Gene 762 (2020) 145040

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

An ensemble approach for CircRNA-disease association prediction based on autoencoder and deep neural network

K. Deepthi[a,b,*], A.S. Jereesh[a]

[a] Bioinformatics Lab, Department of Computer Science, Cochin University of Science and Technology, Kochi 682022, Kerala, India
[b] Department of Computer Science, College of Engineering, Vadakara, Kozhikkode 673104, Kerala, India

ARTICLE INFO

Keywords:
Circular RNA
Autoencoder
Neural network
Association
Disease

ABSTRACT

Circular RNAs (circRNA) are a special kind of covalently closed single-stranded RNA molecules. They have been shown to control and coordinate various biological processes. Recent researches show that circRNAs are closely associated with numerous chronic human diseases. Identification of circRNA-disease associations will contribute towards diagnosing the pathogenesis of diseases. Experimental methods for finding the relation between the diseases and their causal circRNAs are difficult and time-consuming. So computational methods are of critical need for predicting the associations between circRNAs and various human diseases. In this study, we propose an ensemble approach AE-DNN, which relies on autoencoder and deep neural networks to predict new circRNA-disease relationships. We utilized circRNA sequence similarity, disease semantic similarity, and Gaussian interaction profile kernel similarities of circRNAs and diseases for feature construction. The constructed features are fed to a deep autoencoder, and the extracted compact, high-level features are fed to the deep neural network for association prediction. We conducted 5-fold and 10-fold cross-validation experiments to assess the performance; AE-DNN could achieve AUC scores of 0.9392 and 0.9431, respectively. Experimental results and case studies indicate the robustness of our model in circRNA-disease association prediction.

1. Introduction

CircRNAs are single-stranded RNAs in which the $3'$ and $5'$ termini are covalently bound by back-splicing of exons from a single pre-mRNA. This back -splicing gives circular structure to the circRNA molecule (Greene et al., 2017; Yu and Kuo, 2019). CircRNAs were discovered in the 1970s, but they got attention in most recent years (Sanger et al., 1976; Chen et al., 2016; Zhang et al., 2018). A considerable number of circRNAs have been identified with the advancements in RNA sequencing technologies. CircRNAs are involved in various biological processes (Fang, 2018). Recently, they have been identified as biomarkers and treatment targets for various acute diseases. CircRNAs are associated with multiple chronic diseases, like lung cancer, Alzheimer's disease, diabetes mellitus, cardiovascular diseases, etc. (Lukiw, 2013; Greene et al., 2017). Recently, databases depicting circRNA-disease relationships have been proposed, CircR2Disease (Fan et al., 2018a), circRNADisease (Zhao et al., 2018), and Circ2Disease (Yao et al., 2018), which facilitates research relating to circRNAs and diseases.

Numerous approaches have been proposed to find non-coding RNA-disease relationships (Esteller, 2011; Wang et al., 2013; Nacher and Akutsu, 2019). Several studies have been conducted to discover associations of diseases with long non-coding RNAs and micro RNAs (Zhang et al., 2017; Lu et al., 2018; Fu and Peng, 2017; Peng et al., 2019; Chen et al., 2019). However, very few studies are conducted to reveal potential associations between circRNAs and diseases. The studies pertained to circRNA-disease association predictions can be classified as network-oriented and machine-learning oriented approaches, considering the methods they used.

Network-based approaches construct heterogeneous networks by considering circRNA and disease similarities and existing associations between circRNA and diseases. A path-based computational method for circRNA-disease association prediction, was proposed by (Lei et al., 2018). The method calculates a score for each circRNA and disease pairs, depending on the paths associated with them in the constructed network. The method proposed by (Fan et al., 2018b), constructed a heterogeneous network by integrating circRNA and disease similarities and existing circRNA-disease relations. They used KATZ measures (Katz, 1953) for predicting circRNA-disease associations. A

Abbreviations: circRNA, Circular RNA; ROC, Receiver Operating Characteristic Curve; AUC, Area Under the ROC curve; AUPR, Area Under the Precision Recall curve

[*] Corresponding author.
E-mail addresses: deepthi523@gmail.com (K. Deepthi), jereesh@cusat.ac.in (A.S. Jereesh).

https://doi.org/10.1016/j.gene.2020.145040
Received 26 May 2020; Received in revised form 28 June 2020; Accepted 4 August 2020
Available online 07 August 2020

computational approach based on matrix-factorization, was proposed by (Wei and Liu, 2019), to identify novel circRNA-disease associations. They measured circRNA similarity and disease similarity by considering semantic disease data, existing circRNA-gene, gene-disease, and circRNA-disease associations. (Li et al., 2020) applied an inductive matrix completion-based algorithm for circRNA-disease association prediction. The model proposed by (Ge et al., 2020), predicted circRNA-disease associations with the help of confirmed circRNA-disease relations, circRNA, disease semantic similarity networks and reconstructed circRNA and disease similarity networks. They applied Locality-Constrained Linear Coding (LLC) on the verified interaction matrix and cosine similarities of circRNAs and diseases for reconstruction.

The machine learning-based approaches constructed features for classification, utilizing circRNA, disease similarities, and existing associations. The model proposed by (Yan et al., 2018), used regularized least-squares of Kronecker product kernel for finding out novel circRNA-disease relationships. They used a decreasing weight k-nearest neighbour approach to find the association score for novel circRNAs and diseases. (Wang et al., 2019) proposed a convolutional neural network-based method for circRNA-disease association prediction. They extracted hidden features from multi-source circRNA and disease similarities with the convolutional neural network's help. These extracted features were then fed to the extreme learning machine classifier (ELM), for predicting novel circRNA-disease relations. Depending on random walk with restart and k-nearest neighbors (KNN), (Lei and Bian, 2020) proposed a computational method, for predicting new relations among circRNAs and diseases. They constructed weighted features with the help of random walk with restart method and circRNA, disease similarity information. With the constructed features, they trained the KNN classifier for circRNA-disease relation prediction.

