www.nature.com/scientificreports

scientific reports

OPEN

Inferring pseudogene–MiRNA associations based on an ensemble learning framework with similarity kernel fusion

Chunyan Fan[1]✉ & Mingchao Ding[2]

Accumulating evidence shows that pseudogenes can function as microRNAs (miRNAs) sponges and regulate gene expression. Mining potential interactions between pseudogenes and miRNAs will facilitate the clinical diagnosis and treatment of complex diseases. However, identifying their interactions through biological experiments is time-consuming and labor intensive. In this study, an ensemble learning framework with similarity kernel fusion is proposed to predict pseudogene–miRNA associations, named ELPMA. First, four pseudogene similarity profiles and five miRNA similarity profiles are measured based on the biological and topology properties. Subsequently, similarity kernel fusion method is used to integrate the similarity profiles. Then, the feature representation for pseudogenes and miRNAs is obtained by combining the pseudogene–pseudogene similarities, miRNA–miRNA similarities. Lastly, individual learners are performed on each training subset, and the soft voting is used to yield final decision based on the prediction results of individual learners. The *k*-fold cross validation is implemented to evaluate the prediction performance of ELPMA method. Besides, case studies are conducted on three investigated pseudogenes to validate the predict performance of ELPMA method for predicting pseudogene–miRNA interactions. Therefore, all experiment results show that ELPMA model is a feasible and effective tool to predict interactions between pseudogenes and miRNAs.

Non-coding RNAs (ncRNAs) refer to the RNA molecules that could not translate into proteins, which composed up to about 98% of the human genome. These ncRNAs play an essential role in epigenetic regulation of gene expression at transcriptional and post-transcriptional levels. Pseudogenes are defined as incomplete copies of genes that code for proteins, but lack of coding function. However, pseudogenes could be transcribed into ncRNAs and be considered as regulators in organisms. MicroRNAs (miRNAs) are a class of small, single stranded, non-coding RNAs, which are involved gene expression at post-transcriptional level[1]. By binding to targeting mRNAs, miRNAs cause degradation and translation repression of mRNAs[2]. The fine-tuning of gene regulation by pseudogenes and miRNAs has attracted attentions in many biological processes.

Pseudogenes and miRNA are essential components of competing endogenous RNAs (ceRNAs) network. ceRNA hypothesis is proposed to describe the interactions among ceRNAs members and miRNAs[3]. The ceRNAs members include pseudogenes, long noncoding RNAs (lncRNAs), circular RNA (circRNAs), and protein-coding RNAs, etc. The ceRNAs could form a ceRNA network modulate mRNA expression and regulate protein levels. Recent experimental results show that abnormal expression and dysregulations of both pseudogenes and miRNAs are related to complex diseases. For example, pseudogene GBAP1 contributes to the development and progression of gastric cancer by sequestering the miR-212-3p from binding to GBA[4]. Therefore, pseudogenes and miRNAs can interact with each other, which jointly associated the occurrence of human diseases. However, it is very laborious and time-consuming to verify the associations between pseudogenes and miRNAs through biological experiments. So reasonable and effective computational methods is urgently need to mine the associations between pseudogenes and miRNAs.

Identifying pseudogene–miRNA associations contribute to discover more biological mechanisms in biological process and disease states. Compared with biological methods, the computational approaches are less time consumption. In the area of miRNA research, mining the potential miRNA-disease associations is a high hop topic[5–8]. For example, RWRMMDA model is proposed to predict the miRNA-disease associations by integrating

[1]School of Computer Science and Engineering, Xi'an Technological University, Xi'an 710021, China. [2]School of Computer Science, Hubei University of Technology, Wuhan 430068, China. ✉email: cyfan@xatu.edu.cn

multiple similarities, which also used improved extended random walk with restart algorithm based on miRNA similarity-based and disease similarity-based heterogeneous networks[9]. Zhou et al.[10] proposed GBDT-LR method to prioritize miRNA candidates for diseases by combining gradient boosting decision tree with logistic regression. Besides, a large number of computational models are also developed to forecast other ncRNA associations and disease-biomolecule associations, for example, predicting the lncRNA–miRNA[11,12], circRNA–miRNA[13,14], lncRNA–disease[15,16], circRNA–disease[17–19], drug–disease[20] interactions. Motived by these ncRNA interaction prediction, Zhou et al.[21] incorporates feature fusion and graph auto-encoder to predict pseudogene–miRNA associations. In the model, various perspective attribute information for pseudogenes and miRNAs is obtained as their similarity features, and graph auto-encoder is used to obtain the low-dimensional representation of nodes. Then, the low-dimensional vector is fed into Extreme Gradient Boosting (XGBoost) to predict the pseudogene–miRNA associations. Compared with these ncRNA-miRNA and ncRNA-disease association prediction, only one computational model is developed to predict pseudogene–miRNA associations. Therefore, it still exists some limitations for further improvement. Especially, there is an urgent need to develop more accurate and efficient computational methods to infer associations between pseudogenes and miRNAs.

In this study, an ensemble learning framework with similarity kernel fusion (SKF) method is developed to mine the pseudogene–miRNA associations, named ELPMA. First, GIP kernel similarity, hamming profile similarity, cosine similarity for pseudogenes and miRNAs is calculated based on the known pseudogene–miRNA associations. Then, pseudogene expression similarity and miRNA function similarity are computed based on the pseudogene expression profiles and miRNA–target information, respectively. Besides, the pseudogene similarities and miRNA similarities are fused using SKF method. Then, the feature representation of pseudogene–miRNA interactions is constructed by combing the pseudogene–pseudogene similarity, miRNA–miRNA similarity, and experimentally validated pseudogene–miRNA associations. Next, resampling method is used to build multiple different balanced pseudogene–miRNA association training subsets, which could reduce the bias of small-scale samples. Finally, individual learners are performed on each subset to obtain the primitive outcomes, and the soft voting is used to yield final decision based on the prediction results of individual learners. To assess the effectiveness of ELPMA model, five-fold cross validation is implemented applied to assess the prediction performance of our proposed method. As a result, the mean area under the ROC curve (AUC) and mean area under the precision-recall curve (AUPR) of ELPMA method achieved 0.9896 and 0.9913, respectively. According to comparison with other four methods, assessment results shown that ELPMA method obtain comparable performance. In the case studies, the predicted miRNAs for the three investigated pseudogenes are also used to validate the prediction performance of ELPMA method. All the results shown that our proposed model could serve as a recommendable tool for predicting pseudogene–miRNA associations.

