biology

MDPI

Article

SMMDA: Predicting miRNA-Disease Associations by Incorporating Multiple Similarity Profiles and a Novel Disease Representation

Bo-Ya Ji [1], Liang-Rui Pan [1], Ji-Ren Zhou [2], Zhu-Hong You [2,*] and Shao-Liang Peng [1,*]

[1] College of Computer Science and Electronic Engineering, Hunan University, Changsha 410200, China; byj@hnu.edu.cn (B.-Y.J.); lip141772@gmail.com (L.-R.P.)
[2] College of Computer Science, Northwestern Polytechnic University, Xi'an 710072, China; zhoujiren@mail.nwpu.edu.cn
* Correspondence: zhuhongyou@gmail.com (Z.-H.Y.); slpeng@hnu.edu.cn (S.-L.P.)

Simple Summary: Predicting possible associations between miRNAs and diseases would provide new perspectives on disease diagnosis, pathogenesis, and gene therapy. In this work, considering the limited accessibility, high time consumption and high cost in traditional biological researches, we presented a novel computational method called SMMDA by incorporating multiple similarity profiles and a novel disease rep-resentation to accelerate the identification of potential miRNA-disease associations. SMMDA was intended to be useful for the prediction of associations between miRNAs and diseases, and to be effective for prevention, diagnosis, treatment and prognosis of Human diseases.

Abstract: Increasing evidence has suggested that microRNAs (miRNAs) are significant in research on human diseases. Predicting possible associations between miRNAs and diseases would provide new perspectives on disease diagnosis, pathogenesis, and gene therapy. However, considering the intrinsic time-consuming and expensive cost of traditional Vitro studies, there is an urgent need for a computational approach that would allow researchers to identify potential associations between miRNAs and diseases for further research. In this paper, we presented a novel computational method called SMMDA to predict potential miRNA-disease associations. In particular, SMMDA first utilized a new disease representation method (MeSHHeading2vec) based on the network embedding algorithm and then fused it with Gaussian interaction profile kernel similarity information of miRNAs and diseases, disease semantic similarity, and miRNA functional similarity. Secondly, SMMDA utilized a deep auto-coder network to transform the original features further to achieve a better feature representation. Finally, the ensemble learning model, XGBoost, was used as the underlying training and prediction method for SMMDA. In the results, SMMDA acquired a mean accuracy of 86.68% with a standard deviation of 0.42% and a mean AUC of 94.07% with a standard deviation of 0.23%, outperforming many previous works. Moreover, we also compared the predictive ability of SMMDA with different classifiers and different feature descriptors. In the case studies of three common Human diseases, the top 50 candidate miRNAs have 47 (esophageal neoplasms), 48 (breast neoplasms), and 48 (colon neoplasms) are successfully verified by two other databases. The experimental results proved that SMMDA has a reliable prediction ability in predicting potential miRNA-disease associations. Therefore, it is anticipated that SMMDA could be an effective tool for biomedical researchers.

Keywords: miRNA-disease associations prediction; deep neural network; ensemble learning; XGBoost

Citation: Ji, B.-Y.; Pan, L.-R.; Zhou, J.-R.; You, Z.-H.; Peng, S.-L. SMMDA: Predicting miRNA-Disease Associations by Incorporating Multiple Similarity Profiles and a Novel Disease Representation. *Biology* **2022**, *11*, 777. https://doi.org/10.3390/biology11050777

Academic Editors: Alper Kucukural and Stefano Volinia

Received: 14 April 2022
Accepted: 17 May 2022
Published: 20 May 2022

1. Introduction

MicroRNAs (miRNAs) constitute a group of about 22 nucleotide long noncoding RNAs, prevalent in flora and fauna [1]. It acts as an essential regulatory factor of gene

expressions that participate in degradation or post-transcriptional repression by supplementarily binding to corresponding 3′untranslated regions of their mRNA [2].

By targeting multiple transcripts, miRNAs play pivotal roles in biological processes, such as cell development [3–5], apoptosis [6], metabolism [7] and so on. Recently, an increasing amount of researches have revealed the effectiveness of microRNAs as prognostic biomarkers or important diagnostic and promising therapeutic targets for the treatment of malignant tumors [8]. The expression of hsa-miR-17-3p is altered in lung cancer from smokers and the methylation levels of hsa-miR-124-2 were reduced in SiHa cells [9]. The critical role of miRNAs in humans has attracted the attention of many researchers, and traditional in vitro experimental methods have been used to investigate the association between miRNAs and human diseases, and many significant results have been achieved. However, biological in vitro experiments require high human and financial costs and are not destined to study large-scale miRNA and disease data. In recent years, machine learning, deep learning, and other methods have improved and integrated bioinformatics problems. Accordingly, more and more researchers are trying to use methods such as machine learning to conduct miRNA-human disease studies.

Based on the hypothesis that interacting miRNA-disease pairs are more functionally similar and tend to be associated with the same miRNAs or diseases [10–12], computational models for predicting miRNA–disease associations have emerged in recent years. For example, Chen et al. [13] developed a heterogeneous label propagation method (HLPMDA) by propagating a heterogeneous label in the multiple networks of miRNAs, diseases, and lncRNAs to predict miRNA-disease associations. Ji et al. [10] focused on constructing a human biological association network using the association between miRNAs and diseases, and other biomolecules in the human body for predicting potential associations between miRNAs and diseases. In addition, this work also introduces graph representation learning methods and deep stacked autoencoder methods to obtain excellent prediction performance. Chen et al. [14] invented a bipartite network projection method (BNPMDA) by fusing integrated miRNA and disease similarity to predict miRNA-disease associations. In this work, a bipartite network recommendation method was applied to predict the potential associations between miRNAs and diseases.

In addition, machine learning approaches have been widely investigated in bioinformatics for predicting potential associations between miRNAs and diseases [15]. For example, Ji et al. [16] used a typical integrated learning approach, random forest, for the potential association of miRNAs with human diseases. They designed an attribute network embedding approach to construct a model with mighty predictive power by considering both the attribute features and network features using a typical integrated learning approach, random forest, for the potential association of miRNAs with human diseases. Zheng et al. utilized deep auto-encoder neural network (AE) and random forest classifier to predict potential miRNA-disease associations (MLMDA). Xu et al. [17] proposed a novel-method-based miRNA target–dysregulated network. Based on the changes and features in miRNA expression, they used SVM classifier to general predictive accuracy. Zhang et al. [18] utilized a variational auto-encoder approach for miRNA-disease association prediction, called VAEMDA. They constructed two spliced matrices by combining the integrated miRNA similarity and the integrated disease similarity with known miRNA–disease associations, respectively. This method prevents the noise created by the random selection of negative instances and shows miRNA-disease associations from the viewpoint of data distribution.