Through this study, we propose an ensemble machine learning approach relied on autoencoders and deep neural networks to predict novel circRNA-disease relationships. Ensemble approaches use more than one learning algorithms to achieve better results. To the best of our knowledge, autoencoder-based feature selection is applying the first time for circRNA-disease association predictions. We utilized circRNA, disease pairwise similarities, and experimentally confirmed circRNA-disease relationships for novel circRNA-disease relation predictions. CircRNA pairwise similarities include sequence and Gaussian interaction profile kernel similarities of circRNAs. Disease pairwise similarities include semantic and Gaussian interaction profile kernel similarities of diseases. Depending on these similarity measurements, features are constructed, and the features are fed to a deep autoencoder. The high-level features with reduced dimensions, from autoencoder output, are used to train a three-layer deep neural network. The trained deep neural network predicts new circRNA-disease associations with corresponding probabilities as classification results. When the predicted probability is above the predefined threshold, the association is said to exist. Since the approach combines the results from multiple learning algorithms, the model performance is robust. The sequence similarities represent circRNAs more deeply. The proposed method could achieve AUC scores of 0.9392 and 0.9431 in 5-fold and 10-fold cross-validation experiments, respectively. We compared our model with previous studies and different classifiers such as SVM, decision tree, and deep neural network without autoencoder. We implemented the method with different datasets. The experimental results and case studies showed our method outperformed previous approaches.

2. Materials and methods

2.1. Datasets

A short description of the datasets used in our study is described below.

CircRNA-disease associations: We acquired the circRNA-disease association data from the CircR2Disease database http://bioinfo.snnu.edu.cn/CircR2Disease/, which is used as the benchmark dataset in circRNA-disease association predictions. The database contains 739 experimentally confirmed circRNA-disease associations between 661 circRNAs and 100 diseases. The circRNAs without sequence information and diseases without DOID (Disease Ontology ID) were removed. Finally, after preprocessing and restricting the data to humans, we obtained 445 associations between 389 circRNAs and 61 diseases. Based on these associations, we constructed a circRNA-disease association matrix where rows represent circRNAs, and columns represent diseases. If there exists an experimentally confirmed association between the circRNA and disease, the corresponding entry in the matrix is set to 1, otherwise it is set to 0.

CircRNA-CircRNA similarities: The circRNA sequence information used in our study has been obtained from the database circBase, which can be downloaded at http://www.circbase.org/. The sequence similarities between the two circRNAs were calculated based on the Levenshtein distance (Levenshtein, 1966), which is a widely used measure of string metric. The Levenshtein distance between two strings is measured as the minimum cost for single-character changes (insertions, deletions, or replacements) needed to convert one string to the other. We set the editing costs for insertion and deletion to 1 and the substitution cost to 2 in our study. The sequence similarity $CSS(c_i, c_j)$ between two circRNAs c_i and $c_j$is calculated based on the Eq. (1).

$$CSS(c_i, c_j) = 1 - \frac{x}{l(c_i) + l(c_j)} \tag{1}$$

Here × denotes the minimum cost required to change circRNA sequence c_i to c_j, and *l(.)* indicates the length of the sequence.

Disease-disease similarities: Diseases can be represented as directed acyclic graphs (DAG), relied on its semantic associations of terms in the ontology. The DAG nodes denote diseases, and edges indicate the relations between the diseases. The semantic similarity score between two diseases depends on their relative positions in the disease DAG. Diseases with more shared parts in the DAG tend to be semantically more similar.

The semantic similarity scores were calculated based on the method suggested by (Wang et al., 2007). The semantic similarity score $DSS(d_i, d_j)$, between two diseases *di*and d_j can be calculated using Equation (2).

$$DSS(d_i, d_j) = \frac{\sum_{x \in N_{d_i} \cap N_{d_j}} (S_{d_i}(x) + S_{d_j}(x))}{\sum_{x \in N_{d_i}} S_{d_i}(x) + \sum_{x \in N_{d_j}} S_{d_j}(x)} \tag{2}$$

where N_{d_i} represents the diseases in the graph of disease d_i and N_{d_j} represent the diseases in the graph of disease d_j. $S_{d_i}(x)$ represents the semantic value of disease × $\in N_{d_i}$ and $S_{d_j}(x)$ represents the semantic value of disease × $\in N_{d_j}$, compared to disease d_i and d_j, respectively. The semantic value of disease d is calculated based on Eq. (3).

$$\begin{cases} S_d(x) = \max\{\mu * S_d(d')|d' \in \quad children \quad of \quad (d) \quad if x \neq d \\ \qquad S_d(x) = 1 \qquad\qquad othewise \end{cases} \tag{3}$$

Where μ, the semantic contribution factor is set to 0.5 depending on the previous study (Wang et al., 2010).

The circRNA sequence and disease semantic, similarities were obtained from (Li et al., 2020).

