IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 22, 2019, accepted September 6, 2019, date of publication September 11, 2019,
date of current version September 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2940470

CGMDA: An Approach to Predict and Validate MicroRNA-Disease Associations by Utilizing Chaos Game Representation and LightGBM

KAI ZHENG[1,2], LEI WANG[3], AND ZHU-HONG YOU[4], (Member, IEEE)
[1]School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China
[2]Mine Digitization Engineering Research Center of Ministry of Education of the People's Republic of China, Xuzhou 221116, China
[3]College of Information Science and Engineering, Zaozhuang University, Zaozhuang 277100, China
[4]Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Science, Ürümqi 830011, China

Corresponding authors: Lei Wang (leiwang@ms.xjb.ac.cn) and Zhu-Hong You (zhuhongyou@ms.xjb.ac.cn)

This work was supported in part by the Double-First Class Discipline Construction Project under Grant 2018ZZCX14.

ABSTRACT Recent studies have shown that microRNAs (miRNAs) play an important role in complex human diseases. Identifying potential miRNA-disease associations is useful for understanding the pathogenesis. However, there are currently only a few methods proposed to predict miRNA-disease association based on sequence information. And these methods can only quantify nonlinear sequence relationships without taking linear sequence information into account. In this work, we designed a computational method for predicting miRNA-disease association based on chaos game representation, called CGMDA, to overcome these problems. CGMDA combines association information with miRNA sequence information, miRNA functional information and disease semantic information to improve prediction accuracy. In particular, we use chaos game representation (CGR) technology for the first time to transform miRNA sequence information into image information and extract its features. In the cross-validation experiment, CGMDA achieved a mean the area under the receiver operating characteristic curve (AUC) of 0.9099 on the HMDD v3.0 data set. To better evaluate the performance of CGMDA, we compared it to different classifiers and related prediction methods. In addition, CGMDA is applied to three human complex diseases. The results showed that of the top 40 disease-related miRNAs predicted, 39 (Breast Neoplasm), 39 (Lymphoma) and 38 (Colon Neoplasm) were validated by experiments in case studies. These experimental results show that CGMDA is a reliable tool and has potential application prospects in assisting early diagnosis and treatment of prognosis.

INDEX TERMS miRNAs, chaos game representation, disease, heterogenous information, LightGBM.

I. INTRODUCTION

MicroRNAs (miRNAs) are small RNAs that are 20 to 25 nucleotides in length [1], [2]. Line-4 and let-7 are the first two miRNAs discovered in the past two decades [3], [4]. Since then, many miRNAs have been revealed and identified by using different biological experimental methods, which gives new insights into the functions and regulatory mechanisms of miRNAs. The experiment proves that many miRNAs are specifically expressed in certain types of diseases, including arthritis, adenoid cystic, arteriosclerotic occlusive disease, immune thrombocytopenic purpura, and idiopathic pulmonary hypertension [5]–[10]. For example, Bang *et al.* found that miR-23, miR-27 and miR-24 cluster have the relation to angiogenesis and endothelial apoptosis with the progress of cardiac ischemia and retinal vascular, and are also the key of cardiovascular angiogenesis [11]. For the above reasons, exploring the potential association between miRNA and disease is gradually concerned by scholars [12]–[15]. However, the high experimental cost, long experimental cycle and sensitivity of noise may hinder the validation of potential miRNA-disease associations through biological experiments. Therefore, it is necessary to find more effective calculation methods to assist biological experiments to provide effective association candidates to promote the development of biomedicine.

Since existing methods cannot accurately measure miRNA attribute information based on incompletely correlated biological information or only nonlinear sequence relationships are considered. In this study, we introduce a new

The associate editor coordinating the review of this manuscript and approving it for publication was Chintan Amrit.

computational approach of Chaos Game Representation for predicting miRNA-Disease Association called CGMDA to try to overcome the above problems. The proposed method integrates manifold sources including miRNA sequence information, miRNA functional similarity information, disease semantic similarity information, and known miRNA-disease association information. The three advantages of our approach are as follows: (1) miRNA sequence information can accurately measure miRNA property information. (2) The introduction of chaos game representation method can provide new ideas for extracting sequence features. (3) Imaging the sequence information can align the features.

In order to better verify the robustness and reliability of the method, the experiment was designed. 5-fold cross-validation was used to evaluate the performance of CGMDA on the HMDD V3.0 dataset, resulting in the AUC of 90.99%. In addition, CGMDA is applied to Breast Neoplasms, Lymphoma and Colon Neoplasms, and the accuracy of the first 40 predicted miRNAs in other databases was 97.7%, 97.5% and 95%, respectively. The above experimental results prove that the proposed method is reliable and robust. We hope that the introduction of chaos game representation can provide a new perspective for extracting sequence feature research. In particular, we introduce chaos game representation into miRNA disease prediction models for the first time and hope to provide a new perspective for extracting nucleic acid sequence features.

II. RELATED WORK

In recent years, more and more prediction methods for underlying disease-miRNA associations have been discovered. There have been two types of classical calculation methods, similarity-based measures methods and machine learning-based methods. Shi *et al.* established a computational method based on a random walk algorithm to identify unknown miRNA-disease associations [16]. Xu *et al.* established a prioritization method that does not require miRNA-disease association information to prioritize disease-related miRNAs [17]. Chen *et al.* Inferred potential miRNA-disease interactions by implementing random walks on miRNA-miRNA functional similarity networks using global network similarity measures [18]. Li *et al.* proposed an algorithm for predicting potential disease-related miRNA by updating adjacency matrices more efficiently based only on known miRNA-disease association information [19]. Next, machine learning-based approaches were introduced. Xu *et al.* developed a predictive method based on support vector machine (SVM) for candidate miRNAs in prostate cancer [20]. Wang *et al.* proposed a new method of Logistic Model Tree for predicting miRNA-Disease Association and extracted miRNA sequence information for the first time using natural language processing techniques [21].

III. MATERIALS AND METHODS

A. HUMAN MIRNA-DISEASE ASSOCIATIONS

The HMDD (Human MicroRNA Disease Database) dataset provided by Li *et al.* provides experimental support for human miRNA and disease association [22]. The latest version of the HMDD dataset now collects 32,281 miRNA disease associations, including 1,102 miRNAs and 850 diseases from 17,412 papers. The home page of the data set is http://www.cuilab.cn/hmdd. When preprocessing the dataset, we excluded some miRNAs that cannot find the corresponding sequence information in the public database miRBase [23]. After screening, we selected miRNA-disease association pairs constructed by 1057 miRNAs and 850 diseases as positive set in the experiment. When disease $d(i)$ and miRNA$m(j)$ that have association are verified in the HMDD v3.0 database, the element $X(i,j)$ of the adjacency matrix X is equal to 1, otherwise it is equal to 0, where $d(i)$ is the i-th disease and $m(j)$ is the j-th miRNA [24].