In this work, we presented a novel computational method called SMMDA by incorporating multiple similarity profiles and a novel disease representation to accelerate the identification of potential miRNA-disease associations. The flowchart of SMMDA to predict potential miRNA-disease associations was shown in Figure 1. In summary, the main contributions of this paper are as follows below.

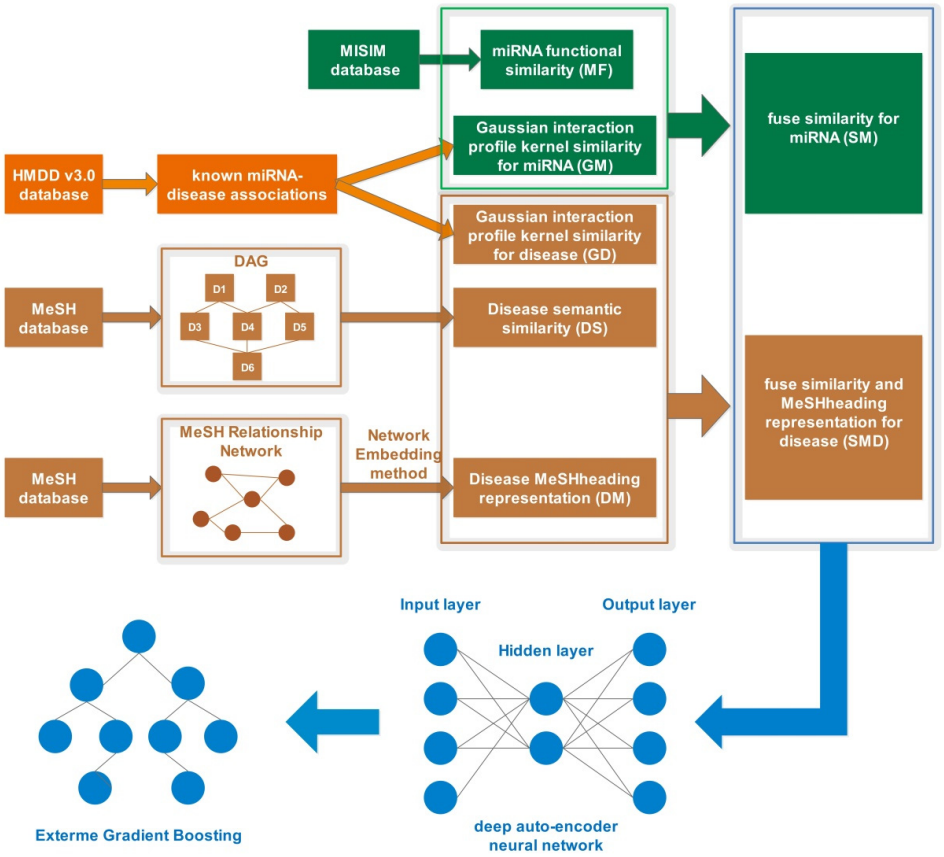

Figure 1. Flowchart of SMMDA to predict potential miRNA-disease associations.

Considering the limited accessibility, high time consumption, and high cost of traditional biological research, a novel computational model called SMMDA was proposed to accelerate the identification of potential associations between miRNAs and diseases.

The multiple similarity profiles of miRNAs and diseases and a novel disease representative feature were incorporated to predict potential miRNA-disease associations, enhancing predictive accuracy.

Deep learning is used for high-quality extraction of integrated features, and the gradient boosting method is used for fast and highly accurate training and prediction.

Compared with previous related works, the experiment results have proved the superior performance of SMMDA for predicting potential miRNA-disease associations.

2. Materials and Methods

2.1. Human miRNA-Disease Associations

The HMDD v3.0 database (Human MicroRNA Disease Database) [19] contains 1102 miRNAs and 850 diseases and 32,281 associations in 17,412 papers. In our experiments, the positive dataset contains 1057 miRNAs, 850 diseases and 32,226 associations. What was removed were association data considered unreliable by the public database miRBase. In addition, we randomly selected 32,226 unrelated associations as the negative dataset, and it should be noted that these associations have been removed from the positive dataset.

2.2. miRNA Functional Similarity

Functional similarity between various miRNAs is a critical feature used for miRNA-disease association prediction, derived from the calculations of Wang et al. [20] They constructed a miRNA functional similarity score matrix (*MF*), available in http://www.cuilab.cn/files/images/cuilab/misim.zip (accessed on 1 March 2022), based on the principle that miRNAs with similar functions are more likely to be associated with diseases with

similar phenotypes. Finally, the similarity score between miRNA m_1 and miRNA m_2 can be expressed as $MF(m_1, m_2)$.

2.3. Gaussian Interaction Profile Kernel Similarity

Since miRNAs with similar functions are more likely to be associated with diseases with similar phenotypes and vice versa, we further calculated Gaussian interaction profile kernel similarity (GIP) for miRNAs and diseases [21]. In particular, an 850 rows and 1057 columns adjacency matrix was first constructed, with the rows in the matrix representing the number of miRNAs and the columns representing the number of diseases. The values of the elements in the matrix depend on whether there is an miRNA m_i and disease d_j association in the HMDD database; if it does, $MD(m_i, d_j)$ is equal to 1, otherwise it is equal to 0. The i-row vector of the adjacency matrix MD can be expressed as the binary vector $MD(m_i)$, denoting the interaction profiles of miRNA m_i. Based on the above definition, the GIP feature between miRNA m_i and m_j, $GM(m_i, m_j)$, is defined as follows:

$$\mathrm{GM}(m_i, m_j) = \exp\left(-\delta_m \|MD(m_i) - MD(m_j)\|^2\right) \tag{1}$$

where δ_m can be obtained by normalizing original parameter, which is the kernel bandwidth, as shown below:

$$\delta_m = \frac{1}{m}\sum_{i=1}^{m}\|MD(m_i)\|^2 \tag{2}$$

where m denotes the number of rows of the MD.

