Interdisciplinary Sciences: Computational Life Sciences
https://doi.org/10.1007/s12539-023-00590-y

ORIGINAL RESEARCH ARTICLE

DCDA: CircRNA–Disease Association Prediction with Feed-Forward Neural Network and Deep Autoencoder

Hacer Turgut[1] · Beste Turanli[2] · Betül Boz[1]

Received: 20 March 2023 / Revised: 13 October 2023 / Accepted: 15 October 2023

Abstract

Circular RNA is a single-stranded RNA with a closed-loop structure. In recent years, academic research has revealed that circular RNAs play critical roles in biological processes and are related to human diseases. The discovery of potential circRNAs as disease biomarkers and drug targets is crucial since it can help diagnose diseases in the early stages and be used to treat people. However, in conventional experimental methods, conducting experiments to detect associations between circular RNAs and diseases is time-consuming and costly. To overcome this problem, various computational methodologies are proposed to extract essential features for both circular RNAs and diseases and predict the associations. Studies showed that computational methods successfully predicted performance and made it possible to detect possible highly related circular RNAs for diseases. This study proposes a deep learning-based circRNA–disease association predictor methodology called DCDA, which uses multiple data sources to create circRNA and disease features and reveal hidden feature codings of a circular RNA–disease pair with a deep autoencoder, then predict the relation score of the pair by a deep neural network. Fivefold cross-validation results on the benchmark dataset showed that our model outperforms state-of-the-art prediction methods in the literature with the AUC score of 0.9794.

✉ Hacer Turgut
hacertilbecturgut@gmail.com

✉ Betül Boz
betul.demiroz@marmara.edu.tr

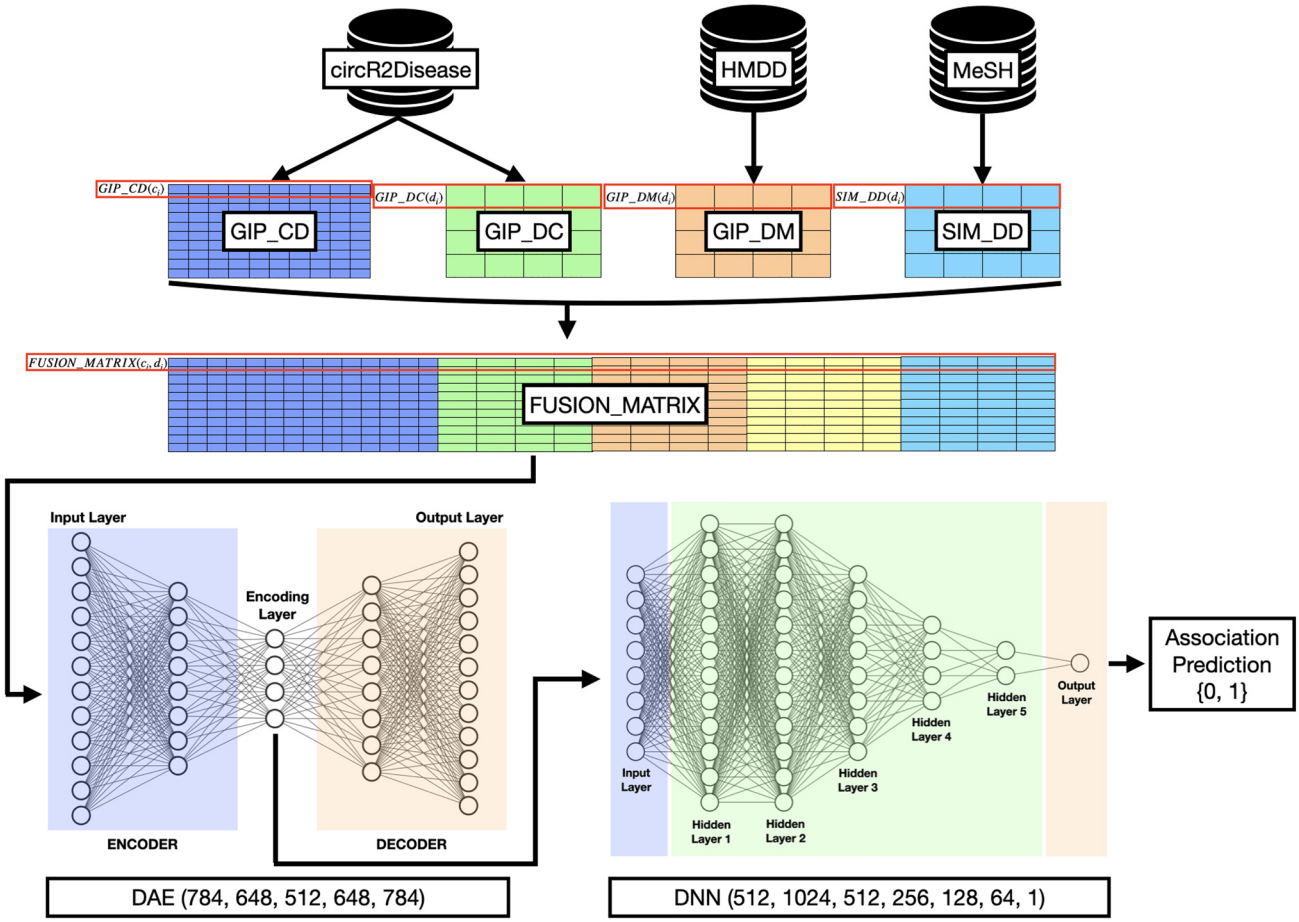
Beste Turanli
beste.turanli@marmara.edu.tr

[1] Computer Engineering Department, Marmara University, 34854 Istanbul, Türkiye

[2] Bioengineering Department, Marmara University, 34854 Istanbul, Türkiye

Published online: 17 November 2023

Graphical abstract



Keywords CircRNA · CircRNA–disease association · Deep learning · Autoencoder · Neural network

1 Introduction

Circular RNA (circRNA) is a single-stranded RNA with a continuous loop structure resulting from its covalently joined 5′ and 3′ ends. Due to its closed-loop structure, circRNAs are generally stable [1]. Even though researchers discovered circRNAs in the 1970s [2, 3], their significance was unrecognized and considered as an RNA splicing error. Thanks to improvements in RNA sequencing technology, bioinformatics, and many studies, many circRNAs have been discovered. Researchers established various functionalities of circRNAs in biological processes such as acting as sponges for microRNAs, regulating gene transcription, and expression by interacting RNA binding proteins (RBPs) [4]. In recent years, circRNAs have been seen as potential biomarkers of several diseases and treatment targets. For example, Wang et al. [5] found the correlation between tumor size and the level of expression of hsa_circ_0012673, hsa_circ_0000064, circPUM1, hsa_circ_0007385, and circ_000073558 in lung carcinoma tissue.

Several researchers collected circRNA–disease associations from previous research papers and created circRNA-related databases, like circR2Disease [6], Circad [7], circ2Disease [8], and CircFunBase [9]. As the number of expressed associations between diseases and circRNAs increases, the possibility of predicting novel circRNA–disease relationships by computational algorithms becomes viable.