B. MIRNA FUNCTIONAL SIMILARITY

Based on the hypothesis that pathologically similar diseases are affected by functionally similar miRNAs and vice versa, Wang et al. proposed an algorithm for calculating the similarity of miRNA functions [25]. The MiRNA functional similarity score was uploaded to http://www.cuilab.cn/files/images/cuilab/misim.zip. In this method, we download and construct a 495-line × 495-column miRNA functional similarity matrix $\boldsymbol{FS}$ as miRNA functional similarity information, where the entity $\boldsymbol{FS}(\boldsymbol{a},\boldsymbol{b})$ represents the similarity score between miRNA $\boldsymbol{m(a)}$ ***and*** $\boldsymbol{m(b)}$. $\boldsymbol{a}$ and $\boldsymbol{b}$ are the serial numbers of miRNA. The data used to calculate the functional similarity of miRNAs comes from the HMDD database and there is a possibility that the inclusion of label information in the feature would result in inaccurate results. Therefore, this data is only added to the feature in the case study.

C. DISEASE SEMANTIC SIMILARITY MODEL

The medical subject term (MeSH) is a disease descriptor to rigorously classify diseases which can be downloaded from the medical library (http://www.nlm.nih.gov/), and its hierarchical information can reflect the relationship between miRNA-related diseases. Diseases can be described as a directed acyclic graph (DAG) based on MeSH, where the edge is the relationship between the diseases and the node is the disease [26]. If the MeSH of disease $d(j)$ is a subset of the MeSH of disease $d(i)$, then $d(j)$ is the parent node of $d(i)$ and $d(i)$ is the child node of $d(j)$. For example, ''Neoplasms by histologic type'' (C04.577) is the parent node of ''Neoplasms, glandular and epithelial'' (C04.557.470). Therefore, the disease $d(i)$ can be represented by $\mathrm{DAG}_{d(i)} = (d(i), T_{d(i)}, W_{d(i)})$, where $T_{d(i)}$ is the set of ancestor nodes including $d(i)$ and $W_{d(i)}$ is the set of edges between diseases. The disease semantic information calculated by the above DAG can reflect the attributes of the disease and enrich the information contained in the features. In addition, $d(i)$ is the i-th disease of all 850 diseases used. Here, the previous method provided by Xuan *et al.* based on the MeSH disease descriptor was used to calculate the semantic similarity of the disease [27]. In particular, the semantic value $D_{d(i)}(t)$ is considered to be the contribution of disease t to disease $d(i)$, as follows:

$$\begin{cases} \boldsymbol{D_{d(i)}\,(t) = 1 \quad if\; t = d(i)} \\ \boldsymbol{D_{d(i)}\,(t) = max\left\{\Delta * D_{d(i)}\left(t'\right) \mid t' \in \; children\; of\; t\right\}} \\ \qquad\qquad \boldsymbol{if\; t \neq d(i)} \end{cases} \tag{1}$$

where Δ is the semantic contribution attenuation factor, which we set to 0.5 according to previous research [26]. In addition, we defined the semantic value $DV(d)$ as follows:

$$DV(d) = \sum\nolimits_{t \in T_{d(i)}} D_{d(i)}(t) \tag{2}$$

If the diseases $d(i)$ and $d(j)$ have more common parts of their DAG maps, then the two diseases are semantically similar. We can calculate the semantic similarity values based on this conjecture, as defined below:

$$Sim1(d(i), d(j)) = \frac{\sum_{t \in T_{d(i)} \cap T_{d(j)}} \left(D_{d(i)}(t) + D_{d(j)}(t)\right)}{DV\,(d(i)) + DV(d(j))} \tag{3}$$

where $Sim1$ is semantic hierarchical information with 850 rows and 850 columns, and the elements $Sim1(d(i), d(j))$ are treated as semantic similarities of $d(i)$ and $d(j)$.

We consider the contribution of hierarchical information to semantic values in $Sim1$. However, each disease occurs at a different frequency in each DAG, and the less frequently occurring diseases have higher specificity in general. Therefore, in order to retain the term specific information, we define the second semantic value $D2_{d_{(i)}}(t)$ to quantify the contribution of disease t to disease $d_{(i)}$ as follows:

$$D2_{d_{(i)}}(t) = log(1 + \frac{number\ of\ DAGs\ includingt}{number\ of\ disease}) \tag{4}$$

The semantic similarity score $Sim2$ between disease $d(i)$ and $d(j)$ is defined as follows:

$$Sim2(d(i), d(j)) = \frac{\sum_{t \in T_{d(i)} \cap T_{d(j)}} \left(D2_{d(i)}(t) + D2_{d(j)}(t)\right)}{DV\ (d(i)) + DV(d(j))} \tag{5}$$

where $Sim2$ is semantic specificity information with 850 rows and 850 columns, and the elements $Sim2(d(i), d(j))$ are treated as semantic similarities of $d(i)$ and $d(j)$.

D. GAUSSIAN INTERACTION PROFILE KERNEL (GIPK) SIMILARITY FOR DISEASES AND MIRNA

The HMDD v3.0 dataset provides information about the associations that contains similarity information between diseases and between miRNAs, therefore Gaussian interaction profile kernel (GIPK) similarity are utilized to extract this information [28]. In detail, we describe the interaction profile of disease $d(a)$ with $d(a)$-associated miRNAs as $IP(d(a))$. Among them, the binary vector $IP(d(a))$ is composed of the a-th row vector of the adjacency matrix X. We described disease GIPK similarity between $d(a)$ and $d(b)$ as follow:

$$KD\,(d(a), d(b)) = \exp(-\gamma_d * ||IP\,(d(a)) - IP\,(d(b))\,||^2) \tag{6}$$

where the width parameter γ_d of the function can be calculated by normalizing the original parameters. nd is the number of diseases. The formula is as follows:

$$\boldsymbol{\gamma_d = \frac{1}{\frac{1}{nd}\sum_{i=1}^{nd} ||IP(d(i))||^2}} \tag{7}$$

The Gaussian interaction profile kernel ($GIPK$) similarity between miRNAs is calculated in the same way:

$$KM\,(m(a), m(b)) = exp(-\gamma_m * ||IP\,(m(a)) - IP(m(b))||^2) \tag{8}$$

$$\gamma_m = \frac{1}{\frac{1}{nm}\sum_{i=1}^{nm} ||IP(m(i))||^2} \tag{9}$$

In detail, we describe the interaction profile of miRNA $m(a)$ with $m(a)$-associated miRNAs as $IP(m(a))$. Among them, the binary vector $IP(m(a))$ is composed of the a-th column vector of the adjacency matrix X. nm is the number of miRNAs.

