Hindawi
Journal of Chemistry
Volume 2023, Article ID 9808201, 1 page
https://doi.org/10.1155/2023/9808201

Retraction

Retracted: PLncWX: A Machine-Learning Algorithm for Plant lncRNA Identification Based on WOA-XGBoost

Journal of Chemistry

Received 22 August 2023; Accepted 22 August 2023; Published 23 August 2023

Copyright © 2023 Journal of Chemistry. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] F. Guo, Z. Yin, K. Zhou, and J. Li, "PLncWX: A Machine-Learning Algorithm for Plant lncRNA Identification Based on WOA-XGBoost," *Journal of Chemistry*, vol. 2021, Article ID 6256021, 11 pages, 2021.

Hindawi
Journal of Chemistry
Volume 2021, Article ID 6256021, 11 pages
https://doi.org/10.1155/2021/6256021

RETRACTED

Research Article

PLncWX: A Machine-Learning Algorithm for Plant lncRNA Identification Based on WOA-XGBoost

Fei Guo, Zhixiang Yin, Kai Zhou, and Jiasi Li

School of Mathematics Physics and Statistics, Shanghai University of Engineering Science, Shanghai 201620, China

Correspondence should be addressed to Zhixiang Yin; zxyin66@163.com

Received 6 November 2021; Accepted 8 December 2021; Published 31 December 2021

Academic Editor: Shaohui Wang

Copyright © 2021 Fei Guo et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Long noncoding RNAs (lncRNAs) are a class of RNAs longer than 200 nt and cannot encode the protein. Studies have shown that lncRNAs can regulate gene expression at the epigenetic, transcriptional, and posttranscriptional levels, which are not only closely related to the occurrence, development, and prevention of human diseases, but also can regulate plant flowering and participate in plant abiotic stress responses such as drought and salt. Therefore, how to accurately and efficiently identify lncRNAs is still an essential job of relevant researches. There have been a large number of identification tools based on machine-learning and deep learning algorithms, mostly using human and mouse gene sequences as training sets, seldom plants, and only using one or one class of feature selection methods after feature extraction. We developed an identification model containing dicot, monocot, algae, moss, and fern. After comparing 20 feature selection methods (seven filter and thirteen wrapper methods) combined with seven classifiers, respectively, considering the correlation between features and model redundancy at the same time, we found that the WOA-XGBoost-based model had better performance with 91.55%, 96.78%, and 91.68% of accuracy, AUC, and F_1_score. Meanwhile, the number of elements in the feature subset was reduced to 23, which effectively improved the prediction accuracy and modeling efficiency.

1. Introduction

Noncoding RNA (ncRNA) refers to a functional RNA molecule that cannot be translated into a protein, in which lncRNA is a class of ncRNA, longer than 200 nt previously considered "noise" and ignored. Until 1984, the study of lncRNAs had attracted increasing attention when Pachnis and his colleagues found the H19 gene in mice, which was the first eukaryotic lncRNA, and highly expressed during embryonic development [1]. The current researches on lncRNAs generally focus on lncRNA screening, identification, expression, and localization, so it is very necessary to accurately and efficiently screen out lncRNAs from mRNAs. There have been already several tools, which can be used to analyze the coding potential of transcript sequences. CPC [2], CNCI [3], PLEK [4], lncRScan-SVM [5], and CPC2 [6] classified the sequences using the support vector machine (SVM) algorithm, while CPAT utilized the logistic regression [7]. lncRNA-ID [8], PredLnc-GFStack [9], and CNIT [10] utilized ensemble learning algorithms such as random forest and XGBoost, whereas lncRNA-MFDL [11] and lncRNA-LSTM [12] identified the lncRNAs using deep learning.

Since lncRNAs participated in biological regulatory processes, such as transcriptional level regulation, epigenetic level regulation, and posttranscriptional level regulation, and associated with diseases [13–15], scholars at home and abroad mainly paid attention to lncRNAs of humans, mice, and other vertebrates, while the researches on plant lncRNAs were relatively few. However, many studies have shown that lncRNAs play a key role in the plant immune responses to biotic stress, such as regulating plant flowering, affecting male and female differentiation and pollen development [16], and participating in the plant responses to several abiotic stresses (e.g., drought, salt, and cold) [17, 18]. For the past few years, scholars have committed to use a variety of plants to build specialized plant lncRNA identification models. Urminder Singh et al. [19] developed consensus models for dicots and monocots with ten plant species (*Amborella trichopoda*, *Arabidopsis thaliana*, *Chlamydomonas reinhardtii*, *Glycine* max, *Oryza sativa*,

Physcomitrella patens, *Selaginella moellendorffii*, *Solanum tuberosum*, *Vitis vinifera,* and *Zea mays*). Caitlin M. A. Simopoulos et al. [20] used four different species (*Homo sapiens*, *Arabidopsis thaliana*, *Mus musculus*, and *Oryza sativa*) as negative training data sets. Siyu Han et al. [21] developed an integrated platform named LncFinder with the data sets of humans, mouse, wheat, zebra fish, and chicken. RNAplonc [22] used five plant species (*Arabidopsis thaliana*, *Cucumis sativus*, *Glycine* max, *Populus trichocarpa,* and *Oryza sativa*). Cagirici et al. [23] present a crop-specific, alignment-free coding potential prediction tool, LncMachine, using a set of publicly available lncRNA and mRNA sequences for 18 plant species. Shuwei Yin and his colleagues [24] downloaded the circRNAs and lncRNAs of Arabidopsis and maize from PlantcircBase and GreeNC, respectively, to test the universality of PCirc.

