Functional & Integrative Genomics (2025) 25:9
https://doi.org/10.1007/s10142-025-01527-y

RESEARCH

pmiRScan: a LightGBM based method for prediction of animal pre-miRNAs

Amrit Venkatesan[1] · Jolly Basak[3] · Ranjit Prasad Bahadur[1,2]

Received: 28 October 2024 / Revised: 3 December 2024 / Accepted: 1 January 2025

Abstract
MicroRNAs (miRNA) are categorized as short endogenous non-coding RNAs, which have a significant role in post-transcriptional gene regulation. Identifying new animal precursor miRNA (pre-miRNA) and miRNA is crucial to understand the role of miRNAs in various biological processes including the development of diseases. The present study focuses on the development of a Light Gradient Boost (LGB) based method for the classification of animal pre-miRNAs using various sequence and secondary structural features. In various pre-miRNA families, distinct k-mer repeat signatures with a length of three nucleotides have been identified. Out of nine different classifiers that have been trained and tested in the present study, LGB has an overall better performance with an AUROC of 0.959. In comparison with the existing methods, our method 'pmiRScan' has an overall better performance with accuracy of 0.93, sensitivity of 0.86, specificity of 0.95 and F-score of 0.82. Moreover, pmiRScan effectively classifies pre-miRNAs from four distinct taxonomic groups: mammals, nematodes, molluscs and arthropods. We have used our classifier to predict genome-wide pre-miRNAs in human. We find a total of 313 pre-miRNA candidates using pmiRScan. A total of 180 potential mature miRNAs belonging to 60 distinct miRNA families are extracted from predicted pre-miRNAs; of which 128 were novel and are note reported in miRBase. These discoveries may enhance our current understanding of miRNAs and their targets in human. pmiRScan is freely available at http://www.csb.iitkgp.ac.in/applications/pmiRScan/index.php.

Keywords Non-coding RNA · pre-miRNA · micro-RNA · Machine learning · Taxonomic groups

Introduction

MicroRNAs (miRNAs) play significant roles in post-transcriptional gene regulation. miRNAs are non-coding RNA (ncRNA) molecules that influence several biological processes including tumor growth, cell survival, proliferation and apoptosis (Ganju et al. 2017). Animal microRNAs are typically 22 nucleotides (nts) long with a range between 19 nt and 24 nt. The antecedents of miRNAs are stem-loop intermediates, which are about 60 nt to 70 nt long and are known as precursor miRNAs (pre-miRNAs). pre-miRNAs are generated through the nuclear cleavage of primary miRNAs (pri-miRNAs) by the action of Drosha in complex with its cofactor DGCR8 (DiGeorge syndrome critical region 8) (Bartel 2004). Mature miRNAs are in turn generated by the cleavage of pre-miRNAs by Dicer (Bartel 2004). miRNAs can be categorized as canonical miRNAs and mirtrons. Canonical miRNAs are generated through the cleavage of pri-miRNA and pre-miRNAs by Drosha and Dicer, respectively. Mirtrons are spliced from the intronic regions of the pre-mRNA transcripts by spliceosome (Titov and Vorozheykin 2018). The first discovered pre-miRNA was lin-4, which regulates the lin-14 mRNA (Lee et al. 1993). Several studies are now focussing on the identification of sequences that can serve as potential pre-miRNAs (Lee et al. 1993). Identifying pre-miRNAs and miRNAs is critical to understand their roles in regulating gene expression.

✉ Ranjit Prasad Bahadur
r.bahadur@bt.iitkgp.ac.in

1 Computational Structural Biology Lab, Department of Bioscience and Biotechnology, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal 721302, India

2 Bioinformatics Centre, Department of Bioscience and Biotechnology, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal 721302, India

3 Genomics of Plant Stress Biology Lab, Department of Biotechnology, Visva-Bharati, Santiniketan, West Bengal 731235, India

Published online: 09 January 2025

Experimental methods for identifying pre-miRNA and miRNAs are expensive and time consuming (Yones et al. 2018; Fu et al. 2019). This can be complemented by developing computational methods for their genome-scale identification. Compared to a large number of hairpin-like sequences present in a whole genome, the known pre-miRNAs deposited in databases such as MirGeneDB(Fromm et al. 2022) and miRBase (Griffiths-Jones 2006) are substantially low (Bugnon et al. 2021). At the sequence level, ncRNAs including pre-miRNAs and miRNAs differ from the protein-coding genes. Distinguishing pre-miRNAs from other ncRNAs is challenging since other ncRNAs are also capable of forming similar stem-loop structures (Xue et al. 2005). However, pre-miRNAs form a stable secondary stem-loop structure with higher number of base pair in their stem. On the other hand, ncRNAs including circRNAs and lncRNAs can attain multiple secondary structure conformations due to their increased length. In addition, stem loop in pre-miRNA contains a 2 nts 3′ overhang, which facilitates the binding of DICER (Jouravleva et al. 2022). Consequently, computational pipelines can be used to identify the characteristic features of pre-miRNA for their classification and prediction. Sequence conservation-based methods compare known miRNAs with their orthologs in different species. When considering metazoans, miRNAs of *Fugu rubripes* (puffer fish) and *Danio rerio* (zebra fish) share homology with human and mice miRNAs (Chen et al. 2005). Machine learning-based methods leverage characteristic attributes of pre-miRNAs and miRNAs to identify novel pre-miRNAs and miRNAs. Around 30 miRNA families are shared in all bilaterian animals comprising the major group of metazoans (Praher et al. 2021). Utilizing sequencing data and a probabilistic model of miRNA synthesis, experimental data-driven methodologies assess the compatibility between the position and frequency of sequenced RNA and the secondary structure of the pre-miRNA (Nazarov and Kreis 2021). These techniques are also used to identify miRNA signatures from miRNA-mediated regulatory networks to predict the prognosis of various diseases (Vafaee et al. 2018). Machine learning techniques are frequently used to classify the pre-miRNAs from other pseudo pre-miRNAs of similar length, which are also capable of forming stem-loop structures (Ma et al. 2018). Furthermore, it has been discovered that different categories of ncRNAs share common patterns with hairpin secondary structures (Hertel and Stadler 2006; Batuwita and Palade 2009). Support Vector Machine (SVM) based classifiers including MicroPred, triplet-SVM, YamiPred, HuntMi and miRNAss utilize sequence and structural information for pre-miRNA classification (Batuwita and Palade 2009; Kleftogiannis et al. 2015; Stegmayer et al. 2019). mirExplorer is a miRNA classification tool based on AdaBoost for detecting miRNAs in next-generation sequencing data (Guan et al. 2011). piRNApred server, a SVM based approach to predict piRNAs, uses features including k-mers, thermodynamic parameters and sequence-structure triplet elements, which were first described in triplet-SVM (Xue et al. 2005). The triplet-SVM features influence the folding of RNA and stabilizes their secondary structure configuration. These features are also used widely in pre-miRNA prediction tools such as MicroPred (Batuwita and Palade 2009), HuntMi (Gudyś et al. 2013), MiPred (Jiang et al. 2007) and MiReval (Ritchie et al. 2008). Several machine learning classifiers have been used for the prediction of pre-miRNAs and miRNAs (Parveen et al. 2020). Although SVM and Naive Bayes classifiers are powerful for binary classification, yet their performance tend to decrease when dealing with large-scale datasets in terms of both speed and accuracy. Boosting algorithms including AdaBoost struggle in dealing with data characterized by non-linear relationships. XGBoost, being robust and sclable, uses more memory and becomes computationally intensive with large-scale datasets. LightGBM, a histogram-based learning approach, is capable of efficiently handling biological datasets that possess high complexity and dimensionality (Wang et al. 2017; Liang et al. 2021).

In addition, several Deep learning tools including dnnPremiR-master, cnnMirtronPred-master and DeepMir-master have been developed for the human pre-miRNA classification (Zheng et al. 2020). In recent studies, deep learning classifiers including deepBN, bb-DeepMir, deeSOM and MirDeep gained significant importance and are being widely used for genome-wide classification of pre-miRNAs (Raad et al. 2022). Deep learning algorithms require large scale dataset and extensive training for making a generalized prediction of unknown data. Moreover, the nature of the decision making process of deep neural networks is difficult to comprehend. LightGBM requires less training times and lesser amount of labelled data for efficient classification (Wang et al. 2017; Ponsam et al. 2021).

The present study incorporates sequence and structural features including mononucleotide, dinucleotide, k-mer content, AU content, GC content, normalized base pairing propensity (Npb), minimum folding energy (MFE), normalized base pairing distance (ND) and normalized Shannon entropy (NQ) to effectively classify pre-miRNAs. Two major intrinsic properties are the AU content and the length of miRNA, which influence the ability to fold a secondary structure through canonical and non-canonical base pairing (Barik and Das 2018; Amin et al. 2019; Nithin et al. 2022). We have trained and tested nine distinct classifiers using the selected features. The LGB classifier outperforms all other classifiers with the highest accuracy (0.93), sensitivity (0.86), specificity (0.95), F-score (0.82) and AUC (0.959). The developed classifier will contribute to the advancement

of our understanding of pre-miRNAs by improving their genome scale identification across different animal species. We have implemented the classifier into a web server 'pmiRScan', which is freely accessible at http://www.csb.iitkgp.ac.in/applications/pmiRScan/index.php.