Gaussian interaction profile kernel similarity for circRNA and diseases: Gaussian interaction profile (GIP) kernel similarities of circRNA and disease were computed relied on the presumption that similar circRNAs tend to be associated with similar diseases and vice versa. The GIP similarities were calculated concerning the circRNA-disease association matrix obtained from the CircR2Disease database. The GIP kernel similarity $GKC(c_i, c_j)$ of circRNA c_i $and c_j$can be calculated based on Eqs. (4) and (5).

$$GKC(c_i, c_j) = \exp(-\lambda||A(c_i) - A(c_j)||^2) \tag{4}$$

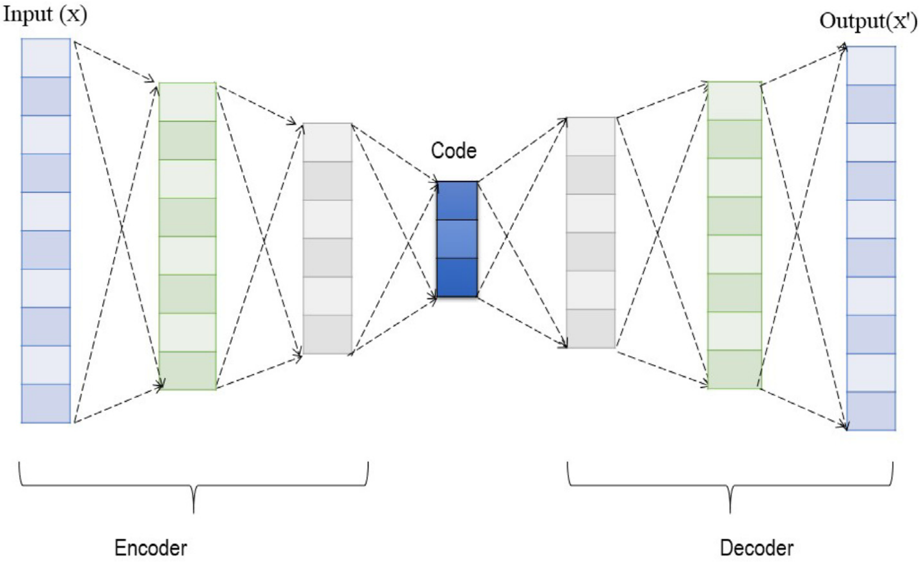

Fig. 1. Schematic diagram of the deep autoencoder.

$$\lambda = \frac{1}{\frac{1}{nc}\sum_{i=1}^{nc} ||A(c_i)||^2} \tag{5}$$

where $A(c_i)$, $A(c_j)$denote the i^{th} and j^{th} rows in the circRNA-disease relationship matrix, λ is the regularization parameter which controls kernel bandwidth, and nc indicates the number of circRNAs.

Similarly, the GIP kernel similarities between two diseases d_i and d_j can be calculated based on Eqs. (6) and (7).

$$GKD(d_i, d_j) = \exp(-\lambda\ ||A(d_i) - A(d_j)||^2) \tag{6}$$

$$\lambda = \frac{1}{\frac{1}{nd}\sum_{i=1}^{nd} ||A(d_i)||^2} \tag{7}$$

where $A(d_i)$, $A(d_j)$ denote the i^{th} and j^{th} columns of the association matrix, λ is the regularization parameter which controls kernel bandwidth and nd, the number of diseases.

2.1.1. Integrated similarity

The integrated similarity matrices of circRNAs (CS) and diseases (DS) were calculated by considering the circRNA sequence similarity, disease semantic similarity, and Gaussian interaction profile kernel similarities of circRNAs and diseases, according to the Eqs. (8) and (9).

$$CS(c_i, c_j) = \frac{CSS(c_i, c_j) + GKC(c_i, c_j)}{2} \tag{8}$$

$$DS(d_i, d_j) = \frac{DSS(d_i, d_j) + GKD(d_i, d_j)}{2} \tag{9}$$

2.2. Method

In this study, we proposed an ensemble classifier AE-DNN, which consists of two parts: Autoencoder and Deep neural network. For each circRNA-disease pair (c_i, d_i) in the dataset, the corresponding similarity vectors were taken from the integrated similarity matrices of circRNAs and diseases. The circRNA similarity vector for the circRNA c_i includes the similarity values of all other circRNAs to c_i. Similarly, the disease similarity vector for the disease d_i, consists of the similarity values of all other diseases to d_i. These similarity vectors were concatenated to make a long feature vector of size $nc + nd$, for the corresponding circRNA-disease pair, where nc and nd indicate the number of circRNAs and diseases respectively in the dataset. Altogether, there were $nc \times nd$ samples, each corresponding to a circRNA-disease pair. For each concatenated feature, the corresponding labels were picked from the association matrix A. If there was an experimentally confirmed association between the corresponding circRNA and disease (positive samples), the label was assigned 1, else it was set to 0. We generated the negative sample set with equal size of the positive sample set, by randomly picking unverified circRNA-disease associations. There can be unverified positive associations in the selected negative sample set, but the probability for this is negligible compared to the whole unverified associations in the dataset. The features of the positive and negative samples (Training set) were fed to a deep autoencoder. The compact high-level features from the output of the autoencoder were used to train the deep neural network. The trained deep neural network predicts the association probability for each unverified circRNA-disease pair as classification result.

2.2.1. Autoencoder based feature selection

Autoencoder is a special kind of neural network structure consisting of two parts: encoder and decoder. This unsupervised neural network model learns the hidden characteristics of input data. They are good at identifying the underlined biological patterns (Chicco et al., 2014; Tan et al., 2016; Jiang et al., 2019). Therefore, we used a deep auto-encoder with two hidden layers, to extract the essential features and to minimize the dimension of the feature vector for each circRNA - disease pair. A deep autoencoder is an autoencoder with more than one hidden layer. The basic arrangement of the deep autoencoder is depicted in Fig. 1.

For achieving dimensionality reduction, the number of neurons in the hidden layers is set less than that of the input layer, while the number of neurons in the output layer is made equal to that of the input layer. As shown in Fig. 1, the autoencoder tries to learn the function:

$$X' = f_{w,b}(X) \approx X \tag{10}$$

Here X represents the input vector, w and b represent the weight and bias variables. When given the input X, the autoencoder first encodes the input sample to a hidden representation by using a mapping function according to Eq. (11).

$$Y = \phi(wX + b) \tag{11}$$

where, $\phi(x) = \frac{1}{1 + \exp(-x)}$ (12)

The resulting low-dimensional representation, Y, is decoded to a vector, X′, with a similar mapping function according to Eq. (13).

$$X' = \phi(w'Y + b') \tag{13}$$

The process is iterated, and the final low-dimensional representation from the encoder is used as the compressed representation of the original input.

