INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research

International Journal of Data Mining and Bioinformatics

ISSN online: 1748-5681 - ISSN print: 1748-5673
https://www.inderscience.com/ijdmb

Identification of disease-related miRNAs based on weighted k-nearest known neighbours and inductive matrix completion

Ahmet Toprak

DOI: 10.1504/IJDMB.2023.10058135

Article History:

Received:	31 January 2023
Last revised:	24 March 2023
Accepted:	19 June 2023
Published online:	17 October 2023

Identification of disease-related miRNAs based on weighted k-nearest known neighbours and inductive matrix completion

Ahmet Toprak

Department of Electricity and Energy,
Selcuk University,
Konya, Turkey
Email: atoprak@selcuk.edu.tr

Abstract: miRNAs, a subtype of non-coding RNAs, have a length of about 18–22 nucleotides. Studies have shown that miRNAs play an important role in the initiation and progression of many human diseases. For this reason, it is very significant to know the miRNAs associated with diseases. Because experimental studies to identify these associations are expensive and time-consuming, many computational methods have been developed to identify disease-related miRNAs. In this study, we propose a calculation method based on nearest known neighbours and matrix completion. ROC curves of our suggested method were plotted using two commonly used cross-validation techniques such as five-fold and LOOCV, and also AUC values were calculated in both validation techniques. Moreover, we carried out case studies on breast cancer, lung cancer, and lymphoma to further demonstrate the predictive accuracy of our method. As a result, our proposed method can be used with confidence to identify possible miRNA-disease associations.

Keywords: ncRNA; miRNA; disease; cancer; miRNA-disease associations.

Reference to this paper should be made as follows: Toprak, A. (2023) 'Identification of disease-related miRNAs based on weighted k-nearest known neighbours and inductive matrix completion', *Int. J. Data Mining and Bioinformatics*, Vol. 27, No. 4, pp.231–251.

Biographical notes: Ahmet Toprak received his Bachelor's and Master's degrees in Electricity-Electronics Engineering from the Fırat University, and Selçuk University, Türkiye, respectively. He obtained his PhD degree from the Selçuk University, Türkiye, in Department of Nanotechnology and Advanced Materials. His research interest focuses on bioinformatics and computational biology. He is currently working as an Assistant Professor at Selçuk University, Türkiye.

1 Introduction

microRNAs (miRNAs) are a single-stranded and endogenous RNA molecule found in humans, animals, plants, and viruses, that regulates gene expression by targeting messenger RNAs (Ambros, 2001; Bartel, 2004). Human cells contain many miRNAs, a subtype of non-coding RNA approximately 18–22 nucleotides long (Toprak and Eryilmaz, 2021; Chandra et al., 2017). miRNAs are involved in numerous biological

processes that are vital to human life. Especially miRNAs synthesise proteins and can regulate many biological processes, including cell division, cell proliferation, and cell death (Wang et al., 2022). Although the first miRNA was discovered by Le et al. in Caenorhabditis elegans in 1993 (Lee et al., 1993), it was not identified as a different biological regulator class up to 2000s, but rather as junk RNA. However, in recent years, an intensive study has been carried out on miRNAs. It has also been confirmed by experimental studies that the dysregulation of miRNAs is associated with the developmental processes of many human diseases (Lynam-Lennon et al., 2009). Research on miRNA has clearly demonstrated that they were responsible for the initiation and development of many human diseases, especially many cancer types such as breast cancer, lung cancer and lymphoma. For instance, miRNA-196 and miRNA-10a, which play a role in the development of breast cancer and are responsible for the levels of malignant properties of cancer cells, are localised in homeobox clusters (Calin et al., 2004; Heneghan et al., 2009). miRNA-17 is under expressed or over expressed in breast tumour tissue and also is regulate BRCA1/2, ATM, PTEN, and CHEK2 (Shenouda and Alahari, 2009). However, it has been observed that CYP1B1 is overexpressed and CYP1B1 and miRNA-27b expression levels are inversely related in breast cancer (Shenouda and Alahari, 2009). Firstly, when the expression levels of miRNA-206 between normal and breast cancer tissues were compared with the miRNA microarray method, it was observed that it suppressed breast cancer (O'Day and Lal, 2010).

In the lung cancer, it has been determined that there is let-7 down regulated, and it has been reported that high expression levels of let-7 inhibit lung cancer (Calin et al., 2004; Johnson et al., 2005). miRNA-17-3p, miRNA-18, miRNA-19a, and miRNA-20 indicating overexpression of the miRNA-17-92 cluster in lung cancer cell tissues (Hayashita et al., 2005). In addition, amplification and overexpression of the miRRNA-17-92 miRNA cluster have been shown to play an important role in the development of lung cancer (Hayashita et al., 2005) and lymphoma (Ota et al., 2004; He et al., 2005).

In lymphoma, it has been revealed by studies that miRNA-155 is overexpressed (Mashima, 2015). In chronic lymphocytic leukemia, miRNA-15a and miRNA-16-1 targeting the BCL2 oncogene act as down-regulated tumour suppressors (Calin and Croce, 2006). In addition, studies have shown that the miRNA-17-92 cluster is amplified in lymphoma (Tagawa and Seto, 2005; Ota et al., 2004). Further research is needed to fully understand the role of miRNAs in lymphoma and how they can be targeted for therapeutic purposes. Table 1 shows some miRNA types with increased and decreased expression levels in breast cancer, lung cancer, and lymphoma.

Table 1 Expression levels of miRNAs

Disease	*Decreased (downregulated)*	*Increased (upregulated)*
Breast (Iorio et al., 2005; Mattie et al., 2006)	miRNA-10b, miRNA-125b, miRNA-145, miRNA-155, miRNA-17-5p, miRNA-27b	miRNA-21, miRNA-29b-2
Lung (Lowery et al., 2008; Lu et al., 2005; Volinia et al., 2006)	let-7 family	miRNA-17-5p, miRNA-17-92 cluster
Lymphoma (Lowery et al., 2008; Volinia et al., 2006)	miRNA-15a	miRNA-10a, miRNA-155, miRNA-17-92 cluster

miRNAs are small non-coding RNAs that play a crucial role of gene expression regulation. It has been seen in studies that miRNAs are responsible for the initiation and progression of various cancer types, such as breast cancer, lung cancer, and lymphoma. Because knowledge of miRNA-disease relationships is so important, many databases such as HMDD (Li et al., 2014), dbDEMC (Yang et al., 2010), miR2Disease (Jiang et al., 2009), deepBase (Yang et al., 2009), miRBase (Kozomara and Griffiths-Jones, 2013), and miRGen (Alexiou et al., 2009) have been created to help researchers examine miRNA-disease relationships.

Due to the high cost and time required for biological experimental studies to identify disease-associated miRNAs, researchers have developed computational models to predict and analyse the relationships between miRNAs and diseases. In recent years, scientists have developed new computational techniques such as MCMDA (Li et al., 2017), RKNNMDA (Chen et al., 2017), IMCMDA (Liu et al., 2021a), GRMDA (Chen et al., 2018e), MDHGI (Chen et al., 2018f), SACMDA (Li et al., 2021), SPLPMDA (Toprak, 2022), LRSSLMDA (Chen and Huang, 2017), MaxFlow (Yu et al., 2017), BNPMDA (Chen et al., 2018d), NDAMDA (Chen et al., 2018c), EGBMMDA (Chen et al., 2018a), NSEMDA (Wang et al., 2019), BRMDA (Zhu et al., 2021), SMALF (Liu et al., 2021b), and DFELMDA (Liu et al., 2022) to predict disease-related miRNAs. In addition, Toprak and Eryilmaz Dogan (2021) obtained successful results in their proposed model for miRNA-disease association prediction. This model uses a combination of similarity-based algorithms and machine learning-based algorithms to identify potential miRNA biomarkers for cancer diagnosis and treatment.

The steps used in this study to predict potential miRNA-disease associations can be described as follows.

1 Functional similarity matrix of miRNA and disease semantic similarity matrix have been generated.

2 Gaussian Interaction Profile (GIP) kernel similarity matrices were formed for miRNAs and diseases. Then, these all-miRNA similarity matrices and all disease similarity matrices were integrated to form the new miRNA matrix and new disease matrix.

