Computational and Structural Biotechnology Journal 20 (2022) 1208–1217

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Identification of piRNA disease associations using deep learning

Syed Danish Ali [a,b], Hilal Tayara [c,*], Kil To Chong [a,d,*]

[a] Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju 54896, South Korea
[b] The University of Azad Jammu and Kashmir, Muzaffarabad 13100, Pakistan
[c] School of International Engineering and Science, Jeonbuk National University, Jeonju 54896, South Korea
[d] Advanced Electronics and Information Research Center, Jeonbuk National University, Jeonju 54896, South Korea

ARTICLE INFO

Article history:
Received 2 December 2021
Received in revised form 24 February 2022
Accepted 26 February 2022
Available online 3 March 2022

Keywords:
Sequence analysis
piRNA disease associations
Positive unlabeled learning
Reliable negative sample
Convolutional Neural Network
Deep learning
Web-server

ABSTRACT

Piwi-interacting RNAs (piRNAs) play a pivotal role in maintaining genome integrity by repression of transposable elements, gene stability, and association with various disease progressions. Cost-efficient computational methods for the identification of piRNA disease associations promote the efficacy of disease-specific drug development. In this regard, we developed a simple, robust, and efficient deep learning method for identifying the piRNA disease associations known as piRDA. The proposed architecture extracts the most significant and abstract information from raw sequences represented in a simplicated piRNA disease pair without any involvement of features engineering. Two-step positive unlabeled learning and bootstrapping technique are utilized to abstain from the false-negative and biased predictions dealing with positive unlabeled data. The performance of proposed method piRDA is evaluated using k-fold cross-validation. The piRDA is significantly improved in all the performance evaluation measures for the identification of piRNA disease associations in comparison to state-of-the-art method. Moreover, it is thus projected conclusively that the proposed computational method could play a significant role as a supportive and practical tool for primitive disease mechanisms and pharmaceutical research such as in academia and drug design. Eventually, the proposed model can be accessed using publicly available and user-friendly web tool at http://nsclbio.jbnu.ac.kr/tools/piRDA/.

1. Introduction

piRNAs are the largest subclass among three distinct classes of regulatory small non-coding RNAs (sncRNAs) along with microRNAs (miRNAs) and small interfering RNAs (siRNAs), which are found in several species including vertebrates and invertebrates, specifically in syntenic genomic locations of humans [1–3]. The three main types of ncRNAs despite differences in their mode of target regulation and biogenesis impart certain common functionalities including the guidance of Argonaute proteins to target the nucleic acids in a sequence-dependent manner [4]. Specifically, there are eight Argonaute proteins in humans, together with four Argonaute subfamily proteins (Ago) and four PIWI subfamilies (PIWI) proteins, respectively [5]. The expressed Ago proteins bind to siRNAs and miRNA, where they are transformed in a dicer-dependent mechanism from double-stranded precursors into mature small RNAs of 20–22 nucleotides (nt) [6]; whereas, the PIWI proteins develop a particular RNA-induced silencing complex (RISC), which is known as piRISCs with a small RNA population termed as piRNAs [7]. The long single strand of primary piRNAs is independent of dicer in biogenesis; however, different nucleases are involved for cutting these strands into each piRNA unit [6,8]. The length of each piRNAs sequence varies from 26 to 32 nt [9]. piRNAs are responsible for the self-renewal of the stem cells as they abundantly exist in spermatogenic cells and play a significant role in maintaining germline and genome veracity by concealing the insertional mutations from transposons [10–13].

The involvement of piRNAs in epigenetic silencing of transposons, regulation of gene transcription, histone modification, heterochromatin modification, and DNA methylation appeals researchers to further explore their associations with specific human diseases [14–16]. Moreover, the aberrant expression of piRNAs is associated with the development of various human diseases such as cardiovascular diseases, neurodegenerative disorders together with Alzheimer's disease, Parkinson's disease, malignant

Fully documented templates are available in the elsarticle package on CTAN.

* Corresponding authors at: School of International Engineering and Science, Jeonbuk National University, Jeonju 54896, South Korea (H. Tayara), and Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju 54896, South Korea (K.T. Chong).

E-mail addresses: mashdi.danish@gmail.com (S.D. Ali), hilaltayara@jbnu.ac.kr (H. Tayara), kitchong@jbnu.ac.kr (K.T. Chong).

https://doi.org/10.1016/j.csbj.2022.02.026

tumors, and hallmarks of cancer like augmented stemness, cell proliferation, inhibited apoptosis, and metastasis [17–22], for example, neurodegenerative disorder considering the differential expression of piRNAs in the healthy human brain in comparison to Alzheimer's disease. The diagnosed brain comprises five more than ten-fold upregulated piRNAs including piR-hsa-25781, piR-hsa-28467, piR-hsa-1177, piR-hsa-26593, and piR-hsa-29114 among the 146 upregulated and 3 downregulated piRNAs, which may act as an effective signature for Alzheimer's disease [23]. In reference to cancers, the expression of piR-651 was upregulated in several gastric, lung, breast, mesothelium, liver, and cervical cancer cell lines [24]. Furthermore, piR-823 was remarkably upregulated in colorectal tumorigenesis where it binds with HSF1 while boosting its transcriptional activity and phosphorylation at Ser326 having an active role as a tumor booster [25]. Therefore, the piRNAs are reliable biomarkers associated with the diagnosis and treatment of diseases, which could be facilitated by identifying piRNAs associated with diseases.

In this regard, several piRNA databases [9,26–28], together with efficient and cost-effective web-server based computational predictors for identifying piRNA and their functions, are available [29–31]; whereas, research regarding human disease-associated piRNAs is in its early stages. Recently, the development of piRDisease v1.0 [32], which is a collection of various experimentally verified piRNA-disease associations, allows researchers to develop robust and cost-efficient computational methods in order to identify piRNA-associated diseases [33–37]. Thus, Wei et. al. proposed computational models for the identification of human disease-associated piRNAs together with iPiDi-PUL [33] and iPiDA-sHN [34]. iPiDi-PUL a random forest-based ensemble learning approach used positive unlabeled learning [38] for predicting piRNA disease association, wherein the features for associations were extracted using three dissimilar biological data sources. The negative data for training of model was randomly selected from unlabeled data consists of samples which were not experimentally verified; thus, there was a possibility of positive associations in unlabeled samples, and those samples employed as a negative data could results in low recall or inappropriate decision boundary of a classifier as illustrated in Fig. 2. Although, iPiDi-PUL utilized positive unlabeled learning to assuage low recall or false-negative problem; however, the selection of random negative samples from the unlabeled piRNA disease associations results in compromising the performance of the predictor due to presence of outliers in negative samples. Recently, to mitigate this false-negative obstruction, Wei et al. [34] proposed iPiDA-sHN. A two-step positive-unlabeled learning technique [39] for selection of reliable negative samples from unlabeled piRNA disease associations. Where the three heterogeneous biological sources were combined to describe the piRNA disease-associated features. Moreover, convolution neural network (CNN) was utilized for feature extraction from the multi-source handcrafted disease-associated features. Finally, a Support Vector Machine (SVM) classifier was employed for predicting the piRNA disease association. The employed biological sources for both of the available computational methods include experimentally verified piRNA-disease associations, disease semantic terms, and piRNA sequence information. The shortfall in the fusion of multiple biological data sources as a feature descriptor introduces irrelevant and noisy information. The performance of the computational method could be compromised due to inadequate description of features without tackling redundancy, irrelevant and noisy information.