Recently, various machine learning and deep learning-based methods have been proposed which are applied to a variety of problems. Deep representative learning features have a high potential to interpret the sequence information in proteomics which is very successful in protein function prediction problems using deep learning-based model Le [10]. Deep learning and machine learning algorithms are used to predict anticancer peptides with

promising performance [11]. A novel Vulture Based Adaboost-Feedforward Neural (VbAFN) scheme is proposed to forecast the COVID-19 severity at an earlier stage to improve health. It recognizes and segments the COVID-19 cell from the Chest X-ray data set [12]. Machine learning is also used to predict lysine lactylation sites in gastric cancer cells. Auto-Kla quickly and accurately predicts Kla sites using automated machine learning [13]. Some studies propose new model architectures for encoder–decoder architectures. Transformer [14] is a model architecture that relies on an attention mechanism to find out dependencies between input and output. It replaces the recurrent layers most commonly used in encoder–decoder architectures with multi-headed self-attention.

There are growing number of studies that use machine learning and deep learning-based methods for circRNA–disease association. In the KATZHCDA model, Fan et al. [15] used similarity scores between circRNA expression profiles for circRNA representations and disease phenotype similarity for disease representations. From known circRNA–disease associations, the Gaussian interaction profile (GIP) kernel is applied, and features are generated for both circRNAs and diseases. A graph-based method, KATZ, is used as the prediction algorithm. Wang et al. [16] also used the GIP kernel to extract features for disease and circRNAs and the semantic similarity between diseases. They used a convolutional neural network (CNN) as a feature extraction method and an extreme learning machine (ELM) as the predictor. In the MSFCNN model, [17] used four circRNA and seven disease similarity features based on topological and biological data from various sources and circRNA–disease associations predicted by two-dimensional CNN. Wang et al. [18] proposed a method that consists of the same features as in the study of Wang et al. [16]. Generative adversarial network (GAN) is used as a feature extraction method and logical model tree (LMT) classifier. Zheng et al. [19] used gene associations of circRNAs and circRNA sequence similarity using chaos game representation and used support vector machine (SVM) as the prediction model. Ge et al. [20] proposed a circRNA representation method that looks at circRNAs' gene targets and creates a circRNA vector based on semantic similarity between other circRNAs' gene targets. Disease semantic similarity between diseases is used to develop disease vectors. CircRNA and disease vectors are reconstructed using locality-constrained linear coding (LLC) to encode vectors so that vectors can preserve local information in previous information. In addition to these features, unlike other research studies, the cosine similarity function is applied to get circRNA and disease similarity scores from circRNA–disease relationships instead of GIP in this study. Label propagation methodology is applied to all similarity networks. The average score between circRNAs and diseases is accepted as the final prediction score.

The AE–DNN model [21] generates circRNA vectors from sequence similarity by edit distance, and disease vectors are generated from semantic similarity using disease ontology relationships. GIP kernel is also used to get additional features for both circRNAs and diseases. circRNA and disease features are concatenated by taking the average. To extract essential features from all features, the autoencoder model is used. A deep neural network model is used to predict association scores between circRNAs and diseases. Although there are similarities between our model and AE–DNN, they also differ in subtle ways. In the context of the AE–DNN model, an additional aspect that warrants attention is the extraction of disease-specific characteristics without considering their interactions with different types of RNA. Our approach endeavors to overcome this limitation by incorporating miRNA–disease relationships into the feature development process, thereby enhancing the similarity-based attributes. Furthermore, the conventional AE–DNN model employs an averaging technique to combine the extracted features. However, acknowledging the potential loss of information inherent in this strategy, we chose to pursue an alternative path by merging the features into a fused vector format, thereby preserving their individuality, and allowing for a more comprehensive analysis. Deepthi and Jereesh [22] introduced a methodology called AE–RF that feeds information gathered from the circRNA–disease association, circRNA functional similarities, and disease semantic similarities to an autoencoder model to extract hidden biological patterns and predict circRNA–disease associations with random forests (RF) classifier. Wei and Liu [23] proposed a model called iCircDA-MF. They constructed features using disease semantic information, circRNA–gene, gene–disease, and circRNA–disease data. To detect and correct false-negative associations, neighbor interaction profiles based on the circRNA similarity and disease similarity were used and circRNA–disease associations were predicted by matrix factorization. Wang et al. [24] formed a unified descriptor of circRNA–disease pairs by disease semantic similarity information and GIP kernel similarity information of diseases and circRNAs. They proposed to use fast learning with graph convolutional networks to reveal high-level features from unified descriptors and then predicted new circRNA–disease pairs with forest by penalizing attributes classifier.

In this study, we propose a deep learning-based methodology called DCDA that consists of two parts: a circRNA–disease association feature extraction model and a circRNA–disease association prediction model. This methodology uses multiple data to generate circRNA–disease association features, which are manually curated circRNA–disease associations, disease–disease semantic similarity information, and disease–miRNA associations. To have a consistent overall dataset, we generated synonym

dictionaries for both diseases and circRNAs. The experimental results revealed that DCDA has the highest AUC score on the CircR2Disease dataset compared to other state-of-the-art methodologies.

2 Materials and Method

2.1 Datasets

The datasets used to extract features in our methodology are described below.

- **CircR2Disease**: The CircR2Disease database is used as the benchmark dataset in our experiments. There are 739 entries which include 661 unique circRNAs and 100 diseases in the database. We selected only human disease-related circRNA–disease pairs in this study. To have a common language between all datasets used in the study, we mapped circRNA names using the circRNA synonyms dictionary. Disease names are also mapped using the disease synonyms dictionary. There was a small number of duplicated circRNA–disease pairs in the data, so we filtered out duplicated circRNA–disease pairs. Eventually, we ended up with 639 entries, 75 unique diseases, and 559 unique circRNAs. These confirmed circRNA–disease pairs are our positive sample. Since we do not have a dataset that shows no associated circRNA–disease pairs, we generated a negative sample set. We designed the negative sample set so that it has the same number of pairs as the positive set to have a balanced dataset. We randomly created circRNA–disease pairs from circRNAs and diseases in the CircR2Disease database and selected pairs that were confirmed. Finally, we have 639 positive samples and 639 negative samples, in total of 1278 pairs, in our benchmark dataset.
- **Circ2Disease**: We used the Circ2Disease dataset to see the performance of our methodology on a different dataset. 273 circRNA–disease pairs were experimentally validated in the data. We changed circRNA and disease names by using our circRNA and disease synonyms dictionaries. After that, we excluded duplicated pairs in the data. Finally, we had 268 pairs, 233 unique circRNAs, and 60 unique diseases. We performed the same steps as explained in the previous section, we created our positive sample from these pairs and created a negative sample set with the same size as the positive sample set. In the end, we ended up with 536 pairs in the data which consists of 268 positive and 268 negative samples.
- **The Human microRNA Disease Database (HMDD)**: The Human microRNA Disease Database (HMDD) [25] is a human miRNA–disease relationship database that is derived from experimentally validated research. HMDD was first published in 2017 and the last version (HMDD v3.2) was released in 2019. In the last version, there are 35,547 miRNA–disease association entries, 1206 unique miRNA genes, and unique 893 diseases. In this study, we used HMDD v3.2.
- **MeSH**: The MeSH database [26] is provided by the National Library of Medicine (NLM). MeSH is a vocabulary for indexing medical and biological publications. The database also has disease-specific information and directed acyclic graphs (DAGs) generated from relationships between diseases. Using these DAGs, tree numbers are assigned to diseases. For example, Fig. 1 illustrates some information from the lung neoplasm disease page. Tree numbers(s) are in numbered format of lung neoplasms' DAGs. If we split a tree number by dot symbols, we will end up with nodes of the DAG where the first node is the top level of the DAG and the last node is the disease node. Diseases can be searched through the https://meshb.nlm.nih.gov/search page.

