molecules

MDPI

Article

Prediction of miRNA–Disease Associations by Cascade Forest Model Based on Stacked Autoencoder

Xiang Hu, Zhixiang Yin *, Zhiliang Zeng and Yu Peng

Center of Intelligent Computing and Applied Statistics, School of Mathematics, Physics and Statistics, Shanghai University of Engineering Science, Shanghai 201620, China; m440121302@sues.edu.cn (X.H.); zlzeng@sues.edu.cn (Z.Z.); py@sues.edu.cn (Y.P.)

* Correspondence: 21190006@sues.edu.cn

Abstract: Numerous pieces of evidence have indicated that microRNA (miRNA) plays a crucial role in a series of significant biological processes and is closely related to complex disease. However, the traditional biological experimental methods used to verify disease-related miRNAs are inefficient and expensive. Thus, it is necessary to design some excellent approaches to improve efficiency. In this work, a novel method (CFSAEMDA) is proposed for the prediction of unknown miRNA–disease associations (MDAs). Specifically, we first capture the interactive features of miRNA and disease by integrating multi-source information. Then, the stacked autoencoder is applied for obtaining the underlying feature representation. Finally, the modified cascade forest model is employed to complete the final prediction. The experimental results present that the AUC value obtained by our method is 97.67%. The performance of CFSAEMDA is superior to several of the latest methods. In addition, case studies conducted on lung neoplasms, breast neoplasms and hepatocellular carcinoma further show that the CFSAEMDA method may be regarded as a utility approach to infer unknown disease–miRNA relationships.

Keywords: miRNA–disease association; multi-source information; stacked autoencoder; cascade forest

check for updates

Citation: Hu, X.; Yin, Z.; Zeng, Z.; Peng, Y. Prediction of miRNA–Disease Associations by Cascade Forest Model Based on Stacked Autoencoder. *Molecules* **2023**, *28*, 5013. https://doi.org/10.3390/molecules28135013

Academic Editor: Vladimir N. Uversky

Received: 30 May 2023
Revised: 23 June 2023
Accepted: 24 June 2023
Published: 27 June 2023

1. Introduction

MicroRNAs (miRNAs) are a class of short non-coding RNA molecules and play critical roles in gene expression programs and biological processes [1–3]. Many scholars have explored and studied the role of miRNAs in the human body [4]. Moreover, some new works in the literature have shown that the differential expression of miRNAs is associated with human disease pathogenesis, such as Alzheimer's disease [5], cardiovascular disease [6], breast cancer [7,8] and many others [9–11]. Therefore, the study of the relationship between miRNA and disease is meaningful for the treatment of diseases and has become a hot topic in the field of bioinformatics.

Some traditional biological experimental methods that are used to verify unknown miRNAs associated with a certain disease, such as reverse transcription polymerase chain reaction [12] and microarray profiling [13], often require a large investment of money and time and are inefficient. Now, the computing efficiency of computers has greatly been improved. We can design some excellent predictive models to identify unknown MDAs, which are efficient and economical. These associations can be used for further experimental verification. The current potential miRNA–disease relationship identification methods can be divided into two types. The first type of method is based on a network algorithm that exploits the similarity of miRNAs and diseases from different perspectives. The pioneering model was designed by Jiang et al. [14]. The model applies a scoring system to rank the probability of disease-related miRNAs. By integrating Gaussian similarity, Chen et al. [15] designed the WBSMDA method to predict MDAs. In addition, due to WBSMDA being a global rank method, this method allows for calculating the connections of miRNA–disease for all diseases simultaneously. Zeng et al. [16] used a bilayer network and applied the structural consistency as a metric of network performance. Then the authors used structural perturbation to infer MDAs in the bilayer network. In [17], via non-negative matrix factorization, the authors proposed a ranking method. In particular, this method introduced the sparseness characteristic to make prediction more reliable. Yu et al. [18] developed KDFGMDA to infer unknown disease-related miRNAs. The KDFGMDA method first obtained triple information of existing databases and massive experimental data, then a graph representation model was trained to complete prediction tasks. Zhang et al. [19] proposed the FLNSNLI method, which is based on fast linear neighborhood similarity. The miRNAs and diseases are expressed in the form association profiles, then the fast linear neighborhood similarity and association profiles are used to measure miRNA–miRNA similarity and disease–disease similarity. Finally, the label propagation algorithm is applied to calculate the probability score. Some other interesting methods based on a complex network can be seen in [20,21].

The second category of methods is machine learning-based methods, which have been indicated to be powerful in many classification tasks, especially in bioinformatics [22,23]. For instance, Zheng et al. [24] developed the MLMDA method to predict microRNA–disease associations. This method used the random forest classifier to predict via integrating heterogeneous information sources. Zhou et al. [25] used K-means clustering to balance the sample set, and used gradient boosting decision trees for feature extraction. Finally, the logical regression is applied to predict labels. Some deep learning methods are also designed for predicting MDAs. Chen et al. [26] inferred novel MDAs by the deep-belief network model. During the pre-training process, this method utilized all of the information. Liu et al. [27] proposed a new feature representation method, and used the random forest method to complete the prediction. Ai et al. [28] designed an end-to-end computational method by integrating the multi-source information and reformulating the score matrix. In [29], based on the deep neural network, the authors developed a new model named NCFM, which integrated the generalization of the neural network and the memorization of matrix factorization. The MDA-CNN method was proposed in [30]. This method first obtains interaction features. Then, an autoencoder is utilized to obtain the critical feature combination. Finally, after inputting the low-dimensional feature representation, MDA-CNN inferred the final label by using a convolutional neural network.Wang et al. [31] fully utilized unlabeled samples and pretrained the stacked autoencoders in an unsupervised manner. This method is suitable for the dataset with a small number of positive data and a large number of negative data.

More and more deep neural networks are used for classification tasks and have achieved good results. Inspired by deep neural networks processing information layer by layer, Zhou et al. [32] developed the deep forest learning model, including the multi-grained scanning module and cascade forest structure. Deep forest methods try to build deep models through non-differentiable modules and have fewer hyperparameters than deep neural networks. Furthermore, it is robust to parameter settings, and thus, even if we do not adjust the parameter settings, it still has excellent performance. Due to this, many various prediction tasks are well solved by using the deep forest method [33]. For instance, Chu et al. [34] developed a cascade forest (DTI-CDF) approach to infer novel interactions between the drug and target. Compared with other methods, such as DDR [35], the DTI-CDF model achieved better predictive performance. Zeng et al. [36] presented the AOPEDF method, an arbitrary-order proximity embedded cascade forest method, for predicting drug–target interactions. The AOPEDF method integrated 15 different types of network information, and a deep cascade forest model with 6 estimators was used to complete the prediction task. In the field of MDAs, Dai et al. [37] proposed the MDA-CF method for predicting unknown MDAs. This method uses multiple types of information to represent the association between miRNA and disease, and applies a cascade forest model with four estimators to infer new MDAs. The prediction efficiency of the MDA-CF method is excellent.

