorithm to extract their features to remove noise information and improve the performance of the model. FastGCN is an efficient algorithm based on the original GCN and realized by importance sampling. It interprets graph convolutions as integral transforms of embedding functions under probability measure. To be specific, FastGCN interprets the graph vertices as independent and identically distributed (i.i.d.) samples of some probability distributions, and integrates loss and each convolution layer as vertex embedding functions. The integrals are then calculated by Monte Carlo approximation to determine the sample loss and sample gradient. Finally, important sampling is used to reduce the approximate variance. FastGCN not only eliminates the reliance on test data but also produces a controllable cost for each batch of computation.

Suppose there is a graph $G$ with the vertex set $V$ associated with a probability space $(V,F,P)$. For the given graph $G$, it is a subgraph of $G$ whose vertices are i.i.d. samples of $V$ obtained from the probability measure $P$. For the probability space, $V$ is used as the sample space, and $F$ can be any event space. The probability measure $P$ defines a sample distribution. Thus, the function generalization can be expressed as:

$$h^{(l+1)}(v) = \int \tilde{A}(v,u)h^{(l)}(u)W^{(l)}dP(u), \quad h^{(l+1)}(v) = \sigma(h^{(l+1)}(v)), \quad l = 0,\ldots,M-1$$

where the function $h^{(l)}$ represents an embedding function from the $l$th layer, $u$ and $v$ are independent random variables that have the same probability measure $P$. The embedding functions of two consecutive layers are correlated by convolution and expressed by an integral transform, where the kernel $\tilde{A}(v,u)$ corresponds to the $(v,u)$ element of the matrix $\tilde{A}$. The loss $L$ is the expected value of $g(h^{(M)})$ that is finally embedded in $h^{(M)}$, and can be expressed as:

$$L = E_{v\sim P}[g(h^{(M)})(v)] = \int g(h^{(M)})(v)dP(v)$$

For the $l$th layer, the $t_i$ i.i.d. sample $u_{t_i}^{(0)},\ldots,u_{t_i}^{(0)} \sim P$ is used to approximatively estimate the integral transformation:

$$h^{(l+1)}_{t_{i+1}}(v) = \frac{1}{t_i} \sum_{j=1}^{t_i} \tilde{A}(v,u_j)h^{(l)}_{t_i}(u_j)W^{(l)}, \quad h^{(l+1)}_{t_{i+1}}(v) = \sigma(h^{(l+1)}_{t_{i+1}}(v)), \quad l = 0,\ldots,M-1$$

Here, $h^{(0)}_{t_0}$ is $h^{(0)}$. Therefore, the loss $L$ is transformed into:

$$L_{t_0,t_1,\ldots,t_M} = \frac{1}{t_M} \sum_{i=1}^{t_M} g(h^{(M)}_{t_i}((u_{t_i}^{(M)})))$$

### 2.7 Prediction by Forest PA Classifier
In the experiment, we send the extracted features into the Forest by Penalizing Attributes (Forest PA) classifier for classification, so as to obtain accurate circRNA-disease association prediction results. Forest PA is a novel decision forest building algorithm recently proposed by Adnan et al. [36]. The Forest PA algorithm uses the complete attribute set to generate decision trees by imposing penalties on attributes participating in the latest decision tree. Besides, the participating attributes obtain random weights from the range of weights associated with the respective levels in the tree, thereby maintaining the decision tree generated by the algorithm with individually accuracy and diversity. The execution steps of the Forest PA algorithm are as follows:

1. The Forest PA first generates a bootstrap sample \( D_i \) from the original training data set \( D \).

2. The Forest PA then uses the weight of attributes to generate decision trees from the bootstrap sample. When choosing the splitting attributes, Forest PA uses the CART algorithm with merit values, whose value is obtained by multiplying its classification ability with its weight.