In the same way, the kernel similarity $GD(d_i, d_j)$ of the GIP similarity feature between disease d_i and d_j is defined as follow:

$$\mathrm{GD}(d_i, d_j) = \exp\left(-\delta_d \|MD(d_i) - MD(d_j)\|^2\right) \tag{3}$$

$$\delta_d = \frac{1}{d}\sum_{i=1}^{d}\|MD(d_i)\|^2 \tag{4}$$

where the total number of columns and i-column vector of the adjacent matrix MD are denoted by d and $MD(d_i)$.

2.4. Disease Semantic Similarity

The U.S. National Library of Medicine classifies all human diseases and has constructed the Medical Subject Headings (MeSH) database. According to this database division, we can use a directed acyclic graph (DAG) to represent each disease. For example, we can use DAG(D) = (D, T(D), E(D)) to represent a disease D, where T(D) denotes node D and all its ancestor nodes, and E(D) denotes the set of edges associated with node D. Further, we defined the contribution of node d in DAG(D) to the semantic value of disease node D as:

$$\mathrm{DV(D)} = \sum_{d \in T(\mathrm{D})} \mathrm{D_D}(d) \tag{5}$$

$$\begin{cases} \mathrm{D_D}(d) = 1 \text{ if } d = \mathrm{D} \\ \mathrm{D_D}(d) = \max\{\Delta_* \mathrm{D_D}(d') | d' \in \text{children of } d\} \text{ if } d \neq \mathrm{D} \end{cases} \tag{6}$$

where Δ is the semantic contribution factor [20,22].

From the above equation, we can get that if two diseases have a larger shared part, then their similarity scores are higher. Therefore, the semantic similarity scores between diseases d_i and d_j are shown below:

$$\mathrm{DS}(d_i, d_j) = \frac{\sum_{t \in T(d_i) \cap T(d_j)} \left(\mathrm{D}_{d_i}(t) + \mathrm{D}_{d_j}(t)\right)}{\mathrm{DV}(d_i) + \mathrm{DV}(d_j)} \tag{7}$$

2.5. MeSHHeading2vec Method

The characterization of diseases is an important part for predicting miRNA-disease associations, which is directly related to the prediction accuracy of the model. More and more researchers are focusing on high-quality feature representation of diseases, and in this section, we utilize a novel computational method, namely MeSHHeading2vec [23]. This new disease representation method compares to traditional GIP similarity features and semantic similarity features of diseases has been shown to have an even better performance. Specifically, a relational network is first constructed which transforms the MeSH tree structure of the diseases, connecting the different disease MeSH headings. In addition, the method calculates the node and edge number in the network and provides a brief analysis of the distribution of labels of nodes and the degree of distribution, where the pattern of tree numbers corresponding to a node determines the label (category) of each node (MeSH heading). Finally, different network representation learning methods including DeepWalk [24], LINE [25], SDNE [26], HOPE [27], and LAP [28] are applied to this relational network thus obtaining high-quality network features of the disease and retainning the raw node related information and network structure. Based on the method, the LINE network representation method was chosen for high-quality disease network feature extraction to enhance the predictive power of SMMDA for potential miRNA-disease associations

2.6. Incorporating Multiple Similarity Profiles and a Novel Disease Representation

In this section, multiple miRNA similarity profile features, disease similarity profile features, and new high-quality disease representation features are incorporating. Specifically, the final matrix MFM(m_i, m_j) of miRNA feature is defined as follows:

$$\mathrm{MFM}(m_i, m_j) = \begin{cases} MF(m_i, m_j), \ if\ m_i\ and\ m_j\ has\ functional\ similarity \\ \mathrm{GM}(m_i, m_j), \ otherwise \end{cases} \tag{8}$$

where GM denotes miRNA GIP similarity and *MF* denotes miRNA functional similarity matrix.

Similarly, the final disease feature matrix DFM(d_i, d_j) is defined:

$$\mathrm{DFM}(d_i,\ d_j) = \begin{cases} \mathrm{DM}(d_i,\ d_j), \ \ if\ d_i\ and\ d_j\ has\ Meshheading\ feature \\ \mathrm{DS}(d_i,\ d_j), \ \ if\ d_i\ and\ d_j\ has\ no\ Meshheading\ feature \\ \mathrm{GD}(d_i,\ d_j), \ \ otherwise \end{cases} \tag{9}$$

where DM denotes the new high-quality disease representation feature, DS denotes the disease semantic similarity feature and GD denotes the disease Gaussian interaction profile kernel similarity feature.

2.7. Deep Auto-Encoder Learning Method

For eliminating noise and reduce dimension of original features, the deep auto-encoder method (DAE) [29] was used for improving prediction accuracy of miRNA-disease associations in our work. Specifically, we constructed the deep learning framework containing 7 fully connected layers as hidden layers, where the number of neurons, respectively, is (2^9, 2^8, 2^7, 2^6, 2^7, 2^8, 2^9), and the activation function for each layer uses the ReLU function. The first 3 hidden layers are the encoding part, the last 3 hidden layers are the decoding part, and the output of the middle layer is the final reduced dimensional feature data. First, the encoding part projects the original features f from the input layer to the hidden layer $h1$ using the mapping function $y1$. Secondly, the decoding part projects the hidden part h to the output layer $h2$ by a mapping function $y2$.

$$h1 = y1(\mathrm{f})\ \mathrm{S}_{y1}(\mathrm{Wf} + \mathrm{p}) \tag{10}$$

$$h2 = y2(h1)\ \mathrm{S}_{y2}(\mathrm{W\prime f} + \mathrm{q}) \tag{11}$$

Furthermore, the ReLU function is chosen as the activation function of AE in our work.

$$S_{y1}(t) = S_{y2}(t) = \max(0, Wt + b) \tag{12}$$

2.8. Exterme Gradient Boosting

In recent years, the Exterme Gradient Boosting (XGBoost) proposed by Chen et al. is widely used by researchers and has yielded satisfactory results. XGBoost is a new classifier based on classification and regression trees integration (CART) and utilizes gradient boosting to optimize trees [30].

Set the output of a tree as shown below:

$$F(x) = W_q(x_i) \tag{13}$$

where W_q is the score of the leaf nodel q and x_i is the input vector. On the basis, the output of the set of K trees is:

$$y_i = \sum_{k=1}^{K} F_k(x_i) \tag{14}$$

The objective function O at step t of XGBoost method is:

$$O(t) = \sum_{i=1}^{n} L\left(y_i, y'^{t-1}_i + F_t(x_i)\right) + \sum_{i=1}^{t} P(F_i) \tag{15}$$

where L is the train loss function between the output y' and real y, the second term in the function is for regularization.