E. INTEGRATED SIMILARITY FOR DISEASES AND MIRNA

The disease similarity matrix SD was built to maximize the use of term hierarchical information, terminology specific information, and $GIPK$ similarity information [29]. The comprehensive similarity $SD\,(d(a), d(b))$ between disease $d(a)$ and $d(b)$ is expressed as follows:

$$SD\,(d(a), d(b)) = \begin{cases} \dfrac{Sim1\,(d(a), d(b)) + Sim2\,(d(a), d(b))}{2} & \\ & if\ \ d(a), d(b)\ in\ Sim1\ and\ Sim2 \\ KD\,(d(a), d(b)) & others \end{cases} \tag{10}$$

$GIPK$ similarity and functional similarity were used to build miRNA similarity. We calculated the similarity between miRNA $m(a)$ and $m(b)$ as follows:

$$SM\,(m(a), m(b)) = \begin{cases} FS\,(m(a), m(b))\ if\ \ m(a), m(b)\ in\ FS \\ KM\,(m(a), m(b))\ others \end{cases} \tag{11}$$

F. CHAOS GAME REPRESENTATION

Gene mutations can change the composition or sequence of amino acids in the polypeptide chain, thus affecting the biological functions of proteins or enzymes and causing abnormalities in the body's phenotype [30]. In general, the nonlinear sequence relationship of a mutated gene does not change significantly when no mutation occurs, and therefore, the expression of the gene is related to linear sequence information. However, most of the sequence comparison algorithms at this stage, like k-mer, only quantify nonlinear sequence relationships [31]. Therefore, a new algorithm for extracting sequence linear information is needed. The chaos game representation (CGR) derived from chaos theory is a mapping method of genome sequences proposed by Jeffrey in 1990 [32], [33]. The CGR has two advantages. One is

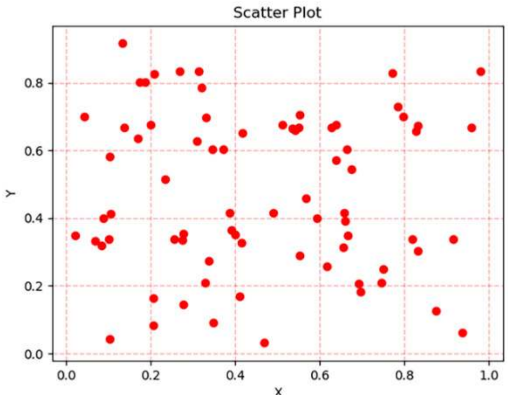

FIGURE 1. **CGR of the miRNA named hsa-mir-4661.**

that it uniquely maps each nucleotide in the sequence to the Euclidean space, and the coordinates of each nucleotide can restore all of the previously mapped nucleotide sequences without loss of information, in other words, the mapped coordinates contain linear relationship information. The second is that the same effect of k-mer can be achieved by the number of nucleotides in each interval, that is, the nonlinear relationship is quantified. Therefore, chaos game representation is introduced to extract linear sequence information from the sequences in this paper. Since only LMTRDA is a k-mer based miRNA-disease association predictor, in order to compare the effects of k-mer and chao game representation, we compare the results of LMTRDA and CGMDA in TABLE 3 [21]. The mapped Euclidean space is confined as four vertices by four possible nucleotides (Figure 1). The positions CGR_i is defined as follow:

$$CGR_i = CGR_{i-1} + \theta * (CGR_{i-1} - g_i) \tag{12}$$

$$g_i \begin{cases} (0,0) & if\ Nucleotide = A \\ (0,1) & if\ Nucleotide = C \\ (1,1) & if\ Nucleotide = G \\ (1,0) & if\ Nucleotide = U \end{cases} \tag{13}$$

where parameter θ is the decay factor. According to previous research, θ is set to 0.5 [33]. And we define $i = 1 \ldots n_G$ and $CGR_0 = (0.5, 0.5)$.

G. MIRNAS SEQUENCE FEATURE

Since microRNAs (miRNAs) are derived from distinct hairpin precursors (pre-miRNAs) that contain more information, we chose the sequences of pre-miRNAs. Firstly, we first downloaded the required 1057 miRNA precursor sequences from miRBase [23]. Secondly, we pigment the CGR of each miRNA and used it to build sequence feature matrixes. After that, sequence feature matrixes are converted into new matrixes whose shape is 640 × 5 by Singular Value Decomposition. Therefore, each miRNA sequence could be described by a 3200-dimensional vector based on reshape the sequence feature matrixes:

$$F_{seq} = \left(f_1, f_2, f_3, \ldots, f_{3199}, f_{3200}\right) \tag{14}$$

H. LIGHTGBM CLASSIFIER

Since traditional boosting algorithms (such as Gradient Boosting Decision Tree and eXtreme Gradient Boosting) need to scan all the sample points for each feature to select the best segmentation point, which makes them less efficient and computationally expensive to meet current needs. In order to reduce the cost of the experiment, lightGBM was selected as the classifier for this experiment [34]–[36]. LightGBM aims to optimize both row and column sub-sampling to improve training speed and prediction accuracy, that is Gradient Based One-side Sampling (GOSS) and Exclusive Feature Bundling (EFB). GOSS improves efficiency by distinguishing samples of different gradients, retaining samples of larger gradients and randomly sampling samples of smaller gradients to reduce the amount of computation. The Exclusive Feature Bundling (EFB) binds mutually exclusive features together in a histogram to form a feature to reduce feature dimensions. Therefore, the complexity of histogram constructing can be decreased from $O(data \times feature)$ to $O(data \times bundle)$ where $feature \gg bundle$.

I. METHOD OVERVIEW

The proposed prediction method includes four steps: 1. Construction of positive and negative sample sets; 2. Fusion of multi-source data into original feature vectors; 3. Abstracting primitive feature vector to get final feature identifier; 4. Building a better predictive method and predict potential association. After that, we will carefully introduce the details of each process. First, building positive and negative sample sets. The positive sample set consists of filtered, experimentally validated miRNA-disease associations in HMDD v3.0. There are summarily three steps of randomly selecting negative sample. Above all, one of the 850 diseases were chosen randomly; then a miRNA from the 1057 miRNAs was selected discretionarily; finally, we constituted a negative sample by using the disease and the miRNA which are not in 32226 known associations. We repeat this step until the same number of negative samples as the positive samples are obtained. Secondly, we fused multi-source data into the original feature vector. Among them, the disease feature vector is composed of the term hierarchical information $Sim1$, term specific information $Sim2$ and GIPK similarity information KD.