3. The incremental values of attribute weights and gradient weight in the latest tree are updated iteratively. Here, the weights of the attributes appear in the latest tree will be updated. The weights of attributes that do not appear in the latest tree remain unchanged. Considering that the weight of attribute is determined by the level \( \lambda \) of test attributes in the latest tree, if an attribute appears on the root node, their value of \( \lambda \) is 1; if an attribute appears on the child node, their value of \( \lambda \) is 2. According to the value of \( \lambda \), the weight of randomly generated attributes within a Weight-Range \( WR \) is defined as follows:

\[
WR^{\lambda} = \begin{cases} 
0.0, & \text{if } \lambda = 1 \\
\left[ e^{-\frac{1}{\lambda - 1}}, e^{\frac{1}{\lambda}} \right], & \text{if } \lambda > 1 
\end{cases}
\]  

(17)

4. Update weights of the applicable attributes with the corresponding weight increment values that do not exist in the latest tree.

3. Results and discussion

3.1 Evaluation Criteria

In this study, we used the 5-fold cross-validation (5-fold CV) method to evaluate the performance of the model. This method can not only reduce over-fitting to a certain extent but also obtain as much effective information as possible from limited data [37]. More concretely, we first randomly divide the initial dataset into five sub-data sets. When the method is executed, a separate sub-data set is reserved for validating the model and the other four sub-data sets are used to train the model. This process is repeated 5 times until each sub-data set is verified once and only verified once. Finally, the average results of these 5 times are used as the performance indicators of the model. General evaluation criteria are used in this study to evaluate the performance of GCNCDA, including accuracy (Accu.), Sensitivity (Sen.), precision (Prec.), F1-Score (F1) and Matthews Correlation Coefficient (MCC). They are defined as:

\[
Accu. = \frac{TP + TN}{TP + FN + FP + FN} 
\]  

(18)

\[
Sen. = \frac{TP}{FP + FN} 
\]  

(19)
\[ \text{Prec.} = \frac{TP}{TP + FP} \quad (20) \]
\[ F1 = \frac{2TP}{2TP + FP + FN} \quad (21) \]
\[ MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (22) \]

Here, TP means true positive, TN means true negative, FP means false positive, and FN means false negative. Furthermore, we also plot the Receiver Operating Characteristic (ROC) \([38,39]\) curves of the 5-fold CV generated by GCNCDA and calculate their average area under the ROC curve (AUC) \([40]\).

### 3.2 Model Performance Evaluation

In the experiment, GCNCDA is implemented on the benchmark dataset circR2Disease to evaluate its ability to predict potential circRNA-disease associations. The detailed results of 5-fold CV are summarized in Table 1. As can be seen from the table, GCNCDA achieved an average accuracy of 91.20% and a standard deviation of 0.74%, of which the accuracy of 5-fold experiments was 91.86%, 91.19%, 90.85%, 90.17% and 91.95%, respectively. In terms of accuracy, sensitivity, precision, F1-Score, Matthews correlation coefficient and area under ROC curve, GCNCDA obtained 92.78%, 90.03%, 91.33%, 82.55% and 90.90%, with standard deviations of 3.03%, 2.37%, 0.78%, 1.60% and 0.81%, respectively. Figure 2 plots the ROC curve generated by GCNCDA using 5-fold CV on the circR2Disease dataset. From the experimental results, we can observe that GCNCDA performs well and can effectively predict the potential disease-related circRNAs.

| Table 1. Results of 5-fold CV generated by GCNCDA on circR2Disease dataset |
| Figure 2. ROC curves of 5-fold CV obtained by GCNCDA on circR2Disease dataset |