Materials and methods

Gold standard data set.

The pseudogene–miRNA associations are obtained from starBase v2.0, in which very high stringency of pseudogene symbol is selected[22]. After screening and removing redundancy, 1570 experimentally supported pseudogene–miRNA associations is sorted out, covering 318 pseudogenes and 260 miRNAs. In this study, a pseudogene–miRNA adjacency matrix $PM(i, j)$ is constructed based on the validated associations between pseudogenes and miRNAs. If there is an association between pseudogenes $p(i)$ and miRNAs $m(j)$, $PM(i, j)$ is assigned as 1, otherwise 0.

Expression similarity for pseudogenes.

The expression level of pseudogenes in various cancers and normal tissues is obtained from dreamBase database[23]. In dreamBase database, expression information of pseudogenes is selected as the characteristic information of pseudogenes. When two pseudogenes have a higher correlation score tend to be more similarity expressed. The pseudogene expression profiles are measures as follows:

$$SP\_EP(m_i, m_j) = \frac{\sum_{k=1}^{N} (x_k - \overline{x})(y_k - \overline{y})}{\sqrt{\sum_{k=1}^{N} (x_k - \overline{x})^2 \sum_{k=1}^{N} (y_k - \overline{y})^2}} \tag{1}$$

where N is the number of properties of the expression profiles, x_k and y_k denote the expression values in different cancers and normal tissues.

Function similarity for miRNAs.

Given that miRNAs targeting more of the same genes tend to be involved in similar biological function. The interactions between miRNA and target gene information are obtained from miRTarBase[24]. The miRNA–target interactions are employed to measure the miRNA function similarity for each pair of miRNAs. If two sets of target genes (say G_i and G_j) respectively have relationship with miRNA M_i and miRNA M_j, the miRNA function similarity is calculated as follows:

$$SM\_FS(m_i, m_j) = \frac{card(G_i \cap G_j)}{\sqrt{card(G_i)} \cdot \sqrt{card(G_j)}} \tag{2}$$

where G_i and G_j represent the sets of target gene that related with miRNAs.

GIP kernel similarity for pseudogenes and miRNAs.

The GIP kernel similarity is applied to calculate the similarity between pseudogenes and miRNAs based on the known pseudogene–miRNA association adjacency matrix[25]. The GIP kernel similarity for pseudogenes can be calculated as follows:

$$SP\_GIP(p(i), p(j)) = exp(-\gamma_p \parallel p(i) - p(j) \parallel^2)$$
$$\gamma_p = \frac{1}{(\frac{1}{n_p}\sum_{i=1}^{n_p} \parallel p(i) \parallel^2)} \tag{3}$$

where $p(i)$ represents the pseudogene interaction profiles, which is a binary vector that encode the interaction between pseudogene i and all miRNAs, i.e., the i-th row of the gold standard pseudogenes-miRNA adjacency matrix PM. The parameter γ_p controls the kernel bandwidth. n_p is the number of pseudogenes.

Similar to pseudogenes, the GIP kernel similarity for miRNAs is defined as:

$$SM\_GIP(m(i), m(j)) = exp(-\gamma_m \parallel m(i) - m(j) \parallel^2)$$
$$\gamma_m = \frac{1}{(\frac{1}{n_m}\sum_{i=1}^{n_m} \parallel m(i) \parallel^2)} \tag{4}$$

where $m(i)$ represents the miRNA interaction profiles, which is a binary vector that encode the interaction between miRNA i and each pseudogene, i.e., the i-th column of adjacency matrix PM. The parameter γ_m is also used to control the kernel bandwidth. n_m is the number of miRNAs.

Hamming profile similarity for pseudogenes and miRNAs.

Given the length for a pair of vectors are same, hamming profile is the number of elements of which corresponding values are different. The higher Hamming profile value represents the two vector has lower similarity. Hamming profile similarity for pseudogenes is calculated as follows:

$$SP\_HP(p_i, p_j) = 1 - \frac{\left|IP(p_i)! = IP(p_j)\right|}{\left|IP(p_i)\right|} \tag{5}$$

where $IP(p_i)$ is the i-th row of the pseudogene–miRNA adjacency matrix PM.

Similarly, the hamming profile similarity for miRNA is defined as follows:

$$SM\_HP(m_i, m_j) = 1 - \frac{\left|IP(m_i)! = IP(m_j)\right|}{|IP(m_i)|} \tag{6}$$

where $IP(m_i)$ is the i-th column of the pseudogene–miRNA adjacency matrix PM.

Cosine similarity for pseudogenes and miRNAs.

Cosine similarity algorithm has been widely used in the collaborative filtering recommendation algorithm. Here, based on known pseudogene–miRNA associations, the similarity of pseudogenes p_i and p_j is defined as follows:

$$SP\_\cos(p_i, p_j) = \frac{MP(p_i) \cdot MP(p_j)}{\left\|MP(p_i)\right\|\left\|MP(p_j)\right\|}$$
$$SP\_\cos = (SP\_\cos(p_i, p_j))^{r*r} \tag{7}$$

where r represents the number of pseudogenes. The binary vector $PM(p_i)$ indicates whether exist an association between pseudogene p_i and each miRNA (the row i of the PM matrix, if p_i is related to miRNA, otherwise 0). Meanwhile, $SP\_cos(p_i, p_j)$ represents the cosine similarity between pseudogene p_i and p_j. SP_cos is the pseudogene cosine similarity matrix.

Similarly, the cosine similarity of miRNA m_i and miRNA m_j is computed as follows:

$$SM\_\cos(m_i, m_j) = \frac{MP(m_i) \cdot MP(m_j)}{\|MP(m_i)\|\left\|MP(m_j)\right\|}$$
$$SM\_\cos = (SM\_\cos(m_i, m_j))^{n*n} \tag{8}$$

where $MP(m_i)$ denotes whether there is an association between miRNA m_i and each pseudogene (the column of MP matrix, if m_j is related to pseudogene, otherwise 0). $SM\_cos(m_i, m_j)$ is the cosine similarity between miRNA m_i and miRNA m_j. The SM_cos is the miRNA cosine similarity matrix. n is the number of miRNAs.