3 The obtained miRNA and disease integrated similarity matrices were used to set a new adjacency matrix by weighted k-nearest known neighbours (WKNKN) method.

4 Then, inductive matrix completion (IMC) method was used to estimate potential miRNA-disease relationships.

The IMC is a graph-based method that uses a partially observed matrix and a graph to complete the missing data. The effectiveness of our proposed model was evaluated using cross validation techniques like five-fold and leave-one-out, and case studies were conducted to further confirm the methodology. The results of the experimental tests showed that the proposed method is efficient in forecasting possible disease related miRNAs.

2 Materials and methods

2.1 Data set of known diseases related miRNAs

The known human miRNA-disease relations data that we used in this study were obtained from the HMDD database (Li et al., 2014). This dataset includes 5,430 confirmed miRNA-disease associations with 495 miRNAs (denoted by set $m = \{m_1, m_2, \ldots, m_{495}\}$ and 383 diseases (denoted by set $d = \{d_1, d_2, \ldots, d_{383}\}$). Adjacency matrix was created by obtaining 1 if an association between miRNA and disease is confirmed, and 0 if there is no relationship. The mathematical expression can be given as follows.

$$\begin{cases} A\big(m(i), d(j)\big) = 1 \text{ miRNA } m(i) \text{ has association with disease } d(j) \\ A\big(m(i), d(j)\big) == 0 \text{ miRNA } m(i) \text{ has no association with disease } d(j)) \end{cases} \quad (1)$$

Consequently, a miRNA-disease matrix was obtained, called adjacency matrix A, consisting of 495 miRNAs and 383 diseases.

2.2 Data set of miRNA functional similarity

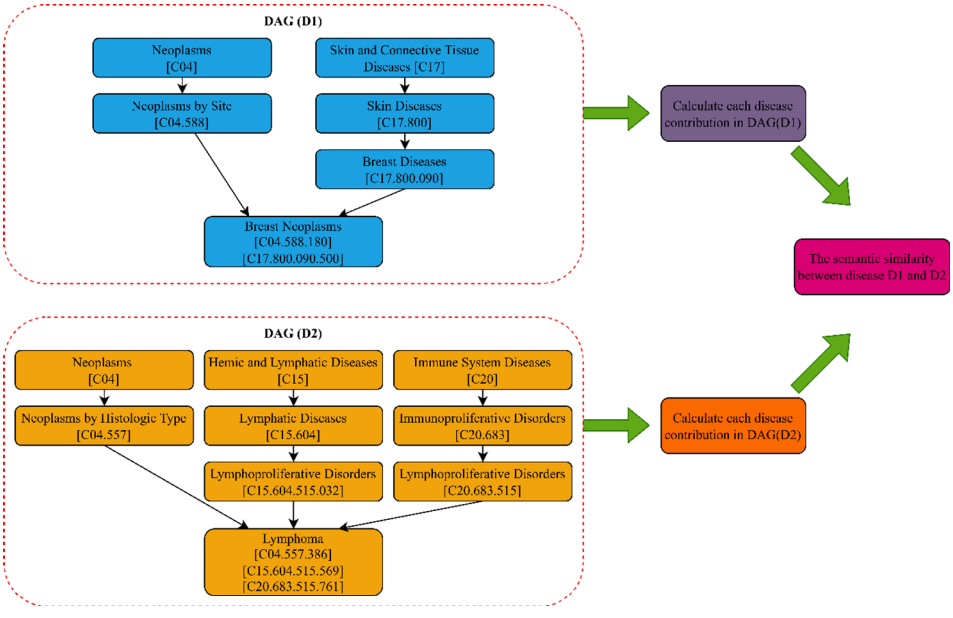
miRNA functional similarity refers to the degree to which two different miRNAs have similar or overlapping targets and biological functions. This can be determined through various computational and experimental methods, such as analysing the sequence complementarity between miRNAs and their targets, comparing their expression patterns in different tissues or under different conditions, and examining the phenotypic consequences of their loss or overexpression. Understanding miRNA functional similarity can be important for predicting the potential consequences of perturbing miRNA expression in different contexts, and for identifying miRNA-based therapies for various diseases.

In 2010, Wang et al. proposed "a method for calculating functional similarity scores of miRNAs, based on the general assumption that miRNAs tend to be associated with similar diseases and vice versa" (Wang et al., 2010). The functional similarities value they have calculated can be obtained from the web site (http://www.cuilab.cn/files/images/cuilab/misim.zip). From these data, we created an FS matrix of 495 × 495 dimensions to represent the functional similarities of miRNAs. Where, $FS(m(i), m(j))$ denotes the functional similarity score between miRNA $m(i)$ and miRNA $m(j)$.

2.3 Disease semantic similarity model 1

Directed acyclic graph (DAG) structure is a type of graphical representation that can be used to capture the relationships between different diseases based on their semantic similarity. In a DAG structure, while each node represents a disease, the edges between nodes show how semantically similar one disease is to another. DAG structures of 'breast neoplasms' and 'lymphoma; are shown in Figure 1.

Figure 1 An example of directed acyclic graph structure (see online version for colours)



The direction of the edges reflects the directionality of the relationships between diseases, with edges pointing from more general or broad concepts to more specific or narrow concepts. One advantage of using a DAG structure to represent disease semantic similarity is that it allows for the representation of hierarchical relationships between diseases, where some diseases are more specific subtypes or subcategories of more general diseases. This can be useful for organising and navigating large datasets of diseases and for making inferences about the relationships between different diseases.

DAG structure of each disease was created by using medical subject headings (MeSH) definitions obtained from the National Library of Medicine (http://www.nlm.nih.gov/) web page. In a DAG, the vertices (also called nodes) represent the entities, and the edges represent the relationships between the entities. For example, the DAG structure of disease D includes D itself, ancestor nodes of D, and all nodes from parent to child nodes. The contribution score of disease t in $DAG(D)$ to disease D can be calculated with the following equation.

$$\begin{cases} D_D(t) = 1, & \text{if } t = D \\ D_D(t) = \max 0.5 * D_D(t') \middle| t' \in \text{children of } t, & \text{if } t \neq D \end{cases} \tag{2}$$

Using the following equation, each disease's semantic value is calculated.

$$V(D) = \sum_{t \in T_D} D_D(t) \tag{3}$$

After calculating the semantic scores for each disease, the semantic similarity values between the diseases are calculated as follows:

$$SS1(D_1, D_2) = \frac{\sum_{t \in T_{D_1 \cap D_2}} (D_1(t) + D_2(t))}{V(D_1) + V(D_2)} \tag{4}$$

These calculations are made for all diseases and as a result, a semantic similarity matrix SS1 with 383 × 383 dimension is obtained. $SS1(d_i, d_j)$ represent the semantic similarity value between disease d_i and disease d_j.

2.4 Disease semantic similarity model 2

In SS1, the contribution values of the diseases in the same layer of $DAG(D)$ structure are the same. However, according to the second assumption, diseases that are less appear in the same layer of DAG structures than other diseases contribute more to disease D. Thus, in SS2, the contribution score of disease t in $DAG(D)$ to disease D can be calculated with equation (5).

$$D_D(t) = -\log\left[\frac{\text{number of DAGs including } t}{\text{number of disease}}\right] \tag{5}$$

With the following equation, each disease's semantic value is calculated.

$$V(D) = \sum_{t \in T_D} D_D(t) \tag{6}$$

Using the following equation, the semantic similarity score between two diseases is calculated.

$$SS2(D_1, D_2) = \frac{\sum_{t \in T_{D_1 \cap D_2}} (D_1(t) + D_2(t))}{V(D_1) + V(D_2)} \tag{7}$$

2.5 Calculation of Gaussian similarity

The Gaussian interaction profile kernel (GIP), has been proposed by Van Laarhoven et al. (2011), is a method that can be used to measure the similarity between two sets of interactions, such as the interactions between miRNA molecules and their target genes or the interactions between proteins and small molecules. The GIP kernel similarity measure is based on the idea that the similarity between two sets of interactions can be determined by measuring the overlap between the interactions and by considering the strength of the interactions. The GIP kernel similarity measure has been applied to various biological systems, including the analysis of miRNA-disease associations. In this context, the GIP kernel similarity measure can be used to identify diseases that are associated with similar sets of miRNA molecules. This can be useful for understanding the role of miRNA in the development and progression of diseases and for identifying potential therapeutic targets for the treatment of diseases in which miRNA expression is dysregulated.