Materials and methods

Dataset construction

pre-miRNA sequences were retrieved from miRbase v22 (Griffiths-Jones 2006). miRBase contains pre-miRNA (hairpin loops) sequences, their genomic location and the mature miRNA sequence contained in them (Kozomara et al. 2019). We have classified animal pre-miRNAs from the pseudo-hairpins. The positive set includes animal pre-miRNAs and the negative set includes sequences obtained from the coding regions of different metazoan genomes. Coding regions serve as reliable negative dataset (Mendes et al. 2009). Both pre-miRNAs and coding regions are capable of forming stem-loop secondary structures. The extent of base pairing in the stem region of a coding sequence is considerably lower in comparison to the stem region of a pre-miRNA. This corresponds to a stable conformation and lower folding energy of pre-miRNA. In addition, pre-miRNAs contain specific motifs such as 'GUG' or 'GGU' near the loop region (Takashima et al. 2022). These motifs facilitate the binding of Drosha to generate pre-miRNAs from pri-miRNAs. Since the study focusses on classifying animal pre-miRNAs, we have considered the metazoan pre-miRNA sequences. Curation of the dataset was done by eliminating ambiguous characters like "N" and degenerate letters for bases. Recurrent sequences were removed from the dataset using CD-HIT with a sequence identity threshold of 0.8 and query coverage of 0.9 (Li and Godzik 2006).

Positive dataset

miRBase v22 contains 38,589 pre-miRNA sequences, of which 29,562 are metazoan pre-miRNAs from 148 different species. After redundancy removal, 16,364 sequences were retained. RNAfold in Vienna 2.5 (Lorenz et al. 2011) was used to obtain the secondary structures of human pre-miRNA. RNAfold is a well-recognized and widely used tool for the prediction of RNA secondary structures. It uses thermodynamic models to generate ensemble of secondary structures from which a consensus secondary structure is obtained. Moreover, the versatility of RNAfold in handling different types of RNA, including mRNA, makes it more useful than the other secondary structure prediction tools (Gardner and Giegerich 2004). We set the temperature at 37 °C and used 'noLP' to generate secondary structures free from dangling ends.

Negative dataset

We used the sequences retrieved from coding regions of six animal species available in NCBI (https://www.ncbi.nlm.nih.gov/refseq). The intronic and the non-coding exonic regions were removed from the coding sequences using BEDtools (Quinlan and Hall 2010). Moreover, we confirmed the absence of any miRNA or pre-miRNA sequences in the coding sequences comprising the negative training set. Sequences with lengths similar to that of pre-miRNAs were extracted followed by the removal of redundant sequences. The final negative dataset contains 22,337 mRNAs.

We have used 80:20 split ratio to construct training and testing datasets. Stratified splitting was used to maintain class balance. The training set contains 12,931 and 17,124 instances of the positive and the negative datasets, respectively. The test set contains 3233 positive and 4281 negative instances. The remaining 201 positive and 932 negative instances were considered as the validation set, which was not used during the training of the different classifiers.

Feature extraction

The length of pre-miRNA is one of the important features used in the prediction model. In our dataset, the length of pre-miRNAs varies from 50 nts to 150 nts, which covers the 99% probability range of the length distribution. The length of the coding sequences was also kept in similar range in order to match the sequence and secondary structure characteristics of pre-miRNAs. This improves the predictive performance and reduces false positives. In the current study, we extracted both sequence and structural features along with triplet-SVM features (Xue et al. 2005) for the classification of pre-miRNAs. We have used four mononucleotides, sixteen di-nucleotides and sixty-four k-mer of length three nucleotides as the sequence features. In order to account for the variability in pre-miRNA length, k-mers were normalised per 100 nucleotides (Nithin et al. 2015, 2017, 2022).

$$R = \frac{n_{kmer}}{L} \times 100 \quad (1)$$

Here, n_{kmer} is the number of k-mer signatures present in the pre-miRNA sequence. L is the length of the pre-miRNA sequence.

Secondary structures of the pre-miRNA sequences were calculated using RNAfold with default parameters. The

genRNAstats program was used for the calculation of the RNA folding measures (Nithin et al. 2015).

$$NMFE = -\frac{(MFE \text{ x } 100)}{L} \tag{2}$$

Here, L is the sequence length. The base pair probability distribution (BPPD) was used for calculating ND and NQ. The base pair probability p_{ij} between the bases of i and j was calculated using the MaCaskill algorithm described by the following equations (Lorenz et al. 2020).

$$p_{ij} = \sum_{S_\alpha \in S(s)} P(S_\alpha)\, \delta_{ij} \tag{3}$$

$$P(s_\alpha) = \frac{e^{\frac{-E}{RT}}}{\sum_{S_\alpha \in S(s)} e^{\frac{-E}{RT}}} \tag{4}$$

$$\delta_{ij} = \begin{cases} 1, x_i \, pairs \, x_j \\ 0, \, otherwise \end{cases} \tag{5}$$

$$NQ = \frac{-1}{L} \sum_{i<j} p_{ij} \cdot log(p_{ij}) \tag{6}$$

$$ND = \frac{-1}{L} \sum_{i<j} p_{ij}(1 - p_{ij}) \tag{7}$$

Features of triplet-SVM are represented by the base pairing of three consecutive nucleotides following a base in a stem. These features are represented by a dot and parentheses such as 'A(.(', where a '.' represents an unpaired base and a '(' represents a paired base. The features are derived from RNAfold notation for secondary structures. These give 32 features in all (Xue et al. 2005). In addition to the above features, ensemble diversity, ensemble energy and the number of base pair in each stem in case of multi-loop pre-miRNAs were also calculated. Finally, a total of 125 features were extracted.

Feature selection

Selecting the best features to distinguish the classes is a key aspect of machine learning. Feature selection methods can be classified as embedded, wrapper and filter-based (Stańczyk 2015; Chen et al. 2020). The embedded method determines the interaction between the features and the target variable subject to the learning process of a model. The feature subsets in this method differ based on the model. Wrapper employs different feature subsets for model training and the evaluation of feature subsets is done based on the model performance. This method is considerably slower than the embedded method. Filter measures the dependency of a variable based on the target variable. This method is much faster and more robust against overfitting. We have used two different feature selection methods: embedded and filter. In embedded feature selection, Least Absolute Shrinkage and Selection Operator (LASSO) regression(Ranstam and Cook 2018) was used. In filter feature selection, mutual information was used (Vergara and Estévez 2014). LASSO shrinks the features in linear combination to a probability space between 0 and 1. This method minimizes the sum of squared residuals using a penalizing factor. The LASSO regression can be written as:

$$L_{lasso}\left(\widehat{\beta}\right) = \sum_{i=1}^{n}\left(y_i - x_j^T \widehat{\beta}\right)^2 + \lambda \sum_{j=1}^{m} \left|\widehat{\beta}_j\right| \tag{8}$$

where, $\left|\widehat{\beta}_j\right|$ is the absolute value of the slope and $\lambda \in (0, \infty)$. Mutual information is calculated using the joint probability to determine the similarity between a feature and the target variable.

$$\sum_{l \in l} \sum_{m \in m} p(l, m)\, log\left[\frac{p(l, m)}{p(l)\, p(m)}\right] \tag{9}$$

Here, l and m refer to the feature and target variable, respectively. Embedded learning approaches other than LASSO include Ridge regression and Elastic Net (Gonzales and De Saeger 2018; Ranstam and Cook 2018). LASSO introduces a 'L1' regularization penalty to the loss function (refer to Eq. 8). It affects the loss function by shrinking the coefficient $\widehat{\beta}_j$ to absolute zero, thereby helping to remove the insignificant features. On the other hand, Ridge uses 'L2' regularization, which reduces coefficients but does not set them to zero, resulting in the retention of unimportant features. In addition, LASSO enhances the model simplicity in comparison to the Elastic Net, which uses a combination of 'L1' and 'L2' penalty (Pudjihartono et al. 2022). Mutual information has the ability to handle a wide range of data types and to capture non-linear relationships between the feature variables. Other filter-based approaches including ANOVA and correlation-based feature selection are limited by their capabilities to handle only specific data type and capture only the linear relationship (Nasiri and Alavi 2022; Pudjihartono et al. 2022). In the present study, application of mutual information generates a subset of features followed by LASSO regression. The feature subset resulting from the above feature selection methods were finally considered for training classifiers. We selected 28 distinguishing features that best classified the animal pre-miRNAs from the pseudo-hairpins.

Class imbalance

One of the most common issues in machine learning is class imbalance. In practice, the methods to remove class imbalance in the dataset are under-sampling and over-sampling (Niaz et al. 2022). Under-sampling refers to a reduction of majority class samples to match with minority class samples. This improves the prediction accuracy for the minority class. However, under-sampling can result in the loss of important information in the majority class. Conversely, over-sampling aims to augment the class size of minority groups in order to prevent the loss of knowledge. Synthetic minority over-sampling technique (SMOTE) constructs synthetic examples of the minority class, thus minimizing the possibility of overfitting (Torgo et al. 2013; Fernandez et al. 2018). In this study, our training set contains 16,364 positive and 22,338 negative instances, thus suffering a class imbalance. SMOTE was applied to remove the class imbalance between positive and negative instances with a random state of 5 and k-neighbour value of 5.

Classifier selection

Several classifiers were trained and tested on the dataset following the feature selection and class imbalance removal. The scikit-learn module in python was used for implementing the classifiers (Bisong 2019). To compare the performance metrics, the following nine classifiers were trained: SVM with RBF kernel, XGBoost, Random Forest, Gaussian Naive Bayes, AdaBoost, LGB, Gradient Boost, Decision Tree and k-Nearest Neighbor (kNN) clustering (Natekin and Knoll 2013; Kotsiantis 2013; Suthaharan 2016; Chen and Guestrin 2016; Rigatti 2017; Zhao et al. 2021; Solomatine and Shrestha 2004). We have also trained boosting classifiers since they are capable of performing binary classification. The nine classifiers tested in this study belong to various classes of machine learning algorithms. Tree-based methods such as LGB, Random Forest, Gradient Boost, XGBoost, Decision Tree and AdaBoost split the data using trees based on the threshold values of features that represent nodes. This allows them to capture feature interactions effectively. Gaussian Naive Bayes (GNB) is a probabilistic approach that simplifies prediction considering features as conditionally independent. GNB can thus handle large number of features. SVM and kNN clustering are distance-based approaches, which requires minimal assumptions. They are capable of dealing with non-linear relationships between the features in a dataset. The training dataset used in this study contains a large number of different class of features that possess non-linear relationships. In order to account for such complexity, the authors have considered the aforementioned nine different classifiers. The performance of each classifier was evaluated using the following metrics.

$$Sensitivity = \frac{TP}{TP + FN} \tag{10}$$

$$Specificity = \frac{TN}{TN + FP} \tag{11}$$

$$F - score = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{12}$$

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{13}$$

The LGB classifier is implemented in the web server since it outperformed all the other classifiers in terms of performance metrics. We have also performed a 10-fold cross-validation, which resulted in an overall accuracy of ~95%.