Moreover, the complexity of the XGBoost method is defined as follows:

$$P(F) = \gamma T + 0.5\lambda \sum_{j=1}^{T} w_j^2 \tag{16}$$

where γ is the pseudo-regularization hyperparameter, T is the total number of leaf nodes and λ is the L2 norm for leaf weights.

For detecting the optimal weights W, the gradient is used to conduct second-order approximation to the loss function, and the optimal value of the objective function is

$$O(t) = -0.5 * \sum_{j=1}^{T} \left(\sum_{i \epsilon I} g_i\right)^2 * \left(\sum_{i \epsilon I} h_i + \lambda\right)^{-1} + \gamma T \tag{17}$$

where I is the set of leaf nodes, g_i and h_i are the gradient statistics on the loss function, given by:

$$g_i = \partial_{y'^{t-1}} L\left(y_i, y'^{t-1}_i\right) \tag{18}$$

$$h_i = \partial^2{}_{y'^{t-1}} L\left(y_i, y'^{t-1}_i\right) \tag{19}$$

3. Results and Discussion

3.1. The Detailed Prediction Performance of SMMDA

To accurately assess the predictive power of SMMDA for potential miRNA-disease associations, the more widely adopted five-fold cross-validation method was utilized. The method was repeated five times by randomly shuffling the samples and dividing them evenly into five parts, with one part as the test dataset and the remaining four groups as the training dataset. The detailed results of the experiments are recorded in Table 1, containing six commonly used predictive metrics, namely accuracy (Acc.), precision (Prec.), sensitivity (Sen.), Mathews correlation coefficient (MCC), and areas under the ROC curve (AUC). From the experimental results, we can see that SMMDA achieved a mean accuracy of 86.68% with a standard deviation of 0.42%, which is a good proof of the excellent performance of SMMDA. For the AUC metric, which is more indicative of the model's predictive power, SMMDA obtained a mean of 94.06% with a standard deviation of 0.23% under five-fold cross-validation.

Table 1. The detailed prediction performance of SMMDA.

Fold	ACC. (%)	Spec. (%)	Sen.(%)	MCC (%)	Prec. (%)	AUC (%)
0	86.82	86.95	86.69	73.64	86.92	94.16
1	86.99	86.45	87.53	73.98	86.60	94.30
2	86.80	86.52	87.08	73.59	86.59	94.02
3	85.94	85.76	86.13	71.89	85.81	93.70
4	86.86	87.01	86.70	73.72	86.97	94.17
Average	**86.68 ± 0.42**	**86.54 ± 0.50**	**86.83 ± 0.52**	**73.36 ± 0.84**	**86.58 ± 0.46**	**94.06 ± 0.23**

3.2. Comparison of Different Feature Combinations

To further assess the capability of our proposed feature descriptors, we compared them with different descriptors. In particular, the feature descriptors in our work is generated by fusing a novel disease representation, miRNA functional similarity, disease semantic similarity, and GIP kernel similarity information of miRNAs and diseases. Furthermore, a different feature descriptor is generated by only fusing miRNA functional similarity, disease semantic similarity, and GIP kernel similarity information of miRNAs and diseases (DescSim). The detailed results of the feature descriptors DescSim under 5-fold cross-validation were shown in Table 2. The results that our feature descriptors have a better performance than the feature descriptors used in many previous methods which only fuse similarity information to predict underlying miRNA-disease associations.

Table 2. Evaluation of our method with different feature combinations.

Fold	ACC. (%)	Spec. (%)	Sen. (%)	MCC (%)	Prec. (%)	AUC (%)
0	86.64	86.61	86.67	73.29	86.62	94.15
1	86.58	86.10	87.06	73.16	86.23	94.10
2	86.32	86.41	86.24	72.65	86.38	93.68
3	87.02	86.72	87.32	74.04	86.80	94.07
4	86.45	86.10	86.81	72.91	86.20	93.84
Average	**86.60 ± 0.26**	**86.39 ± 0.29**	**86.82 ± 0.41**	**73.21 ± 0.52**	**86.45 ± 0.26**	**93.97 ± 0.20**
SMMDA	**86.68 ± 0.42**	**86.54 ± 0.50**	**86.83 ± 0.52**	**73.36 ± 0.84**	**86.58 ± 0.46**	**94.06 ± 0.23**

3.3. Comparison of Different Classifier Methods

In order to select the best predictive classifier method for SMMDA model, we conducted, respectively, the five-fold cross-validation experiment using different classifier methods including decision tree (DT) [31], logistic regression (LR) [32], random forest (RF) [33], and Extreme Gradient Boosting (XGBoost). It is worth noting that all experiments adopt the same environment and different classification methods adopt default training parameters to ensure the fairness and ease of operation of the comparison experiment. The average results of different classifier methods were displayed in Table 3. The AUC values and ROC curves, AUPR values and PR curves was respectively shown in the Figure 2. The comparison experiment demonstrates that XGBoost has a better performance than the other methods. Therefore, it is more suitable for SMMDA models.

Table 3. Comparison of SMMDA with different classifier methods.

Classifier	ACC. (%)	Spec. (%)	Sen. (%)	MCC (%)	Prec. (%)	AUC (%)
DT	84.10 ± 0.15	83.30 ± 0.51	84.89 ± 0.33	68.20 ± 0.29	83.56 ± 0.38	87.53 ± 0.14
LR	82.50 ± 0.22	84.17 ± 0.66	80.82 ± 0.41	65.03 ± 0.45	83.62 ± 0.52	89.91 ± 0.21
RF	85.66 ± 0.36	85.61 ± 0.21	85.71 ± 0.63	71.32 ± 0.72	85.63 ± 0.22	93.05 ± 0.30
XGBoost	**86.68 ± 0.42**	**86.54 ± 0.50**	**86.83 ± 0.52**	**73.36 ± 0.84**	**86.58 ± 0.46**	**94.06 ± 0.23**