In this study, for each circRNA-disease pair (c_i, d_i), we concatenated the feature vectors, $c_i = [p_1, p_2, \ldots, p_{nc}]$ and $d_i = [q_1, q_2, \ldots, q_{nd}]$, where $p_i andq_i$ represent corresponding integrated similarities of circRNA and disease. The feature vectors corresponding to the positive and negative training samples, each of length $nc + nd$, were fed to the autoencoder to obtain a matrix of compact, high-level features. We set the output (encoder) dimension of the autoencoder to 128. We used the mean-squared error (MSE) (Wax and Ziv, 1977) as the loss function and Adam algorithm as the optimizer to reduce the MSE loss.

2.2.2. Deep neural network-based association prediction

The 128-dimensional feature set obtained as output from the autoencoder was a dense, high-level representation of the input samples. These compact feature vectors were then fed to a three-layer feed-forward deep neural network (Schmidt et al., 1992; Svozil et al., 1997; Ripley, 2007), to train the model. The deep neural network was modeled to predict the association probability for every circRNA-disease pair. After training, the model could be used for predicting new relationships between unknown circRNA-disease pairs in the dataset. If the predicted probability for a given sample exceeds the threshold value, the corresponding circRNA-disease association was said to exist.

We used three fully connected layers in the neural network. In the hidden layers, each neuron in layer t is connected to all the neurons in layer t-1. Each hidden layer generates output according to Eq. (14).

$$x_{i+1} = \sigma\left(\sum_{i=1}^{n} (w_i x_i + b_i)\right) \tag{14}$$

where w and b denote weight and bias variables, x denotes the input, and n denotes the number of neurons in the hidden layer. We used the activation function Rectified Linear Unit ReLU (Nair and Hinton, 2010), in the input and hidden layers. The ReLU function, $f(x) = max(0, x)$. In the output layer, we used the Sigmoid function for activation. The mean-squared error (MSE) (Wax and Ziv, 1977) is used as the loss function, and the Adadelta algorithm is used to optimize the MSE loss. Dropout (Srivastava et al., 2014), is used with input and hidden layers for avoiding overfitting. The overall idea of our model is illustrated in Fig. 2.

3. Results and discussion

3.1. Performance evaluation

For assessing the prediction power of our approach, we conducted five-fold as well as ten-fold cross-validation experiments in the following way. In five-fold (ten-fold) cross-validation, we randomly divided the training samples into five (ten) equal-sized partitions. In each fold, one set was used as the test set, and the other four sets were used as the training sets and measured the performance. This process was repeated five (ten) times by treating each set as test set once, and the average was taken as the final result. All the predicted circRNA-disease pairs were sorted based on their score. The unverified circRNA-disease relationships were considered as candidate samples. The predicted scores of the test samples were ranked together with that of the candidate samples. The known relationships with a prediction score higher than the threshold were considered as True Positives (TP), and the number of unknown samples with a prediction score less than the threshold were considered as True Negatives (TN). False Negatives (FN) were the known relations with a prediction score less than the threshold, and False Positives (FP) were the unverified associations with a prediction score above the threshold.

The Receiver Operating Characteristic curve (ROC) (Kumar and Indrayan, 2011; Mandrekar, 2010), was constructed by measuring the true positive rate (TPR, sensitivity) and false-positive rate (FPR, 1-specificity) at various thresholds, Fig. 3. Sensitivity refers to the percentage of samples above the threshold, and specificity refers to the percentage of samples below the threshold. We computed the Area Under the ROC curve (AUC) (Bradley, 1997) for measuring the prediction power of our method. We could achieve high AUC scores of 0.9392 and 0.9431 in five-fold and 10-fold cross-validations, respectively. An AUC value of 1 indicates excellent prediction power, and the AUC value 0.5 suggests the inability of the model to discriminate positives and negatives. We also calculated, Area under the Precision-Recall curve (AUPR) as another type of quality measure (Boyd et al.,