In this work, we present a new method called CFSAEMDA to infer MDAs. In the proposed method, various types of information, including miRNA similarity (functional and Gaussian similarity) and disease similarity (semantic and Gaussian similarity), are integrated to construct a comprehensive feature descriptor of MDAs. Then, the latent feature information is obtained by using the stacked autoencoder. Finally, the modified cascade forest model is trained by inputting reduced features to accurately classify and infer novel MDAs. The average area under the receiver operating characteristic curve (AUC) is 97.67%, and is superior to several of the latest methods and classical machine learning algorithms. In particular, the case studies of lung neoplasms, breast neoplasms and hepatocellular carcinoma show that 49, 50 and 47 of the top 50 predicted relevant miRNAs are confirmed by the latest works in the literature, respectively.

2. Results

2.1. Evaluation Criteria

In this paper, the sample set is divided into the training set and test set in an 8:2 ratio. The training set is used for training the model, and the test set is used for comparing CFSAEMDA with other methods. To obtain a systematic experimental result, we conduct 4-fold cross validation on the training set to evaluate the performance of CFSAEMDA. The training set is randomly divided into four approximately identical parts. Three parts of them are utilized to train the CFSAEMDA model (training subset), and the residual one is used as the test subset. The final result is obtained by averaging the four test subsets.

Our task is binary classification work. To evaluate the performance of CFSAEMDA fairly, we choose several common metrics, including accuracy (Acc), precision (Pre), sensitivity (Sen), specificity (Spe), the Matthews correlation coefficient (MCC), and the F_1-score (F1). The equations of these metrics are given by

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

$$Pre = \frac{TP}{TP + FP}, \tag{2}$$

$$Sen = \frac{TP}{TP + FN}, \tag{3}$$

$$Spe = \frac{TN}{TN + FP}, \tag{4}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{5}$$

$$F1 = \frac{2TP}{2TP + FP + FN}, \tag{6}$$

where TP and TN are the number of miRNA–disease with association and miRNA–disease without association cases that were successfully identified, respectively. FP(FN) represents the number of miRNA–disease with (without) association examples that are incorrectly identified.

Furthermore, we also calculate the area under the receiver operating characteristic curve (AUC) and the precision–recall curve (AUPR), respectively. The range of values for AUC and AUPR is $[0, 1]$. Normally, the higher AUPR and AUC values represent the better performance of the model.

2.2. Feature Evaluation

In this section, we present the results of 4-fold cross validation. From Table 1, CFSAEMDA obtains an average AUC of 96.80%, which is the mean of 96.60%, 97.04%, 97.07% and 96.51%, and achieves an AUPR of 97.14%, which is the average of 96.78%, 97.50%, 97.53% and 96.74%. Figure 1 shows the receiver operating characteristic curves and precision–recall curves of the CFSAEMDA method. It can be observed that the ROC curve can reach the upper left corner of the graph, while the P-R curve almost reaches the upper right corner of the graph, indicating the effectiveness of this method. In addition, the results of the average Acc, Pre, Sen, Spe, MCC and F1 are 91.00%, 91.02%, 91.09%, 90.88%, 81.99% and 91.05%, respectively. These results indicate that the CFSAEMDA method has an excellent ability to infer unknown miRNA–disease associations.

Table 1. Performance of CFSAEMDA.

Fold	Acc (%)	Pre (%)	Sen (%)	Spe (%)	MCC (%)	F1 (%)	AUPR (%)	AUC (%)
1	89.73	89.84	89.08	90.36	79.45	89.46	96.78	96.60
2	91.48	91.86	91.45	91.52	82.96	91.66	97.50	97.04
3	91.94	91.40	93.12	90.69	83.88	92.25	97.53	97.07
4	90.84	90.98	90.73	90.95	81.68	90.85	96.74	96.51
Mean	91.00	91.02	91.09	90.88	81.99	91.05	97.14	96.80

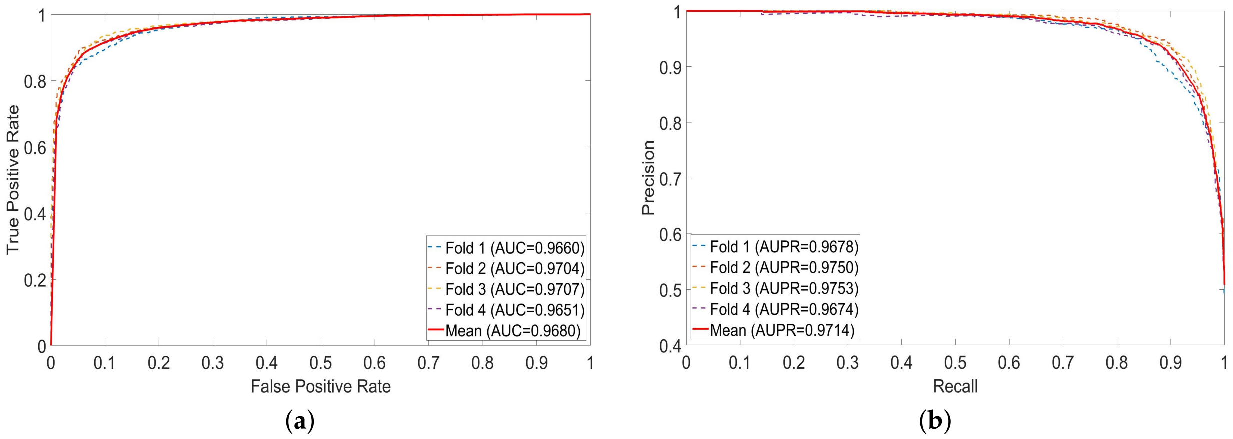


Figure 1. ROC and P-R curves of CFSAEMDA. (**a**) ROC curves; (**b**) P-R curves.

2.3. Ablation Experiments

In this section, we conduct ablation experiments to verify the effectiveness of our modifications of the estimators and predictor. We design two models: One of them is the traditional cascade forest structure without any modifications (Model 1). The other model only has modification to the estimators (Model 2). The training set is applied to train the model, and the test set is used for model comparison. The experimental results are listed in Table 2.