Comparison with the existing classifiers

The existing classifiers including microPred, miPred and triplet-SVM are SVM-based classifiers (Suthaharan 2016). SVMs are known for their capability to handle high-dimensional data, robustness to overfitting and efficient handling of data imbalance. However, they are slower for very large datasets in comparison to other classifiers. On the other hand, classifiers based on deep learning such as deepPremiR, mirDNN and cnnmiRtron are comparatively faster (Zheng et al. 2020; Yones et al. 2021; Tasdelen and Sen 2021).

Computational validation

For computational validation, we retrieved pre-miRNA sequences of four different taxonomical groups of metazoans including mammals, arthropods, molluscs and nematodes from RNAcentral (Petrov et al. 2017). RNAcentral is a ncRNA database comprising sequences from several other databases including miRBase, piRBase, Rfam, Ensembl and NONCODE. We removed the pre-miRNA sequences, which are also present in miRBase. Furthermore, we retrieved the coding sequences from RefSeq for the four taxonomical groups. Finally, we tested our classifier on four different dataset constructed using a subset of positive and negative instances from RNAcentral and RefSeq, respectively.

Prediction of genome-wide human pre-miRNAs

The classifiers constructed in this study were used for the genome-wide prediction of human pre-miRNAs. We

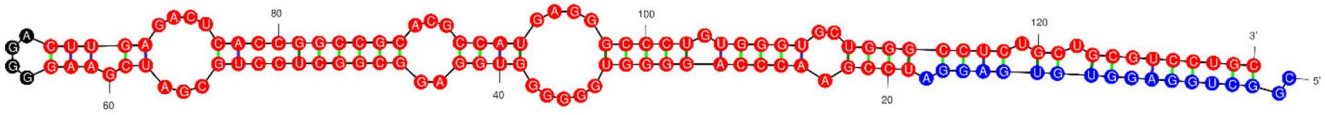

Fig. 1 Secondary structure of human pre-miRNA hsa-mir-3652. The black region shows the loop and the blue region shows the matured miRNA

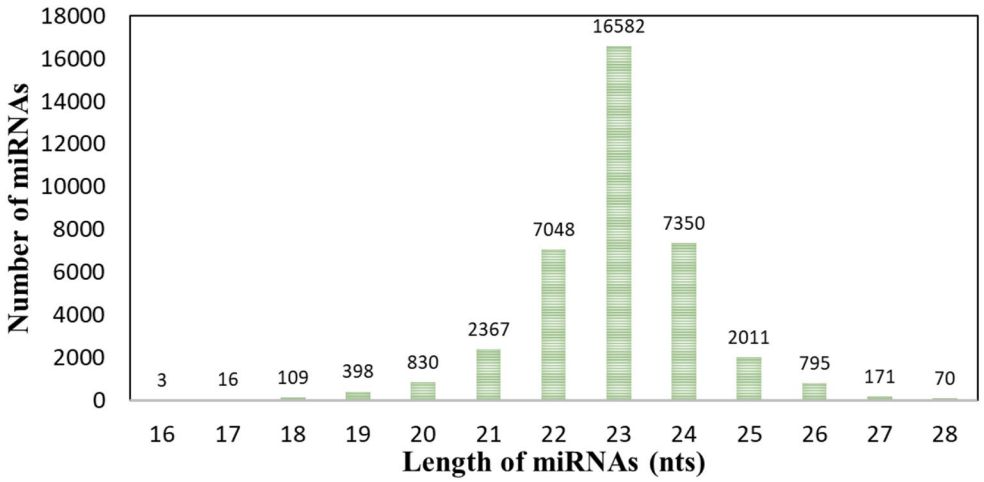

Fig. 2 Distribution of the length of metazoan miRNAs

retrieved 1917 human pre-miRNAs and 2656 human miRNAs from miRbase v22. After removing the redundancy, we retained 1901 pre-miRNAs and 2157 miRNA sequences. Human EST and GSS sequences were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/). The protein sequences were obtained from UniProtKB (https://www.uniprot.org/), which contains 204,063 human protei n sequences. After removing the redundancy, we retained 44,856 sequences. Non-redundant human pre-miRNAs were searched as a query using BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) with the non-redundant human EST and GSS sequences as subjects, followed by the extraction of upstream and downstream regions of the resultant sequences between lengths 50 nt to 150 nt. In order to exclude the protein-coding sequences, an un-gapped BLASTX with a sequence identity cut-off $>80\%$ was performed using extracted sequences as query and non-redundant protein sequences of human as subject. To identify the presence of mature miRNAs, human non-redundant mature miRNAs were searched in the remaining non-coding sequences using BLAST. The resulting non-coding sequences were classified using pmiRScan to predict putative human pre-miRNAs.

Results and discussion

pre-miRNAs are critical biomolecules and mature into miRNAs. Some pre-miRNAs give rise to multiple miRNAs. Several features are used in this study to classify pre-miRNAs. pre-miRNAs are distinguished by a stem-loop structure, which may also include internal loops, bulges and short overhangs. The matured miRNA is located at the arms of a stem-loop structure (Zhu et al. 2011). Figure 1 shows the secondary structure of the human pre-miRNA hsa-mir-3652 having a length of 131 nts.

Sequence length and AU content

Previous studies show that the AU-rich elements (ARE) are organized into several classes and clusters at the 3' untranslated regions (3' UTRs) of mRNAs. In the early response genes such as lymphokines and cytokines, a characteristic ARE "UUAUUAUU" is responsible for inflammatory response and destabilization of mRNA (Garg et al. 2021). A significant number of human protein-coding genes contain AREs. The mRNA decay is initiated by miRNAs and lncRNAs, which recognize ARE in mRNA. The decay proceeds by poly(A) tail shortening followed by the mRNA degradation (Rissland et al. 2017). However, the mechanism by which miRNAs recognize the AREs remains elusive. The length distribution of metazoan miRNA is shown in Fig. 2. Considering the 99% probability range, the AU content in pre-miRNAs varies from 20% to 80% (Fig. 3a). The pre-miRNA length varies from 45 nt to 150 nt (Fig. 3b) when considering the 99% probability range.

Secondary structure

Compared to other ncRNAs or coding RNAs, most of the nucleotides in pre-miRNAs form base pairs leading to the formation of a stem-loop structure. The existence of a matured miRNA in one of the arms of a stem-loop structure

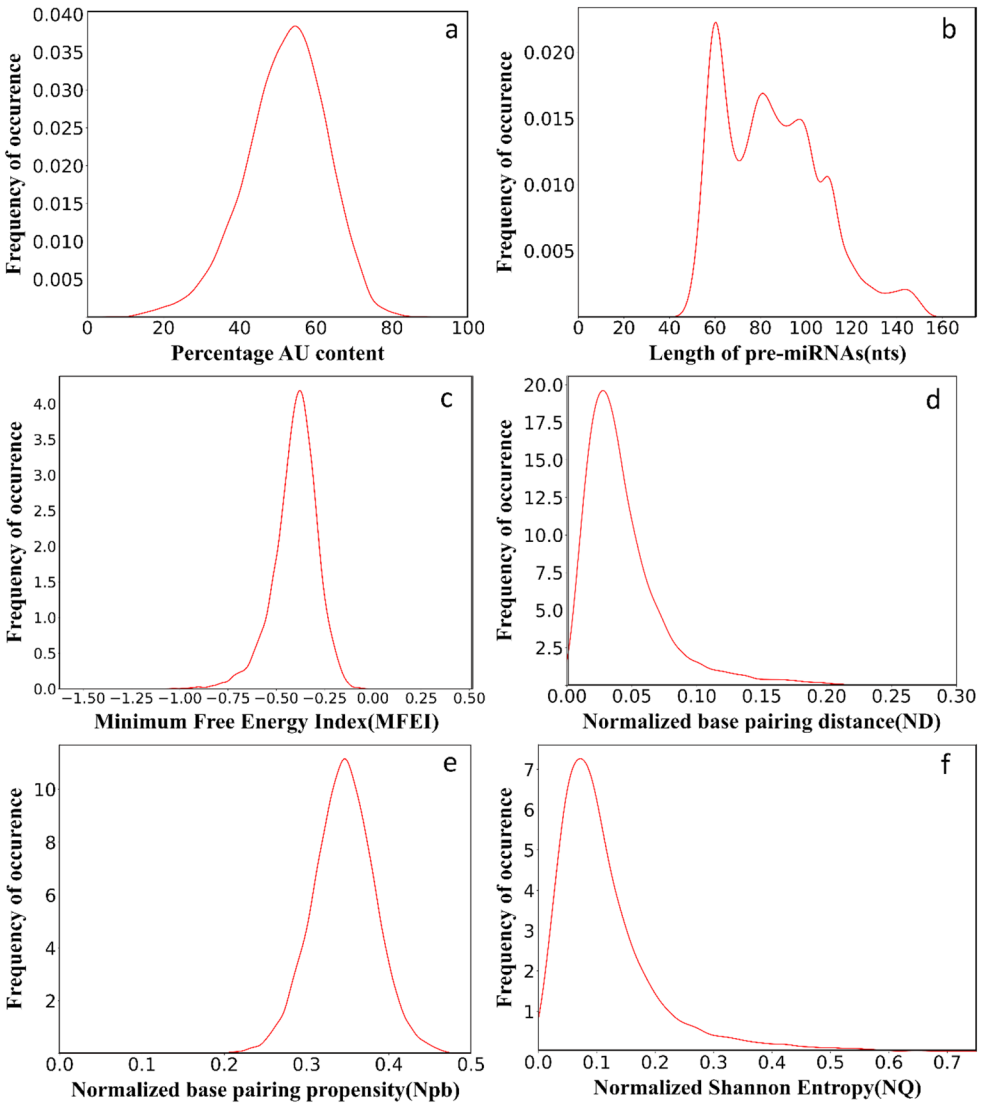

Fig. 3 Probability distributions of (**a**) AU content, (**b**) Length of pre-miRNAs, (**c**) Minimum Folding Energy Index (MFEI), (**d**) Normalized base pairing distance (ND), (**e**) Normalized base pairing propensity (Npb) and (**f**) Normalized Shannon Entropy (NQ)