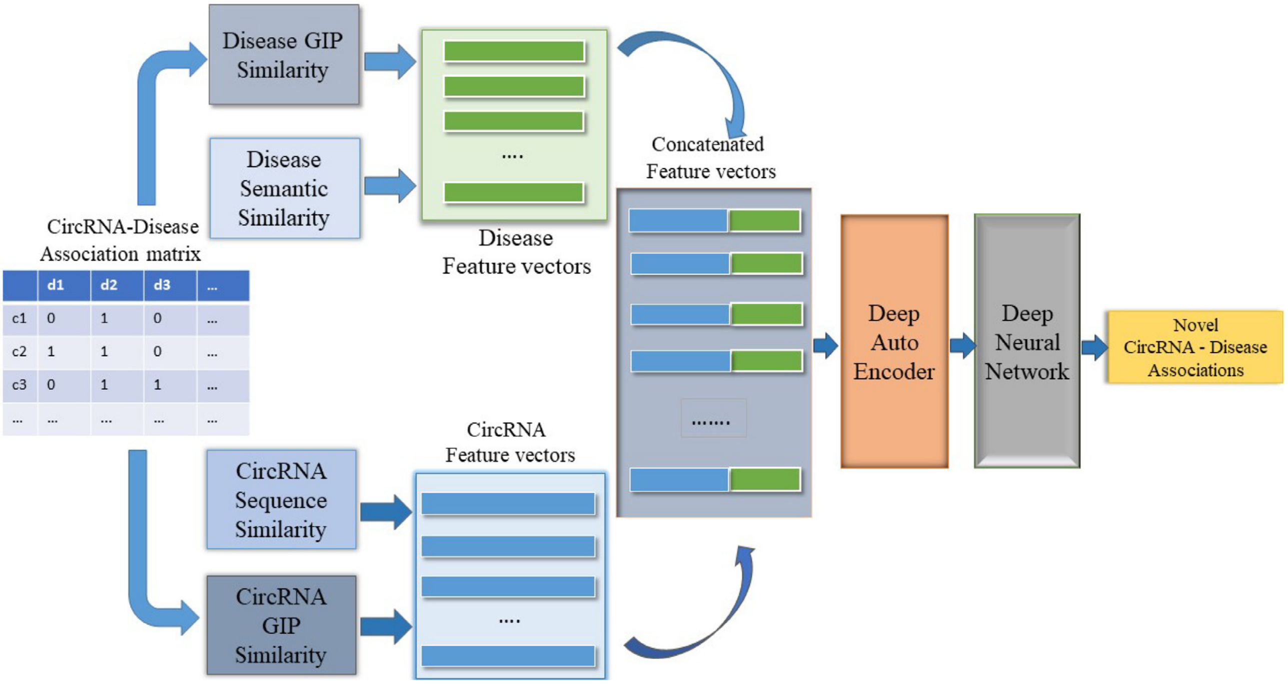

Fig. 2. The flowchart of the proposed approach, AE-DNN, for predicting new circRNA-disease associations. The concatenated circRNA and disease features were fed to the deep autoencoder, and the resulting high-level features were used to train the deep neural network, for novel association prediction.

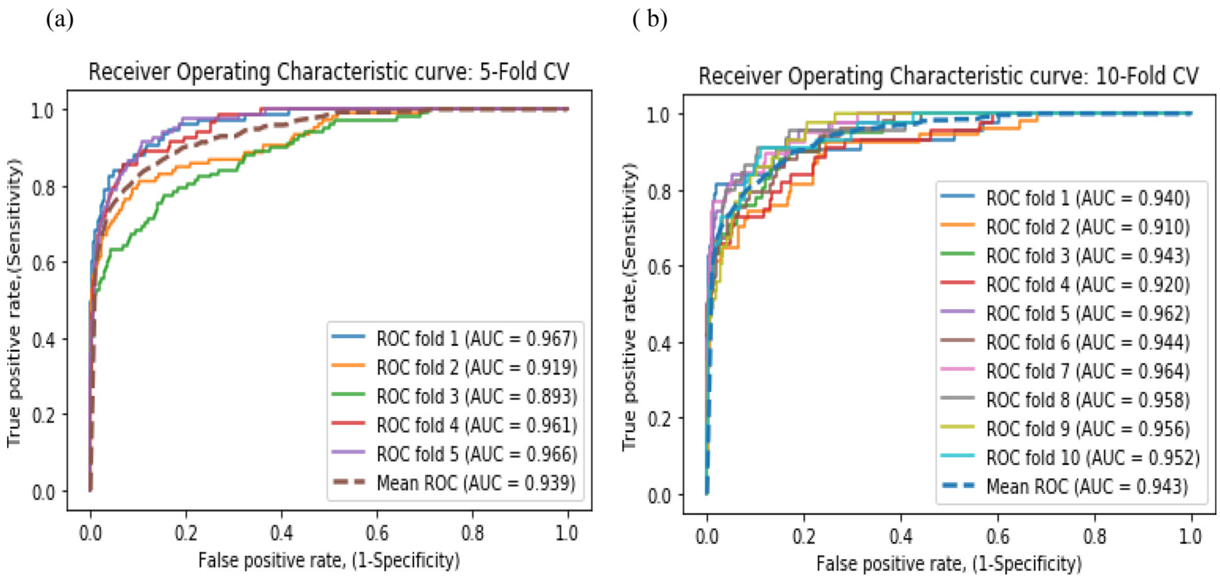

Fig. 3. ROC curves obtained by the proposed approach on the CircR2Disease dataset based on a) 5-fold and b) 10-fold cross-validations.

Table 1
Cross-validation results performed by the proposed model on CircR2Disease dataset.

Method	Accuracy	F1-score	MCC	AUC	AUPR
5-Fold	0.9864	0.5952	0.5836	0.9392	0.6876
10-Fold	0.9852	0.5621	0.6264	0.9431	0.6829

2013). The performance of the method was further measured by the standard statistical parameters such as Accuracy, F1-score, and Matthews correlation coefficient (MCC), Table 1. We conducted five-fold and ten-fold cross-validation ten times and took the average to reduce the variations resulting from random sample partitions.

We also evaluated our model by retaining uncompleted associations such as circRNAs without sequence information and diseases without DOIDs. The new dataset got 516 confirmed associations between 459 circRNAs and 89 diseases. We measured the performance of our model on the new dataset based on 5-fold cross-validations. The experimental results obtained are, Accuracy (0.9891), F1-score (0.5588), MCC (0.6086), AUC (0.9362) and AUPR (0.6170). This result indicates the robustness of our method with noise data.