Regarding the effectiveness of the modification of the estimators, as shown in Table 2, Model 2 can further improve the performance of Acc, Pre, Sen, Spe, MCC, F1, AUPR and AUC. The estimators in Model 2 is more diverse than Model 1, and thus more information can be obtained at each level.

Regarding the effectiveness of the modification of the predictors, as shown in Table 2, CFSAEMDA can further improve the performance of Acc, Pre, Sen, Spe, MCC, F1, AUPR and AUC. Model 1 and Model 2 achieve the final predicted value by averaging the prediction results of the estimators. In the CFSAEMDA method, we modify the predictor to the SVM algorithm, which greatly improves the performance of the model.

Table 2. Performance of ablation experiments.

Method	Acc (%)	Pre (%)	Sen (%)	Spe (%)	MCC (%)	F1 (%)	AUPR (%)	AUC (%)
Model 1	87.52	87.91	85.99	88.96	75.02	86.94	93.81	94.57
Model 2	88.63	89.78	86.27	90.83	77.26	87.99	94.91	95.35
CFSAEMDA	92.27	92.56	91.33	93.14	84.51	91.94	97.87	97.67

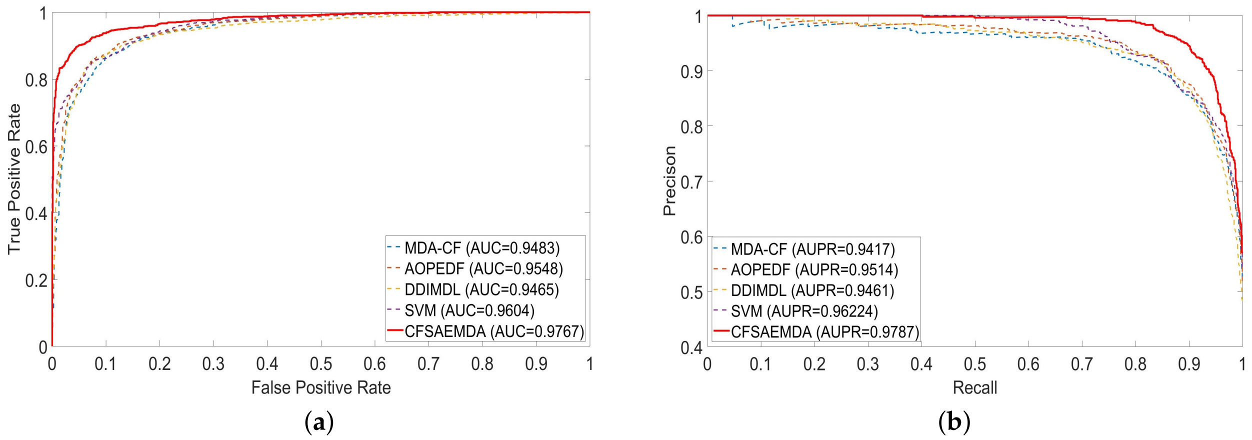
2.4. Method Comparison

In this section, we compare CFSAEMDA with several of the latest prediction methods, including MDA-CF [37], DDIMDL [38] and AOPEDF [36]. One machine learning method, the SVM algorithm, is also considered. These methods have shown excellent performance in classification tasks, especially in the field of bioinformatics. The MDA-CF method achieved prediction by the cascaded forest model. The estimators in the MDA-CF method are set to two XGBoosts and two random forests. The AOPEDF method also used the cascaded forest model to predict potential MDAs, which includes two random forests, two XGBoosts and two extra trees. The DDIMDL method used a deep neural network to obtain the final results. We modified the code of MDA-CF, DDIMDL, AOPEDF and SVM to adapt to the task proposed in this study and compared them with the proposed CFSAEMDA method. We trained the model under identical conditions and used the test set to evaluate its performance.

The predictive performance of each model is shown in Table 3. The CFSAEMDA method obtained an AUC of 97.67%, which is superior to MDA-CF (94.83%), AOPEDF (95.48%), DDIMDL (94.65%) and SVM (96.04%). The value of AUPR of the proposed method is 97.87%, which is superior to MDA-CF (94.17%), AOPEDF (95.14%), DDIMDL (94.61%) and SVM (96.22%). In other metrics, our proposed method also has the best performance. In addition, we also drew the ROC curves and P-R curves of each compared method (Figure 2). It is easy to find that the ROC curve and P-R curve of the proposed method are closer to the upper left and upper right than other methods, which indicates that our method is more efficient.

Table 3. The results of different methods.

Method	Acc (%)	Pre (%)	Sen (%)	Spe (%)	MCC (%)	F1 (%)	AUPR (%)	AUC (%)
MDA-CF	87.75	86.15	88.94	86.64	75.54	87.52	94.17	94.83
AOPEDF	88.86	89.75	86.84	90.74	77.70	88.28	95.14	95.48
DDIMDL	88.35	86.92	89.32	87.44	76.73	88.11	94.61	94.65
SVM	88.21	88.16	87.32	89.05	76.40	87.74	96.22	96.04
CFSAEMDA	92.27	92.56	91.33	93.14	84.51	91.94	97.87	97.67



(**a**) (**b**)

Figure 2. ROC and P-R curves of different methods. (**a**) ROC curves; (**b**) P-R curves.

2.5. Case Study

To further indicate the application of CFSAEMDA in practice, case studies on lung neoplasms, breast neoplasms and hepatocellular carcinoma were performed. Many scholars are very interested in the study of diseases in humans, and effective early diagnosis is crucial to the treatment of diseases. We first removed all the association information of specific diseases, including known and unknown associations. Then we balanced the remaining samples for training. Finally, the trained model was used to output the prediction probability scores and sort them. The prediction results were confirmed using the HMDD v3.0 [39] database.

Firstly, we chose to study the associations between lung neoplasms and miRNAs. Lung neoplasms are abnormal growths of tissue that form in the lungs. Among all cancers, the death rate of lung neoplasms is the highest in men and the second highest in women. Numerous works in the literature have shown that the generation of lung neoplasms is associated with some miRNAs. For instance, let-7 is involved in the pathogenesis, migration and diffusion of lung neoplasms [40]. Thus, it is necessary to conduct the investigation on the lung neoplasms-related miRNAs. According to the experimental results, 49 of the top 50 lung neoplasms-associated miRNAs are verified (see Supplementary Table S1).

We then completed the association study of breast neoplasms. Breast neoplasm is cancer of the breast tissues, most commonly arising from the milk ducts. For women, breast neoplasm is the common type of cancer. Therefore, the early diagnosis of breast neoplasm is crucial to the treatment of the disease. The researchers found that breast neoplasm is related to the overexpression of circulating miRNA-146a [41]. Based on the experimental results, 50 of the top 50 breast neoplasm-associated miRNAs are confirmed by database (see Supplementary Table S2).