2.2 Synonym Dictionaries

Since we used multiple data sources, we needed to create circRNA and disease synonyms dictionaries to have a common terminology between databases. We describe how we created circRNA and disease synonyms dictionaries below.

- **CircRNA synonyms dictionary**: CircRNA synonyms are collected from CircR2Disease, circbase, Circ2Disease, circRNADisease, and CircFunBase databases. We

MeSH Heading	Lung Neoplasms
Tree Number(s)	C04.588.894.797.520 C08.381.540 C08.785.520
Unique ID	D008175
RDF Unique Identifier	http://id.nlm.nih.gov/mesh/D008175
Annotation	coord IM with histol type of neopl (IM)
Scope Note	Tumors or cancer of the LUNG.
Entry Version	LUNG NEOPL
Entry Term(s)	Cancer of Lung Cancer of the Lung Lung Cancer Neoplasms, Lung Neoplasms, Pulmonary Pulmonary Cancer Pulmonary Neoplasms
NLM Classification #	WF 658
See Also	Carcinoma, Non-Small-Cell Lung Carcinoma, Small Cell
Date Established	1966/01/01
Date of Entry	1999/01/01
Revision Date	2012/07/03

Fig. 1 Example search result of "lung cancer" term on MeSH 2021 database. Tree numbers and entry terms are marked with a red box

evaluated every circRNA–disease association information in the benchmark dataset and checked if the same circRNA–disease pair occurs in other datasets. We used PMID information to find common pairs. If a common circRNA–disease pair exists in another database, we collected the circRNA name or the circRNA ID and put them in a dictionary as key, value pairs. The key is the circRNA name mentioned in the CircR2Disease database and the value is a set of synonym names of the circRNA mentioned in other datasets.

- **Disease synonyms dictionary**: A disease synonyms dictionary was created using Circ2Disease, Circ2Disease, MeSH, and HMDD dataset. The MeSH browser was used to search each distinct disease in the dataset; if a disease is discovered in the database, the website would redirect to the disease page. A disease page title is considered as the disease name and any search term (disease names from datasets) is considered as a synonym of the disease.

2.3 Method Overview

In this study, we propose a novel method called DCDA to predict the association of a circRNA–disease pair. As shown in Fig. 2, the DCDA has four steps. Firstly, circRNA and disease features are generated using multiple data sources. Secondly, circRNA–disease pair vectors (fusion vectors) are created by concatenating circRNA and disease features. In the next step, we used a deep autoencoder to extract useful hidden features from fusion vectors in a low-dimensional space. Finally, a deep neural network is trained to predict circRNA–disease associations.

2.3.1 CircRNA–circRNA GIP Kernel Similarity (GIP_CD)

GIP kernel is a way of constructing an interaction kernel to find similarities in a network. Van Laarhoven et al. [27] proposed building a kernel from interaction profiles of

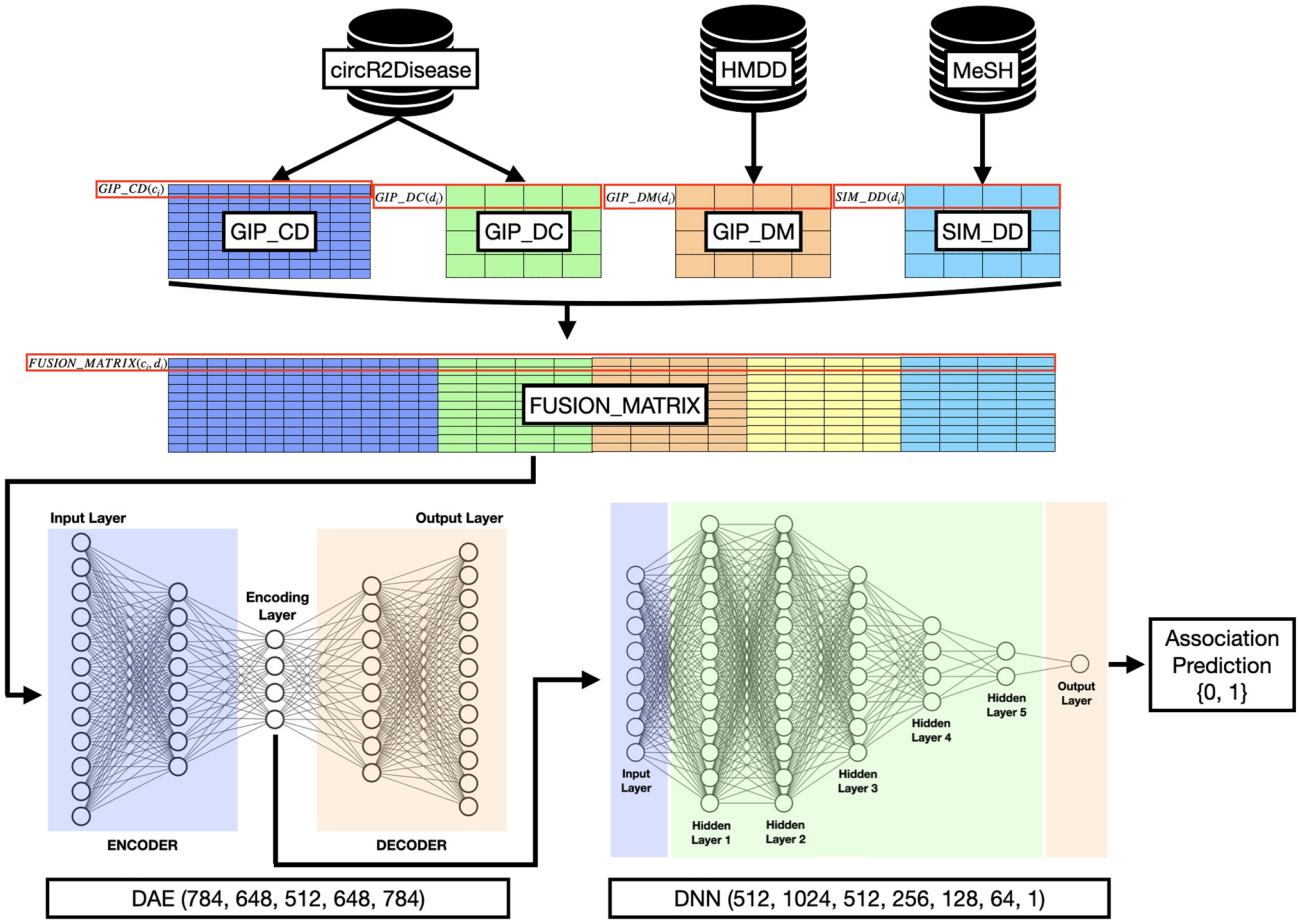


Fig. 2 Flow of the DCDA methodology. GIP_CD, GIP_DC, GIP_DM, and SIM_DD features are created from circR2Disease, HMDD, and MeSH databases. Then FUSION_MATRIX is generated by concatenating circRNA and disease features. Feature extraction is performed by the DAE model. Extracted features from the encoding layer of the DAE are fed to the DNN prediction model. Finally, DNN returns the probability of circRNA–disease pair being associated. By using 0.5 threshold, we decided whether the pair is associated (1) or not (0)