3.2. *Comparison with other methods*

For assessing the prediction power of our approach, we compared it with previous studies. We selected five state-of-the-art methods predicting circRNA-disease relationships. All these methods used the same benchmark dataset CircR2Disease for association predictions. The methods include KATZ measures for human circRNA-disease association prediction, KATZHCDA (Fan et al., 2018b), Locality-Constrained Linear Coding for predicting disease associated circRNAs, LLCDC (Ge et al., 2020), identification of circRNA-disease associations based on matrix factorization, iCircDA-MF (Wei and Liu, 2019), Speedup Inductive Matrix Completion for CircRNADisease Associations prediction, SIMCCDA (Li et al., 2020) and Convolutional neural network and extreme learning machine based approach, CNN-ELM (Wang et al., 2019). The first two methods were network-based, the third and fourth methods were matrix-based, and the latter approach was machine-learning-based. Network and matrix oriented approaches usually achieve good performance in small scale data, but their time complexity increases exponentially with increasing samples. In network-based methods, the network needs to be reconstructed when adding new circRNAs or diseases to the dataset. Most of the network-based methods have the drawback that they cannot be applied to circRNAs for which there are no associated diseases in the interaction matrix. Our approach can be applied to circRNAs /diseases for which no associated diseases/ circRNAs in the interaction matrix. Machine-learning techniques can quickly adapt to changes. We only need to extract features for new diseases and circRNAs, and the model can be trained with the new features together with previous training samples. The only machine-learning based approach, CNN-ELM, considered only Gaussian interaction profile kernel similarity as circRNA similarity for association predictions, did not find sequence or functional similarities of circRNAs. This may be the reason for the slight low performance of the method. Since some of the above methods only reported the AUC scores in the paper, we compared those methods with the reported AUC scores. The experimental results based on 5-fold cross-validations are summarized in Table 2. From the table, it is clear that the proposed approach outperforms other state-of-the-art methods.

3.3. *Comparison with other classifiers*

To further measure the prediction power of the proposed model, we compared it with other machine-learning classifiers such as support vector machines (SVM), deep neural networks without autoencoder (Raw-DNN), and decision tree classifiers. We implemented SVM classifiers on the benchmark dataset (CircR2disease), with the proposed features. We also implemented deep neural networks without autoencoder and decision tree classifiers with the same features. We conducted 5-fold cross-validation and constructed ROC curves for each classifier Fig. 4. The performance was measured in terms of the

Table 2
Comparison of the proposed method with previous studies based on the CircR2disease dataset, using 5-fold cross validation.

Methods	AE-DNN	iCircDA-MF	LLCDC	CNN-ELM	KATZHCDA	SIMCCDA
AUC Scores	0.9392	0.9178	0.9177	0.8667	0.7936	0.7477

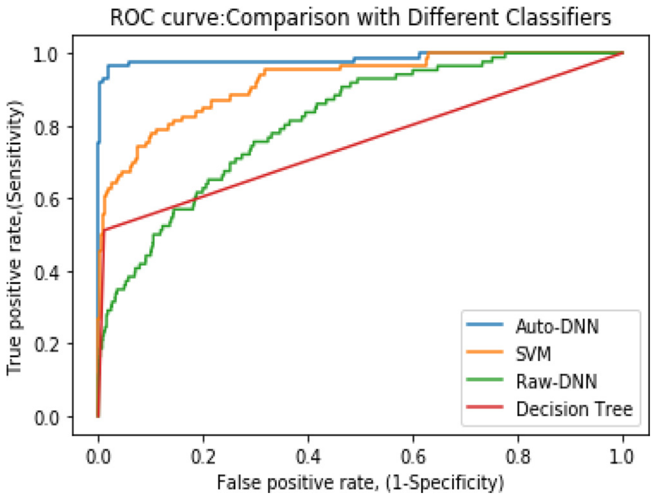

Fig. 4. Comparison of the proposed method, Auto-DNN with different classifiers based on 5-fold cross-validation on the CircR2disease dataset.

Table 3
Comparison of the proposed approach with different classifiers in terms of AUC scores, based on 5-fold cross-validations.

Classifiers	AUC
Auto-DNN	0.9392
SVM	0.8649
Raw-DNN	0.8116
Decision Tree	0.7251

Table 4
Comparison of the proposed approach with different datasets, based on 5-fold cross-validations.

Dataset	Accuracy	F1-score	MCC	AUC	AUPR
circRNAdisease	0.9752	0.4182	0.4883	0.9094	0.6284
Circ2Disease	0.9764	0.4924	0.4222	0.8622	0.5167

Table 5
Top twelve predicted circRNA-disease associations by our approach with their evidence in literature.

Rank	circRNA	Disease	Literature (PMID)
1	hsa_circ_0001946	Glioma	26,683,098
2	hsa_circ_0004458	Gastric cancer	28,544,609
3	hsa_circ_0072088	Liver cancer	28,727,484
4	hsa_circ_0004214	Glioma	28,622,299
5	hsa_circ_0013958	Lung cancer	28,685,964
6	hsa_circ_0037911	Heart disease	Unconfirmed
7	hsa_circ_0054633	Diabetes mellitus	27,878,383
8	hsa_circ_0068087	Coronary Artery	Unconfirmed
9	hsa_circ_0001187	Acute myeloid leukemia	28,282,919
10	hsa_circ_0005870	Diabetic retinopathy	Unconfirmed
11	hsa_circ_0023404	Cervical cancer	31,082,770
12	hsa_circ_0000284	Pancreatic cancer	29,255,366

obtained AUC scores, and the results are listed in Table 3. The results showed that the performance of our model was higher compared to other classifiers.

3.4. Comparison on different datasets

To further evaluate our approach, we implemented our model with two other datasets and conducted cross-validation experiments. The first dataset circRNAdisease contains 330 circRNAs and 48 diseases, with 354 experimentally confirmed circRNA-disease associations. This dataset (Yao et al., 2018) can be obtained from http://cgga.org.cn:9091/circRNADisease/. The second dataset, Circ2Disease (Zhao et al., 2018), contains 249 circRNAs and 61 diseases with 273 known circRNA-disease associations between them. The circRNAs without sequence information and diseases without DOID were removed, and finally, we got 241 associations in the dataset-1 between 223 circRNAs and 34 diseases. Similarly, we obtained 240 associations in dataset-2, between 215 circRNAs and 46 diseases. The experimental results on the two datasets are shown in Table 4. The corresponding ROC curves for the two datasets, based on 5-fold cross-validations, are given in Fig. 5.