In this section, GIP kernel similarities were calculated for both miRNAs and diseases. The GIP kernel similarity scores of diseases indicated by $GM(m_i, m_j)$ are calculated as follows:

$$GM(m_i, m_j) = \exp\left(-\gamma_m \left\| IP(m_i) - IP(m_j) \right\|^2\right) \tag{8}$$

$$\gamma_m = \delta_m \Big/ \left(\frac{1}{n_m} \sum_{i=1}^{n_m} \left\| IP(m_i) \right\|^2 \right) \tag{9}$$

Here, the parameter γ_m obtained by normalising the parameters controls the kernel bandwidth, and the number of miRNAs is indicated by parameter m.

Similarly, the GIP kernel similarity for diseases represented by $GD(d_i, d_j)$ is calculated as follows.

$$GD\left(d_i, d_j\right) = \exp\left(-\gamma_d \left\| IP(d_i) - IP(d_j) \right\|^2\right) \tag{10}$$

Using the equation below, the parameter γ_d that controls the core bandwidth, can be computed.

$$\gamma_d = \delta_d \Big/ \left(\frac{1}{n_d} \sum_{i=1}^{n_d} \left\| IP(d_i) \right\|^2 \right) \tag{11}$$

Here, n parameter is the number of diseases.

2.6 Integration of multisource data

We calculated the GIP kernel similarities of miRNAs and diseases, miRNA functional similarities (FS), and disease semantic similarities ($SS1$ and $SS2$), and integrated these similarities. The similarity ratios between miRNA $m(i)$ and miRNA $m(j)$ are calculated by following equation.

$$SM = \begin{cases} FS\left(m(i), m(j)\right) & m(i) \text{ and } m(j) \text{ has functional similarity} \\ GM\left((m(i), m(j)\right) & \text{otherwise} \end{cases} \tag{12}$$

Similarly, new disease similarity scores created by integrating can be calculated as follows.

$$SD = \begin{cases} \dfrac{SS1\left(d(i), d(j)\right) + SS2\left(d(i), d(j)\right)}{2} & d(i) \text{ and } d(j) \text{ has semantic similarity} \\ GD\left(d(i), d(j)\right) & \text{otherwise} \end{cases} \tag{13}$$

2.7 Weighted k-nearest known neighbours

WKNKN (Ezzat et al., 2017) is a classification algorithm that makes predictions based on the class labels of the K-nearest neighbours of a given instance. In the WKNKN algorithm, the weight assigned to each neighbour is based on its distance from the instance being classified. The farther away a neighbour is, the less weight it is given in the prediction.

To classify an instance using the WKNKN algorithm, you first need to specify the number of neighbours (K) to consider and the distance metric to use. Then, for each instance in the training set, you calculate its distance to the instance being classified using the chosen distance metric. You then sort the training instances by distance and select the k-nearest neighbours. The class label of the instance being classified is then predicted based on the class labels of these K-nearest neighbours, with the weights being applied

based on the distances of the neighbours. WKNKN is a simple and effective classification algorithm that is often used in pattern recognition and machine learning applications. It is often used as a baseline method for comparison with more complex algorithms.

Adjacency matrix A density is very sparse (2.86%), due to the 5430 experimentally confirmed relationships between 495 miRNAs and 383 diseases. Obviously, our aim here is to replace the unknown values in the adjacency matrix A with values between 0 and 1 with the WKNKN algorithm. WKNKN algorithm is explained in detail in Figure 2.

Figure 2 Wknkn algorithm

$$
\begin{aligned}
&function\ WKNKN(A, S_m, S_d, K, n)\\
&\\
&A_m = A_d = 0\\
&\\
&for\ m = 1\ to\ n:\\
&\qquad mnn = KNearestKnownNeighbors(m, S_m, K)\\
&\qquad for\ i > K:\\
&\qquad\qquad w_i = n^{i-1} * S_m(m, mnn_i)\\
&\qquad end\ for\\
&\qquad Z_m = \sum_{i=1}^{K} S_m(m, mnn_i)\\
&\qquad A_m(m) = \frac{1}{z_m}\sum_{i=1}^{K} w_i A(mnn_i)\\
&end\ for\\
&\\
&for\ d = 1\ to\ m:\\
&\qquad dnn = KNearestKnownNeighbors(d, S_d, K)\\
&\qquad for\ j > K:\\
&\qquad\qquad w_j = n^{j-1} * S_d(d, dnn_j)\\
&\qquad end\ for\\
&\qquad Z_d = \sum_{j=1}^{K} d(d, dnn_j)\\
&\qquad A_d(d) = \frac{1}{z_d}\sum_{j=1}^{K} w_j A(dnn_j)\\
&end\ for\\
&\\
&A_{md} = (A_m + A_d)/2\\
&A = max(A, A_{md})\\
&\\
&return\ A
\end{aligned}
$$

2.8 Inductive matrix completion

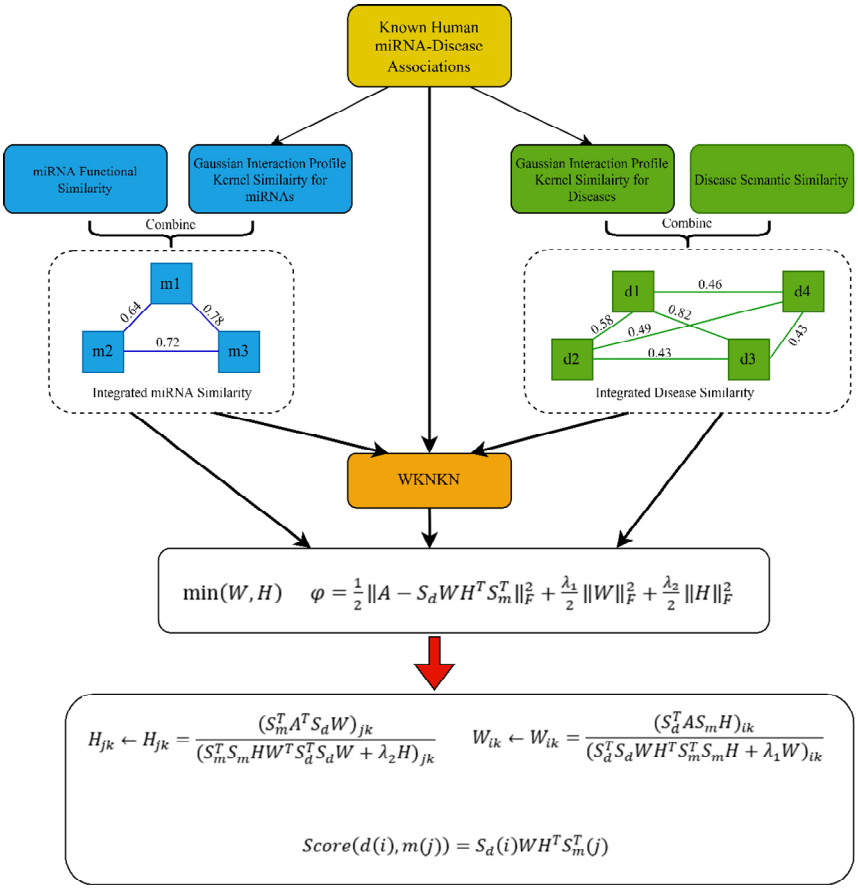
Here we used a new matrix completion-based model to predict potential relationships between miRNAs and diseases, called IMC (Chen et al., 2018b). The advantage of IMC is that it can solve matrix completion problems using small information set. In the IMC model, integrated miRNA similarity $S_m \in R^{nm \times nm}$), integrated disease similarity ($S_d \in R^{nd \times nd}$) and novel miRNA-disease associations ($A \in \mathrm{R}^{nd \times nm}$) calculated with WKNKN were used. Here, the miRNA $m(i)$ feature vector is represented by $S_m(i)$, and the disease $d(j)$ feature vector is represented by $S_d(j)$. Using IMC method, a new matrix $Z = WH^T$ ($Z \in R^{nd \times nm}$) is created, where $W \in R^{nd \times r}$ and $H \in R^{nm \times r}$. Here, the r value is a very small parameter that affects the convergence rate of the IMC algorithm. With the equation (14), W and H matrices can be calculated.

$$\min(W, H) \quad \varphi = \frac{1}{2}\left\|A - S_d W H^T S_m^T\right\|_F^2 + \frac{\lambda_1}{2}\|W\|_F^2 + \frac{\lambda_2}{2}\|H\|_F^2 \tag{14}$$

Here, W and H are set a random matrix, then updates W and H matrices iteratively, while the parameters λ_1 and λ_2 represent the regularisation parameters. The steps of the IMC algorithm are given in Figure 3. The forecasting score of disease-associated miRNAs can be calculated as following.

$$Score\left(m(i), d(j)\right) = S_m(i) W H^T S_d^T(j) \tag{15}$$

Figure 3 IMC model constructed to predict disease-related miRNAs (see online version for colours)



3 Results

3.1 Performance evaluation

To evaluate the performance of a predictive model, five-fold cross-validation and leave-one-out cross-validation (LOOCV) techniques are commonly used.