is a prerequisite for the prediction of pre-miRNAs. The folding of a sequence is determined by several thermodynamic factors. The accuracy in the prediction of the secondary structure of a pre-miRNA is constrained by inadequate knowledge of folding principles. To counter this, several possible structures are constructed for a given sequence. The partition function (Q) is the total of the equilibrium constants of all potential secondary structures, which defines the ensemble thermodynamic properties of a system. An ensemble of secondary structures of pre-miRNAs is generated using RNAfold, which calculates the minimum folding energy structure based on the equilibrium partition function and base pair probability distance. The various RNA folding measures include MFEI, Npb, NQ, ensemble energy and ensemble diversity. Considering 99% probability range, MFEI varies from -1.0 to -0.15, ND varies from 0.0 to 0.2, Npb varies from 0.2 to 0.45 and NQ varies from 0.0 to 0.4 (Fig. 3c-f). The distributions of MFEI and Npb are symmetric, whereas the distributions of ND and NQ are skewed. The average base pairs including AU, GC and GU wobble pair per stem are also calculated. On average, the occurrence of GC pair per stem is the highest, comprising ~ 15% of the paired and the unpaired bases. AU pair comprises ~ 14% and the GU wobble pair being the lowest, comprises ~ 4%.

K-mer frequency

In general, k-mers of less than six nucleobases are considered to curb the high dimensionality of vectors representing the statistical samples. In addition, k-mer location and the distance between the first k-mer and the terminal k-mer are also used as features. k-mer signatures in pre-miRNAs vary between different metazoan species (Yousef and Allmer 2021). Consequently, they can also be employed to categorize pre-miRNA candidates according to their species

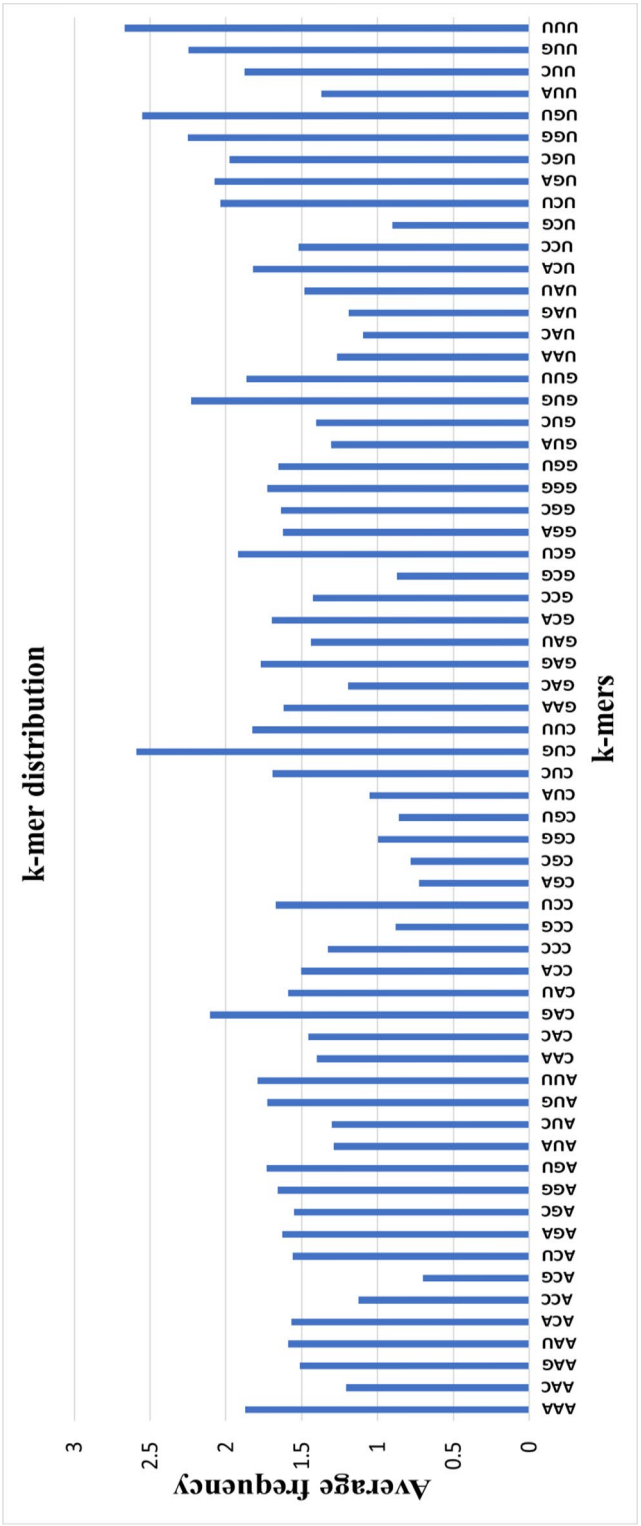

Fig. 4 Distribution of k-mers of length three nt in pre-miRNAs. The k-mers 'CUG', 'UGU' and 'UUU' have the highest frequency of occurrence

of origin. In the present study, the frequency of k-mers of length three nucleotides is calculated for the pre-miRNA sequences with a window size of three. The distribution of the k-mers is shown in Fig. 4. Since the pre-miRNA sequences are of varying lengths, k-mers are normalized per 100 nts. Although the k-mers among the different pre-miRNA families are not conserved, yet they are conserved within the same family (Kozomara et al. 2019). k-mer signatures CAG, CUG, GUG, UCU, UGA, UGG, UGU, UUG and UUU occur at least twice per 100 nts with UUU as the highest repeating with an R-value of 2.67. k-mers having a low occurring frequency are CGA, ACG, CCG, CGC, CGG, CGU, GCG and UCG with ACG being the least occurring with an R-value of 0.70.

Non-redundant features to train the classifiers

Feature selection involves removing redundant characteristics and selecting a subset of features from the original feature set. This facilitates the effective categorization of different classes in a dataset. Feature selection can be performed based on several criteria including correlation, dependence and information measure (refer to Materials and Methods section). We have performed feature selection on 125 extracted features calculated for both pre-miRNAs and coding sequences. Mutual information and LASSO regression are applied for the feature selection (Vergara and Estévez 2014; Ranstam and Cook 2018). Initially, mutual information is applied and the top 30 features with the highest scores are considered. LASSO regression is applied to the features selected from mutual information to obtain the final feature set. Eventually, out of 125 features, we retain 28 features including 11 secondary structural features and 17 sequence features to train the classifiers (Table 1). The selected features correlate less than 80% (Fig. 5). A combination of filter-based and embedded feature selection helps in identifying features that substantially contribute to the prediction capability of the model and have a significant statistical relationship with the target. Moreover, this method also reduces overfitting.

Selection and training of classifiers

Classifier selection is challenging while dealing with biological datasets due to their high complexity and dimensionality. Classification algorithms can be linear, distance-based and tree-based methods (Hemphill et al. 2014). The linear classifier employs a linear function to assign scores to different classes. This is done by calculating the dot product of the feature values and the feature weights that are determined during the training process. The distance-based method calculates the number of closest matching samples

Table 1 List of 28 features obtained after feature selection

Sl no.	Sequence		Structural		
	Dinucleotide features	k-mer features (length = 3nts)	Themodynamic features	Sequence-structure triplet elements	(Percent base-pair)/(No.of stem-loop)
1.	%AA	%AAC	Npb	A.(	%(G-C)/n
2.	%AU	%AAU	MFEI	A…	%(A-U)/n
3.	%GC	%CAA	NQ	U…	%(G-U)/n
4.	%GG	%CAG		G…	
5.	%GU	%CUG		C…	
6.	%UA	%GCU			
7.	%UG	%GGC			
8.		%GUG			
9.		%UGC			
10.		%UGG			