Consequently, the issues related to manual extraction of features that are highly dependent upon field knowledge need to be addressed. While the deep learning algorithms are extremely efficient and effective in extracting the most significant and abstract features from raw data utilizing the general purpose learning [40]. Moreover, deep learning is also capable of identifying and recognizing the patterns in unstructured data with low-level involvement of manual configuration [41]. Thus, deep learning has breakthroughs in the fields of natural language processing [42], speech recognition [43], image recognition [44], precision agriculture [45–47], potential drug molecules [48], post-translation modifications [49,50], RNA binding proteins [51,52], post-transcriptional modifications [53–55], identification of promoters [56–58], DNA modifications [59–62], and prediction of disease association [63–65]. In the present study, we proposed deep learning architecture piRDA consist of CNN and fully connected layers, CNN is the most commonly used deep learning method considering its efficacy and efficiency in various applications. The CNN-based deep learning architecture is a hierarchical model capable of learning the patterns by utilizing the series of convolutional operations [40]. Fully connected layers are utilized for extracting high level features. For construction of reliable negative data from the unlabeled samples, a two-step positive-unlabeled learning technique [39] was employed to reduce the false-negative rate while predicting piRNA disease association. The raw piRNA sequences are encoded as feature vectors by implementing one-hot encoding technique as an input to CNN where the concealed information of raw piRNA sequences is recognized by CNN. Nevertheless, the disease association for each piRNA is represented with one-dimensional feature vector known as disease association one-hot vector (DAOHV). This is then concatenated with piRNA features extracted by CNN and fed into fully connected neural network layers. These layers extract the piRNA disease association patterns (utilizing multiple levels of abstraction) which leads to high perfor-

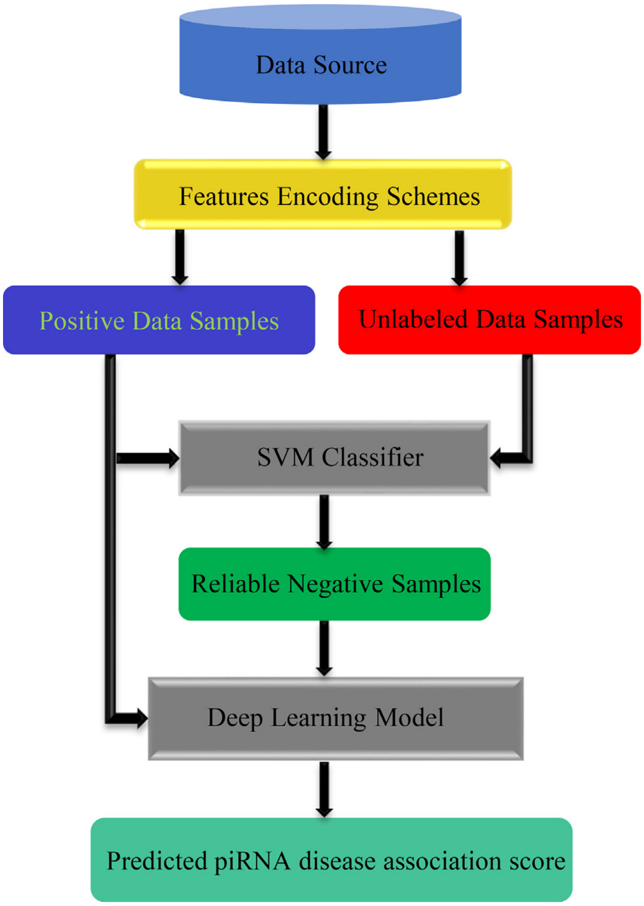
Data Source
Features Encoding Schemes
Positive Data Samples
Unlabeled Data Samples
SVM Classifier
Reliable Negative Samples
Deep Learning Model
Predicted piRNA disease association score

Fig. 1. The overall workflow of proposed Architecture piRDA for identifying piRNA disease associations.

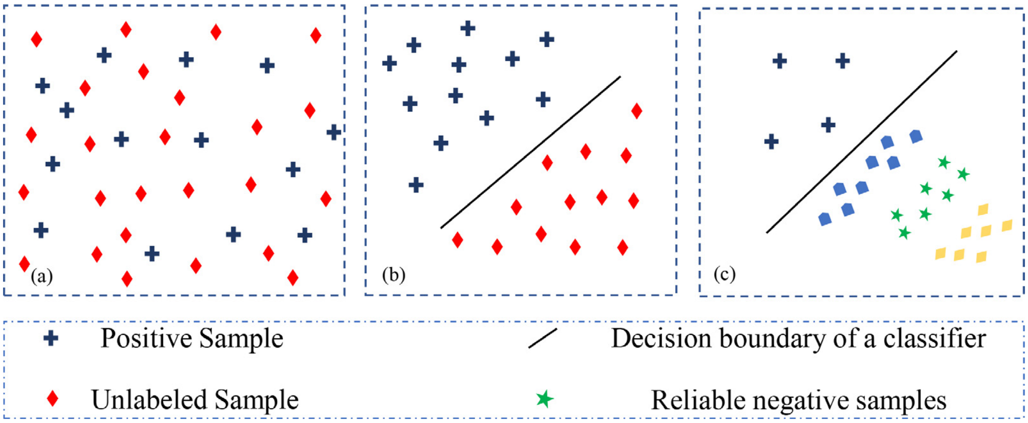

Fig. 2. Illustration of reliable negative selection. (a) Positive and unlabeled data samples. (b) Training with random negative. (c) Unlabeled samples according to their prediction scores.

mance identification of piRNA disease associations without losing any contextual information among piRNAs and diseases. To extenuate the bias of proposed computational method for piRDA toward the majority class in predictions, we used the bootstrapping method [66]. Furthermore, the grid search algorithm was utilized for optimum hyperparameter selection. We utilized the subsampling (k-fold cross-validation) test for comprehensively evaluating the performance, where the proposed architecture of piRDA significantly outperformed the state of the art. Additionally, for the convenience of drug developers, experimental scientists, and considering the importance of webservers in medical sciences research, we developed a publicly available web-server for identifying piRNA associated with disease accessible at http://nsclbio.jbnu.ac.kr/tools/piRDA/. The overall description of the proposed architecture piRDA is illustrated in Fig. 1. The major contributions of the piRDA are enlisted as.

- Novel and simple supervised learning-based representation of sequences and their disease associations.
- Development of a deep learning model for identification of raw piRNA sequences and their associated diseases.
- Achieving significantly high performance in the identification of piRNA disease association.
- Visualization of the feature space learned by piRDA in the prediction of piRNA disease association.
- Development of publicly accessible web-server.

2. Materials and methods

2.1. Dataset construction

piRDisease v1.0 [32] is a manually curated database collection comprising experimentally verified 7939 piRNA disease associations. The redundant and non–human piRNAs were filtered; by extracting the human piRNAs with the piRNA IDs accordingly in piRBase [28]. Eventually, 4350 piRNAs were associated with 21 diseases, thereby providing 5002 experimentally validated disease associations. The benchmark data is the same as introduced and utilized in the literature by Wei et al. [33,34]. Mathematically, the benchmark dataset is described as:

$$D_{all} = D^{P} \cup D^{U} \tag{1}$$

In Eq. (1) D_{all} is the union of all the 4350 piRNAs associated among 21 diseases with 91,350 total numbers of samples. The D^{P} represents positive samples comprising 5002 experimentally validated associations of 4350 piRNAs and 21 diseases, whereas D^{U} represents the unlabeled 86348 piRNA disease pairs among 4350 piRNAs and 21 diseases. The diseases are enlisted in Table 1.

2.2. Proposed methodology

The effective and efficient computational method (piRDA) in terms of computational cost and efficacy is proposed for the identification of disease-associated piRNAs. The overall flow of the proposed study is illustrated in Fig. 1. The flowchart depicts that the proposed computational method comprises three main steps. The first step is simple and effective one-hot feature representation of respective association between piRNAs and diseases. Second, to avoid the false negative rate of the classifier in prediction classes, high-quality reliable negative data samples were selected from the unlabeled dataset. To maintain consistency, fair comparison and generalization the reliable negative data was same as used in the previous study by Wei et al. [34]. Eventually, the features were processed using CNN-based deep learning architecture (piRDA) for identifying the piRNAs associated with the diseases.