Cross-validation is a technique for assessing the performance of a machine learning model by dividing the data into a training set and a test set, training the model on the training set, and evaluating its performance on the test set. This process is typically repeated multiple times with different partitions of the data to obtain a more robust estimate of the model's performance. Although, there are several different types of cross-validation, five-fold cross-validation technique is the most common method of cross validation. In five -fold cross-validation, the data is randomly partitioned into five equal-sized folds. The model is trained on four folds and evaluated on the remaining fold, and this process is repeated five times, with each fold serving as the evaluation set once. five-fold cross-validation is a good choice for many applications because it strikes a balance between bias (due to the limited number of folds) and variance (due to the relatively large number of training instances used in each fold). It is also relatively quick to compute, compared to more thorough cross-validation methods such as leave-one-out cross-validation.

LOOCV is a resampling procedure used to evaluate the performance of a machine learning model. It is a type of cross-validation, where the model is trained on all but one sample of the data and then evaluated on the left-out sample. This process is repeated until every sample in the data has been used as the evaluation set exactly once. The performance measure is then averaged over all of the evaluations. LOOCV is a very thorough cross-validation method because it uses the maximum amount of data for training and the minimum amount of data for evaluation. However, it can be computationally expensive because it requires training the model n times, where n is the size of the data. As a result, it is generally not used for large datasets. It is more commonly used for small datasets or when the goal is to obtain very accurate estimates of model performance.

The relationship between the positive rate (TPR) and the false positive rate (FPR) gives us the receiver operating characteristic (ROC) curve, where the TPR is the proportion of positive samples that are correctly classified as positive and the FPR is the proportion of negative samples that are incorrectly classified as positive.

$$TPR = \frac{TP}{TP + FN} \tag{16}$$

$$FPR = \frac{TN}{TN + FP} \tag{17}$$

where true positive (*TP*), true negative (*TN*), false positive (*FP*), and false negative (*FN*) are an example that is correctly predicted as positive, correctly predicted as negative, incorrectly predicted as positive, and incorrectly predicted as negative, respectively. These terms are often used in evaluating the performance of a classification model using metrics such as precision, recall, and accuracy.

The area under the curve (AUC) of the ROC curve is a measure of the overall performance of the method, and value ranges from 0 to 1. If AUC value is 0.5 it demonstrates an accidental outcome, if 1 it demonstrates excellent classifier. The ROC curve of our proposed method in the five-fold cross validation technique is shown in Figure 4, and the ROC curve in the LOOCV technique is shown in Figure 5. The calculated AUC values in five-fold cross validation and LOOCV are 0.8745 and 0.8786, respectively.

Figure 4 Five-fold (see online version for colours)

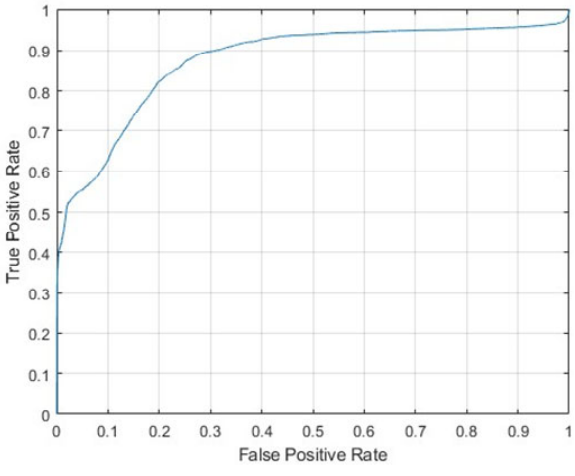
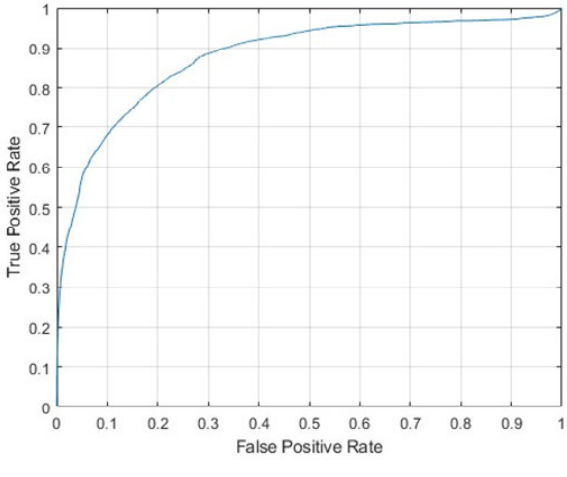
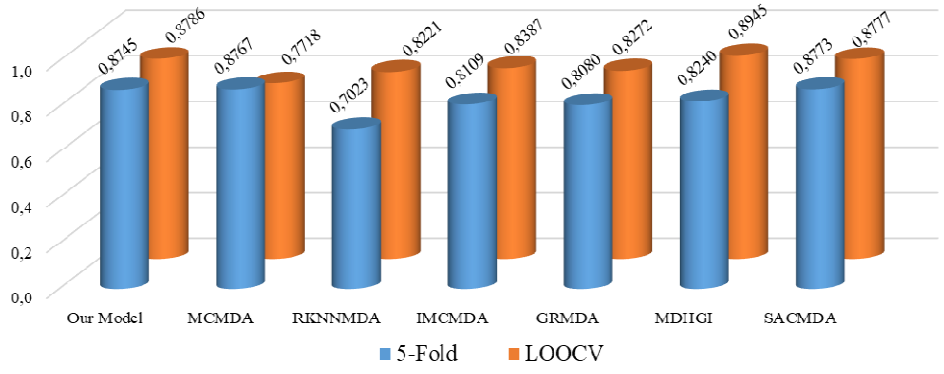


Figure 5 LOOCV (see online version for colours)



We compared our study with six methods such as MCMDA (Li et al., 2017), RKNNMDA (Chen et al., 2017), IMCMDA (Liu et al., 2021a), GRMDA (Chen et al., 2018e), MDHGI (Chen et al., 2018f), and SACMDA (Li et al., 2021). The AUC values calculated in 5-fold CV method of the six methods we compared were 0.8767, 0.7023, 0.8109, 0.8080, 0.8240, and 0.8773, respectively, and the AUC values calculated in the LOOCV technique were 0.7718, 0.8221, 0.8387, 0.8272, 0.8945, and 0.8777, respectively. When the results are examined in detail, the AUC value calculated in the five-fold cross-validation technique of our method is higher than the AUC value of the other methods except the AUC value of the MCMDA and SACMDA methods. The AUC value calculated in the LOOCV technique is more successful than the other methods except MDHGI.

Figure 6 Comparison (see online version for colours)



3.2 Case studies

After predicting the potential miRNA-disease associations, we conducted case studies on three diseases like breast neoplasm, lung neoplasm, and lymphoma that are common today and cause deaths. The forecasted results have been validated by several important databases containing proven miRNA-disease associations such as HMDD v2.0 (Li et al., 2014), dbDEMC (Yang et al., 2010), and miR2Disease (Jiang et al., 2009). The training set we used in this study contains 5430 associations between experimentally proven miRNA and disease and was retrieved from the database of HMDD v2.0. To use this obtained data as a training set, we removed associations of the related disease-associated miRNAs during the case study. A scoreboard of predicted miRNAs for breast cancer, lung cancer, and lymphoma was generated and then first 30 predicted miRNAs for the aforementioned diseases were validated in the databases.