Finally, hepatocellular carcinoma is chosen as the third case study. hepatocellular carcinoma tends to be the main cause of death for patients with cirrhosis. Among all cancer diseases, hepatocellular carcinoma is the third-leading cause of death. Furthermore, the differential expression of miRNAs contributes to the development of hepatocellular carcinoma. Therefore, we chose hepatocellular carcinoma as a case study. The experimental results show that 47 of the top 50 hepatocellular carcinoma-associated miRNAs are confirmed (see Supplementary Table S3).

It should be pointed out that some novel MDAs were found by our proposed method. For example, hsa-mir-92b was not verified to be associated with breast neoplasms in HMDD v2.0. However, the CFSAEMDA method found this MDA, and HMDD v3.0 confirmed it. Other novel MDAs found by our method are highlighted in Supplementary Tables S1–S3. The case studies show the potential of the CFSAEMDA approach in guiding biological experiments.

3. Materials and Methods

3.1. Dataset

The dataset used in this paper is obtained from the HMDD v2.0 database [42], comprising 495 miRNAs, 383 diseases and 5430 MDAs that are experimentally verified, and the number of unknown MDAs is 184,155. The adjacency matrix $A(i,j)$ is utilized to represent the interaction information between miRNAs and diseases. If miRNA m_i is correlated with disease d_j, the value of $A_{(i,j)}$ is equal to 1; otherwise, it is equal to 0. Here, we treat 5430 verified associations as the positive sample set. In order to better reflect the performance of the model, we randomly select 5430 associations from the unknown associations as a negative sample set to balance the sample set (see Table 4).

Table 4. Summary of the samples on the original and balanced datasets.

Data	Known Associations	Unknown Associations
Original dataset	5430	184,155
Balanced Dataset	5430	5430

3.2. Multi-Source Information

3.2.1. Functional Similarity of miRNA

Based on the assumption that different miRNAs associated with similar diseases have similar functions, Wang et al. collected and organized the miRNA functional similarity data [43]. The corresponding data can be downloaded from http://www.cuilab.cn/files/images/cuilab/misim.zip, accessed on 28 May 2023. Thus, in this paper, the matrix $MFS(m_i, m_j)$ is utilized to describe the functional similarity two miRNAs.

3.2.2. Semantic Similarity of Disease

We apply several directed acyclic graphs (DAGs) to estimate the semantic similarity of diseases [44]. Specifically, each disease i can be represented as follows: $DAG_i = (i, G(i), E(i))$. Here, $G(i)$ denotes the node set of all diseases, and $E(i)$ stands for the edges in DAG. Therefore, the semantic contribution value $C1$ of disease d to ancestor node k can be computed by

$$C1_d(k) = \begin{cases} \max\{\mu * C1_d(k') | k' \in \text{children of } k\}, \text{ if } d \neq k \\ 1, \text{ if } d = k \end{cases}, \tag{7}$$

where μ denotes the semantic contribution factor and is equal to 0.5 [43]. Then the disease semantic value $DS1(d)$ is given by

$$DS1(d) = \sum_{k \in G(d)} C1_d(k). \tag{8}$$

The disease semantic similarity $DSS1$ is described by

$$DSS1(d_i, d_j) = \frac{\sum_{k \in G(d_i) \cap G(d_j)} (C1_{d_i}(k) + C1_{d_j}(k))}{DS1(d_i) + DS1(d_j)}, \tag{9}$$

where d_i and d_j represent two diseases.

However, the diseases' semantic contribution value is different at the same level in the $DAGs$. Thus another method is applied to calculate the semantic similarity of diseases [45]; the specific formula is given by

$$C2_d(k) = -\log \frac{NG(k)}{nd}, \tag{10}$$

where nd is the number of diseases, and $NG(k)$ represents the number of DAGs, including k.

Similarly, the disease semantic value $DS2(d)$ and the disease semantic similarity $DSS2$ can be calculated by

$$DS2(d) = \sum_{k \in G(d)} C2_d(k), \tag{11}$$

$$DSS2(d_i, d_j) = \frac{\sum_{k \in G(d_i) \cap G(d_j)} (C2_{d_i}(k) + C2_{d_j}(k))}{DS2(d_i) + DS2(d_j)}. \tag{12}$$

Therefore, the final disease semantic similarity DSS is given by

$$DSS(d_i, d_j) = \frac{1}{2}(DSS1(d_i, d_j) + DSS2(d_i, d_j)). \tag{13}$$

3.2.3. Gaussian Interaction Profile Kernel Similarity of miRNAs and Diseases

According to the literature [46], it can be found that functionally similar miRNAs are more likely to be associated with phenotypically similar diseases. The Gaussian interaction profile (GIP) kernel similarity between miRNAs m_i and m_j can be established by the following equation:

$$KSM(m_i, m_j) = \exp(-\alpha_m||IP(m_i) - IP(m_j)||^2), \tag{14}$$

$$\alpha_m = \frac{\alpha'_m}{\frac{1}{nm}\sum\limits_{i=1}^{nm}||IP(m_i)||^2}, \tag{15}$$

where α_m controls the kernel bandwidth, $IP(m_i)$ and $IP(m_j)$ denote the ith and jth rows in matrix IP, nm is the number of miRNAs, and α'_m is equal to 1 [47].

Similarly, the GIP kernel similarity between diseases d_i and d_j is calculated as follows:

$$KSD(d_i, d_j) = \exp(-\alpha_d||IP(d_i) - IP(d_j)||^2), \tag{16}$$

$$\alpha_d = \frac{\alpha'_d}{\frac{1}{nd}\sum\limits_{i=1}^{nd}||IP(d_i)||^2}. \tag{17}$$

3.3. Integrated Similarity Characteristic

For miRNAs, we collected the functional similarity information and the GIP kernel similarity information. Considering the lack of functional similarity of some miRNAs, the miRNA similarity is calculated by

$$MS(m_i, m_j) = \begin{cases} \frac{KSM(m_i, m_j) + MFS(m_i, m_j)}{2}, \text{ if } MFS(m_i, m_j) \text{ exists} \\ KSM(m_i, m_j), \text{ otherwise} \end{cases}. \tag{18}$$

Similarly, the following formula is applied to compute the disease similarity:

$$DS(d_i, d_j) = \begin{cases} \frac{KSD(d_i, d_j) + DSS(d_i, d_j)}{2}, \text{ if } DSS(d_i, d_j) \text{ exists} \\ KSD(d_i, d_j), \text{ otherwise} \end{cases}. \tag{19}$$

3.4. CFSAEMDA

The present paper introduces a new approach called CFSAEMDA for predicting disease-related miRNAs, which contains three main steps: (i) this paper first construct a comprehensive feature representation of miRNA and disease based on multi-source information. (ii) the stacked autoencoder is applied to extract latent information representation, and (iii) the potential MDAs are predicted by a modified cascade forest structure. The framework of CFSAEMDA is shown in Figure 3.