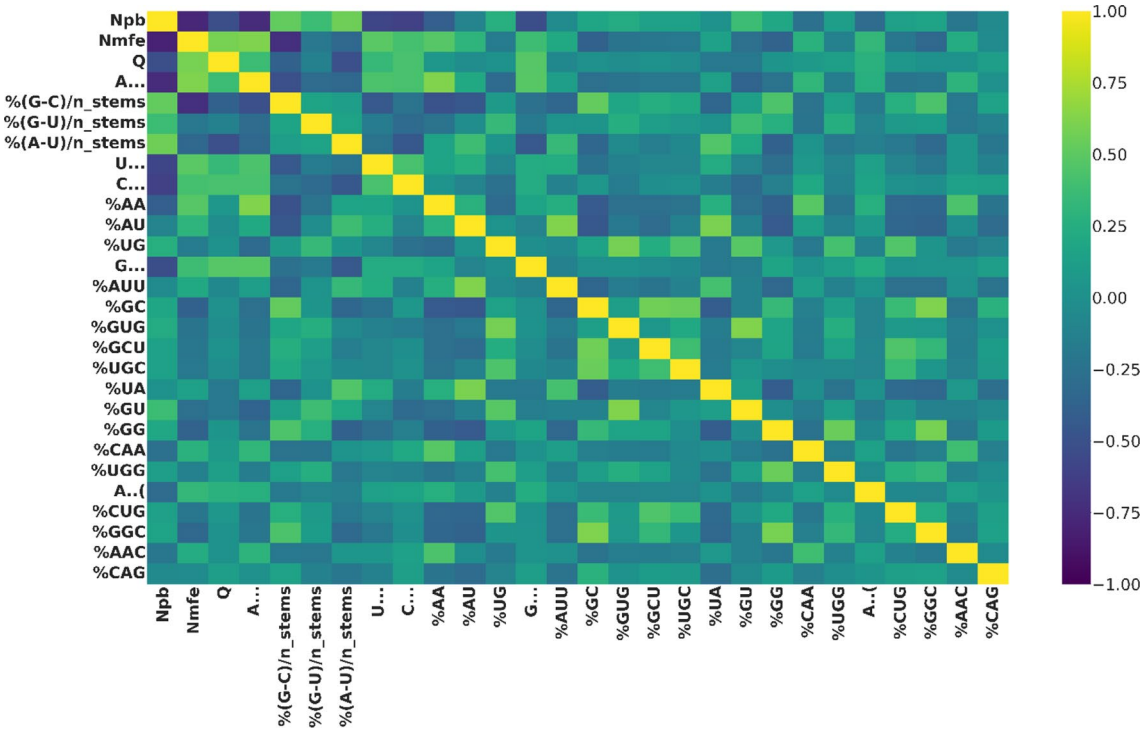

Fig. 5 Heatmap showing the correlation of the twenty-eight selected features with correlation < 80%

Table 2 Performance comparison among the nine different classifiers

	Accuracy	Sensitivity	Specificity	F-score
XGBoost	0.90	0.86	0.91	0.76
Light Gradient Boost	**0.93**	**0.86**	**0.95**	**0.82**
SVM	0.90	0.88	0.90	0.76
Gaussian Naïve Bayes	0.85	0.86	0.85	0.67
Gradient Boost	0.91	0.87	0.92	0.79
AdaBoost	0.81	0.79	0.81	0.60
Random Forest	0.87	0.85	0.88	0.70
Decision Tree	0.85	0.85	0.86	0.68
K-Nearest Neighbor	0.86	0.82	0.88	0.68

and assigns membership to the majority class. The tree-based method constructs multiple trees, and their ensemble is used for making the final decision. We train nine different classifiers belonging to three previously mentioned categories using selected features. The sensitivity, specificity, F-score and AUC are compared. Among the various classifiers, LGB shows overall better performance (Table 2). The confusion matrix and the AUROC for different classifiers are shown in Figs. 6 and 7, respectively. Although Gradient Boost and LGB show similar performance, the latter is faster. Consequently, we select LGB for the classification of pre-miRNAs and implemented in webserver. Finally, a 10-fold cross-validation is performed on the training and the

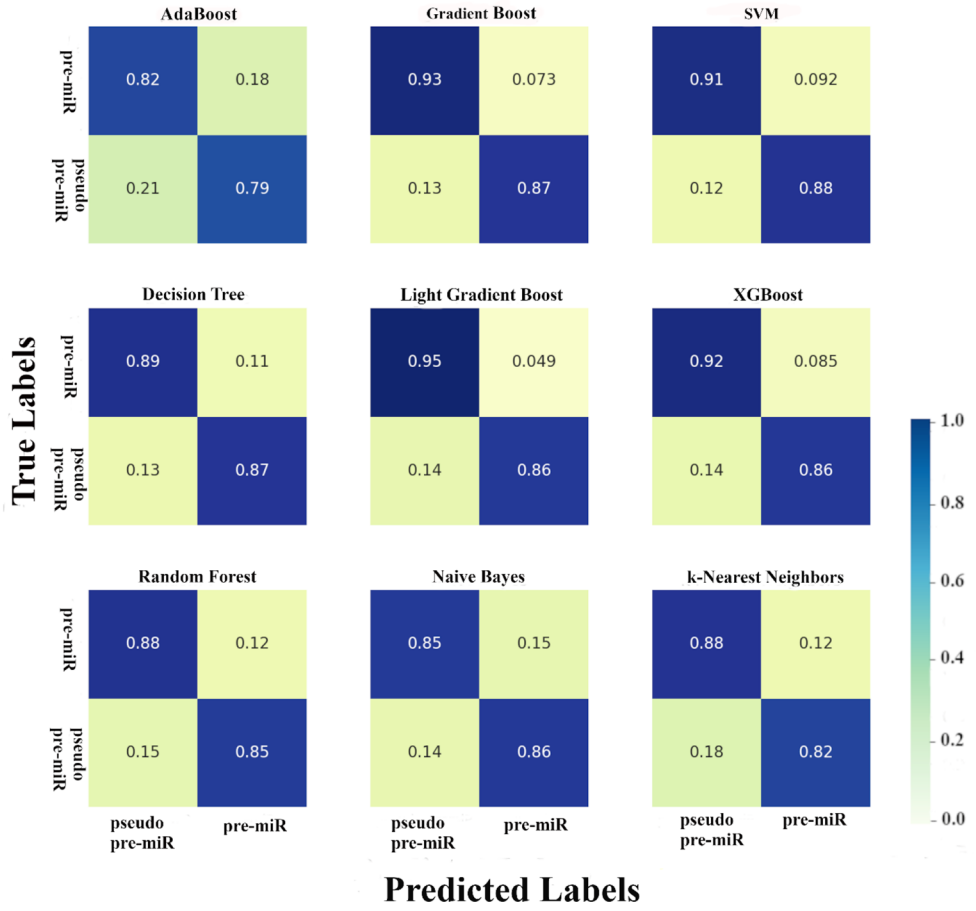

Fig. 6 Confusion matrices showing the performance of nine distinct classifiers

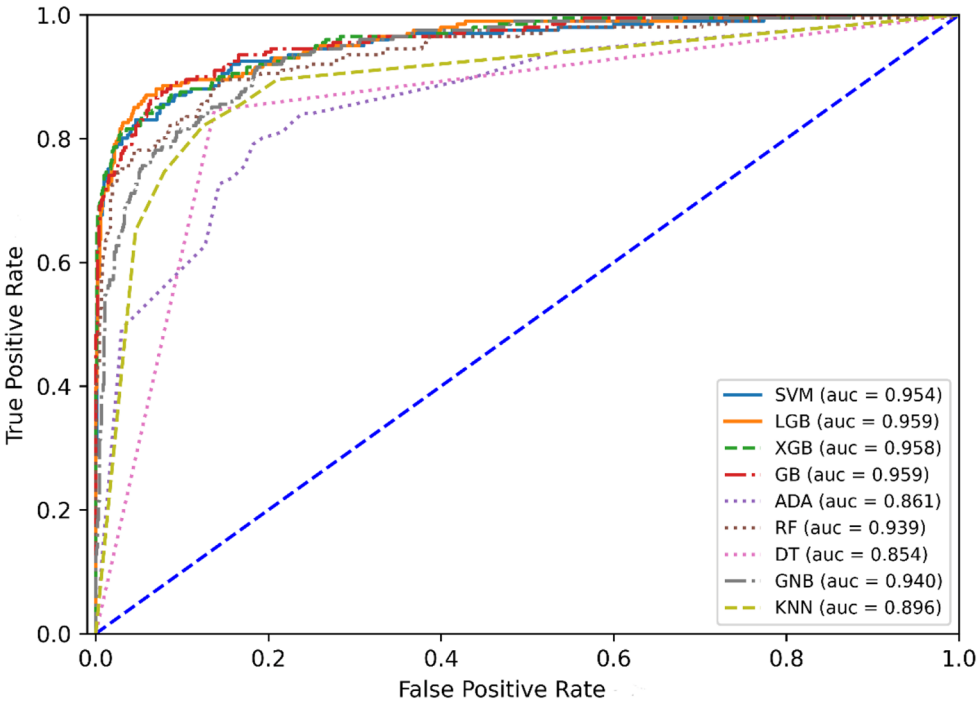

Fig. 7 AUROC curve for the prediction of pre-miRNAs using nine distinct classifiers

Table 3 Performance evaluation of pmiRScan among the different taxonomic groups

Taxon	Accuracy	Specificity	Sensitivity	F-score
Mammals	0.83	0.84	0.82	0.79
Arthropods	0.83	0.81	0.95	0.76
Molluscs	0.92	0.89	0.91	0.90
Nematodes	0.92	0.90	0.93	0.85

test datasets keeping the learning rate as 0.1 and number of estimators to 100. LGB outperforms other classifiers in terms of accuracy (0.93), sensitivity (0.86), specificity (0.95) and F-score (0.82) (Table 2). LGB and Gradient Boost both have the highest AUROC, but LGB is faster, more efficient and robust in handling data with high dimensionality compared to Gradient Boost.

Computational validation

Various performance measures are employed in machine learning to assess the performance of a model. The performance of our classifiers is evaluated on a validation set containing 201 positive and 932 negative instances. They are not included in the training and the test datasets. In addition, we also test the performance of the classifiers on four different taxonomical datasets. The performance evaluation is shown in Table 3 with the maximum specificity of 0.9 for nematodes and the maximum sensitivity of 0.95 for arthropods. The AUROC of the classifier for four different taxonomical groups is shown in Fig. 8. While AUROC evaluates the performance of a classifier over a range of classification threshold, it alone cannot effectively assess the model.

Table 4 Comparison of the performance of pmiRScan with the existing classifiers

Classifier	Accuracy	Sensitivity	Specificity	F-score
microPred	0.79	0.83	0.78	0.58
dnnPreMiR-master	0.79	0.82	0.79	0.58
triplet-SVM	0.44	0.31	0.78	0.44
pmiRScan	**0.93**	**0.86**	**0.95**	**0.82**
HuntMi	0.86	0.90	0.85	0.71
mirDNN	0.84	0.83	0.84	0.39

When two ROC curves intersect, the AUC of one curve can be greater even when the model performance is poor (Robinson et al. 2020). The AUROC can be utilized in conjunction with sensitivity and specificity to effectively assess the performance of the model. In our study, the maximum AUROC is obtained for the nematodes and molluscs.