### 3.3 Comparison of Different Classifier Models

To evaluate the impact of the Forest PA classifier on the overall performance of GCNCDA, we compared different classifier models in this experiment. Specifically, when constructing different classifier models, we keep the other parts of the model unchanged, including the composition of descriptors and feature extraction, and only replace the Forest PA classifier with state-of-the-art Support Vector Machine (SVM) and Random Forest (RF) classifiers, respectively. The SVM model and the RF model are thus constructed and implemented on the circR2Disease dataset using 5-fold CV. Table 2 lists the results of the 5-fold CV experiments performed by these two models. Figure 3 plots the 5-fold CV ROC curves generated by the two models on the circR2Disease dataset. For the convenience of visual comparison, we display these results in the form of a histogram. As can be seen from figure 4, GCNCDA achieved the best results in accuracy, sensitivity, F1, MCC and AUC, and achieved the third result in precision, but only 2.75% lower than the best result. From the overall performance point of view, GCNCDA is better than SVM model and RF model. This result indicates that the Forest PA classifier is suitable for GCNCDA model and contributes to the improvement of the model performance.

| Table 2. Results of 5-fold CV generated by SVM model and RF model on circR2Disease dataset |
| Figure 3. ROC curves of 5-fold CV obtained by SVM model on circR2Disease dataset |
| Figure 4. ROC curves of 5-fold CV obtained by RF model on circR2Disease dataset |
Figure 5. Comparison of results of different classifier models on circR2Disease dataset

3.4 Comparison of Different Feature Extraction Algorithms

In order to evaluate the effect of the FastGCN feature extraction algorithm on the overall performance of GCNCDA, we compared different feature extraction algorithm models in this experiment. Similar to the experiment with different classifiers, when we construct different feature extraction algorithm models, the other parts of the model are unchanged, including the composition of the descriptors and classifier. Only the Auto Covariance (AC) [41] and fast Fourier transform (FFT) [42] extraction algorithms are used instead of the FastGCN algorithm. The AC model and the FFT model are thus constructed and implemented on the circR2Disease dataset using 5-fold CV. Table 3 summarizes the results of the 5-fold CV obtained by the two models. Figure 5 plots the 5-fold CV ROC curves generated by the two models on the circR2Disease dataset. Similarly, we used a histogram to visually compare the results of the three models. As can be seen from figure 6, GCNCDA achieved the best results in all the evaluation criteria, including accuracy, sensitivity, precision, F1, MCC and AUC. The experimental results show that the FastGCN algorithm can effectively extract the advanced features of the fusion descriptor, thus helping to improve the performance of the model. In addition, from the comparison experiments of different classifiers and extraction algorithms, we can also see that the FastGCN algorithm is more helpful to the performance improvement of the model than the Forest PA classifier. This suggests that the FastGCN algorithm is the key to the GCNCDA model and plays an important role in predicting potential disease-associated circRNAs.

Table 3. Results of 5-fold CV generated by AC model and FFT model on circR2Disease dataset

Figure 6. ROC curves of 5-fold CV obtained by AC model on circR2Disease dataset

Figure 7. ROC curves of 5-fold CV obtained by FFT model on circR2Disease dataset

Figure 8. Comparison of results of different classifier models on circR2Disease dataset

3.5 Comparison with other existing methods

At present, some researchers have established models for predicting circRNA-disease associations based on the benchmark dataset circR2Disease, including DWNN-RLS [29], KATZHCDA [30], PWCDA [43], GHICD [43] and RWRHCD [43]. To evaluate the performance of GCNCDA, we compared it to the 5-fold CV AUC results of these models. Table 4 summarizes the 5-fold CV AUC scores generated by the various models on the same benchmark dataset circR2Disease. From the table we can see that GCNCDA is outperforms other existing methods. This indicates that the GCNCDA model, which uses the FastGCN algorithm to extract circRNA and disease fusion information features and combines the Forest PA classifier, can effectively improve the predictive performance of circRNA-disease associations.

Table 4. The 5-fold CV AUC scores generated by the various models on the same benchmark dataset circR2Disease

3.6 Case Studies

To demonstrate the capability of GCNCDA to predict new disease-associated circRNAs based on known circRNA-disease associations, the performance of GCNCDA was further evaluated. In particular, all known circRNA-disease associations in benchmark dataset $\mathbb{R}^+$ were used to train GCNCDA, and the remaining unknown circRNA-disease associations were
considered candidates for testing. Then all the candidates were ranked according to the predicted score of GCNCDA. Finally, the predicted disease-circRNA associations were confirmed in the latest published literature. As a result, 10 of the top 15 circRNA-disease associations with the highest predicted scores were confirmed by relevant literature, and the detailed results are listed in table 5. It is worth noting that although some circRNA-disease associations with high predicted scores are not confirmed by existing literature, there is no denying the possibility of an association between them.