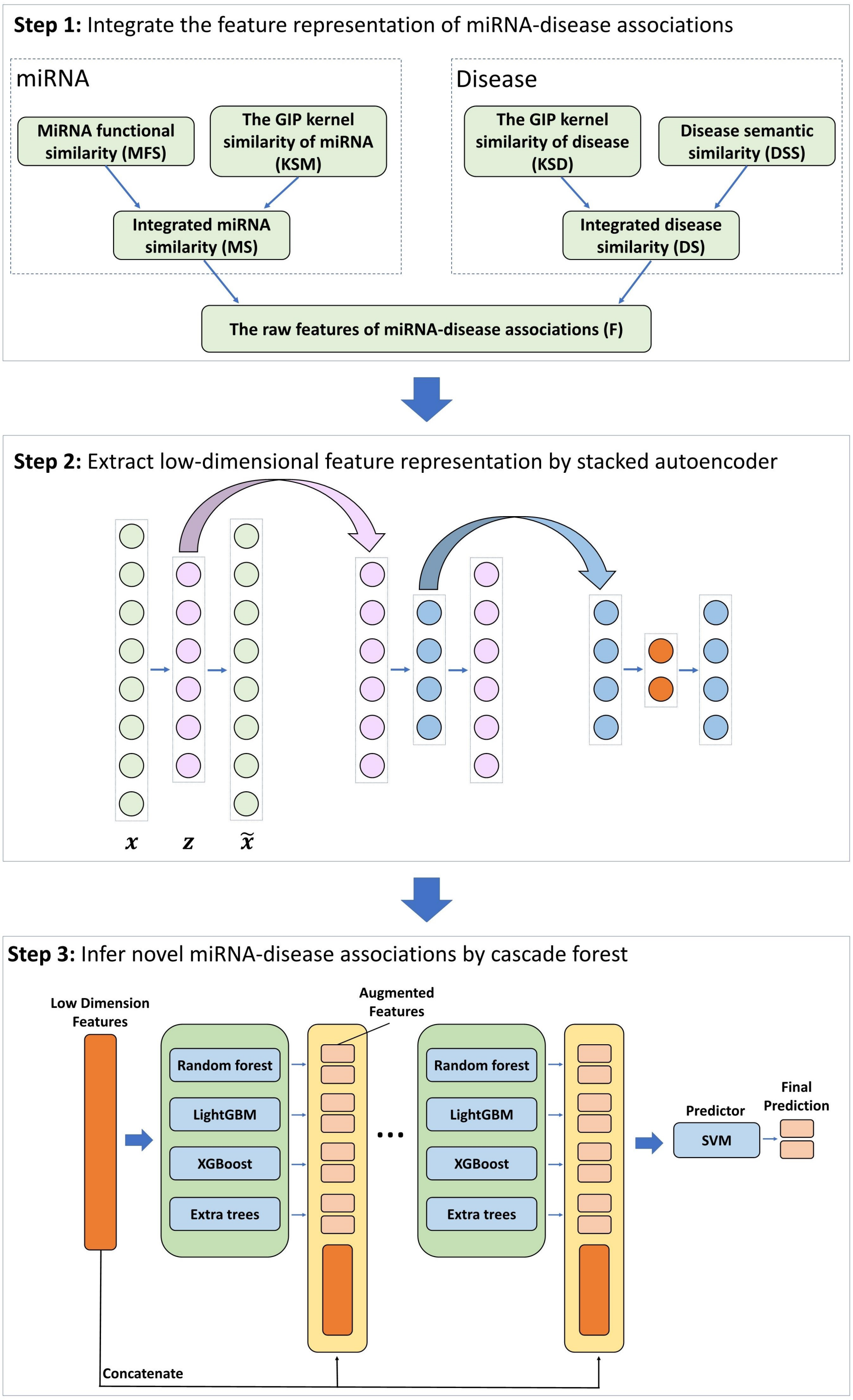

Step 1: Integrate the feature representation of miRNA-disease associations
miRNA
MiRNA functional similarity (MFS)
The GIP kernel similarity of miRNA (KSM)
Integrated miRNA similarity (MS)
Disease
The GIP kernel similarity of disease (KSD)
Disease semantic similarity (DSS)
Integrated disease similarity (DS)
The raw features of miRNA-disease associations (F)
Step 2: Extract low-dimensional feature representation by stacked autoencoder
x
z
$\tilde{x}$
Step 3: Infer novel miRNA-disease associations by cascade forest
Low Dimension Features
Augmented Features
Random forest
LightGBM
XGBoost
Extra trees
Random forest
LightGBM
XGBoost
Extra trees
Predictor
SVM
Final Prediction
Concatenate

Figure 3. The flowchart of CFSAEMDA.

3.4.1. Feature Representation

In this part, inspired by [27], the following strategy is used to construct the high-quality feature representation of MDAs. We obtained the 495-dimensional vector MS_{ij} representing the miRNA similarity network information between two miRNAs. The 383-dimensional vector DS_{ij} for each disease is also obtained. Then, the feature representation of miRNAs (F_m) and disease (F_d) are given by

$$F_m = (MS_1 D'_1, \cdots, MS_1 D'_{383}, \cdots, MS_{495} D'_1, \cdots, MS_{495} D'_{383})^T, \tag{20}$$

$$F_d = (DS_1 D_1, \cdots, DS_1 D_{495}, \cdots, DS_{383} D_1, \cdots, DS_{383} D_{495})^T, \tag{21}$$

where the matrix D is the verified MDA network, D_i is the ith row of D and D'_j is the jth column transpose of D. F_m is the feature representation of miRNAs with 495 × 383 row and 495 + 495 column. F_d is the feature representation of disease with 495 × 383 row and 383 + 383 column. The final high-quality feature representation of MDAs is given by the following formula:

$$F = (F_m, F_d), \tag{22}$$

where F is a matrix with 189,585 (495 × 383) rows and 1756 (495 + 495 + 383 + 383) columns. Then, we randomly select 5430 associations from the unknown associations as a negative sample set to balance the sample set. Therefore, we have 10,860 samples with 1756 dimensional features.