Comparison with the existing classifiers

A comparative analysis is performed between pmiRScan and the existing classifiers microPred, dnnPreMiR, triplet-SVM, miRNAs, HuntMi and mirDNN. The microPred, triplet-SVM, miRNAss and HuntMi use SVM to classify the animal pre-miRNAs. triplet-SVM is only capable of handling single-loop pre-miRNA structures. Other classifiers are capable of dealing with multi-loop pre-miRNA secondary structures (Jiang et al. 2007). A comparison between pmiRScan and the other pre-existing classifiers is shown in Table 4 and Fig. 9. Although HuntMi has higher sensitivity, yet it lacks specificity and has a lower F-score compared to

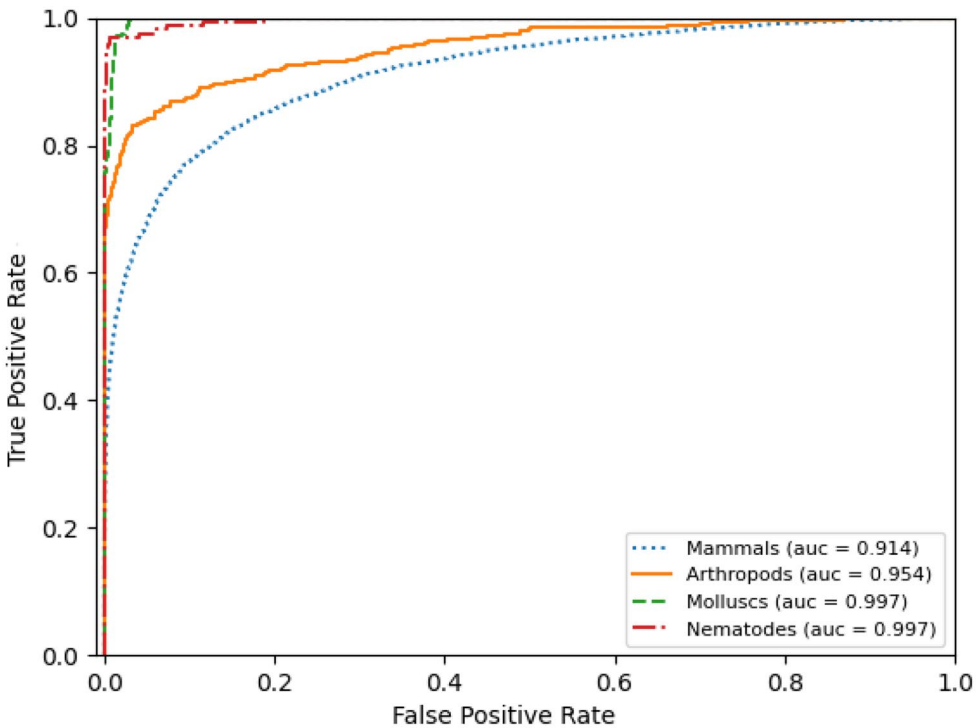

Fig. 8 AUROC curve for the prediction of pre-miRNAs in four different taxonomical groups

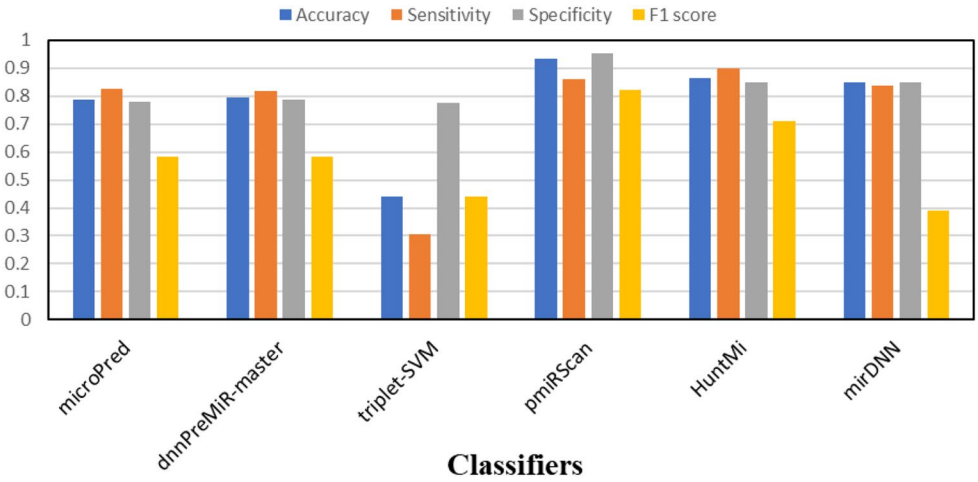

Fig. 9 Performance comparison of pmiRScan with the existing classifiers

Table 5 Comparison of execution time for different classifiers

Classifier	Time(sec)	Log_2(time)
Deepmir_master	42.12	5.39
HuntMi	57911.34	15.82
pmiRScan	62.73	5.97
MicroPred	34505.80	15.07
mirdnn	21.84	4.44
triplet-SVM	121.00	6.91

pmiRScan. pmiRScan also outperforms mirDNN(Yones et al. 2021) in terms of all the performance metrics. pmiRScan has an overall better performance in predicting the pre-miRNAs compared to other existing methods. The time taken for each classifier to classify sequences was evaluated in the present study (Table 5). Figure 10 represents log_2(time) taken by each classifier. microPred and HuntMi take maximum duration while mirDNN takes the least duration for the classification of pre-miRNAs. The execution time for pmiRScan, triplet-SVM and dnnPreMiR-master are almost similar. However, pmiRScan outperforms other classifiers in terms of other performance metrics.

Prediction of putative human pre-miRNAs

The known non-redundant human pre-miRNAs from the miRbase v22 are searched as queries against non-redundant EST and GSS sequences of human. The resulting sequences are further processed using the procedure mentioned in the Materials and Methods section. Finally, a total of 313 sequences are classified as pre-miRNAs by pmiRScan (supplementary Table S1). Extraction of mature miRNA sequences from the pre-miRNAs resulted in 180 miRNAs, of which 128 are novel (supplementary Table S2) and are not reported in miRBase. These mature miRNAs belong to 60 different miRNA families. The highest populated family is MIR-548 with 58 mature miRNAs (Table 6). The number

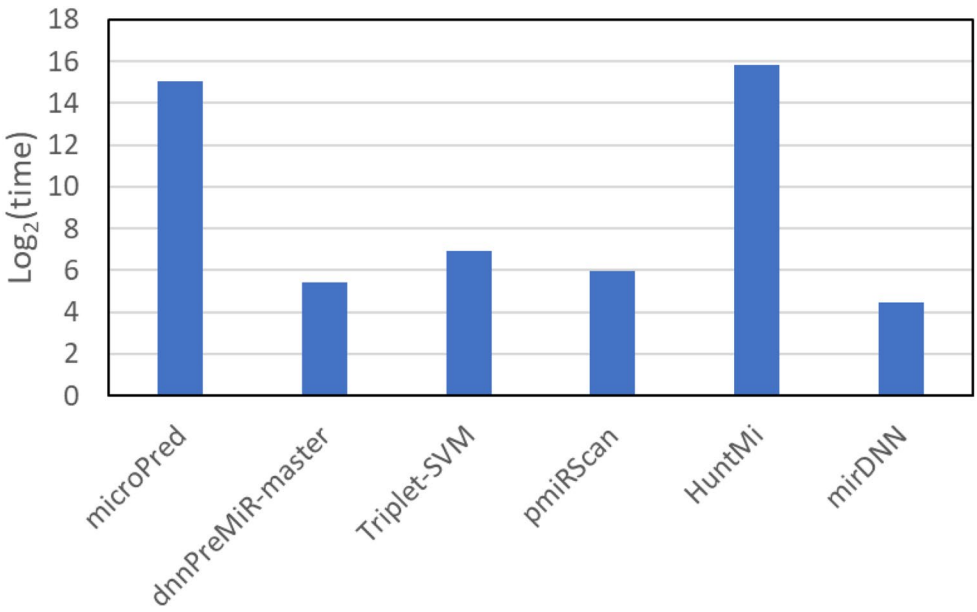

Fig. 10 Plot showing log_2(time) for the different classifiers

Table 6 Population of predicted miRNAs in different families

miRNA family	Number of miRNAs
MIR-7;MIR-10;MIR-95;MIR-126;MIR-142;MIR-146;MIR-147;MIR-151;MIR-297;MIR-302;MIR-466;MIR-499;MIR-511;MIR-516;MIR-523;MIR-526;MIR-576;MIR-590;MIR-644;-MIR-659;MIR-663;MIR-1226;MIR-1279;MIR-1303;MIR-3116;MIR-3149;MIR-3606;MIR-3617;MIR-3714;MIR-4252;MIR-4303;MIR-4325;MIR-4330;MIR-4430;MIR-4466;MIR-4714-;MIR-4778;MIR-4789;MIR-5006;MIR-5011;MIR-5088;MIR-5585;MIR-5739;MIR-6090;MIR-6716;MIR-6845;MIR-6861;MIR-6875;MIR-6878;MIR-7106;MIR-7112;MIR-7851;MIR-10,400;MIR-12,124	1
MIR-518;MIR-570;MIR-1273	2
MIR-519	3
MIR-520	7
MIR-548	58