drug–target protein associations. The idea behind this is that drugs associated with similar target proteins will also interact with similar target proteins. The same idea is used to generate circRNA–disease features in the literature. According to the CircR2Disease dataset, the circRNA–disease interaction profile (CD) is generated where rows are circRNAs and columns are diseases. Suppose an association between circRNA c_i and disease d_i, CD(c_i, d_i) is set to 1, otherwise 0. We construct circRNA–circRNA GIP kernel similarity matrix from circRNA–disease associations (GIP_CD) using the CD. The similarity between circRNA c_i and circRNA c_j is calculated as follows:

$$\text{GIP\_CD}(c_i, c_j) = \exp(-\lambda \|\text{CD}(c_i) - \text{CD}(c_j)\|^2), \tag{1}$$

$$\lambda = \frac{1}{\frac{1}{\text{nc}} \sum_{i=0}^{\text{nc}} \|\text{CD}(c_i)\|^2}, \tag{2}$$

where CD(c_i) and CD(c_j) represent the ith and jth rows in CD. λ is the regularization parameter that is used for controlling kernel bandwidth. nc stands for the number of unique circRNAs.

2.3.2 Disease–Disease GIP Kernel Similarity (GIP_DC)

We also used the Gaussian interaction profile kernel algorithm to create disease features. Disease–disease GIP kernel similarity matrix (GIP_DC) is created using the CircR2Disease database. The similarity between disease d_i and disease d_j is calculated as follows:

$$\text{GIP\_DC}(d_i, d_j) = \exp(-\lambda \|\text{DC}(d_i) - \text{DC}(d_j)\|^2) \tag{3}$$

$$\lambda = \frac{1}{\frac{1}{\text{nd}} \sum_{i=0}^{\text{nd}} \|\text{DC}(d_i)\|^2}. \tag{4}$$

DC is transposed of CD, where DC(d_i) and DC(d_j) represent ith and jth rows in DC. λ is the regularization parameter that is used for controlling kernel bandwidth. nd stands for the number of unique diseases.

2.3.3 Disease–miRNA GIP Kernel Similarity (GIP_DM)

The HMDD database contains experimentally supported human miRNA–disease relationships. In this study, we used HMDD v3.2. There are 35,547 miRNA–disease association entries in this version, 1206 unique miRNA genes, and unique 893 diseases. We used entries related to the benchmark dataset's diseases and attained 13,624 entries, 841 unique miRNAs, and 48 unique diseases. Using the HMDD database, disease–miRNA association matrix DM is generated where rows are diseases and columns are miRNAs.

Disease–miRNA GIP kernel similarity matrix (GIP_DM) is created from disease–miRNA association matrix DM. The similarity between disease d_i and disease d_j is calculated as follows:

$$\text{GIP\_DM}(d_i, d_j) = \exp(-\lambda \|\text{DM}(d_i) - \text{DM}(d_j)\|^2) \tag{5}$$

$$\lambda = \frac{1}{\frac{1}{\text{nd}} \sum_{i=0}^{\text{nd}} \|\text{DM}(d_i)\|^2}, \tag{6}$$

where DM(d_i) and DM(d_j) represent the ith and jth rows in DM. λ is the regularization parameter that is used for controlling kernel bandwidth. nd stands for the number of unique diseases.

2.3.4 Disease–Disease Semantic Similarity (SIM_DD)

In the MeSH database, diseases have tree numbers that describe DAGs containing the disease. We searched unique diseases from CircR2Disease in the MeSH database and gathered tree numbers and entry terms of diseases. From 75 unique diseases from the benchmark dataset, we found 68 diseases in the MeSH database. As proposed in [11], we calculate the semantic similarity between disease d_i and disease d_j as follows:

$$\text{SIM\_DD}(d_i, d_j) = \frac{\sum_{e \in N_{d_i} \cap N_{d_j}} (D_{d_i}(e) + D_{d_j}(e))}{\text{DV}(d_i) + \text{DV}(d_j)}, \tag{7}$$

$$\text{DV}(d) = \sum_{e \in N_d} D_d(e), \tag{8}$$

$$D_d(e) = -\log\left(\frac{\text{num(DAGs}(e))}{\text{num(diseases)}}\right), \tag{9}$$

where num(DAGs(e)) represents the number of DAGs containing disease e, and num(diseases) represents the number of all diseases. If there are no common trees between diseases, their similarity score is set to 0.

2.3.5 Feature Concatenation

The fusion matrix (FUSION_MATRIX) is formed by concatenating feature vectors horizontally for each circRNA–disease pair. For example, if we want to create a fusion vector that represents the association between circRNA c_i and disease d_i using GIP_CD, GIP_DC, GIP_DM, and SIM_DD features, the fusion vector is formed by joining GIP_CD(c_i), GIP_DC(d_i), GIP_DM(d_i), and SIM_DD(d_i) vectors horizontally.

2.3.6 Generating circRNA–Disease Pair Embeddings

Autoencoders are a type of neural network that learns unique encodings of data in an unsupervised way. Several studies apply autoencoder neural networks to reveal unique hidden encodings in biological data [21, 22, 28–30]. In this study, we used a deep autoencoder (DAE) neural network as the feature extraction model. Autoencoders are composed of two symmetrical parts called the encoder and decoder and an encoded input vector in the middle of the encoder and decoder parts. The basic idea behind autoencoders is discovering a representation vector that can store all the critical information in a reduced-size vector such that the encoded input vector can be used to create the original input in the output vector. Deep autoencoders have one or more hidden layers in both the encoder and decoder parts.

In our methodology, a deep autoencoder with three hidden layers where the middle layer is the encoding layer is built. In Fig. 2, the structure of our deep autoencoder model is shown. The input and output layers are represented as I and O, respectively. Hidden layers are represented as H_1, H_2, and H_3. Input, output, and hidden layer sizes are selected based on our experiments. The rule is that layer sizes of H_1 and H_3 are equal and proportional to input and encoding layer sizes. The layer size of H_1 and H_3 is calculated as follows:

$$\text{size}(H_1) = \frac{\text{size}(I) + \text{size}(H_2)}{2}, \tag{10}$$

$$\text{size}(H_3) = \text{size}(H_1), \tag{11}$$

where the size function gets the layer size of the given layer. As an example, if the input layer and encoding layer sizes are 784 and 256, respectively, the layer sizes in the deep autoencoder are 784, 520, 256, 520, and 784, respectively.

The ReLU activation function is used in all layers. Adam optimizer is used as the optimization function.

2.3.7 Prediction of circRNA–Disease Associations

In this study, a feed-forward deep neural network (FF-DNN) model is used to predict circRNA–disease associations. The DNN model has one input layer, one output layer, and five hidden layers. All layers in the model are fully connected. In Fig. 2, the structure of the deep neural network model is shown. ReLU activation function is used in hidden layers and the sigmoid function is used in the output layer. Adam optimizer is used for optimization.