Integrated similarity by similarity kernel fusion method.

In this study, four kinds of pseudogene similarities and five miRNA similarities are calculated. The integrated pseudogene similarity is measured by combining pseudogene expression similarity, pseudogene GIP kernel similarity, pseudogene hamming profile similarity, pseudogene cosine similarity. The integrated miRNA similarity is calculated by combining miRNA function similarity, miRNA GIP kernel similarity, miRNA hamming profile similarity and cosine similarity. Here, similarity kernel fusion method is used to fuse the four pseudogene similarities and five miRNA similarities[26]. Let $S_{p,r}$ ($r = 1,2,\ldots,4$) represents the four pseudogene similarities and $S_{m,n}$ ($n = 1,2,\ldots,5$) represents the five miRNA similarities, respectively.

Firstly, each original kernel for pseudogenes is normalized by Eq. (9).

$$NS_{p,r}(p_i, p_j) = \frac{S_{p,r}(p_i, p_j)}{\sum_{p_k \in P} S_{p,r}(p_k, p_j)} \tag{9}$$

where when $NS_{p,r}$ satisfies $\sum_{c_k \in C} NS_{c,m}(c_k, c_j) = 1$, $NS_{p,r}$ is the normalized pseudogene similarity.

Then, a sparse kernel for each pseudogene similarity is computed by Eq. (10).

$$F_{p,r}(p_i, p_j) = \begin{cases} \frac{S_{p,r}(p_i,p_j)}{\sum_{p_k \in N_i} S_{p,r}(p_i,p_k)} & p_j \in N_i \\ 0 & p_j \notin N_i \end{cases} \tag{10}$$

where $F_{c,m}$ is a sparse kernel and it satisfies $\sum_{c_j \in C} F_{c,m}(c_k, c_j) = 1$. N_i is a set of p_i's neighbors including c_i itself.

Therefore, four pseudogene similarities could be computed as Eq. (11).

$$SP^{t+1}_{p,r} = \alpha(F_{p,r} \times \frac{\sum_{k \neq 1} SP^{t}_{p,k}}{2} \times F^{T}_{p,r}) + (1-\alpha)(\frac{\sum_{k \neq 1} SP^{0}_{p,k}}{2}) \ \alpha \in (0, 1) \tag{11}$$

where $SP^{t+1}_{p,r}$ is the status matrix of r-th pseudogene similarity kernel after $t+1$ iterations.$SP^{0}_{p,k}$ denotes the initial status of $S_{p,k}$.

After $t+1$ steps, the overall kernel for pseudogenes is calculated as Eq. (12).

$$S_p = \frac{1}{4}\sum_{r=1}^{4} SC^{t+1}_{p,r} \tag{12}$$

Finally, a weight matrix w_p is used to remove the noise in the matrix S_p.

$$w_p(p_i,p_j) = \begin{cases} 1 & \text{if } p_i \in N_j \text{ and } p_j \in N_i \\ 0 & \text{if } p_i \notin N_j \text{ and } p_j \notin N_i \\ 0.5 & \text{otherwise} \end{cases} \tag{13}$$

The fused pseudogene similarity is computed as Eq. (14).

$$S^{*}_{p} = w_p \circ S_p \tag{14}$$

Similarly, the integrated miRNA similarity as $S_m{}^{*}$ is computed, in which involved five miRNA similarities to be fused.

Ensemble learning framework with resampling method.

To predict the potential pseudogene–miRNA associations, an ensemble learning framework with similarity kernel fusion method is proposed. Inspired by the previous research[27,28], ELPMA model is proposed through the following steps: (1) using the resampling method to obtain multiple different training subsets, and the diversity of individual learners is increased; (2) to integrate the prediction results of individual learners, soft voting is employed to obtain the final prediction. The process of constructing the ensemble learning framework is shown in Fig. 1.

Resampling strategy.

There are 1570 experimentally confirmed pseudogene–miRNA associations as positive samples, and 81,110 unconfirmed pseudogene–miRNA pairs as unlabeled samples. So only a small part of experimentally confirmed pseudogene–miRNA associations. To settle the problem caused by the imbalanced dataset, the resample strategy is employed to build multiple different balanced training subsets. The negative samples are guaranteed to have the same number with positive samples. When construct a subset, all positive samples are sort out, and same unlabeled samples are randomly selected as negative samples. Then, the negative samples and positive training sample are combined to balance the positive and negative samples. The training set of positive sample P and the unlabeled sample set U are defined as follows:

$$P = \{p(i), m(j) || PM(p(i), m(j)) = 1\} \tag{15}$$

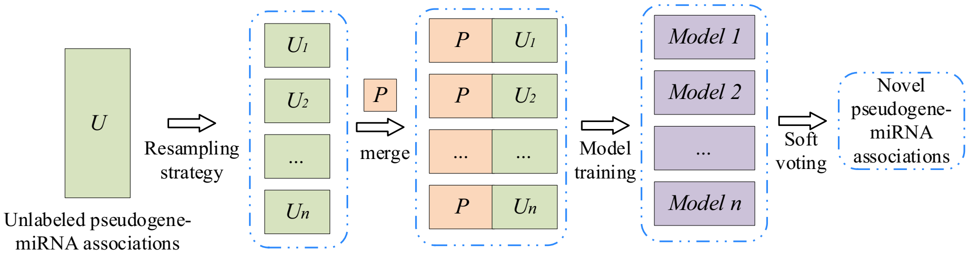

Figure 1. Ensemble learning framework for the pseudogene–miRNA association prediction.

$$U = \{p(i), m(j)||PM(p(i), m(j)) = 0\} \tag{16}$$

where P represents the positive samples, and U denotes the unknown pseudogene–miRNA association samples.

In each training subset, the number of unlabeled pseudogene–miRNA associations is the same as the number of positive samples. The set N ($N \in U$) represents the negative samples selected from U, and the number of N is same as the number of P. The set of $T = P \cup N$ is the training set in base learning.

Sample representation.