Table 5. Top 15 circRNA-disease associations were predicted by GCNCDA based on known circRNA-disease associations in circR2Disease database

4. Conclusion
In this study, we proposed a new computational method called GCNCDA to predict potential circRNA-disease associations. The method makes full use of the disease semantic similarity, disease and circRNA GIP kernel similarity, the known circRNA-disease association information, and extracts the high-level abstract features from them by deep learning FastGCN algorithm. The cross-validation results show that GCNCDA performs well on the benchmark dataset circR2Disease. In comparison with different classifier models, feature extraction algorithm models, and other state-of-the-art methods, GCNCDA has exhibited strong competitiveness. Furthermore, we also predicted new circRNA-disease associations based on known associations. As a result, 10 of the top 15 circRNA-disease associations with the highest predicted scores were confirmed by recently published literature. These experimental results indicate that GCNCDA is an effective method for predicting circRNA-disease associations and can provide highly reliable candidates for biological experiments. In future research, we will improve the FastGCN algorithm to help the model achieve better performance.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This work is supported in part by Awardee of the NSFC Excellent Young Scholars Program, under Grants 61722212, in part by the National Natural Science Foundation of China, under Grants 61702444, 61572506, in part by the Pioneer Hundred Talents Program of Chinese Academy of Sciences, 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-XBQNWXZ-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.

References


Supporting Information Legends

Supplementary Table 1. The benchmark dataset contains 739 pairs of positive samples and 739 pairs of negative samples.

Supplementary Table 2. Names of 661 circRNAs involved in known circRNA-disease associations obtained from CircR2Disease database.

Supplementary Table 3. Names of 100 diseases involved in known circRNA-disease associations obtained from CircR2Disease database.

Supplementary Table 4. The MeSH dataset that provides rigorous disease classification information.

Tables

Table 1. Results of 5-fold CV generated by GCNCDA on circR2Disease dataset

<table>
<thead>
<tr>
<th>Test set</th>
<th>Accu. (%)</th>
<th>Sen. (%)</th>
<th>Prec. (%)</th>
<th>F1 (%)</th>
<th>MCC (%)</th>
<th>AUC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>91.86</td>
<td>88.97</td>
<td>94.16</td>
<td>91.49</td>
<td>83.83</td>
<td>90.93</td>
</tr>
<tr>
<td>2</td>
<td>91.19</td>
<td>93.10</td>
<td>89.40</td>
<td>91.22</td>
<td>82.45</td>
<td>90.54</td>
</tr>
<tr>
<td>3</td>
<td>90.85</td>
<td>92.52</td>
<td>89.47</td>
<td>90.97</td>
<td>81.74</td>
<td>91.24</td>
</tr>
<tr>
<td>4</td>
<td>90.17</td>
<td>91.95</td>
<td>88.96</td>
<td>90.43</td>
<td>80.38</td>
<td>89.80</td>
</tr>
<tr>
<td>5</td>
<td>91.95</td>
<td>97.39</td>
<td>88.17</td>
<td>92.55</td>
<td>84.33</td>
<td>92.00</td>
</tr>
<tr>
<td>Average</td>
<td>91.20±0.74</td>
<td>92.78±3.03</td>
<td>90.03±2.37</td>
<td>91.33±0.78</td>
<td>82.55±1.60</td>
<td>90.90±0.81</td>
</tr>
</tbody>
</table>

Table 2. Results of 5-fold CV generated by SVM model and RF model on circR2Disease dataset