The researchers mostly constructed models directly after extracting only a few features. When the number of extracted features was larger, containing more k-mers, they tended to select features before model construction. For example, Negri et al. [22] extracted 5468 features and selected 16 features with WrapperSubsetEval, InfoGainAttribute, and GainRatioAttributeEval methods before constructing the RNAplonc to prevent the model from overfitting. Here, we built a model specifically used to identify plant lncRNAs. Firstly, we extracted 89 features with *Python* and two online software packages about five different kinds of plant species (*Arabidopsis thaliana*, *Brachypodium distachyon*, *Chlamydomonas reinhardtii*, *Physcomitrella patens*, and *Selaginella moellendorffii*), especially including the algae, moss, and fern. Secondly, we chose the better-performing methods by comparing the accuracy of twenty feature selection methods containing seven filter and thirteen wrapper methods in three single classifiers. Finally, we combined these feature selection methods, respectively, with the meta-learner XGBoost to construct the model. Moreover, we found that the performance of the WOA-XGBoost-based recognition model was the best by comparing it with several selection methods and other classification algorithms.

2. Materials and Methods

2.1. Data Description. The identification of lncRNAs was actually a binary classification problem, and we defined lncRNAs as positive and mRNAs as negative. According to the way of survival, plants can be divided into algae, lichen, fungi, moss, fern, and seed (gymnosperm and angiosperm). To establish a plant lncRNA identification model with strong generalization ability, we used five representative plant species: *Arabidopsis thaliana* (dicotyledon), *Brachypodium distachyon* (monocotyledon), *Chlamydomonas reinhardtii* (algae), *Physcomitrella patens* (moss), and *Selaginella moellendorffii* (fern), hereinafter referred to as AT, BD, CR, PP, and SM. The positive sample data (lncRNAs) were obtained from CANTATAdb 2.0 (https://yeti.amu.edu.pl/CANTATA/), which was an online database of 39 species of plants such as *Arabidopsis thaliana*, *Zea mays*, *Oryza,* and three algae [25]. The negative data were downloaded from RefSeq (https://www.ncbi.nlm.nih.gov/refseq/), including nonredundant gene and protein sequences with biological significance provided by the National Center for Bioinformatics (NCBI), in which we can screen for gene sequences by species, molecular types, source databases, sequence length range, and so on.

As shown in Figure 1, it was the frequency distribution histogram of the sequence length of lncRNAs and mRNAs in five different plants, where all sequence lengths of lncRNAs were below 10,000 nt, and over 98.5% mRNA sequence lengths of five plants were between 200 nt and 10,000 nt (see Table 1). Then, we removed gene sequences below 200 nt and above 10,000 nt. At this time, the positive data were still far smaller than the negative, which was a typical unbalanced data that would have an impact on the performance of the model to some extent, and therefore, we aligned the positive sample with the negative sample data size by stratified sampling.

The conventional stratified sampling was stratified by equal spacing of sequence length, followed by random sampling, so we took 2000 nt as the one unit, and the results are shown in Figure 2, which indicated that the distribution of mRNA sequence lengths of the five plants was uneven. Consequently, we sorted the sequence lengths from small to large, divided them into ten parts, and randomly selected 120 samples from each part. Therefore, 1200 positive and 1200 negative samples for each plant were randomly chosen, and the sizes of original data and used data are shown in Table 2.

2.2. Feature Extraction. Prior to the model construction, we extracted 89 features including sequence features (sequence length, GC content, LORF length and coverage, k-mers) and structural feature (RNA secondary structure), where the LORF length was defined as the longest open reading frame of the three forward frames, starting with a start codon (ATG) and ending with a stop codon (TAG, TAA, or TGA), extracted by the online software ORFfinder. The LORF coverage represented the ratio of LORF length to sequence length [26]. In terms of features like k-mers, we considered $k = 1$, 2, and 3, a total of 84 features. While the sequence features of the lncRNAs represented the surface content of the nucleotide sequence, the secondary structural characteristics might imply some important functional information [21]. We measured the RNA secondary structural feature using the minimum free energy (MFE) of the sequence extracted by the online software RNAfold. As shown in Figure 3, the lncRNAs of all five plants had higher MFE than mRNAs.

2.3. Feature Selection. According to the different combination methods with the learner, the feature selection methods could be divided into three categories: filter, wrapper, and embedded.

The filter feature selection methods are irrelevant to the learning algorithm that the feature selection is the preprocessing process of the latter, while the learning algorithm is the former's verification process. The principle is to score each feature with specific evaluation criteria, to rank features

RETRACTED

Figure 1: Sequence length frequency distribution of different plants. (a) *Arabidopsis thaliana* (AT). (b) *Brachypodium distachyon* (BD). (c) *Chlamydomonas reinhardtii* (CR). (d) *Physcomitrella patens* (PP). (e) *Selaginella moellendorffii* (SM). The horizontal axis is sequence length, and the longitudinal axis represents density.

RETRACTED

Table 1: Numeric description of mRNAs in the five plant data sets.

Species	Original	[0, 200 nt)	(10,000 nt, +∞)	[200 nt, 10,000 nt]
Arabidopsis thaliana	48,347	304	44	47,999 (99.28%)
Brachypodium distachyon	37,816	1	29	37,786 (99.92%)
Chlamydomonas reinhardtii	14,430	125	78	14,227 (98.59%)
Physcomitrella patens	47,897	0	175	47,722 (99.63%)
Selaginella moellendorffii	45,248	0	51	45,197 (99.89%)