To learn the pseudogenes and miRNAs potential feature representation, multiple data source is incorporated to obtain the integrated similarities for pseudogenes and miRNAs. Here, a pseudogene–miRNA pair was taken as a sample. The feature vector of i-th pseudogene, $FP(p(i))$, is defined as follows:

$$FP(p(i)) = (SP(p(i), p(1)), SP(p(i), p(2)), \ldots, SP(p(i), p(N_p))) \tag{17}$$

where N_p represents the number of pseudogenes. Similarly, the feature vector of jth miRNA, $FM(m(j))$, is defined as follows:

$$FM(m(j)) = (SM(m(j), m(1)), SM(m(j), m(2)), \ldots, SM(m(j), m(N_m))) \tag{18}$$

where N_m represents the number of miRNAs. Then, the feature vector of each pseudogene–miRNA pair ($p(i)$,$m(j)$) is defined by combining the $FP(p(i))$ and $FM(m(j))$ as follows:

$$F(p(i), m(j)) = FP(p(i), FM(m(j)) \tag{19}$$

Soft voting for pseudogene–miRNA association prediction.

Ensemble learning combines multiple individual learners to increase the prediction performance compared to individual models. Owing to the training subsets are different and the feature spaces of the subsets are heterogenous, the trained individual learners are also different from each other. In this study, an ensemble learning framework is developed by using the XGBoost as individual learner on the multiple sample subsets. XGBoost is a machine learning algorithm in which regression trees is used as functions in gradient boosting to optimize trees[29].

Set the output of a tree as shown below:

$$f(x) = w_q(x_i) \tag{20}$$

where x_i is the input vector, q represents the structure of each tree and w_q represents the score of the leaf node q. The output of the set of K trees is:

$$\hat{y}_i = \sum\nolimits_{k=1}^{K} f_k(x_i) \tag{21}$$

where K is the number of regression functions, the objective function for learning the set of f_k is shown as follows:

$$\begin{gathered} L(\varphi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \\ where\ \ \Omega(f) = \gamma T + 0.5\lambda\|w\|^2 \end{gathered} \tag{22}$$

where l represents the loss function between the observed value y_i and predict value $\hat{y}_i$. $\Omega(f_k)$ is the regularization term to avoid overfitting. γ is the pseudo-regularization hyperparameter. λ is the L2 norm for leaf weights. T is the total number of leaf nodes.

The optimal objective function value could be written as:

$$\begin{gathered} \hat{L}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda} + \gamma T \\ g_i = \delta_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) \\ h_i = \delta^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) \end{gathered} \tag{23}$$

where I is the set of leaf nodes, g_i is the first derivative of l and h_i is the second derivative of l.

Here, the outputs of XGBoost are taken as primitive results. Then, the soft voting is used to make the final decision. The prediction scores of individual learners are averaged, and confirmed whether the pseudogene is associated with each other. Take an unknown pseudogene–miRNA association as sample input, n individual learners could produce n prediction results, and then the n prediction results are integrated by using the soft voting strategy[30]. Specifically, the output of the i-th sample by soft voting is defined as follows:

$$O(i) = \frac{1}{n}\sum\nolimits_{j=1}^{n} O(i, j) \tag{24}$$

where $O(i,j)$ is the prediction scores of the j-th individual learners for the i-th sample. n represents the number of training subsets. $O(i) > 0.5$ represents the pseudogene–miRNA pair is associated; otherwise, it is considered to be not associated with each other.

Results

Performance evaluation.

In this work, *k*-fold cross validation is employed to evaluate the performance of the ELPMA model. The validated pseudogene–miRNA associations are regarded as the positive set, and equal number of samples are randomly selected from the negative sample set as negative samples. For each cross validation, (*k*-1) positive subsets and the same number of negative subsets took from *k* subsets to train the models; the remaining one positive subset and one negative subset are used for testing to evaluate the prediction performance. Specifically, fivefold and tenfold cross validation are used to evaluate the prediction performance of ELPMA model. Moreover, several metrics are used to measure the prediction performance of ELPMA method, including precision (Pre), sensitivity (Sen), accuracy (Acc), F1-score, AUC (Area under the receiver operating characteristic curve), AUPR (Area under the precision-recall curve), and MCC (Matthews's correlation coefficient). The calculation formulas of these metrics are shown as follows:

$$Pre = \frac{TP}{TP + FP} \tag{25}$$

$$Sen = \frac{TP}{TP + FN} \tag{26}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{27}$$

$$F1 - score = \frac{2 \times Sen \times Pre}{Sen + Pre} \tag{28}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN) * (TP + FP) * (TN + FN) * (TN + FP)}} \tag{29}$$

where *TP* and *TN* represent the number of true positives and true negatives, respectively. *FP* and *FN* represent the number of positives and negatives, respectively, that are wrongly predicted.

Performance analysis of ELPMA method with different individual learners.

To assess the ability of the ELPMA method to predict the associations between pseudogenes and miRNAs, fivefold cross validation is implemented on the gold standard data set. In the ensemble framework, different individual learners could affect the prediction performance. Here, AdaBoost, Random Forest (RF), Extreme Gradient Boosting (XGB) and Extremely Randomized Trees (ERT) are used as the individual learners, respectively. The individual learners are represented as ELPMA-AB, ELPMA-RF, ELPMA-XGB and ELPMA-ERT, respectively. In the ELPMA model, parameter selection are important factors, and the hyper-parameters of each model are tuned. For example, the number of individual learners of ELPMA is range from 2 to 20 with steps of 1. Furthermore, the range of hyper-parameter turning of ELPMA-XGB is as that n_estimators are selected from [50, 100, 200, 300, 400, 500], the learning rate is set from 0.1 to 0.9 with an interval of 0.1. The range of hyper-parameter turning of ELPMA-ERT is as that the value of max_depth is selected from [10, 20, 30, 40, 50] and the n_estimators are selected from [50, 100, 200, 300, 400, 500]. In addition, different hyper-parameters of ELPMA-AB and ELPMA-RF model are selected to obtain optimal performance. Finally, the prediction performance of the ELPMA model that using different individual learners is listed in Table 1. When the number of individual learners, n_estimators, learning rate are respectively set as 10, 400, 0.2, ELPMA-XGB yields the Precision of 0.9716, the Recall of 0.9369, the F1-score of 0.9540, the Acc of 0.9548, the AUC of 0.9897, the AUPR of 0.9914. As shown in Table 1, ELPMA-XGB is higher than other models in these seven metrics.