Table 1
Summary of piRDA performance for identifying piRNA disease associations using independent piRNA IDs.

No.	Disease	DAOHV
1	Renal cell carcinoma	[1]
2	Lung cancer	[0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
3	Breast cancer	[0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
4	Pancreatic carcinoma	[0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
5	Head and neck (squamous cell) carcinoma	[0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
6	Lung cancer (lung adenocarcinoma)	[0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
7	Alzheimer's disease	[0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
8	Cardiovascular diseases (CDC, CF, CCS) cardiac regeneration	[0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0]
9	Head and neck cancer	[0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0]
10	Gastric cancer	[0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0]
11	Colon cancer	[0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0]
12	Non-small cell lung carcinoma (NSCLC)	[0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0]
13	Prostate cancer	[0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0]
14	Dysplastic liver nodules and hepatocellular carcinoma	[0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0]
15	Rheumatoid arthritis	[0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0]
16	Testicular germ cell carcinoma	[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0]
17	Endometrial carcinogenesis	[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0]
18	Male infertility	[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0]
19	Leukemia	[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0]
20	Heart stroke	[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0]
21	Ovarian cancer	[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1]

The 10-fold cross-validation is used for evaluating the performance of proposed architecture.

2.3. Feature representation

The one-hot encoding was utilized for the representation of each piRNA and disease association as an input to proposed model. One-hot encoding is the most prevalent technique because of its simplicity and effectiveness [67]. The piRNA sequences acquired from piRBase v2.0 [28] are inconsistent in lengths. Therefore, the shorter sequences were padded with dummy variable ”N" for making all the sequences to be equal in their lengths for further processing in CNN. Hence, the one input of piRDA was disease-associated raw piRNA sequence, where $P = \{P_1, P_2, P_3, \ldots, P_n\}$, where j in $\{1, 2, 3, \ldots, n\}, P_j \in A, C, G, U$,and $n = 32$. Each raw piRNA sequence was encoded corresponding to A, U, G,and C as 4 four-dimensional feature vectors [1], [0, 1, 0, 0], [0, 0, 1, 0], [0, 0, 0, 1]. The second input of piRDA included information of the diseases associated with piRNAs. Hence, each disease was represented with one-dimensional feature vectors by assigning a distinct unit vector to each associated disease known as disease association one-hot vector (DAOHV). This one-dimensional simple representation of associated disease directly extracts discriminative information of the disease. DAOHV is a 21 elements vector representing 21 diseases, where only one element for the specific disease would have value “1” and all the other 20 elements would be “0”. The description of DAOHV for diseases is presented in Table 1.

2.4. Positive unlabeled learning

A reliable negative dataset was prepared following the same methods as in previous studies [34,51,68,69]. Therefore, a two-step technique was employed for dealing with positive unlabeled sample datasets [39], wherein, the first step is the identification of reliable negative samples and the second step is to create predictors based upon the positive labeled samples and reliable negative samples [70]. For selecting reliable negative samples and to accomplish the first step of the two-step technique, SVM classifier was employed. The SVM classifier was trained by a random selection of unlabeled samples with the same size as positive samples expressed in Eq. (1). Parameters used for training of SVM are C = 1.0, gamma = 1, and kernel = ’rbf’. The trained SVM was utilized for obtaining the prediction scores from all the unlabeled piRNA disease association samples in Eq. (1). The prediction scores corresponding to the unlabeled samples were sorted in descending order and divided into three clusters of nearly the same size. The second cluster of unlabeled piRNA disease association samples was considered as the reliable negative sample having minimum chance to be the false negative. As the selection of difficult examples made the training more effective for yielding substantial performance boost [71]. The selection process of reliable negative samples is illustrated in Fig. 2.

2.5. Bootstrapping technique

Considering training of the proposed architecture piRDA as the number of disease-associated piRNA or positive samples are less than that of the reliable negative samples. The predictor would be biased towards the majority occurring class; to avoid biases in predictions bootstrapping technique [66] was employed similarly used in literature [72,73] to tackle class imbalance. In this technique, we divided the prepared reliable negative samples into chunks of samples, which are approximately equal to the number of samples as of positive samples disease-associated piRNAs, thereby resulting in five sets of data where each dataset comprises disease-associated piRNAs and a reliable negative chunk. Moreover, k-fold cross-validation was employed; where the value of k is equal to 10 as of state of the art for fair comparison, and keeping consistency among the dataset. The results obtained using k-fold cross-validation are rigorous and unbiased as they are evaluated on k numbers of different test sets [74,75]. Furthermore, 10-fold cross-validation employed divides each dataset into 10 sub datasets, where eight folds were used for training, onefold for the validation, and the remaining onefold for testing of the model. This cognitive operation was repeated 10 times so that each fold was considered to be a distinctive test set. The average of these distinctive subsets was considered to be the final outcomes of each of dataset; whereas, the final outcome of the proposed method piRDA was obtained using the average of all the five sets of data.

2.6. Proposed architecture

The proposed architecture is a two inputs deep learning based computational model, as illustrated in Fig. 3. The model can extract more abstract level of features from the raw data. The main components of the proposed model piRDA are convolutional and dense blocks, which reduce noise and acquire high-level features from raw piRNA sequences and their respective association with the disease. The first input of the proposed architecture was the raw piRNA sequence, which was transformed into a single-dimensional four channel vector as an input to the convolution block, comprising a one-dimensional convolution layer (Conv1D), where multiples filters extract features from the input data by preserving the corresponding spatial information. Therefore, each filter in the Conv1D identifies and extracts the most salient patterns and motifs in raw piRNA sequences [76]. Conv1D used in piRDA comprises 24 filters, where the size of each filter was 7. The Rectified Linear Unit (ReLU) [77] was considered as an activation function for Conv1D, whereas the ReLU activation function is responsible for capturing the nonlinearities and interaction among the feature matrix [78]. Following the ReLU activation, a normalization layer was applied, which acts as a regularizer and is responsible for stabilization of training optimization by substantially confiscating the covariate shift [79]. Therefore, for normalization of feature matrix group normalization (GP) was employed, which is an effective alternative to batch normalization while dealing with small batch size [80]. Where the normalization is performed in groups without employing the batch dimension, the group size of 4 was selected. The one-dimensional max-pooling layer is employed for the features from GP layer, which enhances the ability of generalization by eradicating redundancy and dimensionality. The filter size of 2 along with a stride of 2 was utilized for the one-dimensional max-pooling layer. Flatten layer was utilized after max-pooling layer, which collapses the spatial dimensionality of the extracted feature matrix into a one-dimensional vector. The flattened features are concatenated with the disease associated one hot vector, which is the second input of piRDA using the concatenation layer. Thereafter, final high-level features from the disease-associated piRNA pair were extracted using two fully connected layers having 128 and 32 neurons, respectively. The ReLU is utilized as an activation function for both fully connected layers. L2 regularizer on bias and weight is employed for the fully connected layers, which is the most effective and sophisticated technique to mitigate the overfitting by penalizing larger weights of the model [81]. The value assigned to L2 regularization plenty is $1 \times 10^{(-6)}$. The dropout layer, an effective regularization to avoid overfitting by randomly switching off the effects of neurons [82] was used between the two fully connected layers. The dropout probability for the dropout layer is 0.25. Eventually, these high-level features are fed into the output layer where the sigmoid activation was