The first case study was conducted on breast cancer. A breast neoplasm is a growth or tumour that develops in the breast tissue. It could be benign (non-cancerous) or malignant (cancerous). Some common types of breast neoplasms include benign tumours like fibroadenomas and malignant tumours like breast cancer. Breast cancer, which is also seen in men, is the most common type of cancer in women. Breast cancer is the leading cause of cancer-related death among women. Breast cancer occurs in more than 1.3 million people each year, and approximately 465.000 people who are diagnosed die (Garcia et al., 2007).

Lung neoplasms were chosen as the second case study. A lung neoplasm is a growth or tumour that develops in the lung tissue. It could be benign (non-cancerous) or malignant (cancerous). Some common types of lung neoplasms include benign tumours like hamartomas and malignant tumours like lung cancer. Although lung cancer ranks second in women in terms of mortality rate, it ranks first in men (Garcia et al., 2007). Early diagnosis is very important in the success of treatment of lung cancer.

Lymphoma was chosen for the final case study. Lymphoma is a type of cancer that affects the immune system. It occurs when lymphocytes, a type of white blood cell, grow and multiply abnormally. There are two main types of lymphoma: Hodgkin lymphoma and non-Hodgkin lymphoma. The incidence rate in developed countries is approximately 1.6 times higher than in developing countries. About 200.000 Lymphomas are diagnosed each year around the world, and more than half of the patients die after being diagnosed

(Garcia et al., 2007). Lymphoma can be treated with chemotherapy, radiation therapy, and/or targeted therapy, and the choice of treatment depends on the type and stage of the cancer.

Table 2 Breast neoplasms

hsa-mir-21	HMDD; dbDEMC; miR2disease
hsa-mir-145	HMDD; miR2disease
hsa-mir-17	HMDD; dbDEMC; miR2disease
hsa-mir-155	HMDD; miR2disease
hsa-mir-18a	HMDD; dbDEMC; miR2disease
hsa-mir-19b	HMDD; miR2disease
hsa-mir-20a	HMDD; miR2disease
hsa-let-7a	HMDD; dbDEMC; miR2disease
hsa-mir-19a	HMDD; dbDEMC; miR2disease
hsa-mir-125b	HMDD; dbDEMC; miR2disease
hsa-mir-34a	HMDD; miR2disease
hsa-let-7e	HMDD; miR2disease
hsa-mir-223	HMDD; miR2disease
hsa-mir-126	HMDD; miR2disease
hsa-let-7d	HMDD; miR2disease
hsa-let-7c	HMDD; dbDEMC; miR2disease
hsa-let-7b	HMDD; dbDEMC; miR2disease
hsa-mir-143	HMDD; miR2disease
hsa-mir-92a	HMDD; dbDEMC; miR2disease
hsa-let-7f	HMDD
hsa-let-7i	HMDD; miR2disease
hsa-mir-132	HMDD; miR2disease
hsa-mir-125a	HMDD; miR2disease
hsa-mir-200b	HMDD; dbDEMC; miR2disease
hsa-mir-146a	HMDD; miR2disease
hsa-mir-199a	HMDD; dbDEMC
hsa-mir-141	HMDD; dbDEMC; miR2disease
hsa-mir-106a	HMDD; miR2disease
hsa-mir-221	HMDD; miR2disease
hsa-mir-101	HMDD; dbDEMC; miR2disease

In conclusion, all of the top 30 potential miRNAs associated with breast cancer, lung cancer, and lymphoma were validated with the HMDD, dbDEMC, and miR2Disease databases, as seen in Table 2, Table 3, and Table 4. For example, in experimental studies, it has been revealed that miRNA-21 and miRNA-155 can be used as a promising potential biomarker for early diagnosis of breast cancer (Li et al., 2016; Cheng et al., 2016; Lv et al., 2013). In addition, miRNA-145 inhibits breast cancer cell growth by targeting RTKN (Wang et al., 2009). In lung cancer, it was observed that miRNA-17 was

overexpressed (Lum et al., 2007) and miRNA-17 increased cell proliferation (Hayashita et al., 2005). Moreover, miRNA-21, miRNA-146b, and miRNA-155 have been found to be consistently up-regulated in lung cancer patients (Melkamu et al., 2010; Munagala et al., 2016). In lymphoma, it has been in many studies that miRNA-17, miRNA-18a, miRNA-19a, miRNA-19b and miRNA-20a are highly expressed (Zhang et al., 2009). In addition, downregulation of miRNA-145 has been observed in both breast cancer (Dalmay and Edwards, 2006), lung cancer (Dalmay and Edwards, 2006) and adult T-cell leukemia/lymphoma (Xia et al., 2014).

Table 3 Lung neoplasms

hsa-mir-21	HMDD; miR2disease
hsa-mir-155	HMDD; miR2disease
hsa-mir-145	HMDD; miR2disease
hsa-mir-17	HMDD; dbDEMC; miR2disease
hsa-mir-19b	HMDD; miR2disease
hsa-mir-20a	HMDD; dbDEMC; miR2disease
hsa-mir-18a	HMDD; miR2disease
hsa-let-7a	HMDD; miR2disease
hsa-mir-19a	HMDD; miR2disease
hsa-mir-125b	HMDD; miR2disease
hsa-mir-126	HMDD
hsa-mir-34a	HMDD; miR2disease
hsa-let-7e	HMDD; dbDEMC; miR2disease
hsa-mir-143	HMDD; miR2disease
hsa-let-7d	HMDD; miR2disease
hsa-mir-223	HMDD; miR2disease
hsa-let-7c	HMDD; miR2disease
hsa-let-7b	HMDD; miR2disease
hsa-mir-146a	HMDD; miR2disease
hsa-mir-199a	HMDD
hsa-mir-132	HMDD; miR2disease
hsa-let-7f	HMDD
hsa-mir-200b	HMDD; miR2disease
hsa-let-7i	HMDD; miR2disease
hsa-mir-125a	HMDD; dbDEMC; miR2disease
hsa-mir-92a	HMDD; miR2disease
hsa-mir-221	HMDD
hsa-mir-16	HMDD
hsa-mir-141	HMDD; miR2disease
hsa-mir-146b	HMDD

Table 4 Lymphoma

hsa-mir-17	HMDD; miR2disease
hsa-mir-18a	HMDD; miR2disease
hsa-mir-20a	HMDD; dbDEMC; miR2disease
hsa-mir-19b	HMDD; miR2disease
hsa-mir-19a	HMDD; miR2disease
hsa-mir-21	HMDD; dbDEMC; miR2disease
hsa-mir-155	HMDD; dbDEMC; miR2disease
hsa-mir-145	HMDD; dbDEMC; miR2disease
hsa-mir-92a	HMDD; miR2disease
hsa-let-7a	HMDD; dbDEMC
hsa-mir-34a	HMDD; dbDEMC; miR2disease
hsa-mir-125b	HMDD; dbDEMC
hsa-let-7e	HMDD
hsa-let-7d	HMDD
hsa-mir-126	HMDD
hsa-mir-223	HMDD; dbDEMC; miR2disease
hsa-let-7b	HMDD
hsa-let-7c	HMDD
hsa-let-7f	dbDEMC
hsa-mir-200b	HMDD
hsa-let-7i	dbDEMC
hsa-mir-125a	HMDD; dbDEMC
hsa-mir-199a	HMDD; miR2disease
hsa-mir-146a	HMDD; dbDEMC
hsa-mir-92b	HMDD
hsa-mir-221	HMDD; dbDEMC
hsa-mir-9	dbDEMC
hsa-mir-29b	HMDD; dbDEMC; miR2disease
hsa-mir-16	HMDD
hsa-mir-127	HMDD

4 Discussion

In this study, WKNKN technique and IMC method were applied to forecast possible disease related miRNAs. The AUC values of our proposed method in cross validation techniques such as five-fold and LOOCV were calculated, and obtained AUC values of 0.8745 and 0.8786, respectively. When predicted outcomes were examined, it was seen that our method could be used to identify miRNAs associated with possible diseases without costly and time-consuming laboratory tests. The successful prediction performance achieved by this method is attributed to the integration of FS and GIP kernel

similarity of miRNAs and integration of SS and GIP kernel similarity of diseases, as well as the use of nearest neighbour information and IMC method to complete missing data. The predictions were validated with HMDD, dbDEMC, and miR2Disease databases resulting in 100% success rate in case studies of breast cancer, lung cancer, and lymphoma. In conclusion, our proposed method has shown that it could be used identifying possible disease related miRNAs.