3.4.2. Stacked Autoencoder

Machine learning models have high requirements for the dimensionality of the data input. Namely, the high-dimensional features seriously influence the prediction performance. The feature representation obtained in the previous section is of 1756 dimensions, which is highly dimensional. Therefore, in this article, the stacked autoencoder is utilized to learn the latent feature vectors.

The stacked autoencoder is a deep learning model constructed by stacking multiple single autoencoders [48]. An autoencoder consists of an encoder and a decoder. In the encoding phase, the original feature representation is fed to the encoder to achieve feature compression and dimensionality reduction. The corresponding formula is given by

$$z = f_e(wx + b), \tag{23}$$

where x is the original high-dimensional feature input, w and b denote the weights and bias, $f_e(\bullet)$ represents the non-linear activation function of the encoder phase, and z is the output of the encoder.

The purpose of the decoder is to use the latent representation z to reconstruct the input x. The corresponding formula is given by

$$\widetilde{x} = f_d(w'z + b'), \tag{24}$$

where $\widetilde{x}$ is the reconstruction feature representation of input x; w' and b' represent the weight and bias of the decoder phase; and $f_d(\bullet)$ denotes the non-linear activation function.

Finally, the autoencoder is trained to minimize the reconstruction errors, and the formula is given by

$$\ell(x, \widetilde{x}) = \sum_{i=1}^{N} ||x_i - \widetilde{x}_i||^2, \tag{25}$$

where N is the number of sample sets. The autoencoder is trained to minimize the above loss function and update all parameters iteratively.

Here, according to previous literature [49], the stacked autoencoder is constructed by stacking three of the same autoencoders. The latent feature output dimensions of the three autoencoders are set to 1024, 512 and 256, respectively. Namely, the dimension of reduced feature representations is 256. In addition, the Adam optimizer [50], an algorithm for first-order gradient-based optimization of stochastic objective functions based on low order moment adaptive estimation, is used to update the parameters in the training process of the stacked autoencoder. The Adam optimizer is better than the stochastic gradient descent optimizer. In each hidden layer, the activation function is set to the tanh activation function, and the formula is given by

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \tag{26}$$

3.4.3. Modified Cascade Forest Structure

In this article, after extracting the latent vector representation, the cascade forest structure (see Figure 3) is utilized for prediction. This structure includes two parts: estimators and predictor, which can ensure the layer-by-layer processing of information.

Here, at each level of cascade, we set four estimators, including one random forest [51], one completely random tree forest, one XGBoost [52] and one LightGBM [53]. Each estimator has 100 trees, and there are 400 trees in each cascaded level. The diversity of estimators is crucial for ensemble construction [54]. Given an instance, each estimater can produce an estimate of class distribution. For example, the random forest estimator averages the results of each tree to obtain the final class vector as illustrated in Figure 4. In our work, it is a binary classification task. Four estimators will generate eight-dimensional class vectors, which can be regarded as augmented features. Then, the class vectors are concatenated with the original feature vectors. In addition, each estimator will conduct 5-fold cross validation to prevent overfitting. After expanding to a new level, the performance of the entire cascade will be evaluated on the validation set. If there is no significant improvement in the performance, the cascade layer will automatically terminate. Namely, the complexity of the cascade forest model is determined by the datasets.

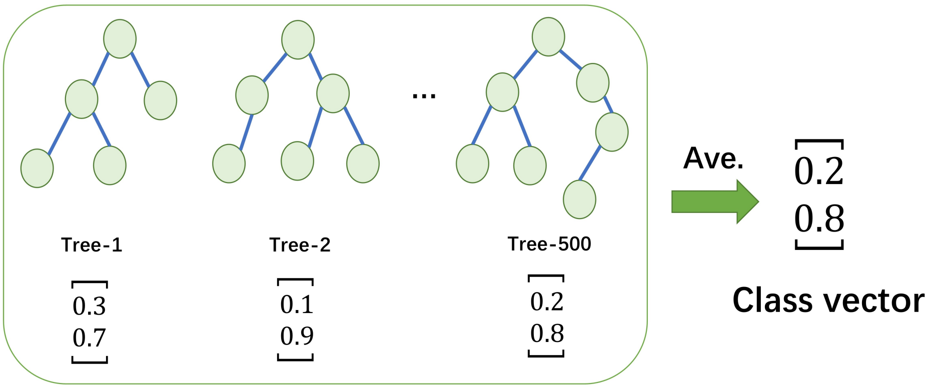


Figure 4. Illustration of class vector generation.

The traditional cascaded forest model obtains the final predicted value by averaging the prediction results of four estimators. In this paper, we modify the predictor to the support vector machine (SVM) algorithm [55], which has excellent predictive ability. The kernel function of SVM is set to the polynomial function. The modification of the predictor greatly improves the performance of the model.

4. Discussion and Conclusions

In this work, we propose a novel computational model, called CFSAEMDA, to infer unknown MDAs. First, the association between diseases and miRNAs can be represented by integrating multi-source similarity features, including functional and GIP kernel similarity of miRNA, semantic and GIP kernel similarity of disease. Then the stacked autoencoder is applied to extract the latent representation of feature. Finally, a modified cascade forest model is employed for predicting the final labels of MDAs. The experimental results demonstrate that our model is excellent and stable. Three case studies further show that

CFSAEMDA can be utilized to guide biological experiments. In conclusion, CFSAEMDA is a valuable and convenient method to infer novel MDAs.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/molecules28135013/s1, Table S1: The prediction results of lung neoplasms; Table S2: The prediction results of breast neoplasms; Table S3: The prediction results of hepatocellular carcinoma.

Author Contributions: X.H.: Writing-original draft, Conceptualization, Software; Z.Y.: Funding acquisition, Supervision; Y.P.: Conceptualization; Z.Z.: Conceptualization. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by National Natural Science Foundation of China (No: 62072296).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: We provided the python data and code for CFSAEMDA at https://github.com/huxiang666666/CFSAEMDA, accessed on 28 May 2023.

Acknowledgments: We would like to thank reviewers for their constructive feedback on this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Sample Availability: Not applicable.