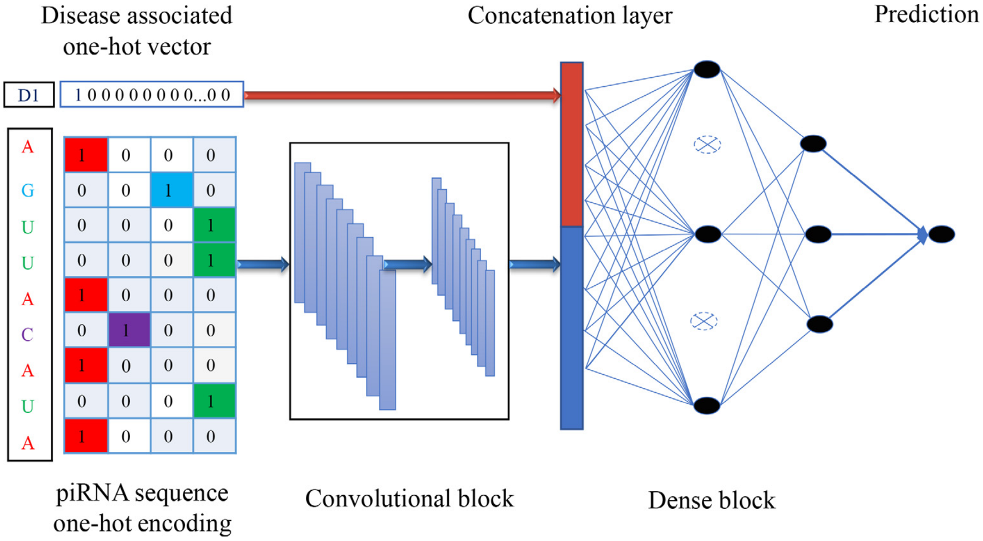

Fig. 3. Illustrating detailed architecture of proposed method piRDA where the convolutional block comprises convolution layer with ReLU as an activation function along with group normalization and max-pooling layers. The dense block consists of two fully connected layers along with dropout probability ReLU as an activation function and sigmoid activation function for prediction associated scores.

employed to assign the prediction scores for the disease-associated piRNA pairs. The mathematical representation of proposed architecture is formulated as:

$$Conv1D(P)_{kl} = ReLU\left(\sum_{s=0}^{S-1}\sum_{n=0}^{N-1} W^{l}_{sn} P_k + s, n\right) \quad (2)$$

Eq. (2) represents one-dimensional convolution layer where Prepresents the raw piRNA sample as an input, $l$is the index of the filter, and $k$is the index of output position. W^lrepresents each of the convolution filters having a weight matrix of $S \times N$dimensions. Sdenotes the size of the filter, whereas Nrepresents the number of input channels.

$$ReLU(x) = max(0, x) \quad (3)$$

Eq. (3) is the representation of ReLU activation function having x as an input.

$$D = ReLU\left(b_{d+1}\sum_{k=1}^{d} m_k w_k z_k\right) \quad (4)$$

Eq. (4) is the representation of a fully connected layer where the additive bias term is denoted by b_{d+1}, m_k is a representation of the dropout operator derived from Bernoulli distribution having with the probability of p, z_k represents the $1 \times$d dimensional feature vector, and w_k represents the previous layer weights of z_k.

$$Sigmoid(x) = \frac{1}{1 + e^{-x}} \quad (5)$$

Eq. (5) denotes the final prediction layer having sigmoid as an activation function and x as an input.

2.7. Model implementation/training

The proposed architecture piRDA was constructed using the Keras framework (https://keras.io/). For optimizing the parameters of the piRDA, the adaptive moment estimation, commonly known as Adam, was used; this is an efficient stochastic optimization method where the magnitude in updates of parameters is unaffected by the rescaling of gradient [83]. The learning rate used for the optimizer was $5 \times 10^{(-4)}$. Moreover, the loss function utilized was binary cross-entropy for computation of the classification loss among the actual labels and the predicted probabilities during training [84]. Furthermore, the early stopping on validation loss was employed to diminish the overfitting. The patience utilized in early stopping was 20, which signifies that the model will stop training if there is no improvement (reduction) in validation loss for 20 epochs. The maximum number of epochs for training were 200 and the batch size was 32. The optimum hyper-parameters of the proposed architecture piRDA are selected using a grid search algorithm known as keras-hypetune (https://github.com/cerlymarco/keras-hypetune). The tuning of hyper-parameters provides a substantial part in selecting an optimal deep-learning model.

3. Results

3.1. Evaluation measures

To evaluate the prediction performance and efficiency of statistical predictors, the most commonly used k-fold cross-validation was utilized. The performance evaluation metrics include accuracy (Acc), sensitivity (Sn), specificity (Sp), Mathew correlation coefficient (Mcc), and rank index (RI). The accuracy is the ratio of correctly classified samples to all samples. The sensitivity and specificity are the proportion of true positive (Tp) and true negative (Tn) respectively. Mcc is a measure of the classifier quality and stability. All of the true positive, false positive (Fp), true negatives, and false negatives (Fn) are considered for evaluating this metric, which results in an effective evaluation in case of class imbalance. Tp is the number of correctly identified positive samples, whereas the number of positive samples predicted as negative samples are known as Fn. Similarly, Tn is the number of accurately identified negative samples and Fp is incorrectly identified negative samples as positive associations. Furthermore, the rank index [33,34,85] is a measure of the identification capacity of the positive association with respect to their ranks in all the piRNA-disease pairs of the test subset. Considering higher values of Acc, Sn, Sp, and Mcc, better the predictor's performance. Conversely, the lower value of the RI metric signifies superior the performance. The evaluation measures can be calculated as follows:

$$Acc = \frac{Tp + Tn}{Tp + Tn + Fn + Fp} \quad (6)$$

$$Sn = \frac{Tp}{Tp + Fn} \tag{7}$$

$$Sp = \frac{Tn}{Tn + Fp} \tag{8}$$

$$Mcc = \frac{Tp \times Tp - Fp \times Fn}{\sqrt{(Tp + Fn)(Tp + Fp)(Tn + Fp)(Tn + Fn)}} \tag{9}$$

$$RI = \frac{1}{|D_{+}^{sub}|} \sum_{as \in D_{+}^{sub}} \frac{R_{as}}{|D_{test}^{sub}|} \tag{10}$$

where $|D_{+}^{sub}|$represents the number of positive test subset associations. R_{as}denotes the positive piRNA disease association rank position among all the pairs of piRNA-disease in the test subset D_{test}^{sub}.

Moreover, the receiver operating characteristic curve (ROC) was utilized for evaluating the success rate of the classifier. ROC is the graphical plot between true positive rate and false positive rate depicting the predictor's performance at all thresholds of classification. Additionally, the precision-recall curve (PRC) is a measure of evaluating the positive class prediction of a classifier. The PRC is plotted between precision and recall on all classification thresholds; where both of these measures, ROC and PRC, are the significant indicators for positive class evaluation. Herein, area under the ROC (AUC) and PRC (AUPRC) signifies the prediction quality of the classifier. Both AUC and AUPRC are the composite metrics of the classifier's success that considers all the potential classification thresholds.

3.2. Model performance

The proposed method piRDA for identifying piRNA-disease association by using the two-step positive unlabeled learning, together with the supervised learning labeling method, where the contextual information of the sequence is contemplated. The piRDA was evaluated by rigorous k-fold cross-validation techniques. The performance of the piRDA by employing the evaluation measures is summarized in Table 2. These outcomes are the average values along with standard deviation error of 50 sub test dataset from piRNA disease and reliable negative sequence datasets, where the values for Acc, Sn, Sp, Mcc, RI, AUC, and AUPRC are 91.32%, 90.89%, 91.80%, 0.827, 0.056, 0.951%, and 0.931%, respectively. Also, the AUC and AUPRC together with standard deviation errors of five folds are illustrated in Fig. 4 and Fig. 5 respectively. Similarly, Fig. 6 and Fig. 7 illustrates the AUC and AUPRC along with standard deviation errors by utilizing 10 sub folds cross-validation.The feature space learned by piRDA was represented using UMAP and is shown in Fig. 8.