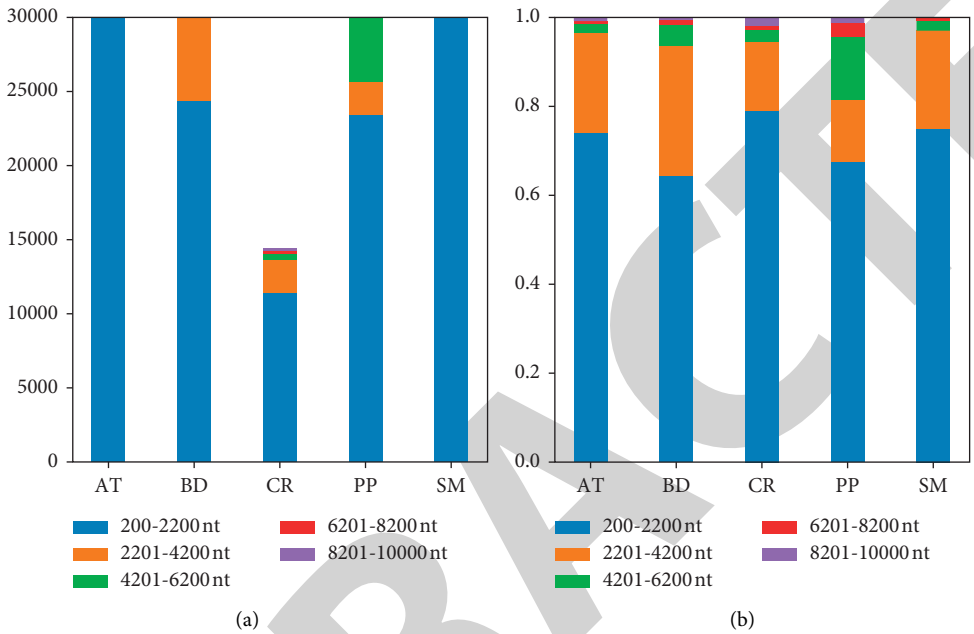

Figure 2: Distribution of mRNA sequence lengths of the five different plants. (a) Number of mRNAs in each interval. (b) Corresponding frequency.

Table 2: Numeric description of the lncRNA and mRNA data sets used to build the model.

Species	lncRNA		mRNA	
	Original data	Used data	Original data	Used data
Arabidopsis thaliana	4373	1200	48,347	1200
Brachypodium distachyon	4945	1200	37,816	1200
Chlamydomonas reinhardtii	3425	1200	14,430	1200
Physcomitrella patens	1498	1200	47,897	1200
Selaginella moellendorffii	2267	1200	45,248	1200

according to the score, then to select the top k features as the subset of features (or to select the features whose values are larger than threshold as the feature subset), and finally to verify the quality of the subset by calculating the accuracy of machine-learning algorithms. Commonly used evaluation criteria include chi-square test and mutual information. Here, we discussed seven evaluation criteria including chi_square, f_score, gini_index, t_score, fisher_score, FCD, and FCQ for feature screening [27,28].

The wrapper feature selection process combines with the learning algorithm to encapsulate the selected learner into a black box, evaluates the excellence of the selected features according to its prediction accuracy of machine learning, and adjusts the subset using the search strategy to finally obtain the approximate optimal subset. The wrapper feature selection methods generally include the sequence search and random search strategies. Sequence search can be divided into three categories: forward search, backward search, and bidirectional search, in which the sequential forward search (SFS) starts with an empty set, and the feature with the highest score is greedily added to the subset of selected features each time [29]. The intelligent optimization algorithms are random search strategies based on biological intelligence or physical phenomena, such as genetic algorithm and particle swarm optimization, which generally do not require the continuous type and convexity of objective functions and constraints, and also have strong adaptability to the data uncertainty in computation. We selected twelve intelligent optimization algorithms and one sequence search strategy-based algorithm as the wrapper feature selection methods, which are genetic algorithm [30], particle swarm optimization [31], differential evolution [32], cuckoo search

RETRACTED

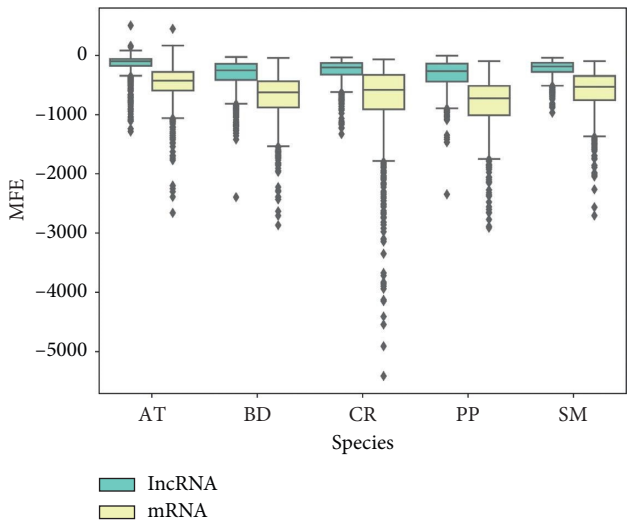

Figure 3: Boxplots of MFE. Green indicated the MFE dispersion of lncRNAs, and yellow indicated the MFE dispersion of mRNAs. Compared with protein-coding transcripts, lncRNAs had higher MFE.

algorithm [33], firefly algorithm [34], bat algorithm [35], flower pollination algorithm [36], grey wolf optimizer [37], sine-cosine algorithm [38], whale optimization algorithm [39], salp swarm algorithm [40], and Harris hawks optimization [41], hereinafter referred to as GA, PSO, DE, CS, FA, BA, FPA, GWO, SCA, WOA, SSA, and HHO.

Embedded methods embed feature selection into the construction of the model, such as Lasso and tree models, with higher accuracy and less computational complexity than filter and wrapper methods, yielding a subset of features when the classification algorithm training process ends. Since embedded methods based on the tree model are also classification models, we did not consider it here.

2.4. Model Construction. After the feature selection, we constructed the model with the meta-learner XGBoost. XGBoost is an abbreviation for extreme gradient boosting, an integrated machine-learning algorithm based on the decision tree, using a gradient boosting framework. The algorithm has a wide range of applications and can be adopted to solve regression, classification, ranking, and user custom prediction problems. XGBoost employs a gradient descent structure to improve the learning of weak learners (CART), implements the sequence tree building process using parallelization, and punishes more complex models via L_1 ridge L_2 regularization to prevent overfitting [42], while the algorithm carries built-in cross-validation methods in each iteration without assigning the exact number of boost iterations.