<table>
<thead>
<tr>
<th>Test set</th>
<th>Accu. (%)</th>
<th>Sen. (%)</th>
<th>Prec. (%)</th>
<th>F1 (%)</th>
<th>MCC (%)</th>
<th>AUC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>86.10</td>
<td>78.62</td>
<td>91.94</td>
<td>84.76</td>
<td>72.87</td>
<td>84.30</td>
</tr>
<tr>
<td>2</td>
<td>86.78</td>
<td>77.93</td>
<td>94.17</td>
<td>85.28</td>
<td>74.56</td>
<td>85.54</td>
</tr>
<tr>
<td>3</td>
<td>87.46</td>
<td>83.67</td>
<td>90.44</td>
<td>86.93</td>
<td>75.12</td>
<td>88.49</td>
</tr>
<tr>
<td>4</td>
<td>87.12</td>
<td>83.22</td>
<td>90.51</td>
<td>86.71</td>
<td>74.50</td>
<td>87.96</td>
</tr>
<tr>
<td>5</td>
<td>87.25</td>
<td>88.24</td>
<td>87.10</td>
<td>87.66</td>
<td>74.48</td>
<td>88.50</td>
</tr>
<tr>
<td>SVM Model</td>
<td>86.94±0.53</td>
<td>82.34±4.20</td>
<td>90.83±2.58</td>
<td>86.27±1.21</td>
<td>74.31±0.84</td>
<td>86.96±1.92</td>
</tr>
<tr>
<td>1</td>
<td>88.14</td>
<td>82.07</td>
<td>92.97</td>
<td>87.18</td>
<td>76.73</td>
<td>87.37</td>
</tr>
<tr>
<td>2</td>
<td>90.17</td>
<td>84.83</td>
<td>94.62</td>
<td>89.45</td>
<td>80.72</td>
<td>89.08</td>
</tr>
<tr>
<td>3</td>
<td>91.19</td>
<td>87.07</td>
<td>94.81</td>
<td>90.78</td>
<td>82.64</td>
<td>90.41</td>
</tr>
<tr>
<td>4</td>
<td>89.15</td>
<td>87.92</td>
<td>90.34</td>
<td>89.12</td>
<td>78.34</td>
<td>88.85</td>
</tr>
<tr>
<td>5</td>
<td>89.26</td>
<td>87.58</td>
<td>91.16</td>
<td>89.33</td>
<td>78.59</td>
<td>89.54</td>
</tr>
<tr>
<td>RF Model</td>
<td>89.58±1.15</td>
<td>85.89±2.46</td>
<td>92.78±2.01</td>
<td>89.17±1.29</td>
<td>79.41±2.30</td>
<td>89.05±1.11</td>
</tr>
<tr>
<td>----------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td>GCNCDA</td>
<td>91.20±0.74</td>
<td>92.78±3.03</td>
<td>90.03±2.37</td>
<td>91.33±0.78</td>
<td>82.55±1.60</td>
<td>90.90±0.81</td>
</tr>
</tbody>
</table>