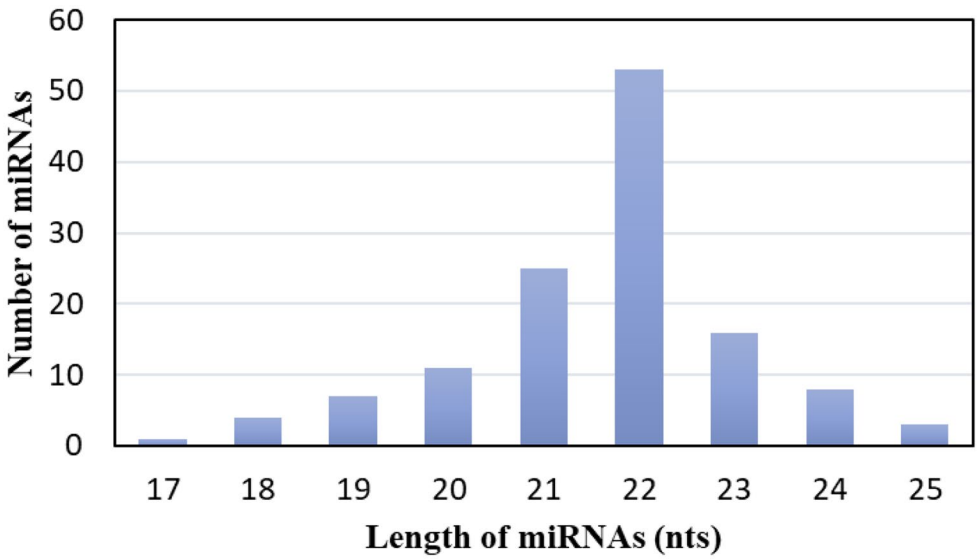

Fig. 11 Distribution of the length of human miRNAs predicted using pmiRScan

of members in five miRNA families varies from two to seven, while 54 miRNA families contain only one miRNA each. The distribution of the length of the predicted miRNA is shown in Fig. 11. The majority of the miRNAs extracted from the predicted pre-miRNAs have a length of 22 nt.

Conclusion

The present study provides a fast and efficient approach to predict the animal pre-miRNAs using various sequence and structural features. We have trained and tested nine different classifiers followed by a performance evaluation. The LGB classifier, implemented in pmiRScan, has the superior performance in comparison to all other classifiers with an accuracy of 0.93, sensitivity of 0.86, specificity of 0.95 and F-score of 0.82. Moreover, pmiRScan outperforms the existing methods in predicting the pre-miRNAs of four different taxonomic groups. Being able to accurately classify the pre-miRNAs of different taxonomic groups, pmiRScan can be implemented for the identification of novel pre-miRNAs in various metazoan species. Enriching our understanding of miRNAs will aid in mapping regulatory pathways and gene co-regulation. Moreover, miRNAs serve as biomarkers in various diseases including cancer and neurological disorders. Hence, their prediction might be useful in the development of diagnostic tools to monitor disease progression. Newly discovered miRNAs may have the potential to be targeted for innovative therapeutic strategies. miRNA mimics can be created to regulate the activity of miRNAs in disorders characterized by their abnormal gene expression. Thus, the identification of pre-miRNAs will enhance our comprehensive understanding of gene regulation in animals, both at the transcriptional and post-translational levels.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10142-025-01527-y.

Acknowledgements Amrit Venkatesan thanks the Indian Institute of Technology Kharagpur for his fellowship; Ranjit Prasad Bahadur and Jolly Basak thank the Department of Biotechnology, India.

Author contributions Conceptualization, A.V., R.P.B.; Methodology, A.V., J.B; Writing-draft and editing, A.V., J.B. and R.P.B.; Supervision, R.P.B.

Funding Department of Biotechnology (DBT), Govt. of India (grant no. BT/PR40175/BTIS/137/41/2022) for the Bioinformatics Centre.

Data availability The datasets used and the source code for pmiRScan are available in the GitHub repository: https://github.com/amrit-debug/pmiRScan.git.

Declarations

Competing interests The authors declare no competing interests.

References

Amin N, McGrath A, Chen Y-PP (2019) Evaluation of deep learning in non-coding RNA classification. Nat Mach Intell 1:246–256. https://doi.org/10.1038/s42256-019-0051-2

Barik A, Das S (2018) A comparative study of sequence- and structure-based features of small RNAs and other RNAs of bacteria. RNA Biol 15:95–103. https://doi.org/10.1080/15476286.2017.1387709

Bartel DP (2004) MicroRNAs Cell 116:281–297. https://doi.org/10.1016/S0092-8674(04)00045-5

Batuwita R, Palade V (2009) *microPred*: effective classification of pre-miRNAs for human miRNA gene prediction. Bioinformatics 25:989–995. https://doi.org/10.1093/bioinformatics/btp107

Bisong E (2019) Introduction to scikit-learn. Building machine learning and deep learning models on google cloud platform. A, Berkeley, CA, pp 215–229

Bugnon LA, Yones C, Milone DH, Stegmayer G (2021) Genome-wide discovery of pre-miRNAs: comparison of recent approaches based on machine learning. Brief Bioinform 22. https://doi.org/10.1093/bib/bbaa184

Chen C, Tsai Y, Chang F, Lin W (2020) Ensemble feature selection in medical datasets: combining filter, wrapper, and embedded feature selection results. Expert Syst 37. https://doi.org/10.1111/exsy.12553

Chen PY, Manninga H, Slanchev K, Chien M, Russo JJ, Ju J, Sheridan R, John B, Marks DS, Gaidatzis D, Sander C, Zavolan M, Tuschl T (2005) The developmental miRNA profiles of zebrafish as determined by small RNA cloning. Genes Dev 19:1288–1293. https://doi.org/10.1101/gad.1310605

Chen T, Guestrin C (2016) XGBoost. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, NY, USA, pp 785–794

Fernandez A, Garcia S, Herrera F, Chawla NV (2018) SMOTE for learning from Imbalanced Data: progress and challenges, marking the 15-year anniversary. J Artif Intell Res 61:863–905. https://doi.org/10.1613/jair.1.11192

Fromm B, Høye E, Domanska D, Zhong X, Aparicio-Puerta E, Ovchinnikov V, Umu SU, Chabot PJ, Kang W, Aslanzadeh M, Tarbier M, Mármol-Sánchez E, Urgese G, Johansen M, Hovig E, Hackenberg M, Friedländer MR, Peterson KJ (2022) MirGeneDB 2.1: toward a complete sampling of all major animal phyla. Nucleic Acids Res 50:D204–D210. https://doi.org/10.1093/nar/gkab1101

Fu X, Zhu W, Cai L, Liao B, Peng L, Chen Y, Yang J (2019) Improved pre-miRNAs identification through mutual information of pre-miRNA sequences and structures. Front Genet 10. https://doi.org/10.3389/fgene.2019.00119

Ganju A, Khan S, Hafeez BB, Behrman SW, Yallapu MM, Chauhan SC, Jaggi M (2017) miRNA nanotherapeutics for cancer. Drug Discov Today 22:424–432. https://doi.org/10.1016/j.drudis.2016.10.014

Gardner PP, Giegerich R (2004) A comprehensive comparison of comparative RNA structure prediction approaches. BMC Bioinf 5:140. https://doi.org/10.1186/1471-2105-5-140

Garg A, Roske Y, Yamada S, Uehata T, Takeuchi O, Heinemann U (2021) PIN and CCCH Zn-finger domains coordinate RNA targeting in ZC3H12 family endoribonucleases. Nucleic Acids Res 49:5369–5381. https://doi.org/10.1093/nar/gkab316

Gonzales GB, De Saeger S (2018) Elastic net regularized regression for time-series analysis of plasma metabolome stability under sub-optimal freezing condition. Sci Rep 8:3659. https://doi.org/10.1038/s41598-018-21851-7

Griffiths-Jones S (2006) MiRBase The MicroRNA sequence database. In: MicroRNA protocols. Humana, New Jersey, pp 129–138

Guan D-G, Liao J-Y, Qu Z-H, Zhang Y, Qu L-H (2011) mirExplorer: detecting microRNAs from genome and next generation sequencing data using the AdaBoost method with transition probability matrix and combined features. RNA Biol 8:922–934. https://doi.org/10.4161/rna.8.5.16026

Gudyś A, Szcześniak MW, Sikora M, Makałowska I (2013) HuntMi: an efficient and taxon-specific approach in pre-miRNA identification. BMC Bioinf 14:83. https://doi.org/10.1186/1471-2105-14-83

Hemphill E, Lindsay J, Lee C, Măndoiu II, Nelson CE (2014) Feature selection and classifier performance on diverse bio-logical datasets. BMC Bioinf 15:S4. https://doi.org/10.1186/1471-2105-15-S13-S4

Hertel J, Stadler PF (2006) Hairpins in a haystack: recognizing microRNA precursors in comparative genomics data. Bioinf 22:e197–e202. https://doi.org/10.1093/bioinformatics/btl257

Jiang P, Wu H, Wang W, Ma W, Sun X, Lu Z (2007) MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Res 35:W339–W344. https://doi.org/10.1093/nar/gkm368

Jouravleva K, Golovenko D, Demo G, Dutcher RC, Hall TMT, Zamore PD, Korostelev AA (2022) Structural basis of microRNA biogenesis by Dicer-1 and its partner protein Loqs-PB. Mol Cell 82:4049–4063e6. https://doi.org/10.1016/j.molcel.2022.09.002

Kleftogiannis D, Theofilatos K, Likothanassis S, Mavroudi S (2015) YamiPred: a novel evolutionary method for predicting pre-miRNAs and selecting relevant features. IEEE/ACM Trans Comput Biol Bioinform 12:1183–1192. https://doi.org/10.1109/TCBB.2014.2388227

Kotsiantis SB (2013) Decision trees: a recent overview. Artif Intell Rev 39:261–283. https://doi.org/10.1007/s10462-011-9272-4

Kozomara A, Birgaoanu M, Griffiths-Jones S (2019) miRBase: from microRNA sequences to function. Nucleic Acids Res 47:D155–D162. https://doi.org/10.1093/nar/gky1141

Lee RC, Feinbaum RL, Ambros V (1993) The C. Elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75:843–854. https://doi.org/10.1016/0092-8674(93)90529-Y