The integrated semantic similarity values stood for each disease as features. For example, we represented disease by a feature vector:

$$SD\left(d(a)\right) = (v_1, v_2, v_3, \ldots, v_{849}, v_{850}) \tag{15}$$

where the integrated similarity value between the diseases $d(a)$ and $d(b)$ is defined as v_b. The miRNA feature vector is composed of functional similarity information FS and GIPK similarity information KM. For example, we represented miRNA by a feature vector:

$$SM\left(m(a)\right) = (w_1, w_2, w_3, \ldots, w_{1056}, w_{1057}) \tag{16}$$

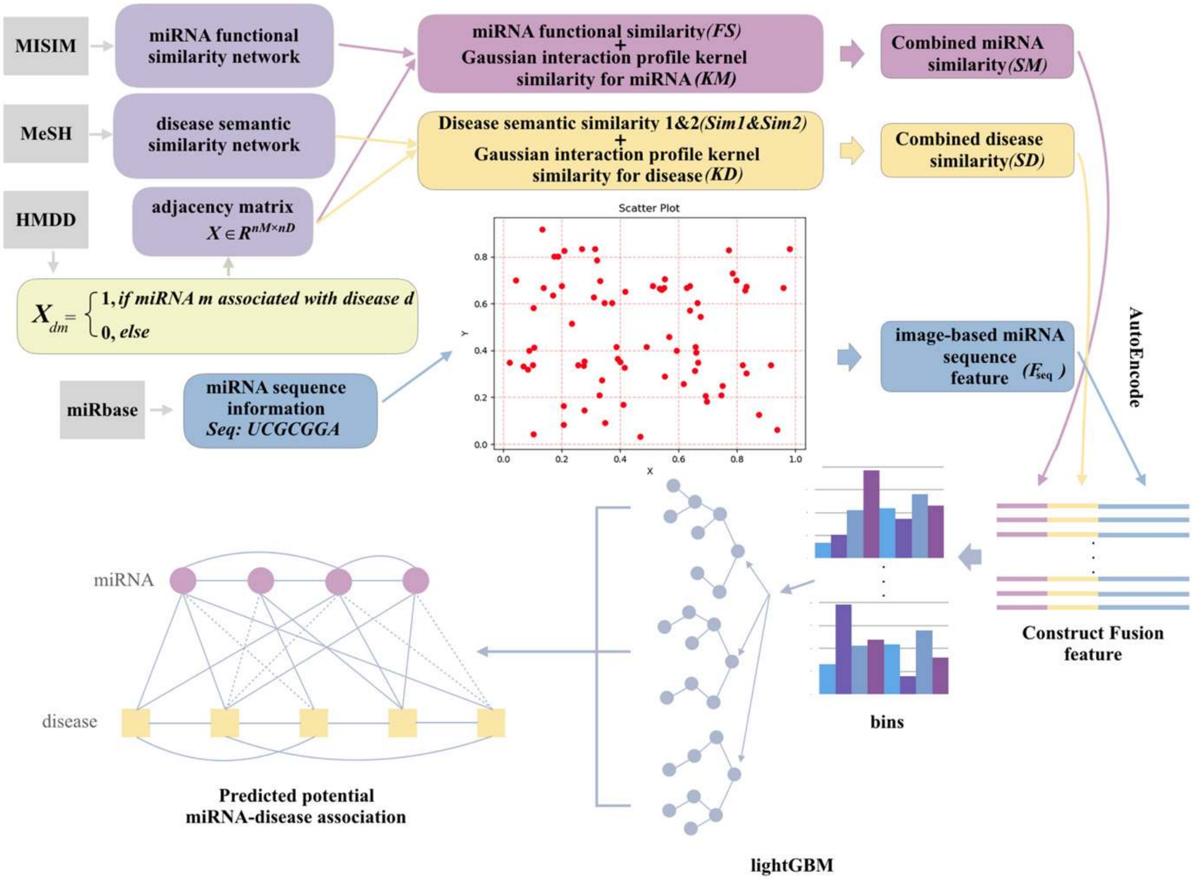
MISIM
miRNA functional similarity network
MeSH
disease semantic similarity network
HMDD
adjacency matrix $X \in R^{nM \times nD}$
$X_{dm} = \begin{cases} 1, \text{if miRNA m associated with disease d} \\ 0, \text{else} \end{cases}$
miRbase
miRNA sequence information
Seq: UCGCGGA
miRNA functional similarity(FS) + Gaussian interaction profile kernel similarity for miRNA (KM)
Disease semantic similarity 1&2(Sim1&Sim2) + Gaussian interaction profile kernel similarity for disease (KD)
Combined miRNA similarity(SM)
Combined disease similarity(SD)
Scatter Plot
0.0
0.2
0.4
0.6
0.8
1.0
X
Y
image-based miRNA sequence feature (F_{seq})
AutoEncode
Construct Fusion feature
bins
lightGBM
miRNA
disease
Predicted potential miRNA-disease association

FIGURE 2. The workflow of CGMDA to predict potential miRNA-disease associations.

TABLE 1. The comparison results of CGMDA and AUCs based on 5-fold cross validation.

Testing set	Accuracy	Recall	Precision	F1-score
1	84.26%	91.14%	80.12%	85.28%
2	84.94%	92.24%	80.49%	85.97%
3	86.04%	88.31%	84.49%	86.36%
4	85.97%	89.96%	83.32%	86.51%
5	85.31%	92.00%	81.15%	86.23%
Average	85.30±0.74%	90.73±1.62%	81.91±1.90%	86.07±0.48%

TABLE 2. Performance comparison among four different classifiers which are LightGBM, SVM, random forest and decision tree.

Method	Accuracy	Recall	Precision	F1-score
SVM	84.42%	85.03%	82.80%	83.90%
RF	80.71%	84.13%	77.41%	80.63%
DT	77.14%	82.93%	72.89%	77.59%
LGBM	85.28%	88.32%	82.17%	85.14%

where the integrated similarity value between the miRNAs $m(a)$ and $m(b)$ is defined as w_b. Each miRNA-disease sample can be described as a 1907-dimensional vector as follow:

$$F_{sim} = (SD\,(d(a))\,,\, SM\,(m(a))) \tag{17}$$

$F_{sim} = (f_1, f_2, f_3, \ldots, f_{1906}, f_{1907})$, where $(f_1, f_2, f_3, \ldots, f_{850})$ stands for the 850 gathered similarity values of the disease and $(f_{851}, f_{852}, f_{853}, \ldots, f_{1907})$ stands for the 1057 gathered similarity values of the miRNAs. The fusion of multi-source information produces noise that can affect the prediction. And the range and dimensions of the data from different sources can make the model easy to overfit. Therefore, we resized F_{sim} from 1907 to 32 by Autoencoder to obtain the new similarity feature $F_{sim}{}'$ and the sequence feature matrixes F_{seq} is resized from 3200 to 32 in the same way to obtain the new sequence feature $F_{seq}{}'$. This operation makes the sequence feature and the similarity information feature weight equivalent. The feature is reduced to 32 dimensions in order to suppress the noise contained in the feature while reducing the computational cost. We defined each miRNA-disease sample as a 64-dimensional vector as follow:

$$F = (F_{sim}{}', F_{seq}{}') \tag{18}$$

TABLE 3. The comparison results of cgmda and related methods.