The predictive performance of our proposed method to find disease-associated miRNAs can be explained as follows. The method we propose uses the advantages of both WKNKN and IMC methods. WKNKN considers not only K-nearest neighbours, but also known K-nearest neighbours with interaction information. Thus, the sparse miRNA-disease association matrix was made denser. The IMC algorithm was used to predict potential miRNA-disease associations. The advantage of IMC is that it can complement missing values in the miRNA-disease relation matrix to improve performance. The method we proposed has the flexibility to combine feature vectors from multiple sources. Also, since IMC is a semi-supervised model, it does not rely on negative samples. This is an advantage for us as there are no negative samples in our data. However, our model also has some limitations. Because the calculation of GIP similarities is based on known miRNA-disease associations, it can lead to an inevitable bias.

References

Alexiou, P., Vergoulis, T., Gleditzsch, M., Prekas, G., Dalamagas, T., Megraw, M., Grosse, I., Sellis, T. and Hatzigeorgiou, A.G. (2009) 'miRGen 2.0: a database of microRNA genomic information and regulation', *Nucleic Acids Research*, Vol. 38, No. suppl_1, pp.D137–D141. https://doi.org/10.1093/nar/gkp888 [online] https://academic.oup.com/nar/article/38/suppl_1/D137/3112243?login=true.

Ambros, V. (2001) 'microRNAs: tiny regulators with great potential', *Cell*, Vol. 107, No. 7, pp.823–826, https://doi.org/10.1016/s0092-8674(01)00616-x [online] https://www.ncbi.nlm.nih.gov/pubmed/11779458.

Bartel, D.P. (2004) 'MicroRNAs: genomics, biogenesis, mechanism, and function', *Cell*, Vol. 116, No. 2, pp.281–297, https://doi.org/10.1016/s0092-8674(04)00045-5 [online] https://www.ncbi.nlm.nih.gov/pubmed/14744438.

Calin, G.A. and Croce, C.M. (2006) 'MicroRNA-cancer connection: the beginning of a new tale', *Cancer Research*, Vol. 66, No. 15, pp.7390–7394, https://doi.org/10.1158/0008-5472.CAN-06-0800 [online] https://www.ncbi.nlm.nih.gov/pubmed/16885332.

Calin, G.A., Sevignani, C., Dumitru, C.D., Hyslop, T., Noch, E., Yendamuri, S., Shimizu, M., Rattan, S., Bullrich, F., Negrini, M. and Croce, C. M. (2004) 'Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers', *Proceedings of the National Academy of Sciences*, Vol. 101, No. 9, pp.2999–3004, https://doi.org/10.1073/pnas.0307323101 [online] https://www.ncbi.nlm.nih.gov/pubmed/14973191.

Chandra, S., Vimal, D., Sharma, D., Rai, V., Gupta, S.C. and Chowdhuri, D.K. (2017) 'Role of miRNAs in development and disease: Lessons learnt from small organisms', *Life Sciences*, Vol. 185, pp.8–14, https://doi.org/10.1016/j.lfs.2017.07.017 [online] https://www.ncbi.nlm.nih.gov/pubmed/28728902.

Chen, X. and Huang, L. (2017) 'LRSSLMDA: Laplacian regularized sparse subspace learning for MiRNA-disease association prediction', *PLoS Computational Biology*, Vol. 13, No. 12, p.e1005912, https://doi.org/10.1371/journal.pcbi.1005912 [online] https://www.ncbi.nlm.nih.gov/pubmed/29253885.

Chen, X., Huang, L., Xie, D. and Zhao, Q. (2018a) 'EGBMMDA: extreme gradient boosting machine for MiRNA-disease association prediction', *Cell Death & Disease*, Vol. 9, No. 1, p.3, https://doi.org/10.1038/s41419-017-0003-x [online] https://www.ncbi.nlm.nih.gov/pubmed/29305594

Chen, X., Wang, L., Qu, J., Guan, N.N. and Li, J.Q. (2018b) 'Predicting miRNA-disease association based on inductive matrix completion', *Bioinformatics*, Vol. 34, No. 24, pp.4256–4265, https://doi.org/10.1093/bioinformatics/bty503 [online] https://www.ncbi.nlm.nih.gov/pubmed/29939227.

Chen, X., Wang, L.Y. and Huang, L. (2018c) 'NDAMDA: network distance analysis for MiRNA-disease association prediction', *Journal of Cellular and Molecular Medicine*, Vol. 22, No. 5, pp.2884–2895, https://doi.org/10.1111/jcmm.13583 [online] https://www.ncbi.nlm.nih.gov/pubmed/29532987.

Chen, X., Wu, Q.F. and Yan, G.Y. (2017) 'RKNNMDA: ranking-based KNN for MiRNA-disease association prediction', *RNA Biology*, Vol. 14, No. 7, pp.952–962, https://doi.org/10.1080/15476286.2017.1312226 [online] https://www.ncbi.nlm.nih.gov/pubmed/28421868.

Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z.H. and Liu, H. (2018d) 'BNPMDA: bipartite network projection for MiRNA-disease association prediction', *Bioinformatics*, Vol. 34, No. 18, pp.3178–3186, https://doi.org/10.1093/bioinformatics/bty333 [online] https://www.ncbi.nlm.nih.gov/pubmed/29701758.

Chen, X., Yang, J.R., Guan, N.N. and Li, J. Q. (2018e) 'GRMDA: graph regression for MiRNA-disease association prediction', *Front Physiol*., Vol. 9, p.92, https://doi.org/10.3389/fphys.2018.00092 [online] https://www.ncbi.nlm.nih.gov/pubmed/29515453.

Chen, X., Yin, J., Qu, J. and Huang, L. (2018f) 'MDHGI: matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction', *PLoS Comput. Biol*., Vol. 14, No. 8, p.e1006418, https://doi.org/10.1371/journal.pcbi.1006418 [online] https://www.ncbi.nlm.nih.gov/pubmed/30142158.

Cheng, F-F., Jiang, N., Li, X., Zhang, L., Hu, L., Chen, X., Jiang, L-P., Abdel-Halim, E. and Zhu, J-J. (2016) 'Target-triggered triple isothermal cascade amplification strategy for ultrasensitive microRNA-21 detection at sub-attomole level', *Biosensors and Bioelectronics*, Vol. 85, pp.891–896, https://doi.org/10.1016/j.bios.2016.06.008 [online] https://pubmed.ncbi.nlm.nih.gov/27311114/.

Dalmay, T. and Edwards, D.R. (2006) 'MicroRNAs and the hallmarks of cancer', *Oncogene*, Vol. 25, No. 46, pp.6170–6175, https://doi.org/10.1038/sj.onc.1209911 [online] https://www.ncbi.nlm.nih.gov/pubmed/17028596.

Ezzat, A., Zhao, P., Wu, M., Li, X.L. and Kwoh, C.K. (2017) 'Drug-target interaction prediction with graph regularized matrix factorization', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 14, No. 3, pp.646–656, https://doi.org/10.1109/TCBB.2016.2530062 [online] https://www.ncbi.nlm.nih.gov/pubmed/26890921.

Garcia, M., Jemal, A., Ward, E.M., Center, M.M., Hao, Y., Siegel, R.L. and Thun, M.J. (2007) 'Global cancer facts & figures 2007', *American Cancer Society*, Vol. 1, No. 3, p.52, Atlanta, GA.

Hayashita, Y., Osada, H., Tatematsu, Y., Yamada, H., Yanagisawa, K., Tomida, S., Yatabe, Y., Kawahara, K., Sekido, Y. and Takahashi, T. (2005) 'A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation', *Cancer Research*, Vol. 65, No. 21, pp.9628–9632, https://doi.org/10.1158/0008-5472.CAN-05-2352 [online] https://www.ncbi.nlm.nih.gov/pubmed/16266980.