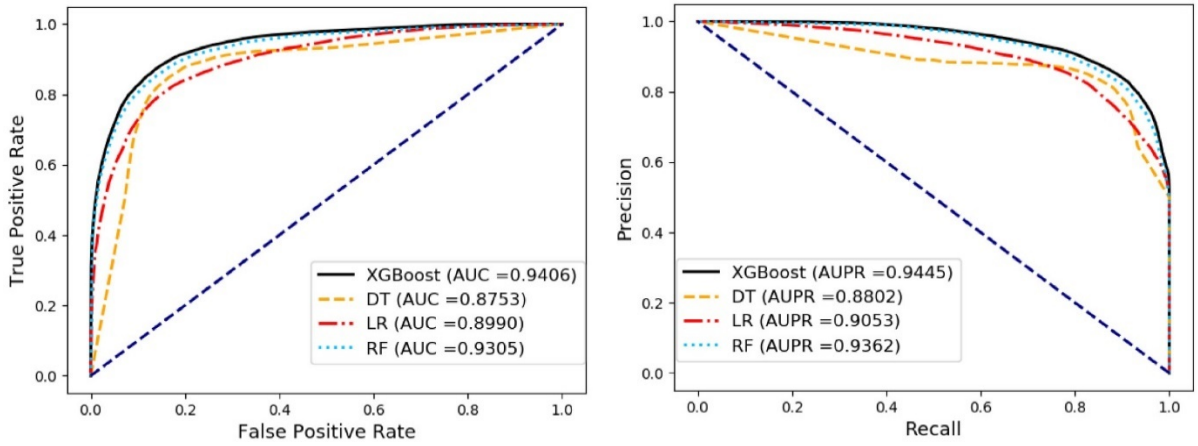

Figure 2. Comparison of SMMDA with random forest, logistic regression, decision tree and XGBoost classifiers.

3.4. Comparison of Previous Related Works

To further demonstrate the good performance of SMMDA, we compared 10 previous start-of-the-art computational models, namely DANE-MDA [16], MLMDA [34], MTDN [17], VAEMDA [18], LMTRDA [35], DBMDA [36], WBSMDA [37], PBMDA [38], HDMP [39], RLSMDA [40]. Furthermore, the data sets used by all these models are from the HMDD database. Here we selected the results of average AUC under five-fold cross-validation experiment as evaluation indicators. As shown in Table 4, SMMDA has a higher mean AUC value in the experiment, which proves its superior performance in the field of miRNA-disease association prediction.

Table 4. Comparison of previous related works under the five-fold cross-validation.

Models	Average AUC (%)
DANE-MDA	92.64
MLMDA	91.72
MTDN	91.89
VAEMDA	90.91
LMTRDA	90.54
RLSMDA	85.69
PBMDA	91.72
WBSMDA	81.85
DBMDA	91.29
HDMP	83.42
SMMDA	**94.07**

3.5. Case Studies

To further evaluate whether SMMDA could perform accurately and robustly, we select three complex Human diseases for case studies including colon neoplasms, breast neoplasms, and esophageal neoplasms. Specifically, the known miRNA-disease associations in HMDD v3.0 [19] are selected as the training samples, and candidate miRNAs for evaluated diseases are ranked in compliance with the predictive scores provided by SMMDA. It is important to note that we have deleted the associations that have been verified in the HMDD v3.0 database to ensure that the validation data set is not correlated with the data set already used for training. Finally, we confirmed the top 50 predicted miRNA-disease associations with the dbDEMC [41] and miR2Disease [42] databases.

Colon neoplasms are cancers that begin in the final part of the digestive tract (colon). It can occur at any age, but the incidence is higher in the elder people. Colon neoplasms usually start as non-cancerous (benign) small cell clumps, called polyps, which form inside the colon. Overtime, a few polyps will become colon cancer. Hence, doctors recommend

regular screening to identify and remove polyps before they become cancer, which can help prevent colon cancer. The SMMDA model was utilized to predict potential miRNA-esophageal-neoplasm associations. In the result, 47 of the top 50 predicted miRNAs are identified in the databases (see Table 5).

Table 5. Top 50 potential colon neoplasms-related miRNAs, 47 were confirmed by dbDEMC and miR2Disease databases.

miRNA	Evidence	miRNA	Evidence
hsa-mir-122	dbDemc	hsa-mir-451	dbDemc; miR2Disease
hsa-mir-146b	dbDemc	hsa-mir-494	dbDemc
hsa-mir-34c	miR2Disease	hsa-mir-10a	dbDemc; miR2Disease
hsa-mir-375	dbDemc	hsa-mir-320a	dbDemc
hsa-mir-9	dbDemc	hsa-mir-19b	dbDemc; miR2Disease
hsa-mir-16	miR2Disease	hsa-mir-139	dbDemc; miR2Disease
hsa-mir-206	dbDemc; miR2Disease	hsa-mir-491	dbDemc
hsa-mir-1	dbDemc; miR2Disease	hsa-mir-26b	dbDemc
hsa-mir-183	dbDemc; miR2Disease	hsa-mir-212	dbDemc
hsa-mir-182	dbDemc; miR2Disease	hsa-mir-193b	dbDemc
hsa-mir-214	dbDemc; miR2Disease	hsa-mir-338	dbDemc
hsa-mir-27b	dbDemc; miR2Disease	hsa-mir-199a-2	miR2Disease
hsa-mir-34b	miR2Disease	hsa-mir-20b	dbDemc; miR2Disease
hsa-mir-26a	miR2Disease	hsa-mir-497	dbDemc; miR2Disease
hsa-mir-199a	miR2Disease	hsa-mir-129	miR2Disease
hsa-mir-429	dbDemc	hsa-mir-130b	dbDemc; miR2Disease
hsa-mir-29c	dbDemc; miR2Disease	hsa-mir-135a	dbDemc
hsa-mir-96	dbDemc; miR2Disease	hsa-mir-328	dbDemc; miR2Disease
hsa-mir-99a	dbDemc; miR2Disease	hsa-mir-503	dbDemc; miR2Disease
hsa-mir-100	dbDemc	hsa-mir-372	dbDemc; miR2Disease
hsa-mir-144	dbDemc	hsa-mir-133a-1	dbDemc
hsa-mir-483	Unconfirmed	hsa-mir-449b	dbDemc
hsa-mir-7	dbDemc; miR2Disease	hsa-mir-29	Unconfirmed
hsa-let-7	Unconfirmed	hsa-mir-98	dbDemc; miR2Disease
hsa-mir-196a-2	dbDemc; miR2Disease	hsa-mir-342	dbDemc; miR2Disease

Breast neoplasms are cancers that occur in the breast cells. It is the most common cancer diagnosed in women in the United States, second only to skin cancer [43–45]. Breast neoplasms can occur in both men and women, but are much more severe in women. In recent years, the survival rates of breast neoplasms have increased largely due to factors such as a better understanding of the disease and earlier detection. In this article, SMMDA was utilized to predict potential miRNA-breast neoplasms associations. Finally, 48 of the top 50 predicted miRNAs are identified in the databases (see Table 6).