We obtained the best parameters of the model using the grid search technique to tune these parameters step by step including n_estimators, max_depth, min_child_weight, gamma, subsample, colsample_bytree, reg_alpha, reg_lambda, and learning_rate (see Table 3), and other hyperparameters were default. To evaluate the performance of the model, we followed the following five evaluation criteria:

$$\begin{aligned}\text{Accuracy} &= \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}},\\ \text{Precision} &= \frac{\text{TP}}{\text{TP}+\text{FP}},\\ \text{Recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}},\\ F_1 &= \frac{2\text{TP}}{2\text{TP}+\text{FN}+\text{FP}},\end{aligned} \tag{1}$$

and area under the receiver operating characteristic curve (AUC), where TP is the number of true positives (long noncoding transcripts correctly classified as lncRNAs), TN is the number of true negatives (coding transcripts correctly classified as mRNAs), FP is the number of false positives (coding transcripts incorrectly classified as lncRNAs), and FN is the number of false negatives (long noncoding transcripts incorrectly classified as mRNAs).

3. Results and Discussion

3.1. Comparison between the Feature Selection Methods. Several feature selection methods with better performance were selected by comparing the classification accuracy of 20 feature selection methods on each single classifier (K-nearest Neighbors, Naive Bayes, and Support Vector Machine, hereinafter referred to as KNN, GaussianNB, and SVM). For seven filter methods, we analyzed the variation tendency of the accuracy under different numbers of features and chose several filter feature selection methods that performed better in all three single classifiers.

Figures 4(a)–4(c) depict the changing characteristics of seven filter feature selection methods under the number of features in three single classifiers including KNN, GaussianNB, and SVM, respectively, and the results are shown in Supplementary Table S1. Figure 4(a) indicates that in KNN, when the number of features was $n \geq 1$, each of the seven filter feature selection methods had higher classification accuracy than using all features. When $n \geq 8$, the classification accuracy of all feature selection methods decreased with the number of features. Among them, the accuracy of the t_score and fisher_score, respectively, achieved the highest (88.85% and 88.92%) at $n = 8$, and these two methods had similar accuracy rate and consistent change trend under different values of n. While the f_score, gini_index, and FCD have already reached the highest accuracy (91.21%, 91.21%, and 91.14%) at $n = 2$, chi_square had a 91.28% accuracy at $n = 4$; the highest accuracy of FCQ was 90.35% at $n = 3$. Therefore, in KNN, chi_square, f_score, gini_index, and FCD performed well.

Figure 4(b) indicates that, in Gaussian NB, all filter methods had higher classification accuracy than without feature selection. Thereinto, the methods of FCD, FCQ, gini_index, and chi_square had better performance (90.05%, 89.87%, 89.69%, and 89.03%) and achieved within $n = 7$. In Figure 4(c), t_score and fisher_score had the same tendency under different n values, but within $n = 30$, no accuracy was larger than the one using all 89 features in SVM. The

RETRACTED

Table 3: Parameter setting by grid search for XGBoost.

Parameter	Range	Used	Description
max_depth	[3, 4, 5, 6, 7, 8, 9, 10]	3	Maximum tree depth for base learners
min_child_weight	[1, 2, 3, 4, 5]	4	Minimum sum of instance weight (hessian) needed in a child
gamma	[0, 0.1, 0.2, 0.3, 0.4, 0.5]	0.2	Minimum loss reduction required to make a further partition on a leaf
subsample	[0.6, 0.7, 0.8, 0.9, 1]	0.8	Subsample ratio of the training instance
colsample_bytree	[0.6, 0.7, 0.8, 0.9, 1]	0.8	Subsample ratio of columns when constructing each tree
reg_alpha	$[1e-5, 1e-4, 1e-3, 1e-2, 0.1, 1, 100]$	$1e-05$	L_1 regularization term on weights
reg_lambda	[0.05, 0.1, 1, 2, 3]	1	L_2 regularization term on weights
learning_rate	[0.01, 0.05, 0.07, 0.1, 0.2]	0.07	Boosting learning rate