Method	AUC
Shi's[16]	75.80%
RWRMDA[18]	86.17%
MCMDA[19]	87.67%
MTDN[20]	88.72%
LMTRDA[21]	90.54%
ABMDA[39]	90.23%
miRGOFS[40]	87.70%
CGMDA	90.99%

Finally, we use an algorithm called lightGBM, which is described in H. LIGHTGBM CLASSIFIER, to build a predictor by training the sample dataset. Specifically, we obtain samples of 64-dimensional vectors in the training set according to steps 2 and 3. The sample label in the positive sample set is assigned a value of 1, otherwise set to 0. Then, the training set is then used to train the lightGBM classifier to obtain predictors that can predict potential miRNA-disease associations. In addition, if a miRNA-disease pair gets a higher predicted score, they are more likely to be associated. The flowchart of CGMDA is shown in figure 2.

J. EVALUATION CRITERIA

For the 5-fold cross-validation, the original samples were randomly divided into five subsets, one of which was retained as the test set, and the remaining four were used as training data. The cross-validation process is repeated 5 times, each subset is taken as a test set, and the resulting results are averaged to produce a single estimate.

IV. RESULTS

A. PERFORMANCE EVALUATION

1) ASSESSMENT OF PREDICTION ABILITY

The area under the receiver operating characteristic curve (AUC) is a common machine learning evaluation criterion for evaluating the training effects of a two-level model. The abscissa of each point on the receiver operating characteristic curve (ROC) is the false positive rate (FPR) obtained under different judgment criteria, and the ordinate is the true positive rate (TPR) corresponding to the false positive rate under the same conditions [37]. The higher the value, the better the classifier effect. CGMDA gained a mean AUC of 0.9099+/−0.0052 which is the average of AUCs of 0.9026 (fold 1), 0.9116 (fold 2), 0.9132 (fold 3), 0.9169 (fold 4)

TABLE 4. Prediction of the top 40 predicted miRNAs associated with Breast Neoplasms based on known associations in dbDEMC v2.0 and miR2Database.

miRNA	dbDEMC	miR2D	miRNA	dbDEMC	miR2D
hsa-mir-29	confirmed	unconfirmed	hsa-mir-183	confirmed	unconfirmed
hsa-mir-483	confirmed	unconfirmed	hsa-mir-142	confirmed	unconfirmed
hsa-mir-16-2	confirmed	unconfirmed	hsa-mir-16	confirmed	unconfirmed
hsa-mir-193a	confirmed	unconfirmed	hsa-mir-29b-2	unconfirmed	confirmed
hsa-mir-193a	confirmed	unconfirmed	hsa-mir-29c	confirmed	confirmed
hsa-mir-29b-1	confirmed	unconfirmed	hsa-let-7a-3	confirmed	confirmed
hsa-mir-125b-2	confirmed	unconfirmed	hsa-mir-92-1	confirmed	unconfirmed
hsa-mir-16-1	confirmed	unconfirmed	hsa-mir-196a-2	confirmed	confirmed
hsa-let-7	confirmed	confirmed	hsa-mir-27b	confirmed	unconfirmed
hsa-mir-23a	confirmed	unconfirmed	hsa-let-7e	confirmed	unconfirmed
hsa-mir-203	confirmed	confirmed	hsa-mir-15b	confirmed	unconfirmed
hsa-mir-222	confirmed	confirmed	hsa-mir-31	confirmed	confirmed
hsa-mir-15a	confirmed	unconfirmed	hsa-mir-146a	confirmed	confirmed
hsa-mir-18	confirmed	unconfirmed	hsa-mir-124	confirmed	unconfirmed
hsa-mir-143	confirmed	confirmed	hsa-mir-101	confirmed	unconfirmed
hsa-mir-92a-1	confirmed	unconfirmed	hsa-mir-429	confirmed	confirmed
hsa-mir-138	confirmed	unconfirmed	hsa-mir-19a	confirmed	unconfirmed
hsa-mir-125b-1	confirmed	unconfirmed	hsa-mir-21	confirmed	confirmed
hsa-mir-19b-1	confirmed	unconfirmed	hsa-mir-192	confirmed	unconfirmed
hsa-mir-106a	confirmed	unconfirmed	hsa-mir-27a	confirmed	confirmed

TABLE 5. Prediction of the top 40 predicted miRNAs associated with Lymphoma based on known associations in dbDEMC v2.0 and miR2Database.

miRNA	dbDEMC	miR2D	miRNA	dbDEMC	miR2D
hsa-mir-24	confirmed	unconfirmed	hsa-mir-196a	confirmed	confirmed
hsa-mir-195	confirmed	unconfirmed	hsa-mir-34b	confirmed	confirmed
hsa-mir-1	confirmed	unconfirmed	hsa-mir-98	confirmed	unconfirmed
hsa-mir-181a	confirmed	unconfirmed	hsa-mir-148a	confirmed	unconfirmed
hsa-mir-182	confirmed	unconfirmed	hsa-mir-192	confirmed	unconfirmed
hsa-let-7b	confirmed	unconfirmed	hsa-mir-26a	confirmed	confirmed
hsa-let-7e	confirmed	confirmed	hsa-mir-27a	confirmed	unconfirmed
hsa-let-7c	confirmed	confirmed	hsa-mir-30a	confirmed	unconfirmed
hsa-mir-106b	confirmed	unconfirmed	hsa-mir-130b	confirmed	unconfirmed
hsa-mir-125a	confirmed	unconfirmed	hsa-mir-130a	confirmed	unconfirmed
hsa-mir-199b	confirmed	unconfirmed	hsa-mir-200c	confirmed	unconfirmed
hsa-mir-30e	confirmed	confirmed	hsa-mir-133b	confirmed	unconfirmed
hsa-mir-107	confirmed	unconfirmed	hsa-mir-429	unconfirmed	unconfirmed
hsa-mir-132	confirmed	unconfirmed	hsa-mir-30b	confirmed	confirmed
hsa-mir-205	confirmed	unconfirmed	hsa-mir-133a	confirmed	unconfirmed
hsa-mir-212	confirmed	unconfirmed	hsa-mir-342	confirmed	unconfirmed
hsa-mir-22	confirmed	unconfirmed	hsa-mir-218	confirmed	unconfirmed
hsa-mir-23b	confirmed	unconfirmed	hsa-mir-150	confirmed	confirmed
hsa-mir-424	confirmed	unconfirmed	hsa-mir-10b	confirmed	unconfirmed
hsa-mir-210	confirmed	confirmed	hsa-mir-214	confirmed	unconfirmed

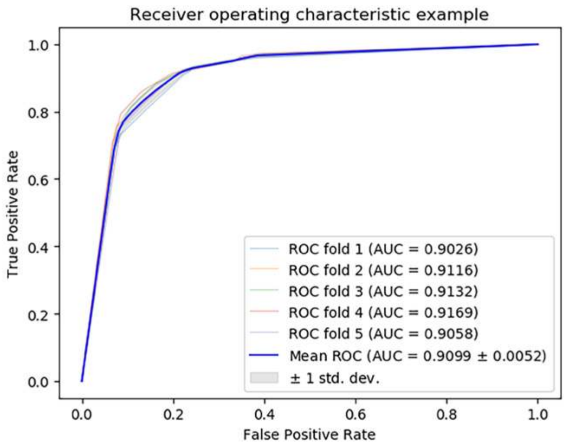

FIGURE 3. The ROCs of CGMDA and AUCs based on 5-fold cross validation.