Liang L, Hu W, Zhang Y, Ma K, Gu Y, Tian B, Li H (2021) An algorithm with LightGBM + SVM fusion model for the assessment of dynamic security region. E3S Web Conferences 256(02022). https://doi.org/10.1051/e3sconf/202125602022

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinf 22:1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL (2011) ViennaRNA Package 2.0. Algorithms Mol Biol 6:26. https://doi.org/10.1186/1748-7188-6-26

Lorenz R, Flamm C, Hofacker I, Stadler P (2020) Efficient computation of base-pairing probabilities in multi-strand RNA folding. In:

proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies. SCITEPRESS - Science and Technology Publications, pp 23–31

Ma Y, Yu Z, Han G, Li J, Anh V (2018) Identification of pre-microRNAs by characterizing their sequence order evolution information and secondary structure graphs. BMC Bioinf 19:521. https://doi.org/10.1186/s12859-018-2518-2

Mendes ND, Freitas AT, Sagot M-F (2009) Current tools for the identification of miRNA genes and their targets. Nucleic Acids Res 37:2419–2433. https://doi.org/10.1093/nar/gkp145

Nasiri H, Alavi SA (2022) A Novel Framework based on deep learning and ANOVA feature selection method for diagnosis of COVID-19 cases from chest X-Ray images. Comput Intell Neurosci 2022:1–11. https://doi.org/10.1155/2022/4694567

Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. Front Neurorobot 7. https://doi.org/10.3389/fnbot.2013.00021

Nazarov PV, Kreis S (2021) Integrative approaches for analysis of mRNA and microRNA high-throughput data. Comput Struct Biotechnol J 19:1154–1162. https://doi.org/10.1016/j.csbj.2021.01.029

Niaz NU, Shahariar KMN, Patwary MJA (2022) Class Imbalance Problems in Machine Learning: A Review of Methods And Future Challenges. In: Proceedings of the 2nd International Conference on Computing Advancements. ACM, New York, NY, USA, pp 485–490

Nithin C, Mukherjee S, Basak J, Bahadur RP (2022) NCodR: a multiclass support vector machine classification to distinguish noncoding RNAs in viridiplantae. Quant Plant Biology 3:e23. https://doi.org/10.1017/qpb.2022.18

Nithin C, Patwa N, Thomas A, Bahadur RP, Basak J (2015) Computational prediction of miRNAs and their targets in phaseolus vulgaris using simple sequence repeat signatures. BMC Plant Biol 15:140. https://doi.org/10.1186/s12870-015-0516-3

Nithin C, Thomas A, Basak J, Bahadur RP (2017) Genome-wide identification of miRNAs and lncRNAs in Cajanus cajan. BMC Genomics 18:878. https://doi.org/10.1186/s12864-017-4232-2

Parveen A, Mustafa SH, Yadav P, Kumar A (2020) Applications of machine learning in miRNA discovery and target prediction. Curr Genomics 20:537–544. https://doi.org/10.2174/1389202921666200106111813

Petrov AI, Kay SJE, Kalvari I, Howe KL, Gray KA, Bruford EA, Kersey PJ, Cochrane G, Finn RD, Bateman A, Kozomara A, Griffiths-Jones S, Frankish A, Zwieb CW, Lau BY, Williams KP, Chan PP, Lowe TM, Cannone JJ, Gutell R, Machnicka MA, Bujnicki JM, Yoshihama M, Kenmochi N, Chai B, Cole JR, Szymanski M, Karlowski WM, Wood V, Huala E, Berardini TZ, Zhao Y, Chen R, Zhu W, Paraskevopoulou MD, Vlachos IS, Hatzigeorgiou AG, Ma L, Zhang Z, Puetz J, Stadler PF, McDonald D, Basu S, Fey P, Engel SR, Cherry JM, Volders P-J, Mestdagh P, Wower J, Clark MB, Quek XC, Dinger ME (2017) RNAcentral: a comprehensive database of non-coding RNA sequences. Nucleic Acids Res 45:D128–D134. https://doi.org/10.1093/nar/gkw1008

Ponsam JG, Bella Gracia SVJ, Geetha G, Karpaselvi S, Nimala K Credit Risk Analysis using LightGBM and a comparative study of popular algorithms. In: 2021 4th International Conference on Computing and, Technologies C (2021) (ICCCT). IEEE, pp 634–641

Praher D, Zimmermann B, Dnyansagar R, Miller DJ, Moya A, Modepalli V, Fridrich A, Sher D, Friis-Møller L, Sundberg P, Fôret S, Ashby R, Moran Y, Technau U (2021) Conservation and turnover of miRNAs and their highly complementary targets in early branching animals. Proceedings of the Royal Society B: Biological Sciences 288:20203169. https://doi.org/10.1098/rspb.2020.3169

Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM (2022) A review of feature selection methods for machine learning-based disease risk prediction. Front Bioinf 2. https://doi.org/10.3389/fbinf.2022.927312

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. https://doi.org/10.1093/bioinformatics/btq033

Raad J, Bugnon LA, Milone DH, Stegmayer G (2022) miRe2e: a full end-to-end deep model based on transformers for prediction of pre-miRNAs. Bioinformatics 38:1191–1197. https://doi.org/10.1093/bioinformatics/btab823

Ranstam J, Cook JA (2018) LASSO regression. Br J Surg 105:1348–1348. https://doi.org/10.1002/bjs.10895

Rigatti SJ (2017) Random Forest. J Insur Med 47:31–39. https://doi.org/10.17849/insm-47-01-31-39.1

Rissland OS, Subtelny AO, Wang M, Lugowski A, Nicholson B, Laver JD, Sidhu SS, Smibert CA, Lipshitz HD, Bartel DP (2017) The influence of microRNAs and poly(A) tail length on endogenous mRNA–protein complexes. Genome Biol 18:211. https://doi.org/10.1186/s13059-017-1330-z

Ritchie W, Théodule F-X, Gautheret D (2008) Mireval: a web tool for simple microRNA prediction in genome sequences. Bioinformatics 24:1394–1396. https://doi.org/10.1093/bioinformatics/btn137

Robinson MC, Glen RC, Lee AA (2020) Validating the validation: reanalyzing a large-scale comparison of deep learning and machine learning models for bioactivity prediction. J Comput Aided Mol Des 34:717–730. https://doi.org/10.1007/s10822-019-00274-0

Solomatine DP, Shrestha DL (2004) AdaBoost.RT: a boosting algorithm for regression problems. In: 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541). IEEE, pp 1163–1168

Stańczyk U (2015) Feature Evaluation by Filter, Wrapper, and Embedded Approaches. pp 29–44

Stegmayer G, Di Persia LE, Rubiolo M, Gerard M, Pividori M, Yones C, Bugnon LA, Rodriguez T, Raad J, Milone DH (2019) Predicting novel microRNA: a comprehensive comparison of machine learning approaches. Brief Bioinform 20:1607–1620. https://doi.org/10.1093/bib/bby037

Suthaharan S (2016) Support Vector Machine. pp 207–235

Takashima Y, Murata A, Iida K, Sugai A, Hagiwara M, Nakatani K (2022) Method for identifying sequence motifs in Pre-miRNAs for small-molecule binding. ACS Chem Biol 17:2817–2827. https://doi.org/10.1021/acschembio.2c00452

Tasdelen A, Sen B (2021) A hybrid CNN-LSTM model for pre-miRNA classification. Sci Rep 11:14125. https://doi.org/10.1038/s41598-021-93656-0

Titov II, Vorozheykin PS (2018) Comparing miRNA structure of mirtrons and non-mirtrons. BMC Genomics 19:114. https://doi.org/10.1186/s12864-018-4473-8

Torgo L, Ribeiro RP, Pfahringer B, Branco P (2013) SMOTE for Regression. pp 378–389

Vafaee F, Diakos C, Kirschner MB, Reid G, Michael MZ, Horvath LG, Alinejad-Rokny H, Cheng ZJ, Kuncic Z, Clarke S (2018) A data-driven, knowledge-based approach to biomarker discovery: application to circulating microRNA markers of colorectal cancer prognosis. NPJ Syst Biol Appl 4:20. https://doi.org/10.1038/s41540-018-0056-1

Vergara JR, Estévez PA (2014) A review of feature selection methods based on mutual information. Neural Comput Appl 24:175–186. https://doi.org/10.1007/s00521-013-1368-0

Wang D, Zhang Y, Zhao Y (2017) LightGBM. In: Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics. ACM, New York, NY, USA, pp 7–11

Xue C, Li F, He T, Liu G-P, Li Y, Zhang X (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinformatics 6:310. https://doi.org/10.1186/1471-2105-6-310

Yones C, Raad J, Bugnon LA, Milone DH, Stegmayer G (2021) High precision in microRNA prediction: a novel genome-wide approach with convolutional deep residual networks. Comput Biol Med 134:104448. https://doi.org/10.1016/j.compbiomed.2021.104448

Yones C, Stegmayer G, Milone DH (2018) Genome-wide pre-miRNA discovery from few labeled examples. Bioinformatics 34:541–549. https://doi.org/10.1093/bioinformatics/btx612

Yousef M, Allmer J (2021) Classification of Precursor MicroRNAs from different species based on K-mer Distance features. Algorithms 14:132. https://doi.org/10.3390/a14050132

Zhao D, Hu X, Xiong S, Tian J, Xiang J, Zhou J, Li H (2021) k-means clustering and kNN classification based on negative databases. Appl Soft Comput 110:107732. https://doi.org/10.1016/j.asoc.2021.107732

Zheng X, Fu X, Wang K, Wang M (2020) Deep neural networks for human microRNA precursor detection. BMC Bioinformatics 21:17. https://doi.org/10.1186/s12859-020-3339-7

Zhu S, Jiang Q, Wang G, Liu B, Teng M, Wang Y (2011) Chromatin structure characteristics of pre-miRNA genomic sequences. BMC Genomics 12:329. https://doi.org/10.1186/1471-2164-12-329