Table 2
Summary of performance comparison of piRDA with existing methods for identifying piRNA disease associations.

Metric	piRDA	iPiDA-sHN	iPiDi-PUL
Acc	0.913 ± 0.007	0.736 ±0.020	0.589 ±0.012
Sn	0.909 ± 0.011	0.779 ±0.078	0.281 ±0.027
Sp	0.918 ±0.014	0.694 ±0.080	0.897 ±0.007
Mcc	0.827 ± 0.016	–	–
RI	0.056 ± 0.004	0.307 ± 0.005	0.322 ±0.005
AUC	0.951 ±0.001	0.887 ±0.009	0.856 ±0.009
AUPRC	0.931 ±0.003	0.834 ±0.023	0.764 ±0.014

”-” denotes Not Applicable.

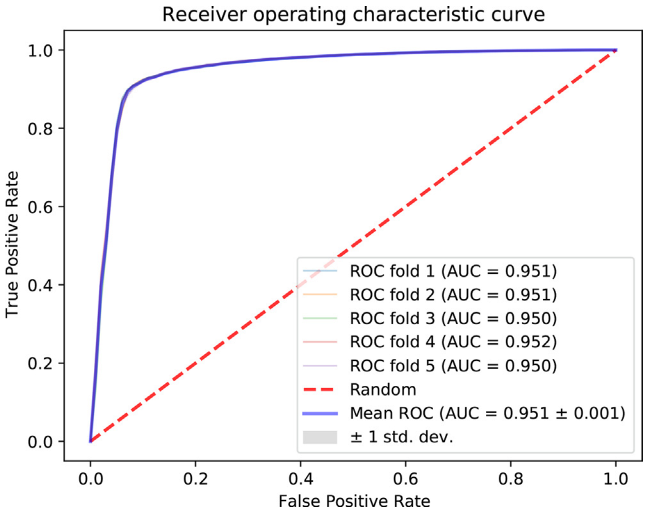

Fig. 4. Illustration of the five folds success rate (ROC), with associated calculation of prediction quality (AUC) and standard deviation error.

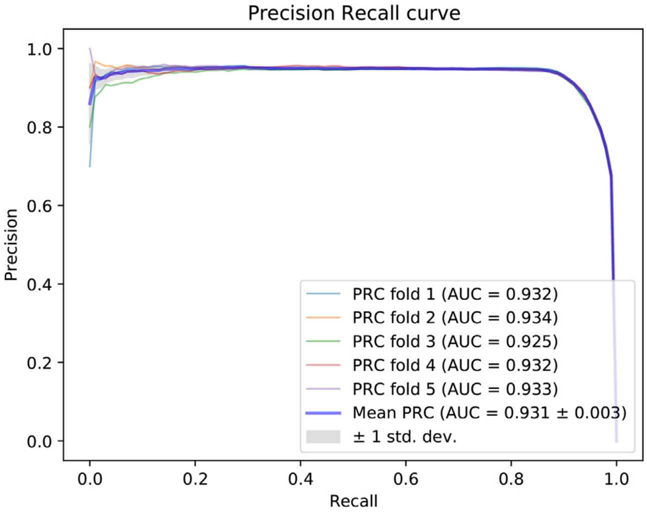

Fig. 5. Illustration of the five folds PRC together with (AUPRC) and standard deviation error.

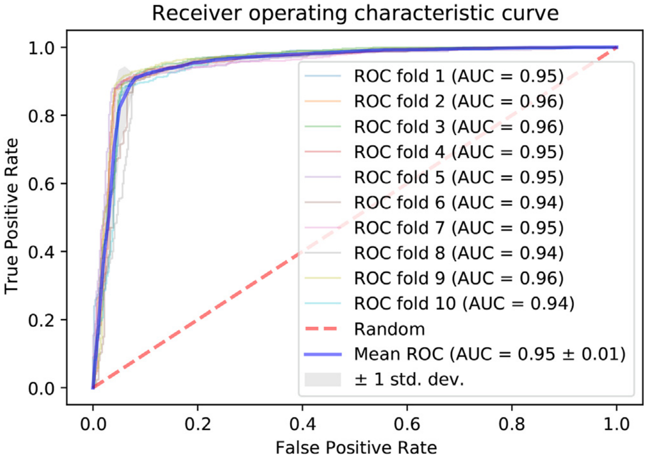

Fig. 6. Illustration of ROC along with AUC and standard deviation error of sub 10-fold cross-validation.

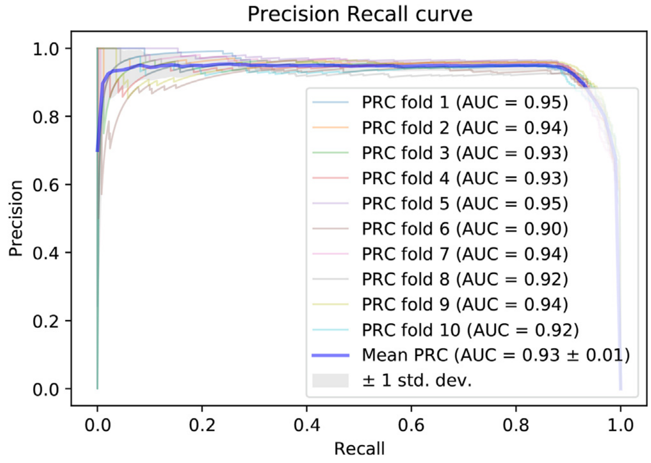

Fig. 7. Illustration of PRC along with AUC and standard deviation error of sub 10-fold cross-validation.

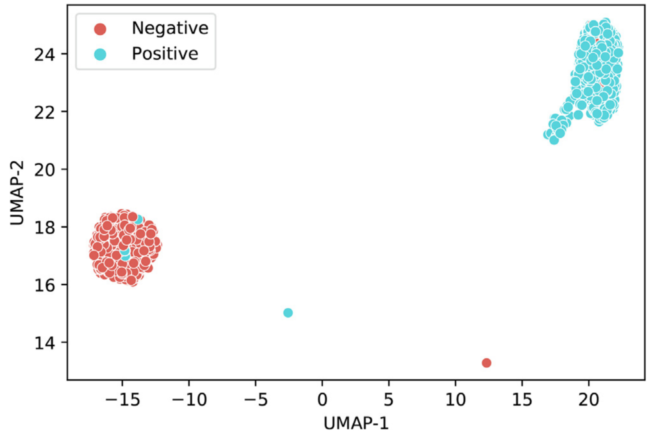

Fig. 8. Clusters of positive and negative piRNA disease associations features of the proposed method obtained from hidden layer activation using UMAP.

3.3. Comparative analysis

To analyze the significance and dominance of the proposed architecture piRDA, we compared the performance with the existing and state-of-the-art methods including iPiDi-PUL [33] and iPiDA-sHN [34] respectively.

- **iPiDi-PUL:** is an ensemble learning-based random forest method where the extracted features include the amalgamation of three biological data sources. The model was trained using positive unlabeled learning, where for the positive set or labeled piRNA disease associations, the equivalent number of negative set was randomly selected from unlabeled piRNA disease associations.
- **iPiDA-sHN:** is an SVM-based classifier where the CNN was used to extract features of computed piRNA similarity and disease similarity from three independent biological sources. Furthermore, SVM-based two-step positive unlabeled learning was employed to construct reliable negative samples from unlabeled data and classification of piRNA disease associations.