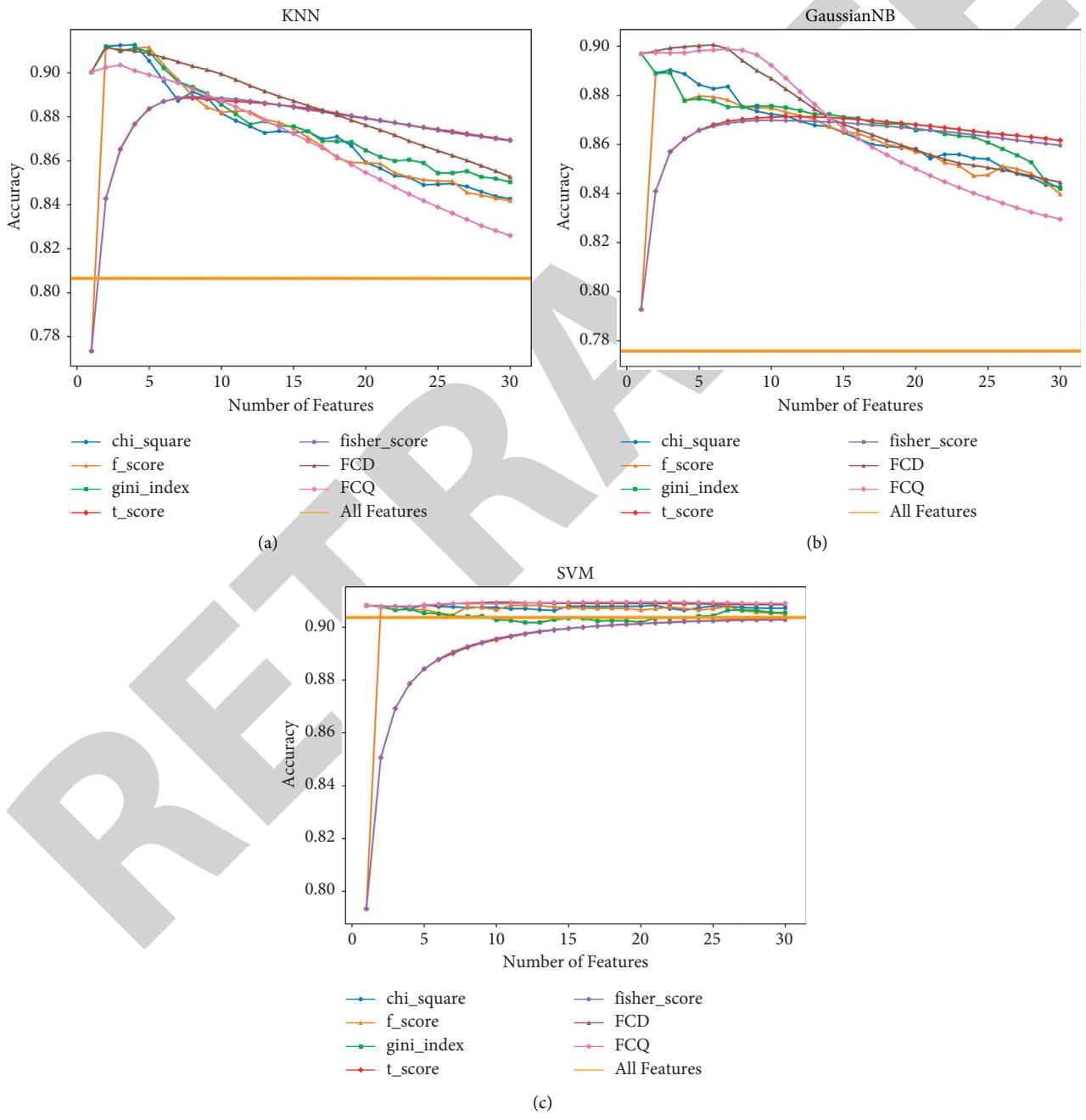

Figure 4: Accuracy of different filter feature selection methods in three classification models.

RETRACTED

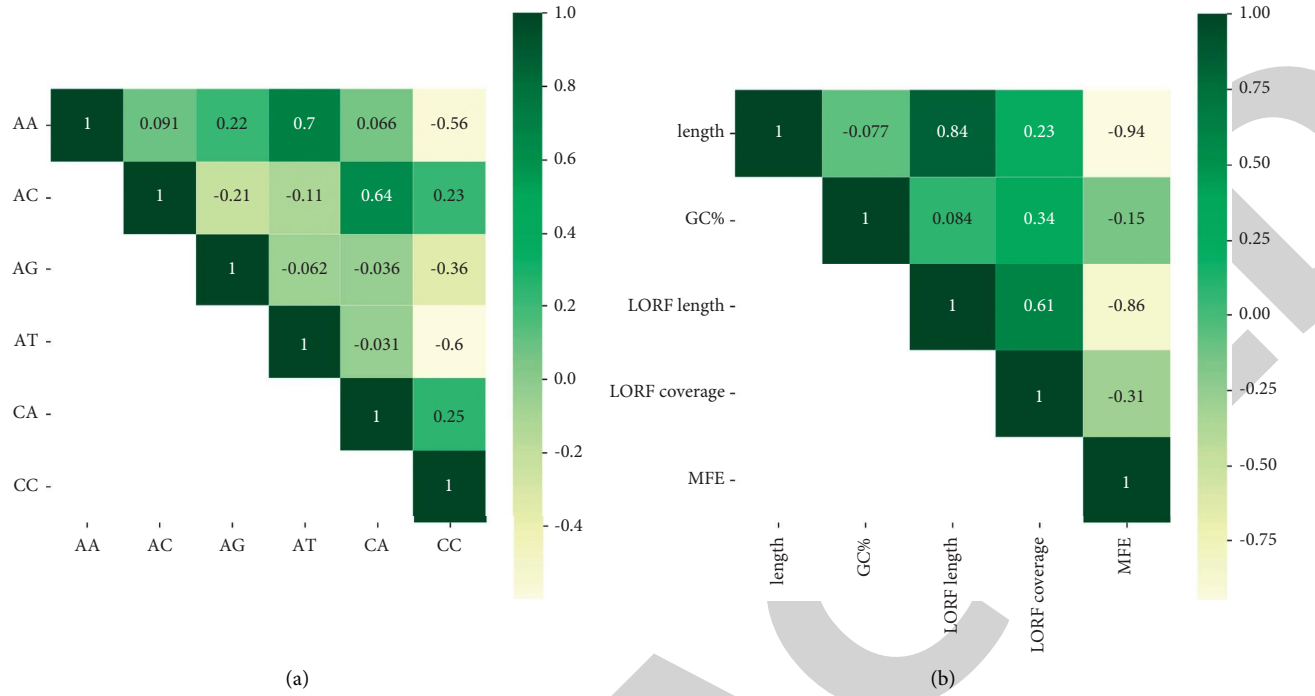

Figure 5: Correlations between the features. The positive numbers indicate a positive correlation between features, the negative numbers indicate a negative correlation, and the larger absolute values indicate a higher correlation.

classification accuracy of feature screening according to gini_index criteria has been floating around the classification accuracy for all features. All the accuracy of FCD was similar to FCQ, and the variation trend was consistent at the same time. In SVM, obtaining better performance was the FCD, FCQ, chi_square, and *f*_score, whose accuracy was 90.94%, 90.94%, 90.84%, and 90.83%, respectively.

In general, the filter feature selection methods with better performance in all three single classifiers were FCD and chi_square. However, the subset of features screening by the chi_square criterion contained both sequence length and LORF length, which had a high correlation coefficient of 0.84 (see Figure 5), easily resulting in model redundancy, so we chose FCD for the next experiment.

For the wrapper feature selection methods, we searched for each intelligent optimization algorithm twenty times to obtain different subsets of features and selected subsets of features with higher accuracy under an individual classifier. The number of elements in the feature subsets was specified for sequence forward search (SFS) from 1 to 30, and then, we chose the feature subset with higher accuracy on all single classifiers. The results are shown in Figure 6.