In addition, the ROC curves of the *k*-fold cross validation are plotted by the proposed ELPMA-XGB method, respectively. The experimental results show that ELPMA-XGB achieves mean AUC values of 0.9897 and 0.9906 for the fivefold and tenfold cross validation (Fig. 2). Therefore, ELPMA-XGB model is appropriate as the individual learners of ELPMA method for the prediction of pseudogene–miRNA associations.

Influence of training data on model performance.

In the task, experimentally validated pseudogene-miRNA associations are selected as the only information source for model construction. The number of known

Model	Precision	Sensitivity	F1-score	Acc	AUC	AUPR	MCC
ELPMA-AB	0.7118	0.7153	0.7128	0.7124	0.7822	0.8000	0.4257
ELPMA-RF	0.9362	0.8592	0.8959	0.9003	0.9568	0.9664	0.8035
ELPMA-ERT	0.9650	0.8962	0.9292	0.9318	0.9793	0.9832	0.8660
ELPMA-XGB	**0.9716**	**0.9369**	**0.9540**	**0.9548**	**0.9897**	**0.9914**	**0.9102**

Table 1. The prediction performance of ELPMA model using different individual learners. Significant values are in bold.

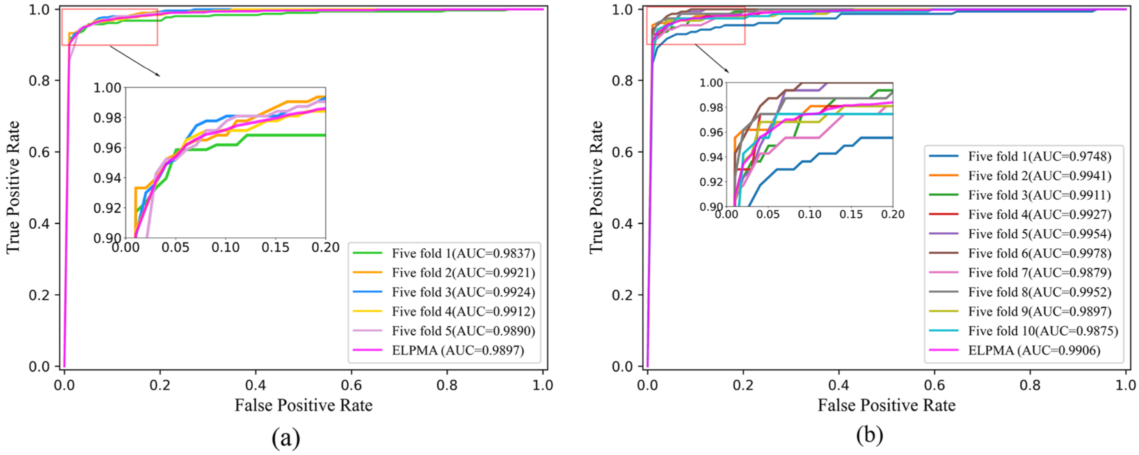

Figure 2. ROC curves under *k*-fold cross validation performed by the ELPMA-XGB framework. (**a**) ROC curves under fivefold cross validation; (**b**) ROC curves under tenfold cross validation.

pseudogene-miRNA associations may influence the prediction of our method ELPMA. To evaluate the impact of the number of training data on the performance, we used different proportions of training data to implement the ELPMA model. The fivefold and tenfold cross-validation results obtained by ELPMA is shown in Table S1. The results shown that the performance of ELPMA model getting better with the training data increasing. Therefore, the size of the training data has a great influence on the prediction performance of ELPMA model. With the number of training data increasing, the prediction performance of is also increased.

Effectiveness of soft voting for the ensemble learning framework. To demonstrate the effectiveness of the soft voting for the ensemble learning method, the soft voting performance is compared with individual learners on ELPMA model. Detailed results of the comparison are shown in Fig. 3. In the figures, the horizontal axis represents the index number of individual learners, and the vertical axis are the AUC values and AUPR values. From the Fig. 3, we also seen that the AUC of individual learners is between 0.9823 and 0.9849, and the AUPR of individual learners is between 0.9849 and 0.9873 under fivefold cross validation. The results indicate that soft voting in the proposed method could improve the prediction performance of ELPMA model. It also indicates that ELPMA is an effective framework to predict the pseudogene–miRNA interactions.

Comparison with other existing methods. To comparatively illustrate the superiority of ELPMA method, GBDT-LR[10], ABMDA[31], CD_LNLP[17], and LAGCN[20] are compared with ELPMA method to predict the pseudogene–miRNA interactions. These five methods are individual evaluated based on gold standard data set with *k*-fold cross validation and recommended hyperparameters. As show in Fig. 4, ELPMA shows the best performance in term of the average AUC values under fivefold and tenfold cross validation. It shows that the ROC curves of ELPMA model is above those of GBDT-LR, ABMDA, CD_LNLP and LAGCN method in most cases. The average AUC scores of ELPMA method are up to 0.9897 and 0.9906 for the fivefold and tenfold cross validation, respectively, which is superior to the other four methods (Fig. 4). In addition, the results of performance evaluation indicators such as F1-score, Acc, MCC are shown in Table 2 for fivefold and tenfold cross validation. Although the Precision of ELPMA is inferior to ABMDA and Acc of ELPMA is inferior to CD_LNLP and LAGCN, the evaluation metrics of ELPMA are higher than others (Table 2). Furthermore, we used the paired

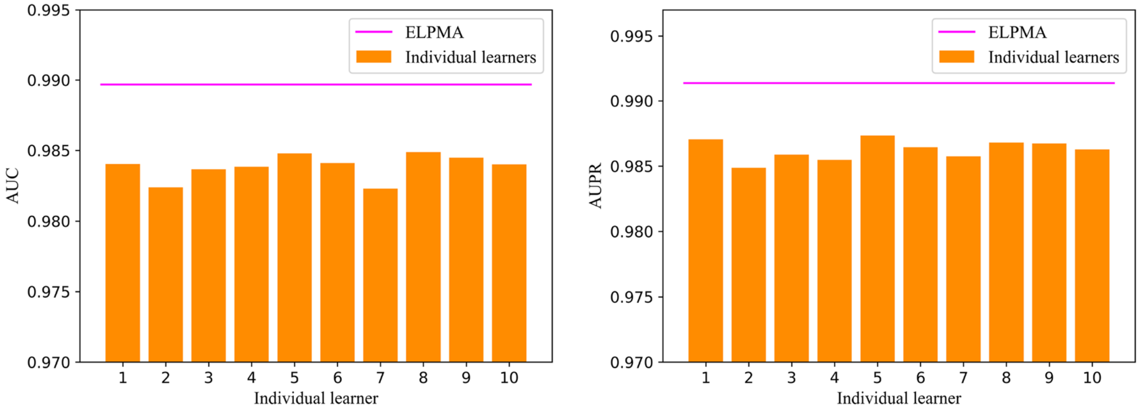

Figure 3. Performance comparison of ELPMA method and individual learners.