3.5. Case studies

Along with the cross-validation experiments, we further measured the performance of our model based on case studies on the CircR2disease dataset. For this, we trained the method with all the

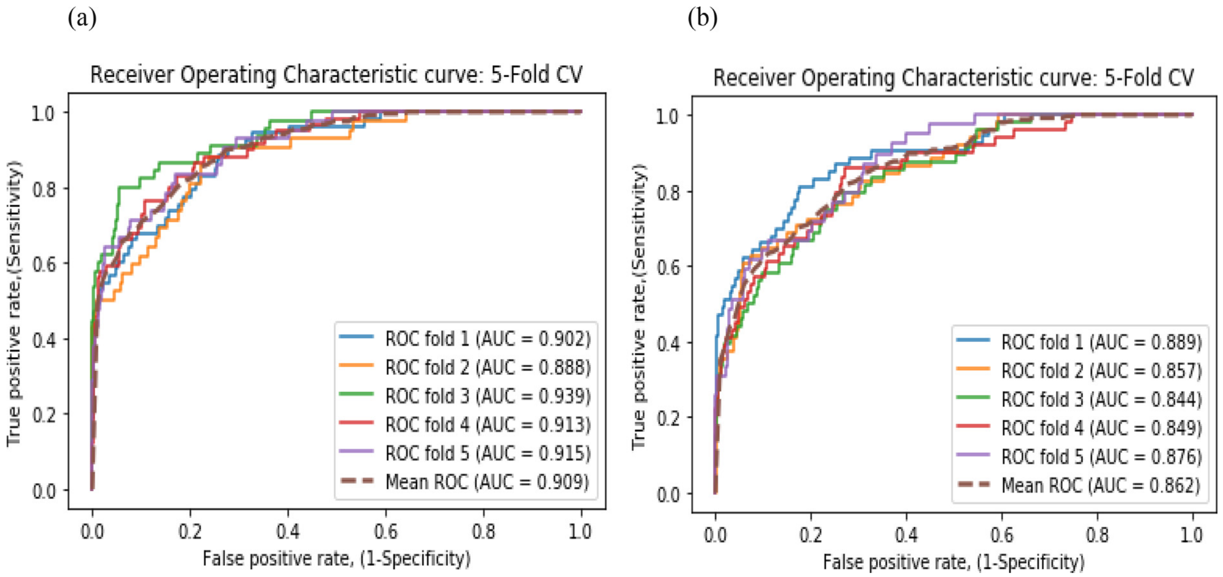

Fig. 5. ROC curves obtained by the proposed approach on the a) circRNAdisease and b) Circ2Disease datasets using 5- fold cross-validation.

samples in the training set. The trained model is used to predict the scores of all the unverified relations in the dataset. These predicted scores, representing the probability of association, were sorted in the descending order with the circRNA-disease associations. Among the top twelve predicted associations, 9 of them were confirmed with the latest literatures. The detailed results are listed in Table 5. These results indicated that there is a high probability for associations between other top predicted circRNA-disease associations.

4. Conclusion

In this study, we implemented a computational approach consisting of a deep autoencoder and deep neural network, to predict potential circRNA-disease relationships. Identification of new circRNA-disease relations is critical for identifying the biomarkers for disease diagnosis. We developed feature vectors for each circRNA-disease pair, utilizing circRNA and disease similarities. The feature vectors were fed to the autoencoder to extract high-level features. With the resulting features, the deep neural network was trained. The built model could be used to predict novel circRNA-disease relationships. Cross validations and case studies indicates the proposed approach outperforms previous models in prediction performance. Ensemble learning algorithms are attractive as they are more accurate than a single classifier. Sequence information can represent circRNAs more deeply, which aids in the performance of our model. The model could adapt to changes very well. So as novel experimentally confirmed, circRNA-disease associations were discovered, they can be easily added to the training set. For new circRNAs and diseases, we only need to calculate their similarities for feature construction and can be added to the dataset. We expect our method could be improved by integrating more experimentally confirmed associations and incorporating more biological information to the features.

CRediT authorship contribution statement

K. Deepthi: Conceptualization, Methodology, Data curation, Software, Validation, Writing - original draft, Visualization, Investigation. **A.S. Jereesh:** Conceptualization, Methodology, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Boyd, K., Eng, K. H., Page, C. D. 2013. Area under the precision-recall curve: point estimates and confidence intervals. In Joint European conference on machine learning and knowledge discovery in databases (pp. 451-466). Springer, Berlin, Heidelberg.

Bradley, A.P., 1997. The use of the Area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. 30 (7), 1145–1159.

Chen, Y., Li, C., Tan, C., Liu, X., 2016. Circular RNAs: a new frontier in the study of human diseases. J. Med. Genet. 53 (6), 359–365.

Chen, X., Xie, D., Zhao, Q., You, Z.H., 2019. MicroRNAs and complex diseases: from experimental results to computational models. Briefings Bioinf. 20 (2), 515–539.

Chicco, D., Sadowski, P., Baldi, P., 2014. September). Deep autoencoder neural networks for gene ontology annotation predictions. In: Proceedings of the 5th ACM conference on bioinformatics, computational biology, and health informatics, pp. 533–540.

Fan, C., Lei, X., Fang, Z., Jiang, Q., Wu, F. X., 2018. CircR2Disease: a manually curated database for experimentally supported circular RNAs associated with various diseases. Database, 2018.

Fan, C., Lei, X., Wu, F.X., 2018b. Prediction of CircRNA-disease associations using KATZ model based on heterogeneous networks. Int. J. Biol. Sci. 14 (14), 1950.

Fang, Y., 2018. Circular RNAs as novel biomarkers with regulatory potency in human diseases. Future Sci. OA 4 (07), FSO314.

Fu, L., Peng, Q., 2017. A deep ensemble model to predict miRNA-disease association. Sci. Rep. 7 (1), 1–13.