The proposed computational method piRDA outperforms all the available relevant computational methods in comparison. The comparison of outcomes in the identification of piRNA disease association along with standard deviation errors are summarized and illustrated in Table 2 and Fig. 9, respectively. The aforementioned comparative results are obtained from the state-of-the-art method iPiDA-sHN [34]. Furthermore, piRDA outstrips state of the art in the performance evaluation measures including Acc, Sn, Sp, RI, AUC, and AUPRC by 17.7, 13.0, 22.4, 25.1, 6.4, and 10.0 percent, respectively.

4. Discussion

The outperformance of proposed method in all evaluation measures signifies that the piRDA is most robust and efficient than the available computational methods in identifying the piRNA disease associations. The efficacy and robustness of the proposed method piRDA are attributable to selection of reliable negative using two-step positive unlabeled learning and DAOHV, a supervised learning representation of the raw piRNAs and their associated disease pairs. This enables the deep learning algorithm to directly extract the most significant and abstract features from the raw inputs without losing the contextual information of the sequences. The multiple levels of abstraction in the deep learning model formulate the possibility to identify the piRNA disease associations more precisely and accurately without being involved in any hand-crafted feature extraction method, whereas the available methods constructed their features matrix by fusing the information of three different biological sources, thereby introducing some noisy information, which leads to misclassification of machine learning based algorithms. Moreover, calculating the similarity matrix for feature representation results in loss of contextual information among the sequences of piRNA and disease pair. Furthermore, biases and false-negative obstruction were diminished by utilizing bootstrapping method, and two steps positive unlabeled learning where the selection of reliable negative associations helps in reducing the false negative problem to the difference of only 1 percent between the Sn and Sp. Which was 8.5 percent in iPiDA-sHN and 61.6 percent following the case of iPiDi-PUL. This drastic difference depicts that piRNA disease associations were inaccurately classified as non-piRNA disease associations. The random selection of negative samples from unlabeled data for training of iPiDi-PUL is responsible to evoke the bias predictions of the classifier.

5. Case study

Evaluation of the proposed method piRDA in reference to the literature regarding piRNAs as potential biomarkers and therapeutic targets of various diseases. We test the proposed method using the experimentally verified piRNAs which were not involved in training of the model. For instance, piRBase ID or NCBI accession number piR-hsa-23317 (DQ593039), piR-hsa-1207 (DQ570956), piR-hsa-27730 (DQ597484), piR-hsa-24016 (DQ593768), piR-hsa-26593 (DQ596377), piR-hsa-29114 (DQ599147) reported in Li et al. [22] showing the highest association for Cardiovascular diseases. piR-hsa-26686 (DQ596470) [86], piR-hsa-20266 (DQ590013) [87] for Renal cell carcinoma, and piR-hsa-25783 (DQ595536), piR-hsa-28467 (DQ598252), piR-hsa-24016 (DQ593768), piR-hsa-2107 (DQ571813), piR-hsa-820 (DQ570540), piR-hsa-515 (DQ570206) [23] for Alzheimer disease. The disease associations for the independent piRNAs is summarized in Table 3.

6. Web-server

The urbanization of a user-friendly and freely accessible web-server accumulating the processes of proposed architecture piRDA is available at http://nsclbio. jbnu.ac.kr/tools/piRDA/. The web ser-

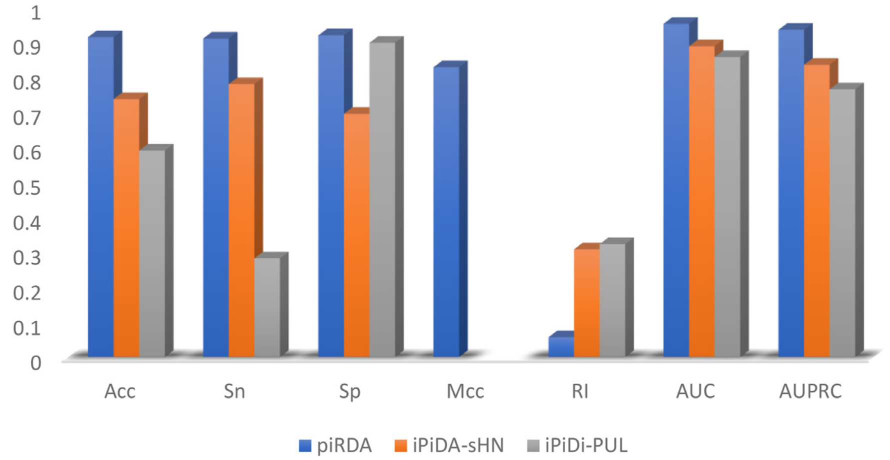

Fig. 9. Illustration of evaluation measures comparision of piRDA with existing methods for identifying piRNA disease associations.

Table 3
Summary of piRDA performance for identifying piRNA disease associations using independent piRNA IDs.

piRNA ID	Association	Reported
piR-hsa-23317 piR-hsa-1207 piR-hsa-24016 piR-hsa-26593 piR-hsa-29114	Cardiovascular diseases	Li et al.[22]
piR-hsa-26686 piR-hsa-20266	Renal cell carcinoma	Wu et al. [86] Fu et al. [87]
piR-hsa-25783 piR-hsa-28467 piR-hsa-24016 piR-hsa-2107 piR-hsa-820 piR-hsa-515	Alzheimer disease	Roy et al. [23]

piRNA ID refers to the piRBase [28].

vers are efficient in maintaining the records of computationally analyzed results. The server is constructed using the python flask web framework. The input in piRNA sequences can be uploaded in FASTA format, whereas the output results in disease associated with the respective piRNA sequence.

7. Conclusion

In this study, we proposed deep learning based computationally efficient and robust algorithm for identifying piRNA disease association. The significantly important features were extracted from disease-associated piRNA without any intervention of hand-designed feature engineering. For constructing a reliable negative dataset and to remove biases of the classifier, two-step positive unlabeled learning and bootstrapping methods were utilized, respectively. The experimental outcomes reveal that the proposed architecture piRDA significantly outperforms the state-of-the-art computational methods for predicting piRNA disease associations. Accurate identification of piRNA disease associations would promote the experimentalists, researchers, and drug developers to further enhance the understanding of mechanism regarding diseases associated with piRNAs. The publicly accessible convenient web tool would be an effective platform to obtain their desired reliable information effectively. Presently, as the research regarding piRNA disease association is in its infancy. Therefore, this model can identify the piRNA disease association of 21 diseases, which could be further enhanced and generalized in future with availability of the verified disease-associated piRNAs.

CRediT authorship contribution statement

Syed Danish Ali: Conceptualization, Methodology, Software, Writing - original draft, Writing - review & editing. **Hilal Tayara:** Conceptualization, Software, Validation, Supervision, Writing - review & editing. **Kil To Chong:** Conceptualization, Validation, Supervision, Writing - review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported in part by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2020R1A2C2005612) and in part by the Brain Research Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. NRF-2017M3C7A1044816).