Figure 6 shows the classification accuracy of 20 feature selection methods on three individual classifiers, where on both Gaussian NB and SVM, the accuracy of wrapper feature selection methods was significantly higher than filter, but wrapper methods had higher computational complexity and were more time-consuming. Figure 6(a) illustrates seven filter feature selection methods showing that FCD had the best performance. Figure 6(b) shows the classification accuracy of the 13 wrapper methods, where methods with the highest accuracy, respectively, on these three classifiers are GWO, HHO, and WOA. Therefore, we used four feature selection methods (FCD, GWO, HHO, and WOA) combined with the ensemble learner XGBoost after tuning parameters to construct the model.

3.2. Performance Comparison with Other Classifiers. We analyzed seven classification algorithms (KNN, Naive Bayes, SVM, Decision Tree, Random Forest, AdaBoost, and XGBoost) and tuned parameters for each model using grid search techniques. We trained the models with fivefold cross-validation techniques, with the performance of each classification model shown in Figure 7.

As shown in Figure 7, XGBoost had the highest AUC value of the seven classifiers with better classification performance. It can be seen from Figure 4 that the accuracy of models after using a filter method was higher than that without feature selection, while from Figure 6, the overall accuracy of models with the wrapper feature selection methods was obviously higher than that of the filter. Therefore, the accuracy of the classifiers combined with a feature selection method was better than the one with all features. The performance of the four feature selection methods combined with seven classifiers is shown in Table 4.

The performance of the four different feature selection methods on seven classifiers is shown in Table 4. When the feature selection method was FCD, Random Forest and AdaBoost had the highest accuracy of 91.47% and the highest AUC is 96.38% on XGBoost, while the SVM had the highest F_1_score of 91.56%. In the intelligent optimization algorithms such as WOA and HHO, the values of accuracy, AUC, and F_1_score were taken on XGBoost at the maximum. Thereinto,

RETRACTED

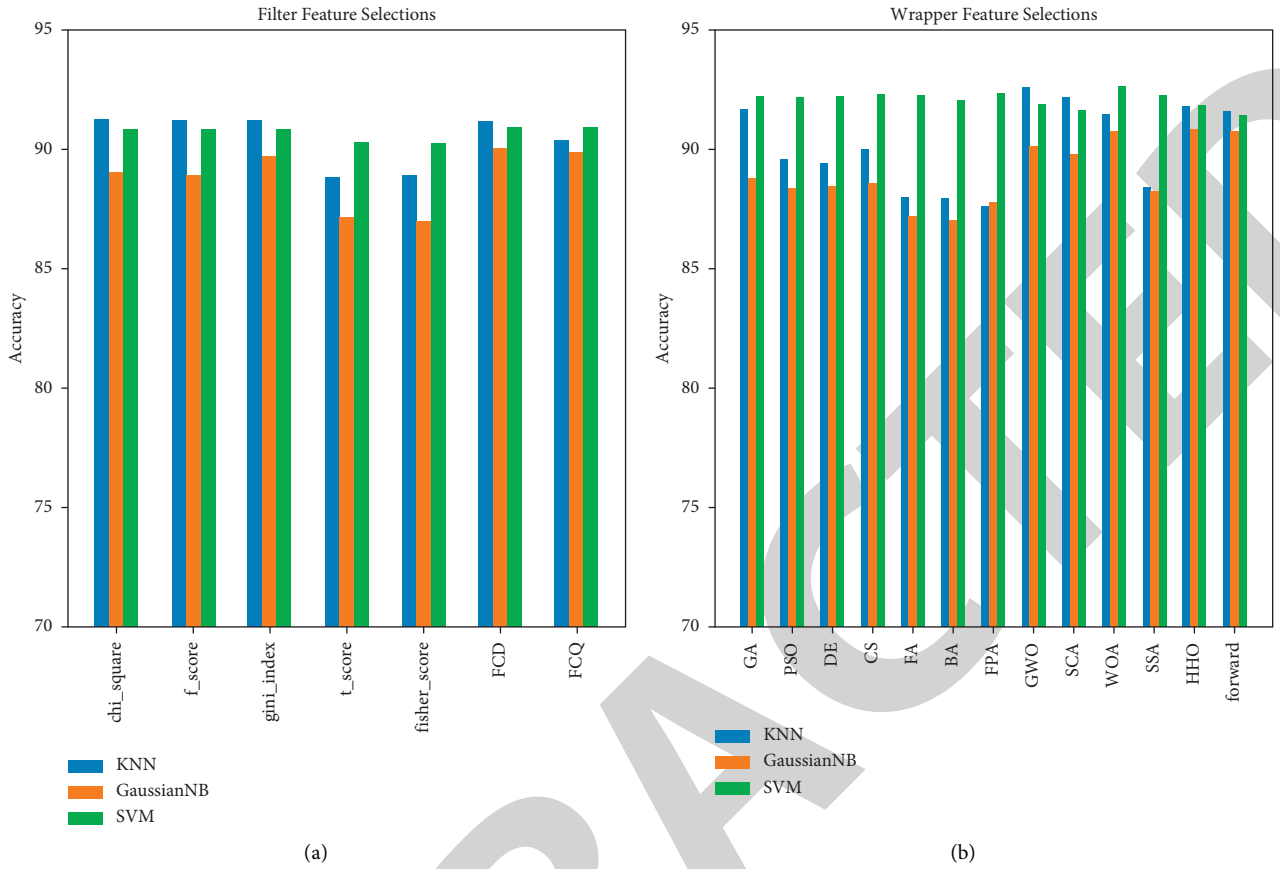

Figure 6: Accuracy of the 20 feature selection methods on the different classifiers. (a) Seven filter feature selection methods. (b) Thirteen wrapper feature selection methods.