**Table 3.** Results of 5-fold CV generated by AC model and FFT model on circR2Disease dataset

<table>
<thead>
<tr>
<th>Test set</th>
<th>Accu. (%)</th>
<th>Sen. (%)</th>
<th>Prec. (%)</th>
<th>F1 (%)</th>
<th>MCC (%)</th>
<th>AUC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>86.44</td>
<td>93.24</td>
<td>82.14</td>
<td>87.34</td>
<td>73.55</td>
<td>85.39</td>
</tr>
<tr>
<td>2</td>
<td>85.08</td>
<td>90.67</td>
<td>81.93</td>
<td>86.08</td>
<td>70.52</td>
<td>85.83</td>
</tr>
<tr>
<td>3</td>
<td>81.36</td>
<td>86.71</td>
<td>77.50</td>
<td>81.85</td>
<td>63.23</td>
<td>81.78</td>
</tr>
<tr>
<td>4</td>
<td>85.76</td>
<td>90.13</td>
<td>83.54</td>
<td>86.71</td>
<td>71.67</td>
<td>85.74</td>
</tr>
<tr>
<td>5</td>
<td>91.95</td>
<td>97.95</td>
<td>87.20</td>
<td>92.26</td>
<td>84.54</td>
<td>93.16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ACModel</th>
<th>86.12±3.81</th>
<th>91.74±4.18</th>
<th>82.46±3.48</th>
<th>86.85±3.71</th>
<th>72.70±7.69</th>
<th>86.38±4.15</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>73.90</td>
<td>76.32</td>
<td>73.89</td>
<td>75.08</td>
<td>47.72</td>
<td>73.59</td>
</tr>
<tr>
<td>2</td>
<td>75.93</td>
<td>75.52</td>
<td>75.00</td>
<td>75.26</td>
<td>51.83</td>
<td>76.38</td>
</tr>
<tr>
<td>3</td>
<td>74.24</td>
<td>68.94</td>
<td>81.02</td>
<td>74.50</td>
<td>49.46</td>
<td>73.96</td>
</tr>
<tr>
<td>4</td>
<td>78.64</td>
<td>76.92</td>
<td>75.19</td>
<td>76.05</td>
<td>56.80</td>
<td>79.21</td>
</tr>
<tr>
<td>5</td>
<td>78.19</td>
<td>82.35</td>
<td>76.83</td>
<td>79.50</td>
<td>56.41</td>
<td>76.74</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FFT Model</th>
<th>76.18±2.19</th>
<th>76.01±4.78</th>
<th>76.38±2.80</th>
<th>76.08±1.99</th>
<th>52.44±4.07</th>
<th>75.98±2.29</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCNCDA</td>
<td>91.20±0.74</td>
<td>92.78±3.03</td>
<td>90.03±2.37</td>
<td>91.33±0.78</td>
<td>82.55±1.60</td>
<td>90.90±0.81</td>
</tr>
</tbody>
</table>

**Table 4.** The 5-fold CV AUC scores generated by the various models on the same benchmark dataset circR2Disease

<table>
<thead>
<tr>
<th>Methods</th>
<th>GCNCDA</th>
<th>DWNN-RLS</th>
<th>KATZHCDAD</th>
<th>PWCDA</th>
<th>GHICD</th>
<th>RWRHCD</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>90.90</td>
<td>88.54</td>
<td>79.36</td>
<td>89.00</td>
<td>72.90</td>
<td>66.60</td>
</tr>
</tbody>
</table>

**Table 5.** Top 15 circRNA-disease associations were predicted by GCNCDA based on known circRNA-disease associations in circR2Disease database

<table>
<thead>
<tr>
<th>circRNA</th>
<th>Disease</th>
<th>Evidence (PMID)</th>
<th>Pub Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa_circ_0004214</td>
<td>cervical cancer</td>
<td>28622299</td>
<td>2017 Sep</td>
</tr>
<tr>
<td>hsa_circ_0004214</td>
<td>glioma</td>
<td>28622299</td>
<td>2017 Sep</td>
</tr>
<tr>
<td>hsa_circ_00072088</td>
<td>liver cancer</td>
<td>28727484</td>
<td>2017 Jul 20</td>
</tr>
<tr>
<td>hsa_circ_0013958</td>
<td>lung cancer</td>
<td>28685964</td>
<td>2017 Jul</td>
</tr>
<tr>
<td>hsa_circ_0021001</td>
<td>Heart disease</td>
<td>unconfirmed</td>
<td>unconfirmed</td>
</tr>
<tr>
<td>circETFA</td>
<td>Glioma</td>
<td>26873924</td>
<td>2016 May</td>
</tr>
<tr>
<td>hsa_circ_0070934</td>
<td>Dilated cardiomyopathy</td>
<td>unconfirmed</td>
<td>unconfirmed</td>
</tr>
<tr>
<td>hsa_circ_0001946</td>
<td>myocardial infarction</td>
<td>26998750</td>
<td>2016 Mar 21</td>
</tr>
<tr>
<td>hsa_circ_0001946</td>
<td>glioma</td>
<td>26683098</td>
<td>2016 Jan 26</td>
</tr>
<tr>
<td>hsa_circ_0001955</td>
<td>Immunosenescence</td>
<td>unconfirmed</td>
<td>unconfirmed</td>
</tr>
<tr>
<td>hsa_circ_0004458</td>
<td>gastric cancer</td>
<td>28544609</td>
<td>2017 Jun</td>
</tr>
<tr>
<td>hsa_circ_0005442</td>
<td>Dilated cardiomyopathy</td>
<td>unconfirmed</td>
<td>unconfirmed</td>
</tr>
<tr>
<td>circRNA_010567</td>
<td>Cardiovascular disease</td>
<td>28412345</td>
<td>2017 Jun</td>
</tr>
<tr>
<td>circZFR</td>
<td>Bladder cancer</td>
<td>27484176</td>
<td>2016 Aug</td>
</tr>
<tr>
<td>hsa_circ_0037911</td>
<td>Heart disease</td>
<td>unconfirmed</td>
<td>unconfirmed</td>
</tr>
</tbody>
</table>
Figure Legend