References

[1] D. Meseure, K.D. Alsibai, Part 1: The piwi-pirna pathway is an immune-like surveillance process that controls genome integrity by silencing transposable elements, in: Chromatin and Epigenetics, IntechOpen, 2018..
[2] Girard A, Sachidanandam R, Hannon GJ, Carmell MA. A germline-specific class of small rnas binds mammalian piwi proteins. Nature 2006;442(7099):199–202.
[3] Tosar JP, Rovira C, Cayota A. Non-coding rna fragments account for the majority of annotated pirnas expressed in somatic non-gonadal tissues. Commun Biol 2018;1(1):1–8.
[4] Ghildiyal M, Zamore PD. Small silencing rnas: an expanding universe. Nat Rev Genet 2009;10(2):94–108.
[5] Sasaki T, Shiohama A, Minoshima S, Shimizu N. Identification of eight members of the argonaute family in the human genome. Genomics 2003;82(3):323–30.
[6] Kim VN, Han J, Siomi MC. Biogenesis of small rnas in animals. Nat Reviews Mol Cell Biol 2009;10(2):126–39.
[7] Iwasaki YW, Siomi MC, Siomi H. Piwi-interacting rna: its biogenesis and functions. Ann Rev Biochem 2015;84:405–33.
[8] Ipsaro JJ, Haase AD, Knott SR, Joshua-Tor L, Hannon GJ. The structural biochemistry of zucchini implicates it as a nuclease in pirna biogenesis. Nature 2012;491(7423):279–83.
[9] Sarkar A, Maji RK, Saha S, Ghosh Z. pirnaquest: searching the pirnaome for silencers. BMC Genom 2014;15(1):1–17.
[10] Kim VN. Small rnas just got bigger: Piwi-interacting rnas (pirnas) in mammalian testes. Genes Develop 2006;20(15):1993–7.

[11] Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, Degnan BM, Rokhsar DS, Bartel DP. Early origins and evolution of micrornas and piwi-interacting rnas in animals. Nature 2008;455(7217):1193–7.

[12] Malone CD, Brennecke J, Dus M, Stark A, McCombie WR, Sachidanandam R, Hannon GJ. Specialized pirna pathways act in germline and somatic tissues of the drosophila ovary. Cell 2009;137(3):522–35.

[13] Siomi MC, Sato K, Pezic D, Aravin AA. Piwi-interacting small rnas: the vanguard of genome defence. Nat Rev Mol Cell Biol 2011;12(4):246–58.

[14] Romano G, Veneziano D, Acunzo M, Croce CM. Small non-coding rna and cancer. Carcinogenesis 2017;38(5):485–91.

[15] Li Y, Wu X, Gao H, Jin JM, Li AX, Kim YS, Pal SK, Nelson RA, Lau CM, Guo C, et al. Piwi-interacting rnas (pirnas) are dysregulated in renal cell carcinoma and associated with tumor metastasis and cancer-specific survival. Mol Med 2015;21(1):381–8.

[16] Assumpção CB, Calcagno DQ, Araújo TMT, Batista dos Santos SE, Ribeiro dos Santos ÂKC, Riggins GJ, Burbano RR, Assumpção PP. The role of pirna and its potential clinical implications in cancer. Epigenomics 2015;7(6):975–84.

[17] Qiu W, Guo X, Lin X, Yang Q, Zhang W, Zhang Y, Zuo L, Zhu Y, Li C-SR, Ma C, et al. Transcriptome-wide pirna profiling in human brains of alzheimer's disease. Neurobiol Aging 2017;57:170–7.

[18] Sun T, Han X. The disease-related biological functions of piwi-interacting rnas (pirnas) and underlying molecular mechanisms. ExRNA 2019;1(1):1–16.

[19] Chalbatani GM, Dana H, Memari F, Gharagozlou E, Ashjaei S, Kheirandish P, Marmari V, Mahmoudzadeh H, Mozayani F, Maleki AR, et al. Biological function and molecular mechanism of pirna in cancer. Practical Labor Med 2019;13:e00113.

[20] Lin Y, Zheng J, Lin D. Piwi-interacting rnas in human cancer. In: Seminars in Cancer Biology Elsevier.

[21] Schulze M, Sommer A, Plötz S, Farrell M, Winner B, Grosch J, Winkler J, Riemenschneider MJ. Sporadic parkinson's disease derived neuronal cells show disease-specific mrna and small rna signatures with abundant deregulation of pirnas. Acta neuropathologica communications 2018;6(1):1–18.

[22] Li M, Yang Y, Wang Z, Zong T, Fu X, Aung LHH, Wang K, Wang J-X, Yu T. Piwi-interacting rnas (pirnas) as potential biomarkers and therapeutic targets for cardiovascular diseases. Angiogenesis 2020:1–16.

[23] J. Roy, A. Sarkar, S. Parida, Z. Ghosh, B. Mallick, Small rna sequencing revealed dysregulated pirnas in alzheimer's disease and their probable role in pathogenesis, Molecular BioSystems 13 (3) 565–576..

[24] Cheng J, Guo J-M, Xiao B-X, Miao Y, Jiang Z, Zhou H, Li Q-N. pirna, the new non-coding rna, is aberrantly expressed in human cancer cells. Clinica chimica acta 2011;412(17–18):1621–5.

[25] Yin J, Jiang X-Y, Qi W, Ji C-G, Xie X-L, Zhang D-X, Cui Z-J, Wang C-K, Bai Y, Wang J, et al. pir-823 contributes to colorectal tumorigenesis by enhancing the transcriptional activity of hsf 1. Cancer Sci 2017;108(9):1746–56.

[26] Sai Lakshmi S, Agrawal S. pirnabank: a web resource on classified and clustered piwi-interacting rnas. Nucl Acids Res 2008;36(suppl_1):D173–7.

[27] Rosenkranz D. pirna cluster database: a web resource for pirna producing loci. Nucl Acids Res 2016;44(D1):D223–30.

[28] Wang J, Zhang P, Lu Y, Li Y, Zheng Y, Kan Y, Chen R, He S. pirbase: a comprehensive database of pirna sequences. Nucl Acids Res 2019;47(D1):D175–80.

[29] Wu W-S, Huang W-C, Brown JS, Zhang D, Song X, Chen H, Tu S, Weng Z, Lee H-C. pirscan: a webserver to predict pirna targeting sites and to avoid transgene silencing in c. elegans. Nucl Acids Res 2018;46(W1):W43–8.

[30] S.D. Ali, W. Alam, H. Tayara, K. Chong, Identification of functional pirnas using a convolutional neural network, IEEE/ACM Transactions on Computational Biology and Bioinformatics..

[31] Liu Y, Li A, Xie G, Liu G, Hei X. Computational methods and online resources for identification of pirna-related molecules. Interdisc Sci: Comput Life Sci 2021;13(2):176–91.

[32] Muhammad A, Waheed R, Khan NA, Jiang H, Song X. pirdisease v1. 0: a manually curated database for pirna associated diseases. Database 2019.

[33] Wei H, Xu Y, Liu B. ipidi-pul: identifying piwi-interacting rna-disease associations based on positive unlabeled learning. Briefings Bioinform 2021;22(3):bbaa058.

[34] Wei H, Ding Y, Liu B. ipida-shn: Identification of piwi-interacting rna-disease associations by selecting high quality negative samples. Comput Biol Chem 2020;88:107361.

[35] Zheng K, You Z-H, Wang L, Li H-Y, Ji B-Y. Predicting human disease-associated pirnas based on multi-source information and random forest. International Conference on Intelligent Computing, Springer 2020:227–38.

[36] Zheng K, You Z-H, Wang L, Wong L, Chen Z-H. Inferring disease-associated piwi-interacting rnas via graph attention networks. International Conference on Intelligent Computing, Springer 2020:239–50.

[37] K. Zheng, Z.-H. You, L. Wang, L. Wong, Z.-H. Zhan, Sprda: a matrix completion approach based on the structural perturbation to infer disease-associated piwi-interacting rnas, bioRxiv..

[38] Mordelet F, Vert J-P. A bagging svm to learn from positive and unlabeled examples. Pattern Recogn Lett 2014;37:201–9.

[39] Liu B, Dai Y, Li X, Lee WS, Yu PS. Building text classifiers using positive and unlabeled examples. In: Third IEEE International Conference on Data Mining IEEE. p. 179–86.

[40] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.

[41] Zhang D, Yin C, Zeng J, Yuan X, Zhang P. Combining structured and unstructured data for predictive models: a deep learning approach. BMC Med Inform Decis Making 2020;20(1):1–11.