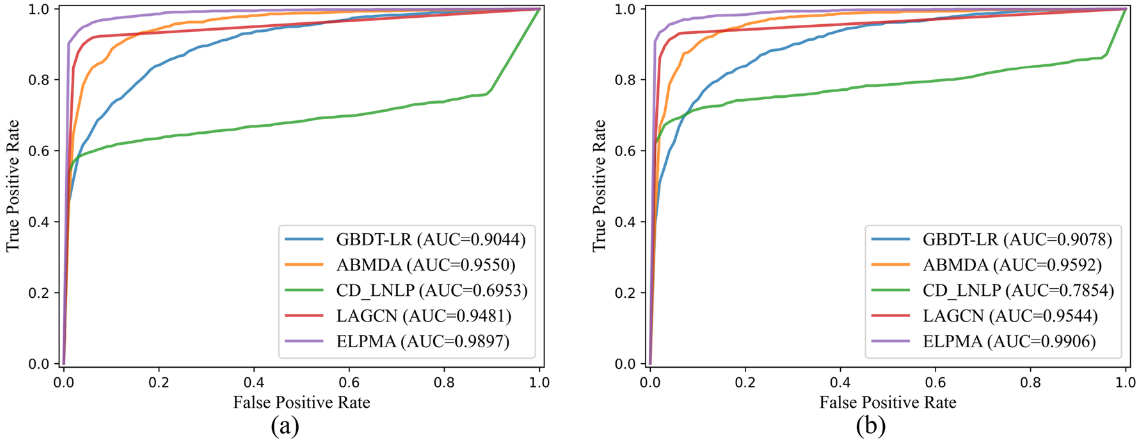

Figure 4. ROC curves of different methods under *k*-fold cross validation. (**a**) ROC curves under fivefold cross validation; (**b**) ROC curves under tenfold cross validation.

	Model	Precision	Sensitivity	F1-score	Acc	AUC	AUPR	MCC
Fivefold cross-validation	GBDT-LR	0.8200	0.8166	0.8179	0.8176	0.9044	0.9144	0.6358
	ABMDA	**0.9832**	0.2834	0.4381	0.6411	0.9550	0.9519	0.3966
	CD_LNLP	0.7780	0.4822	0.5954	**0.9876**	0.6953	0.5216	0.6069
	LAGCN	0.1632	0.8076	0.2712	0.9832	0.9481	0.4847	0.3582
	ELPMA	0.9716	**0.9369**	**0.9540**	0.9548	**0.9897**	**0.9914**	**0.9102**
Tenfold cross-validation	GBDT-LR	0.8278	0.8306	0.8287	0.8275	0.9078	0.9145	0.6558
	ABMDA	**0.9848**	0.3478	0.5078	0.6728	0.9592	0.9551	0.4487
	CD_LNLP	0.8594	0.5605	0.6785	**0.9899**	0.7854	0.6264	0.6895
	LAGCN	0.1007	0.8261	0.1794	0.9853	0.9544	0.4633	0.2852
	ELPMA	0.9727	**0.9414**	**0.9565**	0.9573	**0.9906**	**0.9922**	**0.9155**

Table 2. Comparison with multiple evaluation metrics under fivefold and tenfold cross-validation. Significant values are in bold.

t-test based on 10 runs of fivefold and tenfold cross-validation to test the performance of the ELPMA method and the comparison methods. Table 3 shows that ELPMA is significantly preferred to other computational methods in terms of Sensitivity, F1-score, AUC, AUPR and MCC (Table 3). Therefore, all the above results show that ELPMA method provides a great improvement in predict the pseudogene–miRNA interactions.

ELPMA versus		GBDT-LR	ABMDA	CD_LNLP	LAGCN
Fivefold cross-validation	*p*-value of Precision	3.1222e−19	3.7527e−04	1.0708e−13	1.8361e−34
	p-value of Sensitivity	2.4777e−19	9.1047e−22	6.9457e−27	4.0794e−19
	p-value of F1-score	4.7760e−21	1.8721e−18	1.9262e−26	6.8365e−32
	p-value of Acc	8.4523e−21	2.0776e−21	1.6029e−19	2.0969e−18
	p-value of AUC	8.7304e−19	3.7505e−18	1.3119e−29	2.3623e−17
	p-value of AUPR	5.4014e−19	6.0815e−18	3.2105e−30	3.8908e−34
	p-value of MCC	9.5274e−21	6.4989e−21	3.2757e−23	1.8933e−30
Tenfold cross-validation	*p*-value of Precision	2.7171e−19	0.0018	3.1185e−10	3.7167e−38
	p-value of Sensitivity	4.5501e−20	2.907e−27	1.0613e−26	1.1716e−20
	p-value of F1-score	1.3765e−23	1.7704e−23	2.7238e−26	2.1321e−36
	p-value of Acc	8.2969e−23	1.9452e−26	3.0679e−19	1.0305e−17
	p-value of AUC	1.8574e−20	2.5035e−16	1.1239e−30	2.7548e−23
	p-value of AUPR	7.2994e−18	8.1945e−14	1.8536e−33	4.9185e−38
	p-value of MCC	8.1499e−23	2.0395e−25	3.2985e−22	2.4717e−33

Table 3. The statistical results by paired *t*-test for ELPMA and other comparison methods.