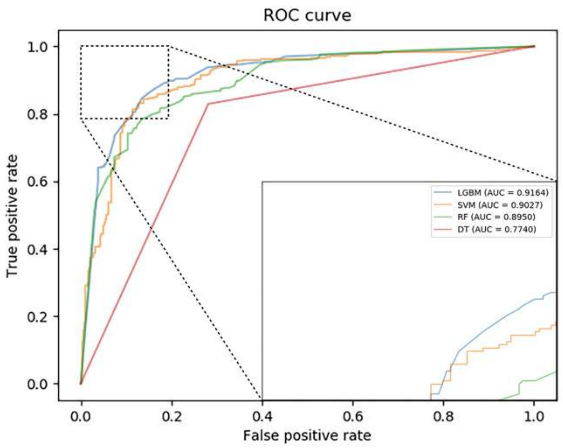

FIGURE 4. The ROCs of four different classifiers which are LightGBM, SVM, random forest and decision tree.

and 0.9058 (fold 5) in 5-fold cross validation as showed in Figure 3 and the yielded averages of accuracy (Acc.), recall (Rec.), precision (Pre.) and f1-score (F1) come to be 87.50%, 85.44%, 89.13% and 87.24% as in TABLE 1.

2) COMPARISON AMONG DIFFERENT CLASSIFIERS

In this part of the experiment, the support vector machine (SVM), random forest (RF) and decision tree (DT) were chosen to compare with the lightGBM used in the proposed method. The accuracy of the four experiments are 85.28% (lightGBM), 84.42% (SVM), 80.71% (RF) and 77.14% (DT). Their AUC are 91.64% (lightGBM), 90.27% (SVM), 89.50% (Random forest) and 77.40% (Decision Tree) shown as Figure 4. The accuracy, sensitivity, precision and f1-score as in TABLE 2. Among the above results, SVM and lightGBM

TABLE 6. Prediction of the top 40 predicted miRNAs associated with Colon Neoplasms based on known associations in dbDEMC v2.0 and miR2Database.

miRNA	dbDEMC	miR2D	miRNA	dbDEMC	miR2D
hsa-mir-125a	confirmed	confirmed	hsa-mir-181a	confirmed	unconfirmed
hsa-mir-186	confirmed	confirmed	hsa-mir-26a	confirmed	confirmed
hsa-mir-10b	confirmed	confirmed	hsa-mir-342	confirmed	unconfirmed
hsa-mir-499	confirmed	unconfirmed	hsa-mir-497	confirmed	unconfirmed
hsa-mir-205	confirmed	unconfirmed	hsa-mir-21	confirmed	unconfirmed
hsa-mir-34b	confirmed	unconfirmed	hsa-mir-29a	confirmed	unconfirmed
hsa-mir-132	confirmed	confirmed	hsa-mir-24	confirmed	unconfirmed
hsa-mir-1	confirmed	unconfirmed	hsa-mir-133a	confirmed	unconfirmed
hsa-mir-30b	confirmed	unconfirmed	hsa-mir-331	unconfirmed	unconfirmed
hsa-mir-218	confirmed	unconfirmed	hsa-let-7b	confirmed	unconfirmed
hsa-mir-200c	confirmed	unconfirmed	hsa-mir-196a	confirmed	unconfirmed
hsa-mir-106b	confirmed	unconfirmed	hsa-mir-133b	confirmed	unconfirmed
hsa-mir-98	confirmed	unconfirmed	hsa-mir-16	confirmed	unconfirmed
hsa-mir-211	confirmed	unconfirmed	hsa-mir-130b	confirmed	unconfirmed
hsa-mir-27a	confirmed	unconfirmed	hsa-mir-212	confirmed	unconfirmed
hsa-mir-126	confirmed	unconfirmed	hsa-mir-330	unconfirmed	unconfirmed
hsa-mir-182	confirmed	unconfirmed	hsa-mir-429	confirmed	unconfirmed
hsa-mir-340	confirmed	unconfirmed	hsa-mir-28	confirmed	unconfirmed
hsa-mir-424	confirmed	unconfirmed	hsa-mir-150	confirmed	unconfirmed
hsa-mir-129	confirmed	unconfirmed	hsa-mir-143	confirmed	unconfirmed

are significantly better than the other two classifiers. In particular, lightGBM has the highest AUC value, and other benchmark parameters are similar to SVM. However, in terms of runtime, lightGBM is significantly better than SVM because SVM is very time consuming when processing big data samples. In addition, AUC is more discriminating and statistically consistent than accuracy [38]. Therefore, when the accuracy performance is similar, in order to more effectively predict the potential miRNA-disease association, we choose the lightGBM with the highest AUC and less running time as the classifier of the proposed method.

3) COMPARISON WITH RELATED METHODS

In recent years, many computational methods have been proposed to identify miRNA-disease associations, and we compare the performance of CGMDA with 7 state-of-the-art methods, as shown in TABLE 3. Most current prediction methods rely only on incompletely related biological information, and we introduce miRNA sequence information to represent attribute information, as we focus on developing characterization of miRNA sequence information. Since these methods do not disclose all the evaluation indicators, the only indicator that can provide comparison are AUC, so in this experiment we only compare the AUC of these methods. From the TABLE 3 we can see that the proposed method gets the highest AUC. The reason that CGMDA is superior to other methods that rely solely on incompletely related biological information is the introduction of miRNA sequences and the quantification of linear sequence information.

B. CASE STUDIES

Based on the hypothesis that only the miRNA-disease association in HMDD v3.0 is known and the remaining associations are unknown, we built the case studies about *Breast Neoplasms, Lymphoma and Colon Neoplasms* to evaluate our approach. In detail, the confirmed miRNA-disease associations in the HMDD v3.0 dataset is utilized as the training set to train the classifier. In addition, we used associations between the three diseases and all possible miRNAs as the test set. When CGMDA obtains the predicted results, we sort the predicted results and select the top 40 candidates with the highest scores based on different diseases, and confirm them in other miRNA-disease associations datasets, dbDEMC v2.0 and miR2Database which manually collected miRNA-disease association entries in the papers [41], [42].