[42] Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. J Mach Learn Res 2011;12(ARTICLE):2493–537.

[43] Hinton G, Deng L, Yu D, Dahl GE, Mohamed A-R, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath TN, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Mag 2012;29(6):82–97.

[44] Tayara H, Soo KG, Chong KT. Vehicle detection and counting in high-resolution aerial images using convolutional regression neural network. IEEE Access 2017;6:2220–30.

[45] Ilyas T, Umraiz M, Khan A, Kim H. Dam: Hierarchical adaptive feature selection using convolution encoder decoder network for strawberry segmentation. Front Plant Sci 2021;12:189.

[46] Bauer A, Bostrom AG, Ball J, Applegate C, Cheng T, Laycock S, Rojas SM, Kirwan J, Zhou J. Combining computer vision and deep learning to enable ultra-scale aerial phenotyping and precision agriculture: A case study of lettuce production. Horticulture Res 2019;6(1):1–12.

[47] Khan A, Ilyas T, Umraiz M, Mannan ZI, Kim H. Ced-net: crops and weeds segmentation for smart farming using a small cascaded encoder-decoder architecture. Electronics 2020;9(10):1602.

[48] Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V. Deep neural nets as a method for quantitative structure–activity relationships. J Chem Inform Modeling 2015;55(2):263–74.

[49] Siraj A, Chantsalnyam T, Tayara H, Chong KT. Recsno: Prediction of protein s-nitrosylation sites using a recurrent neural network. IEEE Access 2021;9:6674–82.

[50] Wang D, Liu D, Yuchi J, He F, Jiang Y, Cai S, Li J, Xu D. Musitedeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization. Nucl Acids Res 2020;48(W1):W140–6.

[51] Cheng Z, Huang K, Wang Y, Liu H, Guan J, Zhou S. Selecting high-quality negative samples for effectively predicting protein-rna interactions. BMC Syst Biol 2017;11(2):1–11.

[52] H. Tayara, K. Chong, Improved predicting of the sequence specificities of rna binding proteins by deep learning, IEEE/ACM transactions on computational biology and bioinformatics..

[53] Tahir M, Tayara H, Chong KT. ipseu-cnn: identifying rna pseudouridine sites using convolutional neural networks. Mol Therapy-Nucl Acids 2019;16:463–70.

[54] Ali SD, Kim JH, Tayara H, Chong K. Prediction of rna 5-hydroxymethylcytosine modifications using deep learning. IEEE Access 2021;9:8491–6.

[55] Alam W, Ali SD, Tayara H, Chong K. A cnn-based rna n6-methyladenosine site predictor for multiple species using heterogeneous features representation. IEEE Access 2020;8:138203–9.

[56] Yang Y, Zhang R, Singh S, Ma J. Exploiting sequence-based features for predicting enhancer–promoter interactions. Bioinformatics 2017;33(14):i252–60.

[57] Shujaat M, Wahab A, Tayara H, Chong KT. pcpromoter-cnn: A cnn-based prediction and classification of promoters. Genes 2020;11(12):1529.

[58] Ali SD, Chong KT. Identification of human promoter using convolutional neural network. In: 2021 International Conference on Artificial Intelligence (ICAI) IEEE. p. 213–6.

[59] Yu H, Dai Z. Snnrice6ma: a deep learning method for predicting dna n6-methyladenine sites in rice genome. Front Genet 2019;10:1071.

[60] Wahab A, Ali SD, Tayara H, Chong KT. iim-cnn: Intelligent identifier of 6ma sites on different species by using convolution neural network. IEEE Access 2019;7:178577–83.

[61] Z. Abbas, H. Tayara, K. Chong, Zayyunet a unified deep learning model for the identification of epigenetic modifications using raw genomic sequences, IEEE/ACM Transactions on Computational Biology and Bioinformatics..

[62] Rehman MU, Chong KT. Dna6ma-mint: Dna-6ma modification identification neural tool. Genes 2020;11(8):898.

[63] G. Yu, Y. Yang, Y. Yan, M. Guo, X. Zhang, J. Wang, Deepida: predicting isoform-disease associations by data fusion and deep neural networks, IEEE/ACM Trans. Comput. Biol. Bioinform..

[64] Lu C, Zeng M, Wu F-X, Li M, Wang J. Improving circrna–disease association prediction by sequence and ontology representations with convolutional and recurrent neural networks. Bioinformatics 2020;36(24):5656–64.

[65] M. Zeng, C. Lu, Z. Fei, F. Wu, Y. Li, J. Wang, M. Li, Dmflda: A deep learning framework for predicting incrna–disease associations, IEEE/ACM transactions on computational biology and bioinformatics..

[66] Yan Y, Chen M, Shyu M-L, Chen S-C. Deep learning for imbalanced multimedia data classification, in, IEEE international symposium on multimedia (ISM). IEEE 2015;2015:483–8.

[67] Berry KJ, Mielke Jr PW, Iyer HK. Factorial designs and dummy coding. Perceptual Motor Skills 1998;87(3):919–27.

[68] Wei L, Liao M, Gao Y, Ji R, He Z, Zou Q. Improved and promising identification of human micrornas by incorporating a high-quality negative set. IEEE/ACM Trans Comput Biol Bioinform (TCBB) 2014;11(1):192–201.

[69] Yang P, Li X-L, Mei J-P, Kwoh C-K, Ng S-K. Positive-unlabeled learning for disease gene identification. Bioinformatics 2012;28(20):2640–7.

[70] Bekker J, Davis J. Learning from positive and unlabeled data: A survey. Mach Learn 2020;109(4):719–60.

[71] Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining, in. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 761–9.

[72] Khan SM, He F, Wang D, Chen Y, Xu D. Mu-pseudeep: A deep learning method for prediction of pseudouridine sites. Comput Struct Biotechnol J 2020;18:1877–83.

[73] Wang D, Zeng S, Xu C, Qiu W, Liang Y, Joshi T, Xu D. Musitedeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. Bioinformatics 2017;33(24):3909–16.

[74] Chou K-C, Shen H-B. Recent progress in protein subcellular location prediction. Anal Biochem 2007;370(1):1–16.

[75] Z.-H. You, Y.-K. Lei, L. Zhu, J. Xia, B. Wang, Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis, in: BMC bioinformatics, Vol. 14, Springer, 2013, pp. 1–11..

[76] Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res 2016;26(7):990–9.

[77] Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. in: Icml; 2010.

[78] M. Tsang, D. Cheng, Y. Liu, Detecting statistical interactions from neural network weights, arXiv preprint arXiv:1705.04977..

[79] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning PMLR. p. 448–56.

[80] Wu Y, He K. Group normalization. In: Proceedings of the European conference on computer vision (ECCV). p. 3–19.

[81] Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning, vol. 1. MIT press Cambridge; 2016.

[82] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15(1):1929–58.

[83] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980..

[84] De Boer P-T, Kroese DP, Mannor S, Rubinstein RY. A tutorial on the cross-entropy method. Ann Oper Res 2005;134(1):19–67.

[85] G. Tsoumakas, I. Katakis, I. Vlahavas, Mining multi-label data, in: Data mining and knowledge discovery handbook, Springer, pp. 667–685..

[86] Wu X, Pan Y, Fang Y, Zhang J, Xie M, Yang F, Yu T, Ma P, Li W, Shu Y. The biogenesis and functions of pirnas in human diseases. Mol Therapy-Nucl Acids 2020;21:108–20.

[87] Fu A, Jacobs DI, Hoffman AE, Zheng T, Zhu Y. Piwi-interacting rna 021285 is involved in breast tumorigenesis possibly by remodeling the cancer epigenome. Carcinogenesis 2015;36(10):1094–102.