Table 6. Top 50 potential breast neoplasms-related miRNAs, 48 were confirmed by dbDEMC and miR2Disease databases.

miRNA	Evidence	miRNA	Evidence
hsa-mir-95	dbDemc	hsa-mir-877	dbDemc
hsa-mir-99b	dbDemc; miR2Disease	hsa-mir-337	dbDemc
hsa-mir-190	dbDemc; miR2Disease	hsa-mir-138-1	miR2Disease
hsa-mir-217	dbDemc; miR2Disease	hsa-mir-650	dbDemc
hsa-mir-206	dbDemc; miR2Disease	hsa-mir-449b	dbDemc
hsa-mir-369	dbDemc	hsa-mir-550a	dbDemc
hsa-mir-19b-3p	dbDemc	hsa-mir-4717	Unconfirmed
hsa-mir-517a	dbDemc	hsa-mir-329	dbDemc
hsa-mir-422a	dbDemc	hsa-mir-639	dbDemc
hsa-mir-133	miR2Disease	hsa-mir-645	dbDemc
hsa-mir-4324	dbDemc	hsa-mir-1308	dbDemc
hsa-mir-378b	dbDemc	hsa-mir-572	dbDemc; miR2Disease
hsa-mir-431	dbDemc	hsa-mir-498	dbDemc; miR2Disease
hsa-mir-1908	dbDemc	hsa-mir-561	dbDemc; miR2Disease
hsa-mir-188	dbDemc	hsa-mir-1321	dbDemc
hsa-mir-658	dbDemc; miR2Disease	hsa-mir-154	dbDemc
hsa-mir-518e	dbDemc	hsa-mir-1825	dbDemc
hsa-mir-636	dbDemc	hsa-mir-504	dbDemc

Table 6. *Cont.*

miRNA	Evidence	miRNA	Evidence
hsa-mir-362	miR2Disease	hsa-mir-147b	dbDemc
hsa-mir-487b	dbDemc	hsa-mir-454	dbDemc
hsa-mir-501	dbDemc; miR2Disease	hsa-mir-208	dbDemc; miR2Disease
hsa-mir-665	dbDemc	hsa-mir-208b	dbDemc
hsa-mir-432	dbDemc	hsa-mir-1236	dbDemc
hsa-mir-30	Unconfirmed	hsa-mir-323	dbDemc
hsa-mir-511	dbDemc; miR2Disease	hsa-mir-186	dbDemc; miR2Disease

Esophageal Neoplasms are a serious digestive disease with a high death rate [46–48]. It is the sixth most common cause of cancer death worldwide. The incidence of it varies from place to place. In some areas, the higher incidence of esophageal neoplasms may be due to smoking and alcohol consumption or special nutritional habits and obesity [49,50]. In this article, SMMDA was utilized to predict potential miRNA-esophageal neoplasms associations. Finally, 48 of the top 50 predicted miRNAs are identified in the databases (see Table 7).

Table 7. Top 50 potential esophageal neoplasms-related miRNAs, 48 were confirmed by dbDEMC and miR2Disease databases.

miRNA	Evidence	miRNA	Evidence
hsa-mir-132	dbDemc	hsa-mir-195	dbDemc
hsa-mir-199a	dbDemc	hsa-mir-339	dbDemc
hsa-mir-29a	dbDemc	hsa-mir-18b	dbDemc
hsa-mir-19b	dbDemc	hsa-mir-101	dbDemc
hsa-mir-23b	dbDemc	hsa-mir-146b	dbDemc
hsa-mir-222	dbDemc	hsa-mir-196a	dbDemc; miR2Disease
hsa-mir-16	dbDemc	hsa-mir-103	dbDemc; miR2Disease
hsa-mir-29b	dbDemc	hsa-mir-215	dbDemc
hsa-mir-429	dbDemc	hsa-mir-224	dbDemc
hsa-mir-182	dbDemc	hsa-mir-137	Unconfirmed
hsa-mir-125a	dbDemc	hsa-mir-24	dbDemc
hsa-mir-181b	dbDemc	hsa-mir-335	dbDemc
hsa-mir-499	dbDemc	hsa-mir-144	dbDemc
hsa-mir-7	dbDemc	hsa-mir-15b	dbDemc
hsa-let-7i	dbDemc	hsa-mir-497	dbDemc
hsa-mir-133a	dbDemc	hsa-mir-106a	dbDemc
hsa-mir-20b	dbDemc	hsa-mir-26a	dbDemc
hsa-mir-221	dbDemc	hsa-mir-218	dbDemc
hsa-mir-204	dbDemc	hsa-let-7f	dbDemc
hsa-mir-181a	dbDemc	hsa-mir-139	dbDemc
hsa-mir-302c	Unconfirmed	hsa-mir-124	dbDemc
hsa-mir-378	dbDemc	hsa-mir-206	Unconfirmed
hsa-mir-1	dbDemc	hsa-mir-372	dbDemc
hsa-mir-18a	dbDemc	hsa-mir-23a	Unconfirmed
hsa-mir-199b	dbDemc	hsa-mir-10a	dbDemc

4. Conclusions

Recently, machine-learning approaches have been widely investigated in the field of bioinformatics including the prediction of potential associations between miRNAs and diseases. In this work, considering the limited accessibility, high time consumption and high cost in traditional biological researches, we presented a novel computational method called SMMDA by incorporating multiple similarity profiles and a novel disease representation to accelerate the identification of potential miRNA-disease associations. The multiple similarity profiles of miRNAs and diseases and a novel disease representative feature were incorporating, thereby enhancing predictive accuracy. The deep learning is used for high-quality extraction of integrated features and gradient boosting method is used for fast and highly accurate training and prediction. Compared with previous related works, the experiment results have proved that the superior performance of SMMDA. The comparison experiment of different classifiers and different feature descriptors further proved that the

good predictive performance of SMMDA. In addition, the results of case studies with three Human diseases, including breast neoplasms, colon neoplasms, and esophageal neoplasms also demonstrated the feasibility of SMMDA in practical applications. Consequently, SMMDA was intended to be useful for the prediction of associations between miRNAs and diseases, and to be effective for prevention, diagnosis, treatment and prognosis of Human diseases.