He, L., Thomson, J.M., Hemann, M.T., Hernando-Monge, E., Mu, D., Goodson, S., Powers, S., Cordon-Cardo, C., Lowe, S.W., Hannon, G.J. and Hammond, S.M. (2005) 'A microRNA polycistron as a potential human oncogene', *Nature*, Vol. 435, No. 7043, pp.828–833, https://doi.org/10.1038/nature03552 [online] https://www.ncbi.nlm.nih.gov/pubmed/15944707.

Heneghan, H.M., Miller, N., Lowery, A.J., Sweeney, K.J. and Kerin, M.J. (2009) 'MicroRNAs as novel biomarkers for breast cancer', *Journal of Oncology*, p.950201, https://doi.org/10.1155/2010/950201 [online] https://www.ncbi.nlm.nih.gov/pubmed/19639033.

Iorio, M.V., Ferracin, M., Liu, C.G., Veronese, A., Spizzo, R., Sabbioni, S., Magri, E., Pedriali, M., Fabbri, M., Campiglio, M., Menard, S., Palazzo, J.P., Rosenberg, A., Musiani, P., Volinia, S., Nenci, I., Calin, G.A., Querzoli, P., Negrini, M. and Croce, C.M. (2005) 'MicroRNA gene expression deregulation in human breast cancer', *Cancer Research*, Vol. 65, No. 16, pp.7065–7070, https://doi.org/10.1158/0008-5472.CAN-05-1783 [online] https://www.ncbi.nlm.nih.gov/pubmed/16103053.

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G. and Liu, Y. (2009) 'miR2Disease: a manually curated database for microRNA deregulation in human disease', *Nucleic Acids Research*, Vol. 37, pp.D98–104, https://doi.org/10.1093/nar/gkn714 [online] https://www.ncbi.nlm.nih.gov/pubmed/18927107.

Johnson, S. M., Grosshans, H., Shingara, J., Byrom, M., Jarvis, R., Cheng, A., Labourier, E., Reinert, K.L., Brown, D. and Slack, F.J. (2005) ') 'RAS is regulated by the let-7 microRNA family', *Cell*, Vol. 120, No. 5, pp.635–647, https://doi.org/10.1016/j.cell.2005.01.014 [online] https://www.ncbi.nlm.nih.gov/pubmed/15766527.

Kozomara, A. and Griffiths-Jones, S. (2013) 'miRBase: annotating high confidence microRNAs using deep sequencing data', *Nucleic Acids Research*, Vol. 42, No. D1, pp.D68–D73, https://doi.org/10.1093/nar/gkt1181 [online] https://academic.oup.com/nar/article/42/D1/D68/1057911?login=true.

Lee, R.C., Feinbaum, R.L. and Ambros, V. (1993) 'The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14', *Cell*, Vol. 75, No. 5, pp.843–854, https://doi.org/10.1016/0092-8674(93)90529-y [online] https://www.ncbi.nlm.nih.gov/pubmed/8252621.

Li, J.Q., Rong, Z.H., Chen, X., Yan, G.Y. and You, Z.H. (2017) 'MCMDA: matrix completion for MiRNA-disease association prediction', *Oncotarget*, Vol. 8, No. 13, pp.21187–21199, https://doi.org/10.18632/oncotarget.15061 [online] https://www.ncbi.nlm.nih.gov/pubmed/28177900.

Li, L., Gao, Z., Wang, Y.T., Zhang, M.W., Ni, J.C., Zheng, C.H. and Su, Y. (2021) 'SCMFMDA: predicting microRNA-disease associations based on similarity constrained matrix factorization', *PLoS Comput Biol*, Vol. 17, No. 7, p.e1009165, https://doi.org/10.1371/journal.pcbi.1009165 [online] https://www.ncbi.nlm.nih.gov/pubmed/34252084.

Li, S., Yang, X., Yang, J., Zhen, J. and Zhang, D. (2016) 'Serum microRNA-21 as a potential diagnostic biomarker for breast cancer: a systematic review and meta-analysis', *Clinical and Experimental Medicine*, Vol. 16, pp.29–35, https://doi.org/10.1007/s10238-014-0332-3 [online] https://pubmed.ncbi.nlm.nih.gov/25516467/.

Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T. and Cui, Q. (2014) 'HMDD v2.0: a database for experimentally supported human microRNA and disease associations', *Nucleic Acids Research*, Vol. 42, pp.D1070-4, https://doi.org/10.1093/nar/gkt1023 [online] https://www.ncbi.nlm.nih.gov/pubmed/24194601.

Liu, B., Zhu, X., Zhang, L., Liang, Z. and Li, Z. (2021a) 'Combined embedding model for MiRNA-disease association prediction', *BMC Bioinformatics*, Vol. 22, No. 1, p.161, https://doi.org/10.1186/s12859-021-04092-w [online] https://www.ncbi.nlm.nih.gov/pubmed/33765909.

Liu, D., Huang, Y., Nie, W., Zhang, J. and Deng, L. (2021b) 'SMALF: miRNA-disease associations prediction based on stacked autoencoder and XGBoost', *BMC Bioinformatics*, Vol. 22, No. 1, p.219, https://doi.org/10.1186/s12859-021-04135-2 [online] https://www.ncbi.nlm.nih.gov/pubmed/33910505.

Liu, W., Lin, H., Huang, L., Peng, L., Tang, T., Zhao, Q. and Yang, L. (2022) 'Identification of miRNA-disease associations via deep forest ensemble learning based on autoencoder', *Briefings in Bioinformatics*, Vol. 23, No. 3, https://doi.org/10.1093/bib/bbac104 [online] https://www.ncbi.nlm.nih.gov/pubmed/35325038.

Lowery, A.J., Miller, N., Mcneill, R.E. and Kerin, M.J. (2008) 'MicroRNAs as prognostic indicators and therapeutic targets: potential effect on breast cancer management', *Clinical Cancer Research*, Vol. 14, No. 2, pp.360–365, https://doi.org/10.1158/1078-0432.CCR-07-0992 [online] https://pubmed.ncbi.nlm.nih.gov/18223209/.

Lu, J., Getz, G., Miska, E.A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B.L., Mak, R.H., Ferrando, A.A., Downing, J.R., Jacks, T., Horvitz, H.R. and Golub, T.R. (2005) 'MicroRNA expression profiles classify human cancers', *Nature*, Vol. 435, No. 7043, pp.834–838,https://doi.org/10.1038/nature03702 [online] https://www.ncbi.nlm.nih.gov/pubmed/15944708.

Lum, A.M., Wang, B.B., Li, L., Channa, N., Bartha, G. and Wabl, M. (2007) 'Retroviral activation of the mir-106a microRNA cistron in T lymphoma', *Retrovirology*, Vol. 4, No. 5, https://doi.org/10.1186/1742-4690-4-5 [online] https://www.ncbi.nlm.nih.gov/pubmed/17442096.

Lv, M., Zhu, X., Chen, W., Zhao, J. and Tang, J. (2013) 'Searching for candidate microRNA biomarkers in detection of breast cancer: a meta-analysis', *Cancer Biomarkers*, Vol. 13, No. 5, pp.395–401, https://doi.org/10.3233/CBM-130379 [online] https://www.ncbi.nlm.nih.gov/pubmed/24440980.

Lynam-Lennon, N., Maher, S.G. and Reynolds, J.V. (2009) 'The roles of microRNA in cancer and apoptosis', *Biological Reviews Cambridge Philosophical Society*, Vol. 84, No. 1, pp.55–71, https://doi.org/10.1111/j.1469-185X.2008.00061.x [online] https://www.ncbi.nlm.nih.gov/pubmed/19046400.

Mashima, R. (2015) 'Physiological roles of miR-155', *Immunology*, Vol. 145, No. 3, pp.323–333, https://doi.org/10.1111/imm.12468 [online] https://www.ncbi.nlm.nih.gov/pubmed/25829072.

Mattie, M.D., Benz, C.C., Bowers, J., Sensinger, K., Wong, L., Scott, G.K., Fedele, V., Ginzinger, D., Getts, R. and Haqq, C. (2006) 'Optimized high-throughput microRNA expression profiling provides novel biomarker assessment of clinical prostate and breast cancer biopsies', *Molecular Cancer*, Vol. 5, No. 1, p.24, https://doi.org/10.1186/1476-4598-5-24 [online] https://www.ncbi.nlm.nih.gov/pubmed/16784538.