2.4 Performance Evaluation

Fivefold cross-validation (CV) is performed to have reliable performance scores of the prediction model. In all experiments, we used a stratified random sampling method for CV. The stratified random sampling method [31] generates subsets of the data with the same proportions of labels in subsets. Each fold in the fivefold CV will have the same proportion of 0 and 1 labeled associations in our example. The area under the receiver operating characteristics curve (AUC) [32] is used to compare the overall performance. Receiver operating characteristic (ROC) [33] curves display the performance based on true-positive and false-positive rates of the model. Accuracy, precision, recall, and F-1 metric formulas are described below.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \tag{12}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{13}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{14}$$

$$F\text{-}1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{15}$$

3 Results and Discussion

In this section, we run some experiments to find out the best parameters in our methodology. We also performed some experiments to measure the performance of our methodology and compared it with the state-of-the-art methodologies in the literature.

In all experiments, the following rules are applied:

- The DAE structure presented in Fig. 2 is used, but the layer sizes can vary according to the experiment.
- DNN structure presented in Fig. 2 is used, but the layer sizes can vary according to the experiment.
- Fivefold cross-validation with a stratified random sampling method is performed to have reliable results.
- AUC score is used as a performance metric for comparisons.

In the model optimization part, we ran our methodology with different DAE and DNN layer sizes to find out the best structure for both models. Secondly, we showed the performance of our methodology on two different datasets.

In the final part, we compared our methodology with state-of-the-art circRNA–disease association prediction methodologies.

3.1 Model Optimization

Selecting the optimal embedding size of circRNA–disease pairs and layer sizes in the prediction model is very critical in model optimization. Embedding size should be fair enough to represent circRNA–disease pairs so that the representation vector has a low vector size to run models efficiently and does not lose information. DNN prediction model also should have an appropriate structure to predict circRNA–disease associations successfully. No best model structure works for every data and problem, so we need to perform experiments with different values to find out optimal parameters. We used default values for some model parameters such as the optimization algorithm and learning rate. Adam optimizer with a 0.001 learning rate is used in models.

In this section, we run our methodology with different encoding sizes and DNN hidden layer sizes. AUC score is considered as the comparison metric. We run the methodology with four different embedding sizes (128, 256, 512, 1024) and three different hidden layer size combinations ((256, 512, 256, 128, 64), (512, 512, 256, 128, 64), (1024, 512, 256, 128, 64)). As shown in Table 1, the best AUC score, 0.9794, is obtained with an embedding size of 512 and a DNN hidden layer size combination of 1024, 512, 256, 128, and 64. Combining GIP_CD, GIP_DC, GIP_DM, and SIM_DD features generates a vector of 784 for a circRNA–disease pair. The autoencoder model generates the best model with an embedding size of 512 which means the autoencoder model finds hidden features from input features and reduced dimension at the same time.

Table 1 Fivefold CV results on different embedding sizes and DNN layer sizes

Embedding size	DNN hidden layer sizes	AUC score ∓ std. dev.
512	(1024, 512, 256, 128, 64)	0.9794 ∓ 0.0078
1024	(256, 512, 256, 128, 64)	0.9792 ∓ 0.0089
1024	(1024, 512, 256, 128, 64)	0.9791 ∓ 0.0077
1024	(512, 512, 256, 128, 64)	0.9791 ∓ 0.0069
512	(512, 512, 256, 128, 64)	0.9776 ∓ 0.0082
128	(256, 512, 256, 128, 64)	0.9742 ∓ 0.0065
128	(512, 512, 256, 128, 64)	0.9736 ∓ 0.0073
256	(1024, 512, 256, 128, 64)	0.9735 ∓ 0.0132
512	(256, 512, 256, 128, 64)	0.9709 ∓ 0.0150
128	(1024, 512, 256, 128, 64)	0.9709 ∓ 0.0032
256	(512, 512, 256, 128, 64)	0.9701 ∓ 0.0126
256	(256, 512, 256, 128, 64)	0.9690 ∓ 0.0092

3.2 Comparison with Alternative Methods

In this part, we created two alternative methods to find out the contribution of the DAE model. Alternative methods are described below:

- Fusion-DNN: To see whether generating embedding vectors with deep autoencoder contributed to the model or not, we created a methodology the same as the DCDA, but without the embedding generation part. Instead of running a deep autoencoder to get embeddings, fusion vectors of circRNA–disease pairs are directly fed to the DNN prediction model.
- PCA–DNN: Another method is built which is similar to DCDA, but circRNA–disease pair embeddings are generated not with DAE, but with principal component analysis (PCA) with the same encoding dimension used in DAE, which is set to 512.

Table 2 presents the results of alternative methodologies. As observed in the table, all alternative methods successfully predict circRNA–disease associations. This indicates that we have generated high-quality features for circRNAs and diseases. When comparing the Fusion-DNN model, which directly utilizes the generated features, with the other two models that employ modeling techniques such as PCA or autoencoder for vector transformation, an approximate 2% improvement is observed. We believe this improvement stems from the fact that the vectors can contain some level of noise due to their high dimensionality, and the discovery of interaction relationships enables the creation of better vectors. Additionally, it is observed that using autoencoder instead of PCA performs similarly in modeling circRNA–disease pairs, but yields better results in terms of AUC score.

Table 2 Fivefold CV performance scores of the DCDA model on CircR2Disease dataset

Model	Accuracy	Precision	Recall	*F*-1	AUC
DCDA	0.9241 ∓ 0.01	0.9262 ∓ 0.01	0.9013 ∓ 0.02	0.9531 ∓ 0.02	0.9794 ∓ 0.01
PCA-DNN	0.9304 ∓ 0.01	0.9326 ∓ 0.01	0.9047 ∓ 0.02	0.9624 ∓ 0.01	0.9713 ∓ 0.01
Fusion-DNN	0.9038 ∓ 0.02	0.9067 ∓ 0.02	0.8797 ∓ 0.03	0.9375 ∓ 0.05	0.9532 ∓ 0.01

Table 3 Fivefold CV performance scores of the DCDA model on CircR2Disease dataset

Fold	Accuracy	Precision	Recall	*F*-1	AUC
1	0.9180	0.9219	0.8794	0.9688	0.9758
2	0.9063	0.9091	0.8824	0.9375	0.9760
3	0.9141	0.9147	0.9077	0.9219	0.9686
4	0.9412	0.9438	0.9065	0.9844	0.9865
5	0.9412	0.9416	0.9308	0.9528	0.9900
Average	0.9241 ∓ 0.01	0.9262 ∓ 0.01	0.9013 ∓ 0.02	0.9531 ∓ 0.02	0.9794 ∓ 0.01

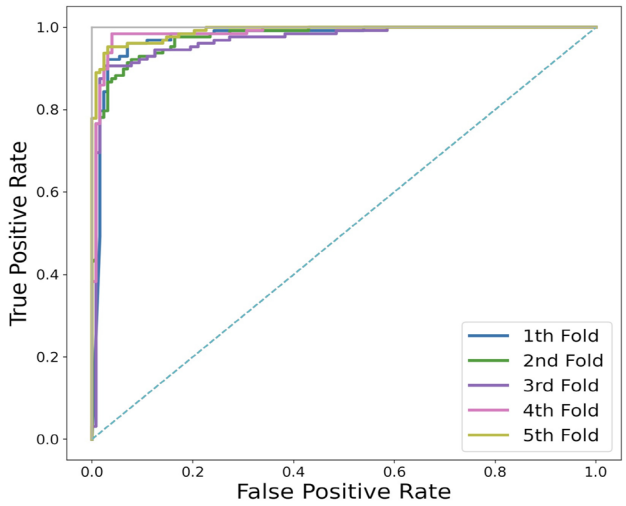


Fig. 3 Fivefold CV ROC curves performed by the DCDA model on CircR2Disease dataset