**Figure 1.** The framework of GCNCD to predict potential circRNA-disease associations

**Figure 2.** ROC curves of 5-fold CV obtained by GCNCD on circR2Disease dataset

**Figure 3.** ROC curves of 5-fold CV obtained by SVM model on circR2Disease dataset

**Figure 4.** ROC curves of 5-fold CV obtained by RF model on circR2Disease dataset

**Figure 5.** Comparison of results of different classifier models on circR2Disease dataset

**Figure 6.** ROC curves of 5-fold CV obtained by AC model on circR2Disease dataset

**Figure 7.** ROC curves of 5-fold CV obtained by FFT model on circR2Disease dataset

**Figure 8.** Comparison of results of different classifier models on circR2Disease dataset
A diagram illustrating the process of combining disease and circRNA similarity networks to predict potential associations. The process involves:

1. **Adjacency Matrix**: $AM \in \mathbb{R}^{C \times D}$

2. **Disease Semantic Similarity**: $SV_1$ + $SV_2$ + $GD$

3. **CircRNA GIP Kernel Similarity**: $GC$

4. **Combined Disease Similarity**: $DSim$

5. **Combined circRNA Similarity**: $RSim$

The diagram also shows the input of predicted potential circRNA-disease associations, forest PA classifier, FastGCN, and fusion descriptor.
Figure 2

The figure shows the Receiver Operating Characteristic (ROC) curves for different folds of the data. The curves represent the sensitivity (true positive rate) against 1-specificity (false positive rate) for each fold. The inset highlights the average Area Under the Curve (AUC) for all folds, which is calculated as 0.9090.
Figure 3

The graph shows the performance of different folds, with the average AUC calculated as 0.8696.
Figure 4
Figure 5

The bar chart compares the prediction results of different models on various metrics. The x-axis represents different metrics: Accuracy (Accu.), Sensitivity (Sen.), Precision (Prec.), F1 score (F1), Matthews Correlation Coefficient (MCC), and Area Under the Curve (AUC). The y-axis shows the prediction results in percentage terms.

- GCNCDA model: Light blue bars
- SVM model: Red bars
- RF model: Yellow bars

The metrics are as follows:
- Accuracy (Accu.)
- Sensitivity (Sen.)
- Precision (Prec.)
- F1 score (F1)
- Matthews Correlation Coefficient (MCC)
- Area Under the Curve (AUC)
Figure 6

Average AUC = 0.8638
Figure 7

The graph shows the Receiver Operating Characteristic (ROC) curves for different folds. The curves indicate the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) for various thresholds. Each curve corresponds to a different fold:
- 1st fold (green)
- 2nd fold (yellow)
- 3rd fold (blue)
- 4th fold (red)
- 5th fold (black)

The inset highlights the average area under the curve (AUC) for all folds, which is given as 0.7598.
Figure 8