References

1. Chen, X.; Xie, D.; Zhao, Q.; You, Z.H. MicroRNAs and complex diseases: From experimental results to computational models. *Brief. Bioinform.* **2019**, *20*, 515–539. [CrossRef]
2. Ambros, V. MicroRNA pathways in flies and worms: Growth, death, fat, stress, and timing. *Cell* **2003**, *113*, 673–676. [CrossRef] [PubMed]
3. Karp, X.; Ambros, V. Encountering microRNAs in cell fate signaling. *Science* **2005**, *310*, 1288–1289. [CrossRef]
4. Xu, P.; Guo, M.; Hay, B.A. MicroRNAs and the regulation of cell death. *TRENDS Genet.* **2004**, *20*, 617–624. [CrossRef] [PubMed]
5. Takousis, P.; Sadlon, A.; Schulz, J.; Wohlers, I.; Dobricic, V.; Middleton, L.; Lill, C.M.; Perneczky, R.; Bertram, L. Differential expression of microRNAs in Alzheimer's disease brain, blood, and cerebrospinal fluid. *Alzheimer's Dement.* **2019**, *15*, 1468–1477. [CrossRef]
6. Taverner, D.; Llop, D.; Rosales, R.; Ferré, R.; Masana, L.; Vallvé, J.C.; Paredes, S. Plasma expression of microRNA-425-5p and microRNA-451a as biomarkers of cardiovascular disease in rheumatoid arthritis patients. *Sci. Rep.* **2021**, *11*, 15670. [CrossRef]
7. Yan, W.; Wu, X.; Zhou, W.; Fong, M.Y.; Cao, M.; Liu, J.; Liu, X.; Chen, C.H.; Fadare, O.; Pizzo, D.P.; et al. Cancer-cell-secreted exosomal miR-105 promotes tumour growth through the MYC-dependent metabolic reprogramming of stromal cells. *Nat. Cell Biol.* **2018**, *20*, 597–609. [CrossRef]
8. Li, D.; Zhao, Y.; Liu, C.; Chen, X.; Qi, Y.; Jiang, Y.; Zou, C.; Zhang, X.; Liu, S.; Wang, X.; et al. Analysis of MiR-195 and MiR-497 Expression, Regulation and Role in Breast CancerMiR-195 and MiR-497 in Breast Cancer. *Clin. Cancer Res.* **2011**, *17*, 1722–1730. [CrossRef] [PubMed]
9. Zhou, W.; Fong, M.Y.; Min, Y.; Somlo, G.; Liu, L.; Palomares, M.R.; Yu, Y.; Chow, A.; O'Connor, S.T.F.; Chin, A.R.; et al. Cancer-secreted miR-105 destroys vascular endothelial barriers to promote metastasis. *Cancer Cell* **2014**, *25*, 501–515. [CrossRef]
10. Morimura, R.; Komatsu, S.; Ichikawa, D.; Takeshita, H.; Tsujiura, M.; Nagata, H.; Konishi, H.; Shiozaki, A.; Ikoma, H.; Okamoto, K.; et al. Novel diagnostic value of circulating miR-18a in plasma of patients with pancreatic cancer. *Br. J. Cancer* **2011**, *105*, 1733–1740. [CrossRef]
11. Wang, G.; Mao, W.; Zheng, S.; Ye, J. Epidermal growth factor receptor-regulated miR-125a-5p—A metastatic inhibitor of lung cancer. *FEBS J.* **2009**, *276*, 5571–5578. [CrossRef] [PubMed]
12. Freeman, W.M.; Walker, S.J.; Vrana, K.E. Quantitative RT-PCR: Pitfalls and potential. *Biotechniques* **1999**, *26*, 112–125. [CrossRef] [PubMed]
13. Baskerville, S.; Bartel, D.P. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* **2005**, *11*, 241–247. [CrossRef]
14. Jiang, Q.; Hao, Y.; Wang, G.; Juan, L.; Zhang, T.; Teng, M.; Liu, Y.; Wang, Y. Prioritization of disease microRNAs through a human phenome-microRNAome network. *BMC Syst. Biol.* **2010**, *4*, S2. [CrossRef]
15. Chen, X.; Yan, C.C.; Zhang, X.; You, Z.H.; Deng, L.; Liu, Y.; Zhang, Y.; Dai, Q. WBSMDA: Within and between score for MiRNA-disease association prediction. *Sci. Rep.* **2016**, *6*, 21106. [CrossRef]