It is well known that malignant breast tumors occur mostly in women with *breast cancer* [43]. In the United States, about 12.5% women has breast cancer and the global breast cancer rate is greater than it in 1970s. A large number of experimental data reveal that many miRNAs have influence on *breast* *neoplasms*. So, in the first case study, we chose *breast neoplasms* and used CGMDA to forecast the potential miRNAs associate about *breast neoplasms*. The results are shown in TABLE 4, 39 out of the top 40 potentially miRNAs which associate with *breast neoplasms* were confirmed by dbDEMC v2.0 and miR2Disease. *Lymphoma* is a malignant tumor that originates from the lymphoid hematopoietic system. It was chosen as a case and the predicted scores of its potentially associated miRNAs were ranked. As a result, the experimental results recorded in miR2Disease confirmed by dbDEMC v2.0, 39 of the top 40 potential miRNAs associated with lymphoma, as shown in TABLE 5. *Colon cancer* is a malignant tumor that can occur at any age, especially in the elderly. It will initially form in the form of polyps inside the colon and may gradually become colon cancer. So, we selected it as an example of the third case study. As a result, 38 out of the top 40 potentially miRNAs which associate with neoplasms were confirmed by experimental findings recorded in dbDEMC v2.0, miR2Disease, as shown in TABLE 6. From the results, the three case analyses obtained 97.5%, 97.5%, and 95% accuracy, respectively, in predicting potential miRNAs associated with disease. This shows that our approach has good ability to predict unknown associations.

V. CONCLUSION

In this paper, we propose a predictive method based on chaos game representation to simultaneously consider linear sequence information and nonlinear relationships. Compared to our method, most sequence comparison algorithms, such as k-mer, can only quantify nonlinear sequence relationships, and gene expression is related to linear sequence information. In addition, CGR is converted to image information to align features due to different sequence lengths. In terms of experimental results, the 5-fold cross-validation showed that the method can reliably predict the potential associations between miRNAs and diseases. CGMDA achieved good results when compared with different classifiers and related prediction methods. What's more, we apply CGMDA to three complex human diseases, Breast Neoplasm, Lung Neoplasm and Esophageal Neoplasm. Experiments have shown that CGMDA is an excellent miRNA-disease association method. In future research, we will continue to explore the application of sequence information in predictions of potential miRNA-disease associations to obtain better predictions.

APPENDIX

Appendixes, if needed, appear before the acknowledgment.

ACKNOWLEDGMENT

The authors would like to thank all anonymous reviewers for their constructive advices.

COMPETING INTERESTS

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

REFERENCES

[1] V. Ambros, "The functions of animal microRNAs," *Nature*, vol. 431, no. 7006, pp. 350–355, Sep. 2004.

[2] D. P. Bartel, "MicroRNAs: Genomics, biogenesis, mechanism, and function," *Cell*, vol. 116, no. 2, pp. 281–297, Jan. 2004.

[3] B. J. Reinhart, F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz, and G. Ruvkun, "The 21-nucleotide let-7 RNA regulates developmental timing in caenorhabditis elegans," *Nature*, vol. 403, no. 6772, pp. 901–906, Feb. 2000.

[4] R. C. Lee, R. L. Feinbaum, and V. Ambros, "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14," *Cell*, vol. 75, no. 5, pp. 843–854, Dec. 1993.

[5] M. Esteller, "Non-coding RNAs in human disease," *Nature Rev. Genet.*, vol. 12, no. 12, pp. 861–874, Nov. 2011.

[6] Y. Liang, L. Duan, J. Xiong, W. Zhu, Q. Liu, D. Wang, W. Liu, Z. Li, and D. Wang, "E2 regulates MMP-13 via targeting miR-140 in IL-1ß-induced extracellular matrix degradation in human chondrocytes," *Arthritis Res. Therapy*, vol. 18, no. 1, p. 105, Dec. 2016.

[7] S. Andreasen, Q. Tan, T. K. Agander, P. Steiner, K. Bjørndal, E. Høgdall, S. R. Larsen, D. Erentaite, C. H. Olsen, B. P. Ulhøi, S. L. von Holstein, I. Wessel, S. Heegaard, and P. Homøe, "Adenoid cystic carcinomas of the salivary gland, lacrimal gland, and breast are morphologically and genetically similar but have distinct microRNA expression profiles," *Mod. Pathol.*, vol. 31, no. 8, pp. 1211–1225, Feb. 2018.

[8] C. Taurino, W. H. Miller, M. W. McBride, J. D. McClure, R. Khanin, M. U. Moreno, J. A. Dymott, C. Delles, and A. F. Dominiczak, "Gene expression profiling in whole blood of patients with coronary artery disease," *Clin. Sci.*, vol. 119, no. 8, pp. 335–343, Oct. 2010.

[9] H. Zhao, Y. Zhang, F. Xue, J. Xu, and Z. Fang, "Has-mir-146a rs2910164 polymorphism and risk of immune thrombocytopenia," *Autoimmunity*, vol. 47, no. 3, pp. 173–176, 2014.

[10] I. Sarrion, L. Milian, G. Juan, M. Ramon, I. Furest, C. Carda, J. C. Gimeno, and M. M. Roig, "Role of circulating miRNAs as biomarkers in idiopathic pulmonary arterial hypertension: Possible relevance of miR-23a," *Oxidative Med. Cellular Longevity*, vol. 2015, Feb. 2015, Art. no. 792846.

[11] C. Bang, J. Fiedler, and T. Thum, "Cardiovascular importance of the MicroRNA-23/27/24 family," *Microcirculation*, vol. 19, no. 3, pp. 208–214, Apr. 2012.

[12] M. Lu, Q. Zhang, M. Deng, J. Miao, Y. Guo, W. Gao, and Q. Cui, "An analysis of human MicroRNA and disease associations," *PloS ONE*, vol. 3, no. 10, Oct. 2008, Art. no. e3420.

[13] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Mol. Syst. Biol.*, vol. 3, no. 1, p. 88, Jan. 2007.

[14] D. Wang, M. Lu, J. Miao, T. Li, E. Wang, and Q. Cui, "Cepred: Predicting the co-expression patterns of the human intronic microRNAs with their host genes," *PLoS ONE*, vol. 4, no. 2, Feb. 2009, Art. no. e4421.

[15] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen, "A new method to measure the semantic similarity of GO terms," *Bioinformatics*, vol. 23, no. 10, pp. 1274–1281, May 2007.

[16] H. Shi, J. Xu, G. Zhang, L. Xu, C. Li, L. Wang, Z. Zhao, W. Jiang, Z. Guo, and X. Li, "Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes," *BMC Syst. Biol.*, vol. 7, no. 1, p. 101, Dec. 2013.

[17] C. Xu, Y. Ping, X. Li, H. Zhao, L. Wang, H. Fan, Y. Xiao, and X. Li, "Prioritizing candidate disease miRNAs by integrating phenotype associations of multiple diseases with matched miRNA and mRNA expression profiles," *Mol. Biosyst.*, vol. 10, no. 11, pp. 2800–2809, 2014.

[18] X. Chen, M.-X. Liu, and G.-Y. Yan, "RWRMDA: Predicting novel human microRNA–disease associations," *Mol. BioSyst.*, vol. 8, no. 10, pp. 2792–2798, 2012.

[19] J.-Q. Li, Z.-H. Rong, X. Chen, G.-Y. Yan, and Z.-H. You, "MCMDA: Matrix completion for MiRNA-disease association prediction," *Oncotarget*, vol. 8, no. 13, pp. 21187–21199, Mar. 2017.