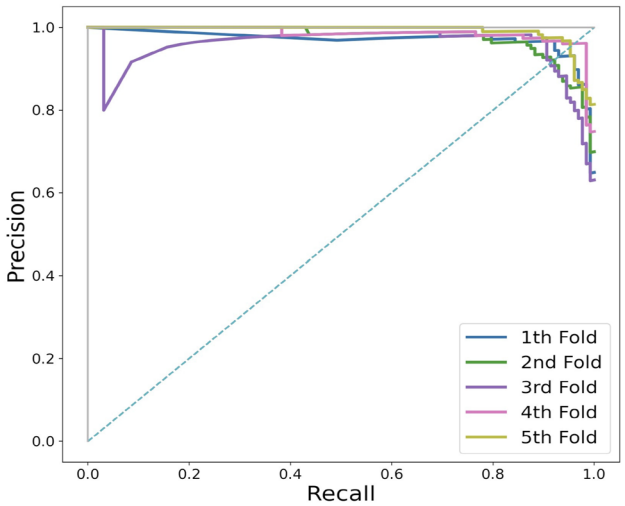


Fig. 4 Fivefold CV PR curves performed by the DCDA model on CircR2Disease dataset

3.3 Performance on the Benchmark Dataset

To evaluate DCDA's prediction performance, we performed a fivefold cross-validation on the benchmark dataset. We obtained 0.9241, 0.9262, 0.9013, 0.9531, and 0.9794 average scores of accuracy, *F*-1, precision, recall, and AUC, respectively. Table 3 shows performance scores in each fold and the average score with standard deviation (std). Moreover, we calculated true-positive rates and false-positive rates in each fold and plotted the ROC curves as shown in Fig. 3. PR curves of each fold are represented in Fig. 4.

3.4 Performance on circ2Disease

To see the performance of the DCDA, we run our methodology on a different dataset. We selected the Circ2Disease dataset in this experiment. We performed fivefold cross-validation on the data and obtained 0.8506, 0.8639, 0.8038, 0.9367, and 0.9425 average scores of accuracy, *F*-1, precision, recall, and AUC, respectively. Table 4 shows the performance scores in each fold and the average score with standard deviation (std. dev.). Figure 5 shows the ROC curves and Fig. 6 shows the PR curves of each fold.

3.5 Performance Comparison of Various Methods

We compared the DCDA model's performance with five state-of-the-art studies including MSFCNN [17], AE–RF [22], AE–DNN [21], iCircDA-MF [23], and KATZHCDA [15].

To make a reliable comparison, we compared studies based on fivefold CV results on the same dataset, CircR2Disease. All compared methodologies used only human-related circRNA–disease associations, but applied different filtering processes. Unlike DCDA, these five studies did not use circRNA or disease synonyms to have a consistent dataset. MSFCNN, AE–RF, and AE–DNN are based on machine learning prediction models, KATZHCDA is a graph-based prediction model, and iCircDA-MF's prediction model is based on matrix factorization. Like in our methodology, AE–DNN and AE–RF use an autoencoder model for feature extraction. As shown in Table 5, DCDA is superior to the other state-of-the-art methods with an AUC score of 0.9794.

Table 4 Fivefold CV performance scores of the DCDA model on Circ2Disease dataset

Fold	Accuracy	*F*-1	Precision	Recall	AUC
1	0.9167	0.9189	0.8947	0.9444	0.9688
2	0.8224	0.8348	0.7869	0.8889	0.9203
3	0.7757	0.8095	0.7083	0.9444	0.9261
4	0.8411	0.8522	0.7903	0.9245	0.9345
5	0.8972	0.9043	0.8387	0.9811	0.9630
Average	0.8506 ∓ 0.05	0.8639 ∓ 0.04	0.8038 ∓ 0.06	0.9367 ∓ 0.03	0.9425 ∓ 0.02

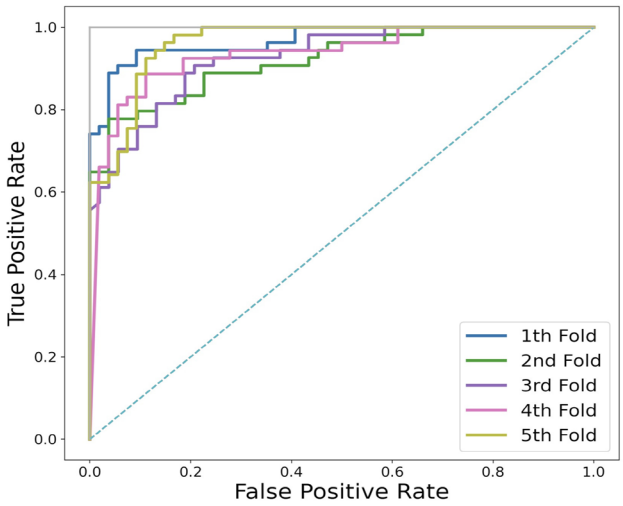


Fig. 5 Fivefold CV ROC curves performed by the DCDA model on the Circ2Disease dataset

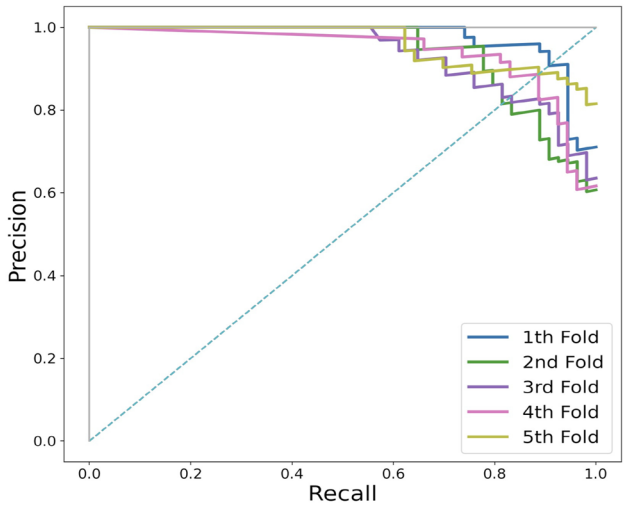


Fig. 6 Fivefold CV PR curves performed by the DCDA model on the Circ2Disease dataset

3.6 Case Study

To verify the prediction performance of the DCDA methodology, we investigated the novel circRNA–disease pairs predicted by the DCDA methodology. For training, the benchmark dataset is used to train DAE and DNN. Then, we created a test set by taking all possible combinations between unique circRNAs and unique diseases in the benchmark dataset and excluded validated pairs from the benchmark dataset. As a result, we had 41,286 pairs.