16. Zeng, X.; Liu, L.; Lü, L.; Zou, Q. Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics* **2018**, *34*, 2425–2432. [CrossRef]
17. Zhong, Y.; Xuan, P.; Wang, X.; Zhang, T.; Li, J.; Liu, Y.; Zhang, W. A non-negative matrix factorization based method for predicting disease-associated miRNAs in miRNA-disease bilayer network. *Bioinformatics* **2018**, *34*, 267–277. [CrossRef]
18. Yu, S.; Wang, H.; Liu, T.; Liang, C.; Luo, J. A knowledge-driven network for fine-grained relationship detection between miRNA and disease. *Brief. Bioinform.* **2022**, *23*, bbac058. [CrossRef]
19. Zhang, W.; Li, Z.; Guo, W.; Yang, W.; Huang, F. A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2019**, *18*, 405–415. [CrossRef]
20. Ma, Y.; He, T.; Ge, L.; Zhang, C.; Jiang, X. MiRNA-disease interaction prediction based on kernel neighborhood similarity and multi-network bidirectional propagation. *BMC Med. Genom.* **2019**, *12*, 185. [CrossRef] [PubMed]
21. Che, K.; Guo, M.; Wang, C.; Liu, X.; Chen, X. Predicting MiRNA-disease association by latent feature extraction with positive samples. *Genes* **2019**, *10*, 80. [CrossRef] [PubMed]
22. Guo, F.; Yin, Z.; Zhou, K.; Li, J. PLncWX: A machine-learning algorithm for plant lncRNA identification based on WOA-XGBoost. *J. Chem.* **2021**, *2021*, 6256021. [CrossRef]
23. Chen, M.; Yin, Z. Classification of Cardiotocography based on Apriori algorithm and multi-model ensemble classifier. *Front. Cell Dev. Biol.* **2022**, 844. [CrossRef] [PubMed]
24. Zheng, K.; You, Z.H.; Wang, L.; Zhou, Y.; Li, L.P.; Li, Z.W. MLMDA: A machine learning approach to predict and validate MicroRNA–disease associations by integrating of heterogenous information sources. *J. Transl. Med.* **2019**, *17*, 260. [CrossRef] [PubMed]
25. Zhou, S.; Wang, S.; Wu, Q.; Azim, R.; Li, W. Predicting potential miRNA-disease associations by combining gradient boosting decision tree with logistic regression. *Comput. Biol. Chem.* **2020**, *85*, 107200. [CrossRef]
26. Chen, X.; Li, T.H.; Zhao, Y.; Wang, C.C.; Zhu, C.C. Deep-belief network for predicting potential miRNA-disease associations. *Brief. Bioinform.* **2021**, *22*, bbaa186. [CrossRef]
27. Liu, W.; Lin, H.; Huang, L.; Peng, L.; Tang, T.; Zhao, Q.; Yang, L. Identification of miRNA–disease associations via deep forest ensemble learning based on autoencoder. *Brief. Bioinform.* **2022**, *23*, bbac104. [CrossRef]
28. Ai, N.; Liang, Y.; Yuan, H.L.; Ou-Yang, D.; Liu, X.Y.; Xie, S.L.; Ji, Y.H. MHDMF: Prediction of miRNA–disease associations based on Deep Matrix Factorization with Multi-source Graph Convolutional Network. *Comput. Biol. Med.* **2022**, *149*, 106069. [CrossRef]
29. Liu, Y.; Wang, S.L.; Zhang, J.F.; Zhang, W.; Li, W. A neural collaborative filtering method for identifying miRNA-disease associations. *Neurocomputing* **2021**, *422*, 176–185. [CrossRef]
30. Peng, J.; Hui, W.; Li, Q.; Chen, B.; Hao, J.; Jiang, Q.; Shang, X.; Wei, Z. A learning-based framework for miRNA-disease association identification using neural networks. *Bioinformatics* **2019**, *35*, 4364–4371. [CrossRef]
31. Wang, C.C.; Li, T.H.; Huang, L.; Chen, X. Prediction of potential miRNA–disease associations based on stacked autoencoder. *Brief. Bioinform.* **2022**, *23*, bbac021. [CrossRef] [PubMed]
32. Zhou, Z.H.; Feng, J. Deep Forest: Towards An Alternative to Deep Neural Networks. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, Melbourne, Australia, 19–25 August 2017; pp. 3553–3559.
33. Lin, W.; Wu, L.; Zhang, Y.; Wen, Y.; Yan, B.; Dai, C.; Liu, K.; He, S.; Bo, X. An enhanced cascade-based deep forest model for drug combination prediction. *Brief. Bioinform.* **2022**, *23*, bbab562. [CrossRef]
34. Chu, Y.; Kaushik, A.C.; Wang, X.; Wang, W.; Zhang, Y.; Shan, X.; Salahub, D.R.; Xiong, Y.; Wei, D.Q. DTI-CDF: A cascade deep forest model towards the prediction of drug-target interactions based on hybrid features. *Brief. Bioinform.* **2021**, *22*, 451–462. [CrossRef] [PubMed]
35. Olayan, R.S.; Ashoor, H.; Bajic, V.B. DDR: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches. *Bioinformatics* **2018**, *34*, 1164–1173. [CrossRef] [PubMed]
36. Zeng, X.; Zhu, S.; Hou, Y.; Zhang, P.; Li, L.; Li, J.; Huang, L.F.; Lewis, S.J.; Nussinov, R.; Cheng, F. Network-based prediction of drug–target interactions using an arbitrary-order proximity embedded deep forest. *Bioinformatics* **2020**, *36*, 2805–2812. [CrossRef]
37. Dai, Q.; Chu, Y.; Li, Z.; Zhao, Y.; Mao, X.; Wang, Y.; Xiong, Y.; Wei, D.Q. MDA-CF: Predicting miRNA-disease associations based on a cascade forest model by fusing multi-source information. *Comput. Biol. Med.* **2021**, *136*, 104706. [CrossRef] [PubMed]
38. Deng, Y.; Xu, X.; Qiu, Y.; Xia, J.; Zhang, W.; Liu, S. A multimodal deep learning framework for predicting drug–drug interaction events. *Bioinformatics* **2020**, *36*, 4316–4322. . [CrossRef]
39. Huang, Z.; Shi, J.; Gao, Y.; Cui, C.; Zhang, S.; Li, J.; Zhou, Y.; Cui, Q. HMDD v3. 0: A database for experimentally supported human microRNA–disease associations. *Nucleic Acids Res.* **2019**, *47*, D1013–D1017. [CrossRef]
40. Johnson, S.M.; Grosshans, H.; Shingara, J.; Byrom, M.; Jarvis, R.; Cheng, A.; Labourier, E.; Reinert, K.L.; Brown, D.; Slack, F.J. RAS is regulated by the let-7 microRNA family. *Cell* **2005**, *120*, 635–647. [CrossRef]
41. Kumar, S.; Keerthana, R.; Pazhanimuthu, A.; Perumal, P. Overexpression of circulating miRNA-21 and miRNA-146a in plasma samples of breast cancer patients. *Indian J. Biochem. Biophys.* **2013**, *50*, 210–214.
42. Li, Y.; Qiu, C.; Tu, J.; Geng, B.; Yang, J.; Jiang, T.; Cui, Q. HMDD v2. 0: A database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res.* **2014**, *42*, D1070–D1074. [CrossRef] [PubMed]
43. Wang, D.; Wang, J.; Lu, M.; Song, F.; Cui, Q. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* **2010**, *26*, 1644–1650. [CrossRef] [PubMed]

44. Schriml, L.M.; Arze, C.; Nadendla, S.; Chang, Y.W.W.; Mazaitis, M.; Felix, V.; Feng, G.; Kibbe, W.A. Disease Ontology: A backbone for disease semantic integration. *Nucleic Acids Res.* **2012**, *40*, D940–D946. [CrossRef]
45. Xuan, P.; Han, K.; Guo, M.; Guo, Y.; Li, J.; Ding, J.; Liu, Y.; Dai, Q.; Li, J.; Teng, Z.; et al. Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors. *PLoS ONE* **2013**, *8*, e70204. [CrossRef]
46. Van Laarhoven, T.; Nabuurs, S.B.; Marchiori, E. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* **2011**, *27*, 3036–3043. [CrossRef]
47. Chen, X.; Yan, G.Y. Novel human lncRNA–disease association inference based on lncRNA expression profiles. *Bioinformatics* **2013**, *29*, 2617–2624. [CrossRef]
48. Bengio, Y.; Lamblin, P.; Popovici, D.; Larochelle, H. Greedy layer-wise training of deep networks. In *Advances in Neural Information Processing Systems 19 (NIPS 2006)*; MIT Press: Cambridge, MA, USA, 2007; pp. 153–160.
49. Bahi, M.; Batouche, M. Deep semi-supervised learning for DTI prediction using large datasets and H_2O-spark platform. In Proceedings of the 2018 International Conference on Intelligent Systems and Computer Vision (ISCV), Fez, Morocco, 2–4 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–7.
50. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
51. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
52. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
53. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. Lightgbm: A highly efficient gradient boosting decision tree. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 3147–3155.
54. Zhou, Z.H.; Feng, J. Deep forest. *Natl. Sci. Rev.* **2019**, *6*, 74–86. [CrossRef]
55. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.