Case studies. To illustration the prediction performance of ELPMA method in screening pseudogene–miRNA interactions, case studies of three pseudogene related miRNA are conduct for further validation. Given the investigated pseudogene–miRNA interaction to be unknown in all known associations. In this section, the pseudogene *MSTO2P*, *MTND4P12* related miRNAs are removed in the known associations, and then use other associations to train the model and predict the probability of all miRNAs associated with the investigated pseudogenes. Through the calculation of ELPMA method, the candidate associations between pseudogene and miRNAs are sorted in descending order. Then, the top 10 rank results are selected with high probability scores for the three investigated pseudogenes, and the predicted associations are verified with the starBase database.

Pseudogene *MSTO2P* is found to be implicated in several diseases including lung cancer[32], colorectal cancer[33], etc. *MSTO2P* could function as a miR-128-3p sponge in non-small cell lung cancer cells (NSCLC), and *MSTO2P*/miR-128-3p to regulate coptisine sensitivity of NSCLC cells via TGF-β pathway. In addition, *MSTO2P* related top 10 miRNAs, in which 9 of the top10 is proved by starBase (Table 4).

MTND4P12 is considered as an oncogenic pseudogene upregulated in skin cutaneous melanoma, and it can upregulate the expression of oncogene *AURKB* by serving as ceRNA[34]. Hsa-let-7e-5p is also identified as candidate miRNA that regulated by *MTND4P12,* hsa-let-7e-5p and *MTND4P12* is co-expression in skin cutaneous melanoma. As shown in Table 4, the *MTND4P12* related top 10 miRNAs is supported by starBase.

Conclusion

Increasing evidences show that both pseudogenes and miRNAs play oncogenic or tumor-suppressive roles in disease progression. Predicting pseudogene–miRNA associations will contribute to understanding the pathological mechanisms, diagnosis, and treatment of diseases. In this work, a computational method is proposed to infer the associations between pseudogenes and miRNAs, which employed an ensemble learning framework with similarity kernel fusion, named ELPMA. By comparing with other four models, the prediction performance of our proposed method is powerful to predict the pseudogene–miRNA interactions. The case study of investigated *MSTO2P* and *MTND4P12* related miRNAs also proved the ELPMA method is reliable and effective.

The good performance of ELPMA method is attributed to three main factors: (1) ELPMA integrates the biological information including pseudogene expression profiles and miRNA–targets interactions. (2) ELPMA introduces the resampling method to settle the problem caused by the imbalanced pseudogene–miRNA dataset. (3) The application of XGBoost as individual learner of the ensemble learning framework guarantees the effectiveness of learning the meaning of combinations of features from feature representation.

There are also some limitations in the ELPMA method. First, the gold standard pseudogene-miRNA associations may have nosy, and the negative samples are randomly selected from the unconfirmed associations, limiting the prediction performance. In addition, the ELPMA method relies on the known pseudogene–miRNA interaction network, and it could not predict novel pseudogene-miRNA interactions without any known associations. Therefore, developing more effective framework is essential to infer the associations between pseudogenes and miRNAs.

Pseudogene	Rank	miRNA	Evidence
MSTO2P	1	hsa-miR-20a-5p	starBase
	2	hsa-miR-106b-5p	starBase
	3	hsa-miR-93-5p	starBase
	4	hsa-miR-519d-3p	starBase
	5	hsa-miR-20b-5p	starBase
	6	hsa-miR-17-5p	starBase
	7	hsa-miR-106a-5p	starBase
	8	hsa-miR-128-3p	starBase
	9	hsa-miR-448	starBase
	10	hsa-miR-373-3p	Unconfirmed
MTND4P12	1	hsa-let-7b-5p	starBase
	2	hsa-miR-98-5p	starBase
	3	hsa-let-7e-5p	starBase
	4	hsa-let-7d-5p	starBase
	5	hsa-let-7a-5p	starBase
	6	hsa-let-7c-5p	starBase
	7	hsa-miR-4500	starBase
	8	hsa-miR-4458	starBase
	9	hsa-let-7g-5p	starBase
	10	hsa-let-7f-5p	starBase

Table 4. The top 10 associated miRNAs for pseudogene MSTO2P, MTND4P12.

Data availability

The data will be made available on request from the corresponding author.