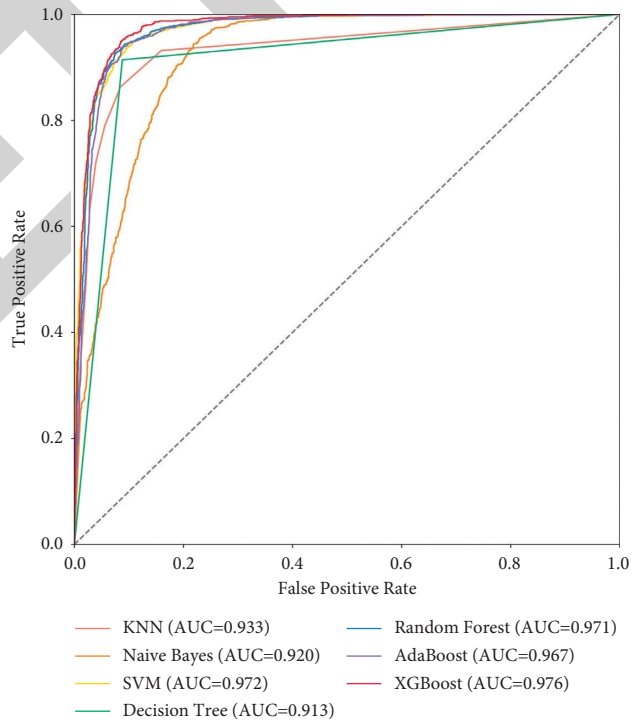

Figure 7: ROC curves for the different classifiers. The AUC values for each model have been given.

RETRACTED

Table 4: Comparative performance among four feature selection methods combined with KNN, GaussianNB, SVM, Decision Tree, Random Forest, AdaBoost, and XGBoost, respectively.

Methods	Models	Accuracy	Precision	Recall	AUC	F_1_score
FCD	KNN	91.14	90.46	95.17	94.54	91.27
	GaussianNB	90.21	87.29	**96.13**	95.50	90.83
	SVM	91.21	88.99	94.82	96.17	**91.56**
	Decision Tree	91.37	**91.34**	91.70	91.37	91.34
	Random Forest	**91.47**	91.17	93.30	95.88	91.48
	AdaBoost	**91.47**	**91.34**	92.80	96.05	91.54
	XGBoost	91.33	91.27	92.58	**96.38**	91.35
GWO	KNN	91.12	89.98	92.93	94.97	91.28
	GaussianNB	87.73	84.38	93.37	94.52	88.48
	SVM	90.76	88.36	**94.57**	95.92	91.16
	Decision Tree	91.37	91.34	91.70	91.37	91.34
	Random Forest	91.75	91.06	92.87	96.35	91.95
	AdaBoost	**91.89**	**91.71**	92.48	**96.51**	**91.90**
	XGBoost	91.56	90.73	93.05	96.50	91.69
WOA	KNN	84.79	**91.02**	77.55	91.91	82.70
	GaussianNB	79.73	79.27	86.52	90.17	81.42
	SVM	86.78	87.02	87.53	94.35	86.78
	Decision Tree	79.69	82.75	76.07	79.69	78.26
	Random Forest	90.54	88.08	**94.22**	96.48	90.80
	AdaBoost	89.46	88.51	92.05	95.57	89.78
	XGBoost	**91.55**	90.46	93.33	**96.78**	**91.68**
HHO	KNN	83.58	**91.79**	74.35	91.43	80.70
	GaussianNB	74.87	78.36	81.58	88.04	76.98
	SVM	86.33	87.19	86.42	94.39	86.17
	Decision Tree	79.69	82.75	76.07	79.69	78.26
	Random Forest	89.71	87.35	**94.23**	96.05	90.69
	AdaBoost	89.23	88.59	91.31	95.52	89.47
	XGBoost	**91.41**	90.65	92.77	**96.79**	**91.48**

The bold values represent the maximum value in each column of evaluation indicators under each feature selection method.

values on the WOA were 91.55%, 96.78%, and 91.68%, while values on the HHO were 91.41%, 96.79%, and 91.48%. It can be seen that the WOA-XGBoost-based algorithm, with higher, AUC, and F_1_score. Similarly, performed a little better than HHO-XGBoost and FCD. the intelligent optimization algorithm GWO achieved the highest accuracy, precision, AUC, and F_1_score values on AdaBoost of 91.89%, 91.71%, 96.51%, and 91.90%, respectively. Although the GWO-AdaBoost-based algorithm performed better than the WOA-XGBoost-based algorithm, the feature subset of the former contained sequence length and LORF length, which had a high correlation (see Figure 5) and would improve the redundancy of the model. Therefore, the performance based on the WOA-XGBoost algorithm was best among twenty feature selection methods in seven classifiers. In brief, 89 features were reduced to 23 features with the WOA-XGBoost algorithm, effectively cutting down computation complexity, and the selected feature subset was C, CA, CT, GC, AGT, CAA, CCT, CGA, CTT, GAC, GCA, GCC, GCG, GCT, GGA, GTC, TCA, TGA, TGG, TGT, sequence length, LORF coverage, and MFE.

4. Conclusions

To establish a specialized plant lncRNA recognition model, we used five plant data sets of different species. Furthermore, after extracting 89 features, we compared the accuracy of 20 feature selection methods on three single classifiers, respectively, and then combined FCD, GWO, HHO, and WOA, respectively, with XGBoost while comparing the accuracy, precision, recall, AUC, and F_1_score score of the other six models (parameters of all classification models were tuned). The results showed that the performance based on the WOA-XGBoost algorithm was best among the twenty feature selection methods and seven classifiers. The identification model we developed was specifically for plants, using not only the dicot (*Arabidopsis thaliana*) but also adding the monocot (*Brachypodium distachyon*), algae (*Chlamydomonas reinhardtii*), moss (*Physcomitrella patens*), and fern (*Selaginella moellendorffii*) to enrich the diversity of the data sets. Moreover, the traditional recognition models generally either did not carry on feature selection containing a few features, directly using all the extracted features, which easily lead to overfitting of the model, or used some or some class of feature selection methods before building the model. While we compared seven filter and thirteen wrapper methods, the redundancy of the model was further considered In short, 89 features were reduced to 23 features using the WOA-XGBoost algorithm. In future work, we plan to extract more features and use the identification algorithm to verify its performance in other plant data sets.