Prediction results are sorted by the prediction score of circRNA–disease pairs and the highest top-10 novel associations are examined. If the predicted association is found in the literature, the PMID of the study is listed in the table. As shown in Table 6, seven out of ten novel circRNA–disease associations are confirmed in publications. For example, novel associations between circRNA CDR1as with breast neoplasms, urinary bladder neoplasms, and glioma diseases are predicted by our methodology and research confirmed these associations.

3.7 Biological Insights

Recently, machine learning and artificial intelligence have been applied in many different fields including biology. When current experimental methods are insufficient, machine learning applications are prioritized to provide potential outcomes in general. In light of this, deep learning techniques are applied in various applications compromising feature representation, functional annotations, classifications, and novel predictions of circRNAs in terms of exposing biological insights [34].

Feature representation is one of the research topics. DNNs can learn complex feature representations from raw sequence data, enabling them to capture important patterns and motifs present in circRNA sequences. By training on known circRNA data, DNNs can extract informative features that contribute to circRNA prediction, potentially uncovering novel sequence characteristics associated with circRNA formation. DNNs can also be used for classification and prediction. Algorithms are trained to classify sequences as either circRNAs or non-circRNAs [35]. By learning from a large dataset of known circRNAs and non-circRNAs, DNN

Table 5 AUC scores of different circRNA–disease association prediction methodologies on CircR2Disease dataset

Method	DCDA	MSFCNN[a]	AE–RF[b]	AE–DNN[c]	iCircDA-MF[d]	KATZHCDA[e]
AUC	0.9794	0.9525	0.9486	0.9392	0.9178	0.7936

[a]Results obtained by the model reported by Fan et al. [17]

[b]Results obtained by the model reported by Deepthi and Jereesh [22]

[c]Results obtained by the model reported by Deepthi and Jereesh [21]

[d]Results obtained by the model reported by Wei and Liu [23] with parameter $k = 2, r = 70, \alpha = 2 \times 10 - 3, \beta = 1 \times 10 - 3$

[e]Results obtained by the model reported by Fan et al. [15]

Table 6 Top ten novel circRNA–disease association predictions by the DCDA methodology

Rank	Disease	circRNA	Evidence[a]
1	Breast neoplasms	CDR1as	31245927
2	Urinary bladder neoplasms	CDR1as	29694981
3	Glioma	CDR1as	26683098
4	Glioma	circBRAF	33650075
5	Glioma	Cir-ITCH	29887952
6	Esophageal neoplasms	Cir-ITCH	25749389
7	Carcinoma, pancreatic ductal	hsa_circ_0001649	29969694
8	Carcinoma, basal cell	hsa_circ_0000284	Unconfirmed
9	Urinary bladder neoplasms	hsa_circ_0001649	Unconfirmed
10	Colorectal neoplasms	hsa_circ_0067934	Unconfirmed

[a]The PMID of literature in PubMed that is evidence of the corresponding circRNA–disease association

models can identify common features or patterns that distinguish circRNAs from other RNA molecules. This can help in accurately predicting new circRNAs in genomic sequences. CircRNAs are generated through back-splicing events, where a downstream splice site is joined with an upstream splice site, forming a circular structure. DNNs can learn to recognize splicing patterns associated with circRNA formation by training on known circRNA sequences and their corresponding linear RNA counterparts. This can provide insights into the splicing mechanisms and factors involved in circRNA biogenesis [36].

Another important point is to accomplish a proper functional annotation process. DNNs can help to predict the functions and potential roles of circRNAs based on their sequence features and associations with other genomic elements. By integrating information from various genomic and transcriptomic datasets, DNN models can uncover potential interactions between circRNAs and RNA-binding proteins, microRNAs, or genomic loci. This can provide insights into the regulatory networks and molecular functions of circRNAs [37].

In addition to highlighting the biological mechanism behind it, DNNs can also assist in identifying novel circRNAs by analyzing large-scale RNA sequencing data as we have focused on in this study. By leveraging their ability to learn complex patterns from data, DNN models can detect potential circular structures that might have been missed by traditional bioinformatics approaches. This can lead to the discovery of new circRNA candidates and expand our understanding of the circRNA landscape [38].

Overall, the application of DNNs in circRNA prediction provides a powerful tool for uncovering the underlying biology of circRNAs in diseases such as cancer as a complex disease [39]. By integrating large-scale genomic and transcriptomic data and learning from known circRNA examples, DNN models can reveal important sequence features, splicing patterns, and functional associations, and even identify novel circRNAs, contributing to a deeper understanding of circRNA biology.

4 Conclusion

In recent years, the importance of circRNAs in diagnoses of diseases and disease treatments come to light. Improvement of prediction algorithms, especially deep learning models, stimulated using machine learning algorithms while discovering disease-related circRNAs. In this thesis, a deep learning-based circRNA–disease association prediction methodology, called DCDA, is presented. We generated circRNA and disease feature vectors using different sources of information such as circRNA–miRNA associations, disease–miRNA associations, circRNA sequence similarity, and disease semantic similarities. We created circRNA and disease synonym dictionaries to have a common language between various datasets. We generated several feature vectors for circRNAs and diseases to find out optimal features. As a consequence of feature selection experimentation, features derived from known circRNA–disease interactions and disease–miRNA associations are selected to generate circRNA and disease feature vectors. We represented circRNA–disease pairs by concatenating circRNA and disease feature vectors and generated representative pair vectors in lower-dimensional space through a deep autoencoder.

Finally, we built a deep neural network to predict whether the pair is associated or not.

To identify the predictive performance of our methodology, we performed several experiments. We tested DCDA on the CircR2Disease dataset, which is a widely used dataset in the literature, and showed that the performance of DCDA has undeniable performance with an AUC score of 0.9794. We also tested DCDA on another widely used dataset, Circ2Disease, and achieved an AUC score of 0.9425. Besides these experimentations, we compared our methodology with state-of-the-art methodologies in the literature. Our model outperformed the top-ranking methodology, MSFCNN, with a 2.82% AUC score. Relying on DCDA's prospering performance in all experimentations, we looked at what DCDA predicts about unknown circRNA–disease pairs. When we investigated circRNA–disease pairs that have the highest predictive association score, we saw that seven out of the top ten pair associations are confirmed in the literature.

This research is useful in several ways. First, the circRNA–disease pair embeddings generated by the deep autoencoder model can be used in other related studies. Based on the successful performance of DCDA, unconfirmed predicted circRNA–disease pairs are worthy of analysis and study by biologists. Starting the study with the most likely associated circRNAs can save time and resources for unlikely circRNA–disease pairs.

Code Availability The source code of DCDA methodology and used datasets are available at https://github.com/hacertilbec/DCDA.