Esteller, M., 2011. Non-coding RNAs in human disease. Nat. Rev. Genet. 12 (12), 861.

Ge, E., Yang, Y., Gang, M., Fan, C., Zhao, Q., 2020. Predicting human disease-associated circRNAs based on locality-constrained linear coding. Genomics 112 (2), 1335–1342.

Greene, J., Baird, A.M., Brady, L., Lim, M., Gray, S.G., McDermott, R., Finn, S.P., 2017. Circular RNAs: biogenesis, function and role in human diseases. Front. Mol. Biosciences 4, 38.

Jiang, H.J., Huang, Y. A., You, Z.H., 2019. Predicting Drug-Disease Associations via Using Gaussian Interaction Profile and Kernel-Based Autoencoder. BioMed research international, 2019.

Katz, L., 1953. A new status index derived from sociometric analysis. Psychometrika 18 (1), 39–43.

Kumar, R., Indrayan, A., 2011. Receiver operating characteristic (ROC) curve for medical researchers. Indian Pediatr. 48 (4), 277–287.

Levenshtein, V.I., 1966. Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady, (Vol. 10, No. 8, pp. 707-710).

Li, M., Liu, M., Bin, Y., Xia, J., 2020. Prediction of circRNA-disease associations based on inductive matrix completion. BMC Med. Genomics 13 (5), 1–13.

Lei, X., Fang, Z., Chen, L., Wu, F.X., 2018. PWCDA: path weighted method for predicting circRNA-disease associations. Int. J. Mol. Sci. 19 (11), 3410.

Lei, X., Bian, C., 2020. Integrating random walk with restart and k-Nearest Neighbor to identify novel circRNA-disease association. Sci. Rep. 10 (1), 1–9.

Lu, C., Yang, M., Luo, F., Wu, F.X., Li, M., Pan, Y., Wang, J., 2018. Prediction of lncRNA–disease associations based on inductive matrix completion. Bioinformatics 34 (19), 3357–3364.

Lukiw, W., 2013. Circular RNA (circRNA) in Alzheimer's disease (AD). Front. Genet. 4, 307.

Mandrekar, J.N., 2010. Receiver operating characteristic curve in diagnostic test assessment. J. Thoracic Oncol. 5 (9), 1315–1316.

Nacher, J.C., Akutsu, T., 2019. Controllability methods for identifying associations between critical control ncrnas and human diseases. In: Computational Biology of Non-Coding RNA (pp. 289-300). Humana Press, New York, NY.

Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10) (pp. 807-814).

Peng, J., Hui, W., Li, Q., Chen, B., Hao, J., Jiang, Q., Wei, Z., 2019. A learning-based framework for miRNA-disease association identification using neural networks. Bioinformatics 35 (21), 4364–4371.

Ripley, B.D., 2007. Pattern recognition and neural networks. Cambridge University Press.

Sanger, H.L., Klotz, G., Riesner, D., Gross, H.J., Kleinschmidt, A.K., 1976. Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures. Proc. Natl. Acad. Sci. 73 (11), 3852–3856.

Schmidt, W.F., Kraaijveld, M.A., Duin, R.P., 1992. Feed forward neural networks with random weights. In: International Conference on Pattern Recognition (pp. 1-1). IEEE COMPUTER SOCIETY PRESS.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.

Svozil, D., Kvasnicka, V., Pospichal, J., 1997. Introduction to multi-layer feed-forward neural networks. Chemometrics Intelligent Lab. Syst. 39 (1), 43–62.

Tan, J., Hammond, J.H., Hogan, D.A., Greene, C.S., 2016. Adage-based integration of publicly available pseudomonas aeruginosa gene expression data with denoising autoencoders illuminates microbe-host interactions. MSystems 1 (1).

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

Yao, D., Zhang, L., Zheng, M., Sun, X., Lu, Y., Liu, P., 2018. Circ2Disease: a manually curated database of experimentally validated circRNAs in human disease. Sci. Rep. 8 (1), 1–6.

Yu, C.Y., Kuo, H.C., 2019. The emerging roles and functions of circular RNAs and their generation. J. Biomed. Sci. 26 (1), 29.

Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S., Chen, C.F., 2007. A new method to measure the semantic similarity of GO terms. Bioinformatics 23 (10), 1274–1281.

Wang, C., Wei, L., Guo, M., Zou, Q., 2013. Computational approaches in detecting non-coding RNA. Curr. Genomics 14 (6), 371–377.

Wang, D., Wang, J., Lu, M., Song, F., Cui, Q., 2010. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. Bioinformatics 26 (13), 1644–1650.

Wang, L., You, Z.H., Huang, Y.A., Huang, D.S., Chan, K.C., 2019. An efficient approach based on multi-sources information to predict CircRNA-disease associations using deep convoltional neural network. Bioinformatics.

Wax, M., Ziv, J., 1977. Improved bounds on the local mean-square error and the bias of parameter estimators (corresp.). IEEE Trans. Inf. Theory 23 (4), 529–530.

Wei, H., Liu, B., 2019. iCircDA-MF: identification of circRNA-disease associations based on matrix factorization. Briefings Bioinformatics.

Zhang, J., Zhang, Z., Chen, Z., Deng, L., 2017. Integrating multiple heterogeneous networks for novel lncRNA-disease association inference. IEEE/ACM Trans. Comput. Biol. Bioinf. 16 (2), 396–406.

Zhang, Z., Yang, T., Xiao, J., 2018. Circular RNAs: promising biomarkers for human diseases. EBioMedicine 34, 267–274.

Zhao, Z., Wang, K., Wu, F., Wang, W., Zhang, K., Hu, H., Jiang, T., 2018. circRNA disease: a manually curated database of experimentally supported circRNA-disease associations. Cell Death Dis. 9 (5), 1–2.