[20] J. Xu, C.-X. Li, J.-Y. Lv, Y.-S. Li, Y. Xiao, T.-T. Shao, X. Huo, X. Li, Y. Zou, Q.-L. Han, X. Li, L.-H. Wang, and H. Ren, "Prioritizing candidate disease miRNAs by topological features in the miRNA target–dysregulated network: Case study of prostate cancer," *Mol. Cancer Therapeutics*, vol. 10, no. 10, pp. 1857–1866, Oct. 2011.

[21] L. Wang, Z.-H. You, X. Chen, Y.-M. Li, Y.-N. Dong, L.-P. Li, and K. Zheng, "LMTRDA: Using logistic model tree to predict MiRNA-disease associations by fusing multi-source information of sequences and similarities," *PLoS Comput. Biol.*, vol. 15, no. 3, 2019, Art. no. e1006865.

[22] Y. Li, C. Qiu, J. Tu, B. Geng, J. Yang, T. Jiang, and Q. Cui, "HMDD v2.0: A database for experimentally supported human microRNA and disease associations," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D1070–D1074, Jan. 2014.

[23] S. Griffiths-Jones, H. K. Saini, S. van Dongen, and A. J. Enright, "miRBase: Tools for microRNA genomics," *Nucleic Acids Res.*, vol. 36, no. 1, pp. D154–D158, Jan. 2008.

[24] L. Chen, B. Liu, and C. Yan, "DPFMDA: Distributed and privatized framework for miRNA-Disease association prediction," *Pattern Recognit. Lett.*, vol. 109, pp. 4–11, Jul. 2018.

[25] D. Wang, J. Wang, M. Lu, F. Song, and Q. Cui, "Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases," *Bioinformatics*, vol. 26, no. 13, pp. 1644–1650, 2010.

[26] C. Pesquita, D. Faria, A. O. Falcao, P. Lord, and F. M. Couto, "Semantic similarity in biomedical ontologies," *PLoS Comput. Biol.*, vol. 5, no. 7, Jul. 2009, Art. no. e1000443.

[27] P. Xuan, K. Han, M. Guo, Y. Guo, J. Li, J. Ding, Y. Liu, Q. Dai, J. Li, and Z. Teng, "Prediction of microRNAs associated with human diseases based on weighted *k* most similar neighbors," *PloS ONE*, vol. 8, no. 8, Aug. 2013, Art. no. e70204.

[28] T. van Laarhoven, S. B. Nabuurs, and E. Marchiori, "Gaussian interaction profile kernels for predicting drug–target interaction," *Bioinformatics*, vol. 27, no. 21, pp. 3036–3043, Nov. 2011.

[29] X. Chen, C. C. Yan, X. Zhang, Z.-H. You, L. Deng, Y. Liu, Y. Zhang, and Q. Dai, "WBSMDA: Within and between score for MiRNA-disease association prediction," *Sci. Rep.*, vol. 6, p. 21106, Feb. 2016.

[30] M. G. Dunlop, S. M. Farrington, A. D. Carothers, A. H. Wyllie, L. Sharp, J. Burn, B. Liu, K. W. Kinzler, and B. Vogelstein, "Cancer risk associated with germline DNA mismatch repair gene mutations," *Human Mol. Genet.*, vol. 6, no. 1, pp. 105–110, Jan. 1997.

[31] S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and M. A. Phillippy, "Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation," *Genome Res.*, vol. 27, no. 5, pp. 722–736, 2017.

[32] Y. Bar-Yam, *Dynamics of Complex Systems*. Reading, MA, USA: Addison-Wesley, 1997.

[33] H. J. Jeffrey, "Chaos game representation of gene structure," *Nucleic Acids Res.*, vol. 18, no. 8, pp. 2163–2170, Apr. 1990.

[34] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3146–3154.

[35] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[36] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[37] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[38] C. X. Ling, J. Huang, and H. Zhang, "AUC: A statistically consistent and more discriminating measure than accuracy," in *Proc. 18th Int. Joint Conf. Artif. Intell.*, Aug. 2003, pp. 519–524.

[39] Y. Zhao, X. Chen, and J. Yin, "Adaptive boosting-based computational model for predicting potential miRNA-disease associations," *Bioinformatics*, vol. 1, p. 9, Apr. 2019.

[40] Y. Yang, X. Fu, W. Qu, Y. Xiao, and H.-B. Shen, "MiRGOFS: A GO-based functional similarity measurement for miRNAs, with applications to the prediction of miRNA subcellular localization and miRNA–disease association," *Bioinformatics*, vol. 34, no. 20, pp. 3547–3556, Oct. 2018.

[41] Z. Yang, F. Ren, C. Liu, S. He, G. Sun, Q. Gao, L. Yao, Y. Zhang, R. Miao, Y. Cao, Y. Zhao, Y. Zhong, and H. Zhao, "dbDEMC: A database of differentially expressed miRNAs in human cancers," *BMC Genomics*, vol. 11, no. 4, p. S5, Dec. 2010.

[42] Q. Jiang, Y. Wang, Y. Hao, L. Juan, M. Teng, X. Zhang, M. Li, G. Wang, and Y. Liu, "miR2Disease: A manually curated database for microRNA deregulation in human disease," *Nucleic Acids Res.*, vol. 37, no. 1, pp. D98–D104, Jan. 2009.

[43] Z. Tao, A. Shi, C. Lu, T. Song, Z. Zhang, and J. Zhao, "Breast Cancer: Epidemiology and Etiology," *Cell Biochem. Biophys.*, vol. 72, no. 2, pp. 333–338, Jun. 2015.

KAI ZHENG received the B.E. degree in computer science and technology from Central South University, Changsha, China, in 2017. He is currently pursuing the master's degree with the China University of Mining and Technology. His research interests include data mining, pattern recognition, recommender systems, machine learning, deep learning, intelligent information processing, and its applications in bioinformatics.

LEI WANG received the Ph.D. degree from the School of Computer Science and Technology, China University of Mining and Technology, Jiangsu, China, in 2018. He is currently a Postdoctoral with the Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Science, Ürümqi, China. His research interests include data mining, pattern recognition, machine learning, deep learning, computational biology, and bioinformatics. He acted as Reviewer for many international journals, such as *Scientific Reports*, *Current Protein & Peptide Science*, *Computational Biology and Chemistry*, *Soft Computing*, and the *Journal of Computational Biology*.

ZHU-HONG YOU (M'14) received the B.E. degree in electronic information science and engineering from Hunan Normal University, Changsha, China, in 2005, and the Ph.D. degree in control science and engineering from the University of Science and Technology of China (USTC), Hefei, China, in 2010. From June 2008 to November 2009, he was a Visiting Research Fellow with the Center of Biotechnology and Information, Cornell University. He is currently a Professor with the Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Science, Ürümqi, China. His current research interests include neural networks, intelligent information processing, sparse representation, and its applications in bioinformatics.