Declarations

Conflict of Interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Holdt L, Kohlmaier A, Teupser D (2018) Molecular roles and function of circular RNAs in eukaryotic cells. Cell Mol Life Sci 75(6):1071–1098. https://doi.org/10.1007/s00018-017-2688-5
2. Sanger H, Klotz G, Riesner D et al (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. https://doi.org/10.1073/pnas.73.11.3852
3. Hsu M, Coca-Prados M (1979) Electron microscopic evidence for the circular form of RNA in the cytoplasm of eukaryotic cells. Nature 280(5720):339–340. https://doi.org/10.1038/280339a0
4. Lu M (2020) Circular RNA: functions, applications and prospects. ExRNA 2(1):1–7. https://doi.org/10.1186/s41544-019-0046-5
5. Wang X, Zhu X, Zhang H et al (2018) Increased circular RNA hsa_circ_0012673 acts as a sponge of miR-22 to promote lung adenocarcinoma proliferation. Biochem Biophys Res Commun 496(4):1069–1075. https://doi.org/10.1016/j.bbrc.2018.01.126
6. Fan C, Lei X, Fang Z et al (2018) CircR2Disease: a manually curated database for experimentally supported circular RNAs associated with various diseases. Database. https://doi.org/10.1093/database/bay044
7. Rophina M, Sharma D, Poojary M et al (2020) Circad: a comprehensive manually curated resource of circular RNA associated with diseases. Database. https://doi.org/10.1093/database/baaa019
8. Yao D, Zhang L, Zheng M et al (2018) Circ2Disease: a manually curated database of experimentally validated circRNAs in human disease. Sci Rep 8(1):1–6. https://doi.org/10.1038/s41598-018-29360-3
9. Meng X, Hu D, Zhang P et al (2019) CircFunBase: a database for functional circular RNAs. Database. https://doi.org/10.1093/database/baz003
10. Le N (2022) Potential of deep representative learning features to interpret the sequence information in proteomics. Proteomics. https://doi.org/10.1002/pmic.202100232
11. Yuan Q, Chen K, Yu Y et al (2023) Prediction of anticancer peptides based on an ensemble model of deep learning and machine learning using ordinal positional encoding. Brief Bioinform. https://doi.org/10.1093/bib/bbac630
12. Mary S, Kumar V, Venkatesan K et al (2022) Vulture-based AdaBoost-feedforward neural frame work for COVID-19 prediction and severity analysis system. Interdiscip Sci: Comput Life Sci 14:582–595. https://doi.org/10.1007/s12539-022-00505-3
13. Lai F, Gao F (2023) Auto-Kla: a novel web server to discriminate lysine lactylation sites using automated machine learning. Brief Bioinform. https://doi.org/10.1093/bib/bbad070
14. Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. Adv Neural Inf Process Syst. https://doi.org/10.48550/arXiv.1706.03762
15. Fan C, Lei X, Wu FX (2018) Prediction of CircRNA-disease associations using KATZ model based on heterogeneous networks. Int J Biol Sci 14(14):1950. https://doi.org/10.7150/ijbs.28260
16. Wang L, You ZH, Huang YA et al (2020) An efficient approach based on multi-sources information to predict circRNA-disease associations using deep convolutional neural network. Bioinformatics 36(13):4038–4046. https://doi.org/10.1093/bioinformatics/btz825
17. Fan C, Lei X, Pan Y (2020) Prioritizing CircRNA-disease associations with convolutional neural network based on multiple similarity feature fusion. Front Genet 11:1042. https://doi.org/10.3389/fgene.2020.540751
18. Wang L, You ZH, Li LP et al (2019) Predicting circRNA-disease associations using deep generative adversarial network based on multi-source fusion information. In: 2019 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE, pp 145–152. https://doi.org/10.3389/fgene.2019.00832
19. Zheng K, You ZH, Li JQ et al (2020) iCDA-CGR: identification of circRNA-disease associations based on chaos game representation. PLOS Comput Biol 16(5):e1007872. https://doi.org/10.1371/journal.pcbi.1007872
20. Ge E, Yang Y, Gang M et al (2020) Predicting human disease-associated circRNAs based on locality-constrained linear coding. Genomics 112(2):1335–1342. https://doi.org/10.1016/j.ygeno.2019.08.001
21. Deepthi K, Jereesh A (2020) An ensemble approach for CircRNA-disease association prediction based on autoencoder and deep neural network. Gene 762:145040. https://doi.org/10.1016/j.gene.2020.145040
22. Deepthi K, Jereesh A (2021) Inferring potential CircRNA-disease associations via deep autoencoder-based classification. Mol Diagn Therapy 25(1):87–97. https://doi.org/10.1007/s40291-020-00499-y
23. Wei H, Liu B (2020) iCircDA-MF: identification of circRNA-disease associations based on matrix factorization. Brief Bioinform 21(4):1356–1367. https://doi.org/10.1093/bib/bbz057

24. Wang L, You ZH, Li YM et al (2020) GCNCDA: a new method for predicting circRNA-disease associations based on graph convolutional network algorithm. PLOS Comput Biol 16(5):e1007568. https://doi.org/10.1371/journal.pcbi.1007568
25. Huang Z, Shi J, Gao Y et al (2019) HMDD v3.0: a database for experimentally supported human microRNA–disease associations. Nucleic Acids Res 47(D1):D1013–D1017. https://doi.org/10.1093/nar/gky1010
26. Lipscomb CE (2000) Medical subject headings (mesh). Bull Med Libr Assoc 88(3):265
27. Van Laarhoven T, Nabuurs S, Marchiori E (2011) Gaussian interaction profile kernels for predicting drug–target interaction. Bioinformatics 27(21):3036–3043. https://doi.org/10.1093/bioinformatics/btr500
28. Chen L, Cai C, Chen V et al (2016) Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model. In: BMC bioinformatics, BioMed Central, pp 97–107. https://doi.org/10.1186/s12859-015-0852-1
29. Chicco D, Sadowski P, Baldi P (2014) 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. https://doi.org/10.1145/2649387.2649442
30. Tan J, Hammond JH, Hogan DA et al (2016) Adage-based integration of publicly available pseudomonas aeruginosa gene expression data with denoising autoencoders illuminates microbe-host interactions. MSystems. https://doi.org/10.1128/mSystems.00025-15
31. de Vries PG (1986) Stratified random sampling. In: Sampling theory for forest inventory. Springer, Berlin, pp 31–55. https://doi.org/10.1007/978-3-642-71581-5_2
32. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1):29–36
33. Fan J, Upadhye S, Worster A (2006) Understanding receiver operating characteristic (ROC) curves. Can J Emerg Med 8(1):19–20. https://doi.org/10.1017/s1481803500013336
34. Lasantha D, Vidanagamachchi S, Nallaperuma S (2023) Deep learning and ensemble deep learning for circRNA-RBP interaction prediction in the last decade: a review. Eng Appl Artif Intell. https://doi.org/10.1016/j.engappai.2023.106352
35. Rebolledo C, Silva J, Saavedra N et al (2023) Computational approaches for circRNAs prediction and in silico characterization. Brief Bioinform. https://doi.org/10.1093/bib/bbad154
36. Qu S, Yang X, Li X et al (2015) Circular RNA: a new star of noncoding RNAs. Cancer Lett. https://doi.org/10.1016/j.canlet.2015.06.003
37. Jiao S, Wu S, Huang S et al (2021) Advances in the identification of circular RNAs and research into circRNAs in human diseases. Front Genet. https://doi.org/10.3389/fgene.2021.665233
38. Hansen T, Venø M, Damgaard C et al (2016) Comparison of circular RNA prediction tools. Nucleic Acids Res. https://doi.org/10.1093/nar/gkv1458
39. Bach DH, Lee SK, Sood AK (2019) Circular RNAs in cancer. Mol Ther Nucleic Acids 16:118–129. https://doi.org/10.1016/j.omtn.2019.02.005