Melkamu, T., Zhang, X., Tan, J., Zeng, Y. and Kassie, F. (2010) 'Alteration of microRNA expression in vinyl carbamate-induced mouse lung tumors and modulation by the chemopreventive agent indole-3-carbinol', *Carcinogenesis*, Vol. 31, No. 2, pp.252–258, https://doi.org/10.1093/carcin/bgp208 [online] https://www.ncbi.nlm.nih.gov/pubmed/19748927.

Munagala, R., Aqil, F. and Gupta, R. C. (2016) 'Exosomal miRNAs as biomarkers of recurrent lung cancer', *Tumour Biol*., Vol. 37, No. 8, pp.10703–10714, https://doi.org/10.1007/s13277-016-4939-8 [online] https://www.ncbi.nlm.nih.gov/pubmed/26867772.

O'day, E. and Lal, A. (2010) 'MicroRNAs and their target gene networks in breast cancer', *Breast Cancer Research*, Vol. 12, No. 2, pp.1–10, https://doi.org/10.1186/bcr2484 [online] https://pubmed.ncbi.nlm.nih.gov/20346098/.

Ota, A., Tagawa, H., Karnan, S., Tsuzuki, S., Karpas, A., Kira, S., Yoshida, Y. and Seto, M. (2004) 'Identification and characterization of a novel gene, C13orf25, as a target for 13q31-q32 amplification in malignant lymphoma', *Cancer Research*, Vol. 64, No. 9, pp.3087–3095, https://doi.org/10.1158/0008-5472.can-03-3773 [online] https://www.ncbi.nlm.nih.gov/pubmed/15126345.

Shenouda, S.K. and Alahari, S.K. (2009) 'MicroRNA function in cancer: oncogene or a tumor suppressor?', *Cancer and Metastasis Reviews*, Vol. 28, Nos. 3–4, pp.369–378, https://doi.org/10.1007/s10555-009-9188-5 [online] https://www.ncbi.nlm.nih.gov/pubmed/20012925.

Tagawa, H. and Seto, M. (2005) 'A microRNA cluster as a target of genomic amplification in malignant lymphoma', *Leukemia*, Vol. 19, No. 11, pp.2013–2016, https://doi.org/10.1038/sj.leu.2403942 [online] https://www.ncbi.nlm.nih.gov/pubmed/16167061.

Toprak, A. (2022) 'A novel method for miRNA-disease association prediction based on space projection and label propagation (SPLPMDA)', *International Journal of Engineering Research and Development*, Vol. 14, No. 3, pp.234–243, https://doi.org/10.29137/umagd.1217754.

Toprak, A. and Eryilmaz Dogan, E. (2021) 'Prediction of potential MicroRNA-disease association using kernelized bayesian matrix factorization', *Interdisciplinary Sciences: Computational Life Sciences*, Vol. 13, No. 4, pp.595–602, https://doi.org/10.1007/s12539-021-00469-w [online] https://www.ncbi.nlm.nih.gov/pubmed/34370220.

Toprak, A. and Eryilmaz, E. (2021) 'Prediction of miRNA-disease associations based on weighted k-nearest known neighbors and network consistency projection', *Journal of Bioinformatics and Computational Biology*, Vol. 19, No. 1, p.2050041, Article 2050041, https://doi.org/10.1142/S0219720020500419 [online] https://www.ncbi.nlm.nih.gov/pubmed/33148093.

Van Laarhoven, T., Nabuurs, S.B. and Marchiori, E. (2011) 'Gaussian interaction profile kernels for predicting drug-target interaction', *Bioinformatics*, Vol. 27, No. 21, pp.3036–3043, https://doi.org/10.1093/bioinformatics/btr500 [online] https://www.ncbi.nlm.nih.gov/pubmed/21893517.

Volinia, S., Calin, G.A., Liu, C.G., Ambs, S., Cimmino, A., Petrocca, F., Visone, R., Iorio, M., Roldo, C., Ferracin, M., Prueitt, R.L., Yanaihara, N., Lanza, G., Scarpa, A., Vecchione, A., Negrini, M., Harris, C.C. and Croce, C.M. (2006) 'A microRNA expression signature of human solid tumors defines cancer gene targets', *PNAS*, Vol. 103, No. 7, pp.2257–2261, https://doi.org/10.1073/pnas.0510565103 [online] https://www.ncbi.nlm.nih.gov/pubmed/16461460.

Wang, C.C., Chen, X., Yin, J. and Qu, J. (2019) 'An integrated framework for the identification of potential miRNA-disease association based on novel negative samples extraction strategy', *RNA Biology*, Vol. 16, No. 3, pp.257–269, https://doi.org/10.1080/15476286.2019.1568820 [online] https://www.ncbi.nlm.nih.gov/pubmed/30646823.

Wang, C.C., Li, T.H., Huang, L. and Chen, X. (2022) 'Prediction of potential miRNA-disease associations based on stacked autoencoder', *Briefings in Bioinformatics*, Vol. 23, No. 2, pp.1–11, https://doi.org/10.1093/bib/bbac021 [online] https://www.ncbi.nlm.nih.gov/pubmed/35176761.

Wang, D., Wang, J., Lu, M., Song, F. and Cui, Q. (2010) 'Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases', *Bioinformatics*, Vol. 26, No. 13, pp.1644–1650, https://doi.org/10.1093/bioinformatics/btq241 [online] https://www.ncbi.nlm.nih.gov/pubmed/20439255.

Wang, S., Bian, C., Yang, Z., Bo, Y., Li, J., Zeng, L., Zhou, H. and Zhao, R.C. (2009) 'miR-145 inhibits breast cancer cell growth through RTKN', *International Journal of Oncology*, Vol. 34, No. 5, pp.1461–1466, https://pubmed.ncbi.nlm.nih.gov/19360360/.

Xia, H., Yamada, S., Aoyama, M., Sato, F., Masaki, A., Ge, Y., Ri, M., Ishida, T., Ueda, R., Utsunomiya, A., Asai, K. and Inagaki, H. (2014) 'Prognostic impact of microRNA-145 down-regulation in adult T-cell leukemia/lymphoma', *Human Pathology*, Vol. 45, No. 6, pp.1192–1198, https://doi.org/10.1016/j.humpath.2014.01.017 [online] https://www.ncbi.nlm.nih.gov/pubmed/24745613.

Yang, J-H., Shao, P., Zhou, H., Chen, Y-Q. and Qu, L-H. (2009) 'deepBase: a database for deeply annotating and mining deep sequencing data', *Nucleic Acids Research*, Vol. 38, No. suppl_1, pp.D123–D130, https://doi.org/10.1093/nar/gkp943.

Yang, Z., Ren, F., Liu, C., He, S., Sun, G., Gao, Q., Yao, L., Zhang, Y., Miao, R., Cao, Y., Zhao, Y., Zhong, Y. and Zhao, H. (2010) 'dbDEMC: a database of differentially expressed miRNAs in human cancers', *BMC Genomics*, Vol. 11, No. Suppl 4, p.S5, https://doi.org/10.1186/1471-2164-11-S4-S5 [online] https://www.ncbi.nlm.nih.gov/pubmed/21143814.

Yu, H., Chen, X. and Lu, L. (2017) 'Large-scale prediction of microRNA-disease associations by combinatorial prioritization algorithm', *Scientific Reports*, Vol. 7, p.43792, https://doi.org/10.1038/srep43792 [online] https://www.ncbi.nlm.nih.gov/pubmed/28317855.

Zhang, Z.W., An, Y. and Teng, C.B. (2009) 'The roles of miR-17-92 cluster in mammal development and tumorigenesis', *Yi Chuan*, Vol. 31, No. 11, pp.1094–10100, https://doi.org/10.3724/sp.j.1005.2009.01094 [online] https://www.ncbi.nlm.nih.gov/pubmed/19933089.

Zhu, C.C., Wang, C.C., Zhao, Y., Zuo, M. and Chen, X. (2021) 'Identification of miRNA-disease associations via multiple information integration with Bayesian ranking', *Briefings in Bioinformatics*, Vol. 22, No. 6, pp.1–10, https://doi.org/10.1093/bib/bbab302 [online] https://www.ncbi.nlm.nih.gov/pubmed/34347021.

Websites

http://www.cuilab.cn/files/images/cuilab/misim.zip (accessed 5 January 2023).

http://www.nlm.nih.gov/ (accessed 5 January 2023).