Author Contributions: B.-Y.J., J.-R.Z. and L.-R.P. conceived the experiment, prepared the data set and wrote the manuscript. Z.-H.Y. and S.-L.P. performed and analyzed the experiment and checked the manuscript. All the authors approved the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Key R\&D Program of China 2017YFB0202602, 2018YFC0910405, 2017YFC1311003, 2016YFC1302500, 2016YFB0200400, 2017YFB0202104; NSFC Grants U19A2067, 61772543, U1435222, 61625202, 61272056, 62102427, 61762031; Science Foundation for Distinguished Young Scholars of Hunan Province (2020JJ2009); Science Foundation of Changsha kq2004010; JZ20195242029, JH20199142034, Z202069420652; The Funds of Peng Cheng Lab, State Key La-boratory of Chemo/Biosensing and Chemometrics; the Fundamental Research Funds for the Central Universities, and Guangdong Provincial Department of Science and Technology under grant No. 2016B090918122.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets analyzed during the current study are available from the cor-responding author on reasonable request.

Acknowledgments: The authors would like to thank all anonymous reviewers for their constructive advice.

Conflicts of Interest: The authors declare that they have no competing interests.

References

1. Ambros, V. The functions of animal microRNAs. *Nature* **2004**, *431*, 350–355. [CrossRef] [PubMed]
2. Bartel, D.P. MicroRNAs: Target recognition and regulatory functions. *Cell* **2009**, *136*, 215–233. [CrossRef] [PubMed]
3. Cheng, A.M.; Byrom, M.W.; Shelton, J.; Ford, L.P. Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis. *Nucleic Acids Res.* **2005**, *33*, 1290–1297. [CrossRef] [PubMed]
4. Griffiths-Jones, S. miRBase: microRNA sequences and annotation. *Curr. Protoc. Bioinform.* **2010**, *29*, 12.9.1–12.9.10. [CrossRef]
5. Karp, X.; Ambros, V. Encountering microRNAs in cell fate signaling. *Science* **2005**, *310*, 1288–1289. [CrossRef]
6. Xu, P.; Guo, M.; Hay, B.A. MicroRNAs and the regulation of cell death. *TRENDS Genet.* **2004**, *20*, 617–624. [CrossRef]
7. Alshalalfa, M.; Alhajj, R. Using context-specific effect of miRNAs to identify functional associations between miRNAs and gene signatures. *BMC Bioinform.* **2013**, *14*, S1. [CrossRef]
8. Mathur, P.; Rani, V. MicroRNAs: A critical regulator and a promising therapeutic and diagnostic molecule for diabetic cardiomyopathy. *Curr. Gene Ther.* **2021**, *21*, 313–326. [CrossRef]
9. Wang, R.; Tian, S.; Wang, H.-B.; Chu, D.-P.; Cao, J.-L.; Xia, H.-F.; Ma, X. MiR-185 is involved in human breast carcinogenesis by targeting Vegfa. *FEBS Lett.* **2014**, *588*, 4438–4447. [CrossRef]
10. Ji, B.-Y.; You, Z.-H.; Cheng, L.; Zhou, J.-R.; Alghazzawi, D.; Li, L.-P. Predicting miRNA-disease association from heterogeneous information network with GraRep embedding model. *Sci. Rep.* **2020**, *10*, 6658. [CrossRef]
11. Guo, Z.-H.; You, Z.-H.; Wang, Y.-B.; Huang, D.-S.; Yi, H.-C.; Chen, Z.-H. Bioentity2vec: Attribute-and behavior-driven representation for predicting multi-type relationships between bioentities. *GigaScience* **2020**, *9*, giaa032. [CrossRef] [PubMed]
12. Guo, Z.-H.; You, Z.-H.; Huang, D.-S.; Yi, H.-C.; Chen, Z.-H.; Wang, Y.-B. A learning based framework for diverse biomolecule relationship prediction in molecular association network. *Commun. Biol.* **2020**, *3*, 118. [CrossRef] [PubMed]
13. Chen, X.; Zhang, D.-H.; You, Z.-H. A heterogeneous label propagation approach to explore the potential associations between miRNA and disease. *J. Transl. Med.* **2018**, *16*, 348. [CrossRef] [PubMed]
14. Chen, X.; Xie, D.; Wang, L.; Zhao, Q.; You, Z.-H.; Liu, H. BNPMDA: Bipartite Network Projection for MiRNA–Disease Association prediction. *Bioinformatics* **2018**, *34*, 3178–3186. [CrossRef] [PubMed]
15. Ji, B.-Y.; You, Z.-H.; Wang, L.; Wong, L.; Su, X.-R.; Zhao, B.-W. Predicting miRNA-Disease Associations via a New MeSH Headings Representation of Diseases and eXtreme Gradient Boosting. In Proceedings of the International Conference on Intelligent Computing, Shenzhen, China, 12–15 August 2021; pp. 49–56.
16. Ji, B.-Y.; You, Z.-H.; Wang, Y.; Li, Z.-W.; Wong, L. DANE-MDA: Predicting microRNA-disease associations via deep attributed network embedding. *Iscience* **2021**, *24*, 102455. [CrossRef]