Received: 29 March 2023; Accepted: 28 May 2023
Published online: 31 May 2023

References

1. Bartel, D. P. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297. https://doi.org/10.1016/s0092-8674(04)00045-5 (2004).
2. Bartel, D. P. MicroRNAs: Target recognition and regulatory functions. *Cell* **136**, 215–233. https://doi.org/10.1016/j.cell.2009.01.002 (2009).
3. Salmena, L., Poliseno, L., Tay, Y., Kats, L. & Pandolfi, P. P. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?. *Cell* **146**, 353–358. https://doi.org/10.1016/j.cell.2011.07.014 (2011).
4. Ma, G. *et al.* A genetic variation in the CpG island of pseudogene GBAP1 promoter is associated with gastric cancer susceptibility. *Cancer* **125**, 2465–2473. https://doi.org/10.1002/cncr.32081 (2019).
5. Huang, L., Zhang, L. & Chen, X. Updated review of advances in microRNAs and complex diseases: Taxonomy, trends and challenges of computational models. *Brief. Bioinform.* https://doi.org/10.1093/bib/bbac358 (2022).
6. Huang, L., Zhang, L. & Chen, X. Updated review of advances in microRNAs and complex diseases: Towards systematic evaluation of computational models. *Brief. Bioinform.* https://doi.org/10.1093/bib/bbac407 (2022).
7. Huang, L., Zhang, L. & Chen, X. Updated review of advances in microRNAs and complex diseases: Experimental results, databases, webservers and data fusion. *Brief. Bioinform.* https://doi.org/10.1093/bib/bbac397 (2022).
8. Chen, X., Xie, D., Zhao, Q. & You, Z. H. MicroRNAs and complex diseases: From experimental results to computational models. *Brief. Bioinform.* **20**, 515–539. https://doi.org/10.1093/bib/bbx130 (2019).
9. Nguyen, V. T., Le, T. T. K., Than, K. & Tran, D. H. Predicting miRNA–disease associations using improved random walk with restart and integrating multiple similarities. *Sci. Rep.* **11**, 21071. https://doi.org/10.1038/s41598-021-00677-w (2021).
10. Zhou, S., Wang, S., Wu, Q., Azim, R. & Li, W. Predicting potential miRNA-disease associations by combining gradient boosting decision tree with logistic regression. *Comput. Biol. Chem.* **85**, 107200. https://doi.org/10.1016/j.compbiolchem.2020.107200 (2020).
11. Xu, M. *et al.* SPMLMI: Predicting lncRNA-miRNA interactions in humans using a structural perturbation method. *PeerJ* **9**, e11426. https://doi.org/10.7717/peerj.11426 (2021).
12. Wang, M. N., Lei, L. L., He, W. & Ding, D. W. SPCMLMI: A structural perturbation-based matrix completion method to predict lncRNA-miRNA interactions. *Front. Genet.* **13**, 1032428. https://doi.org/10.3389/fgene.2022.1032428 (2022).
13. Guo, L. X. *et al.* A novel circRNA-miRNA association prediction model based on structural deep neural network embedding. *Brief. Bioinform.* https://doi.org/10.1093/bib/bbac391 (2022).
14. Wang, X. F. *et al.* KGDCMI: A new approach for predicting circRNA-miRNA interactions from multi-source information extraction and deep learning. *Front. Genet.* **13**, 958096. https://doi.org/10.3389/fgene.2022.958096 (2022).
15. Xie, G. B. *et al.* Predicting lncRNA-disease associations based on combining selective similarity matrix fusion and bidirectional linear neighborhood label propagation. *Brief. Bioinform.* https://doi.org/10.1093/bib/bbac595 (2023).
16. Du, X.-X., Liu, Y., Wang, B. & Zhang, J.-F. lncRNA–disease association prediction method based on the nearest neighbor matrix completion model. *Sci. Rep.* **12**, 21653. https://doi.org/10.1038/s41598-022-25730-0 (2022).
17. Zhang, W., Yu, C., Wang, X. & Liu, F. Predicting CircRNA-disease associations through linear neighborhood label propagation method. *IEEE Access* https://doi.org/10.1109/ACCESS.2019.2920942 (2019).
18. Lei, X. & Bian, C. Integrating random walk with restart and k-nearest Neighbor to identify novel circRNA-disease association. *Sci. Rep.* **10**, 1943. https://doi.org/10.1038/s41598-020-59040-0 (2020).
19. Deng, L., Zhang, W., Shi, Y. & Tang, Y. Fusion of multiple heterogeneous networks for predicting circRNA-disease associations. *Sci. Rep.* **9**, 9605. https://doi.org/10.1038/s41598-019-45954-x (2019).
20. Yu, Z., Huang, F., Zhao, X., Xiao, W. & Zhang, W. Predicting drug-disease associations through layer attention graph convolutional network. *Brief. Bioinform.* https://doi.org/10.1093/bib/bbaa243 (2021).
21. Zhou, S., Sun, W., Zhang, P. & Li, L. Predicting pseudogene-miRNA associations based on feature fusion and graph auto-encoder. *Front. Genet.* **12**, 781277. https://doi.org/10.3389/fgene.2021.781277 (2021).
22. Li, J. H., Liu, S., Zhou, H., Qu, L. H. & Yang, J. H. starBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic acids Res.* **42**, D92–D97. https://doi.org/10.1093/nar/gkt1248 (2014).
23. Zheng, L. L. *et al.* dreamBase: DNA modification, RNA regulation and protein binding of expressed pseudogenes in human health and disease. *Nucleic Acids Res.* **46**, D85-d91. https://doi.org/10.1093/nar/gkx972 (2018).
24. Huang, H. Y. *et al.* miRTarBase update 2022: An informative resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res.* **50**, D222-d230. https://doi.org/10.1093/nar/gkab1079 (2022).
25. van Laarhoven, T., Nabuurs, S. B. & Marchiori, E. Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics (Oxford, England)* **27**, 3036–3043. https://doi.org/10.1093/bioinformatics/btr500 (2011).
26. Jiang, L., Ding, Y., Tang, J. & Guo, F. MDA-SKF: Similarity kernel fusion for accurately discovering miRNA-disease association. *Front. Genet.* **9**, 618. https://doi.org/10.3389/fgene.2018.00618 (2018).
27. Chen, X., Zhu, C. C. & Yin, J. Ensemble of decision tree reveals potential miRNA-disease associations. *PLoS Comput. Biol.* **15**, e1007209. https://doi.org/10.1371/journal.pcbi.1007209 (2019).
28. Wei, Z., Yao, D., Zhan, X. & Zhang, S. A clustering-based sampling method for miRNA-disease association prediction. *Front. Genet.* **13**, 995535. https://doi.org/10.3389/fgene.2022.995535 (2022).
29. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* 785–794. https://doi.org/10.1145/2939672.2939785 (2016).
30. Dai, Q. *et al.* Predicting miRNA-disease associations using an ensemble learning framework with resampling method. *Brief. Bioinform.* https://doi.org/10.1093/bib/bbab543 (2022).
31. Zhao, Y., Chen, X. & Yin, J. Adaptive boosting-based computational model for predicting potential miRNA-disease associations. *Bioinformatics (Oxford, England)* **35**, 4730–4738. https://doi.org/10.1093/bioinformatics/btz297 (2019).
32. Gu, M. & Wang, X. Pseudogene MSTO2P interacts with miR-128-3p to regulate coptisine sensitivity of non-small-cell lung cancer (NSCLC) through TGF-β signaling and VEGFC. *J. Oncol.* **2022**, 9864411. https://doi.org/10.1155/2022/9864411 (2022).
33. Guo, M. & Zhang, X. LncRNA MSTO2P promotes colorectal cancer progression through epigenetically silencing CDKN1A mediated by EZH2. *World J. Surg. Oncol.* **20**, 95. https://doi.org/10.1186/s12957-022-02567-5 (2022).
34. Guo, Y. *et al.* Inhibition of AURKB, regulated by pseudogene MTND4P12, confers synthetic lethality to PARP inhibition in skin cutaneous melanoma. *Am. J. Cancer Res.* **10**, 3458–3474 (2020).

Acknowledgements

This research was funded by the Scientific research plan projects of Shaanxi Education Department (Grant No. 21J K0674).

Author contributions

C.F. conceptualized the study, C.F. and M.D. performed the data collection, designed the method, C.F. drafted the manuscript. All authors read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-36054-y.

Correspondence and requests for materials should be addressed to C.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.