17. Xu, J.; Li, C.-X.; Lv, J.-Y.; Li, Y.-S.; Xiao, Y.; Shao, T.-T.; Huo, X.; Li, X.; Zou, Y.; Han, Q.-L. Prioritizing candidate disease miRNAs by topological features in the miRNA target–dysregulated network: Case study of prostate cancer. *Mol. Cancer Ther.* **2011**, *10*, 1857–1866. [CrossRef] [PubMed]
18. Zhang, L.; Chen, X.; Yin, J. Prediction of potential mirna–disease associations through a novel unsupervised deep learning framework with variational autoencoder. *Cells* **2019**, *8*, 1040. [CrossRef]
19. 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]
20. 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]
21. van Laarhoven, T.; Nabuurs, S.B.; Marchiori, E. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* **2011**, *27*, 3036–3043. [CrossRef]
22. Chen, X.; Clarence Yan, C.; Luo, C.; Ji, W.; Zhang, Y.; Dai, Q. Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. *Sci. Rep.* **2015**, *5*, 11338. [CrossRef] [PubMed]
23. Guo, Z.-H.; You, Z.-H.; Huang, D.-S.; Yi, H.-C.; Zheng, K.; Chen, Z.-H.; Wang, Y.-B. MeSHHeading2vec: A new method for representing MeSH headings as vectors based on graph embedding algorithm. *Brief. Bioinform.* **2020**, *22*, 2085–2095. [CrossRef] [PubMed]
24. Perozzi, B.; Al-Rfou, R.; Skiena, S. Deepwalk: Online learning of social representations. In Proceedings of the 20th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 27 June 2014; pp. 701–710.
25. Tang, J.; Qu, M.; Wang, M.; Zhang, M.; Yan, J.; Mei, Q. Line: Large-scale information network embedding. In Proceedings of the 24th International Conference on world Wide Web, Florence, Italy, 18–22 May 2015; pp. 1067–1077.
26. Wang, D.; Cui, P.; Zhu, W. Structural deep network embedding. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1225–1234.
27. Ou, M.; Cui, P.; Pei, J.; Zhang, Z.; Zhu, W. Asymmetric transitivity preserving graph embedding. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1105–1114.
28. Belkin, M.; Niyogi, P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In Proceedings of the Advances in Neural Information processing Systems, Vancouver, BC, Canada, 3 January 2001; pp. 585–591.
29. Lange, S.; Riedmiller, M. Deep auto-encoder neural networks in reinforcement learning. In Proceedings of the 2010 International Joint Conference on Neural Networks (IJCNN), Barcelona, Spain, 18–23 July 2010; pp. 1–8.
30. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
31. Friedl, M.A.; Brodley, C.E. Decision tree classification of land cover from remotely sensed data. *Remote Sens. Environ.* **1997**, *61*, 399–409. [CrossRef]
32. Hosmer, D.W., Jr.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*; John Wiley & Sons: Hoboken, NJ, USA, 2013; Volume 398.
33. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J.C.; Sheridan, R.P.; Feuston, B.P. Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958. [CrossRef]
34. 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]
35. Wang, L.; You, Z.-H.; Chen, X.; Li, Y.-M.; Dong, Y.-N.; Li, L.-P.; Zheng, K. LMTRDA: Using logistic model tree to predict MiRNA-disease associations by fusing multi-source information of sequences and similarities. *PLoS Comput. Biol.* **2019**, *15*, e1006865. [CrossRef]
36. Zheng, K.; You, Z.-H.; Wang, L.; Zhou, Y.; Li, L.-P.; Li, Z.-W. Dbmda: A unified embedding for sequence-based mirna similarity measure with applications to predict and validate mirna-disease associations. *Mol. Ther. -Nucleic Acids* **2020**, *19*, 602–611. [CrossRef]
37. 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]
38. You, Z.-H.; Huang, Z.-A.; Zhu, Z.; Yan, G.-Y.; Li, Z.-W.; Wen, Z.; Chen, X. PBMDA: A novel and effective path-based computational model for miRNA-disease association prediction. *PLoS Comput. Biol.* **2017**, *13*, e1005455. [CrossRef]
39. Xuan, P.; Han, K.; Guo, M.; Guo, Y.; Li, J.; Ding, J.; Liu, Y.; Dai, Q.; Li, J.; Teng, Z. Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors. *PLoS ONE* **2013**, *8*, e70204. [CrossRef] [PubMed]
40. Chen, X.; Yan, G.-Y. Semi-supervised learning for potential human microRNA-disease associations inference. *Sci. Rep.* **2014**, *4*, 5501. [CrossRef] [PubMed]
41. Yang, Z.; Ren, F.; Liu, C.; He, S.; Sun, G.; Gao, Q.; Yao, L.; Zhang, Y.; Miao, R.; Cao, Y. dbDEMC: A database of differentially expressed miRNAs in human cancers. In *BMC Genomics*; BioMed Central: London, UK, 2010; p. S5.
42. Jiang, Q.; Wang, Y.; Hao, Y.; Juan, L.; Teng, M.; Zhang, X.; Li, M.; Wang, G.; Liu, Y. miR2Disease: A manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.* **2008**, *37*, D98–D104. [CrossRef] [PubMed]
43. Kelsey, J.L.; Horn-Ross, P.L. Breast cancer: Magnitude of the problem and descriptive epidemiology. *Epidemiol. Rev.* **1993**, *15*, 7. [CrossRef] [PubMed]
44. Tao, Z.; Shi, A.; Lu, C.; Song, T.; Zhang, Z.; Zhao, J. Breast cancer: Epidemiology and etiology. *Cell Biochem. Biophys.* **2015**, *72*, 333–338. [CrossRef] [PubMed]

45. Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA A Cancer J. Clin.* **2018**, *68*, 394–424. [CrossRef]
46. Kano, M.; Seki, N.; Kikkawa, N.; Fujimura, L.; Hoshino, I.; Akutsu, Y.; Chiyomaru, T.; Enokida, H.; Nakagawa, M.; Matsubara, H. miR-145, miR-133a and miR-133b: Tumor-suppressive miRNAs target FSCN1 in esophageal squamous cell carcinoma. *Int. J. Cancer* **2010**, *127*, 2804–2814. [CrossRef]
47. He, B.; Yin, B.; Wang, B.; Xia, Z.; Chen, C.; Tang, J. MicroRNAs in esophageal cancer. *Mol. Med. Rep.* **2012**, *6*, 459–465.
48. Dragovich, T.; Campen, C. Anti-EGFR-targeted therapy for esophageal and gastric cancers: An evolving concept. *J. Oncol.* **2009**, *2009*, 804108. [CrossRef]
49. Xie, Z.; Chen, G.; Zhang, X.; Li, D.; Huang, J.; Yang, C.; Zhang, P.; Qin, Y.; Duan, Y.; Gong, B. Salivary microRNAs as promising biomarkers for detection of esophageal cancer. *PLoS ONE* **2013**, *8*, e57502. [CrossRef]
50. Wan, J.; Wu, W.; Che, Y.; Kang, N.; Zhang, R. Insights into the potential use of microRNAs as a novel class of biomarkers in esophageal cancer. *Dis. Esophagus* **2016**, *29*, 412–420. [CrossRef] [PubMed]