Data Availability

The data used to support the findings of this study can be obtained free of charge from CANTATAdb 2.0 and RefSeq.

RETRACTED

The positive sample data (lncRNAs) can be obtained from CANTATAdb 2.0: https://yeti.amu.edu.pl/CANTATA/, and the negative data (mRNAs) can be downloaded from RefSeq: https://www.ncbi.nlm.nih.gov/refseq/.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this study.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (nos. 62072296 and 61672001) and Sub-Project of CST Forward Innovation Project (no. 18163ZT00500901).

Supplementary Materials

Table S1: the accuracy of seven filter feature selection methods under the different numbers of features in three single classifiers including KNN, GaussianNB, and SVM, respectively. (*Supplementary Materials*)

References

[1] V. Pachnis, A. Belayew, and S. M. Tilghman, "Locus unlinked to alpha-fetoprotein under the control of the murine raf and Rif genes," *Proceedings of the National Academy of Sciences*, vol. 81, no. 17, pp. 5523–5527, 1984.

[2] L. Kong, Y. Zhang, Z.-Q. Ye et al., "CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine," *Nucleic Acids Research*, vol. 35, no. 2, pp. W345–W349, 2007.

[3] L. Sun, H. Luo, D. Bu et al., "Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts," *Nucleic Acids Research*, vol. 41, no. 17, Article ID e166, 2013.

[4] A. Li, J. Zhang, and Z. Zhou, "PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme," *BMC Bioinformatics*, vol. 15, no. 1, pp. 311–410, 2014.

[5] L. Sun, H. Liu, L. Zhang, and J. Meng, "lncRScan-SVM: a tool for predicting long non-coding RNAs using support vector machine," *PLoS One*, vol. 10, no. 10, Article ID e0139654, 2015.

[6] Y.-J. Kang, D.-C. Yang, L. Kong et al., "CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features," *Nucleic Acids Research*, vol. 45, no. W1, pp. W12–W16, 2017.

[7] L. Wang, H. J. Park, S. Dasari, S. Wang, J.-P. Kocher, and W. Li, "CPAT: coding-potential assessment tool using an alignment-free logistic regression model," *Nucleic Acids Research*, vol. 41, no. 6, Article ID e80, 2013.

[8] R. Achawanantakun, J. Chen, Y. N. Sun, and Y. Zhang, "LncRNA-ID: long non-coding RNA IDentification using balanced random forests," *Bioinformatics*, vol. 31, no. 24, pp. 3897–3905, 2015.

[9] S. Liu, X. Zhao, G. Zhang et al., "PredLnc-GFStack: a global sequence feature based on a stacked ensemble learning method for predicting lncRNAs from transcripts," *Genes*, vol. 10, no. 9, pp. 672–684, 2019.

[10] J.-C. Guo, S.-S. Fang, Y. Wu et al., "CNIT: a fast and accurate web tool for identifying protein-coding and long non-coding transcripts based on intrinsic sequence composition," *Nucleic Acids Research*, vol. 47, no. W1, pp. W516–W522, 2019.

[11] X.-N. Fan and S.-W. Zhang, "lncRNA-MFDL: lncRNA-MFDL: identification of human long non-coding RNAs by fusing multiple features and using deep learning," *Molecular BioSystems*, vol. 11, no. 3, pp. 892–897, 2015.

[12] J. Meng, Z. Chang, P. Zhang, W. Shi, and Y. Luan, "lncRNA-LSTM: prediction of plant long non-coding RNAs using long short-term memory based on p-nts encoding," in *Proceedings of the 15th International Conference on Intelligent Computing (ICIC 2019)*, pp. 347–357, Nanchang, China, August 2019.

[13] J. Sun, H. Shi, Z. Wang et al., "Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network," *Molecular BioSystems*, vol. 10, no. 8, pp. 2074–2081, 2014.

[14] Y. Xiao, Y. Lv, H. Zhao et al., "Predicting the functions of long noncoding RNAs using RNA-seq based on Bayesian network," *BioMed Research International*, vol. 2015, Article ID 839590, 15 pages, 2015.

[15] X. Chen, C. C. Yan, X. Zhang, and Z. H. You, "Long non-coding RNAs and complex diseases: from experimental results to computational models," *Briefings in Bioinformatics*, vol. 18, no. 4, pp. 558–576, 2017.

[16] E. A. Kiegle, A. Garden, E. Lacchini, and M. M. Kater, "A genomic view of alternative splicing of long non-coding RNAs during rice seed development reveals extensive splicing and lncRNA gene families," *Frontiers in Plant Science*, vol. 9, p. 115, 2018.

[17] T. Qin, H. Zhao, P. Cui, N. Albesher, and L. Xiong, "A nucleus-localized long non-coding RNA enhances drought and salt stress tolerance," *Plant Physiology*, vol. 175, no. 3, pp. 1321–1336, 2017.

[18] Z. Wang, B. Li, Y. Li et al., "Identification and characterization of long noncoding RNA in Paulownia tomentosa treated with methyl methane sulfonate," *Physiology and Molecular Biology of Plants*, vol. 24, no. 2, pp. 325–334, 2018.

[19] U. Singh, N. Khemka, M. S. Rajkumar, R. Garg, and M. Jain, "PLncPRO for prediction of long non-coding RNAs (lncRNAs) in plants and its application for discovery of abiotic stress-responsive lncRNAs in rice and chickpea," *Nucleic Acids Research*, vol. 45, no. 22, Article ID e183, 2017.

[20] C. M. A. Simopoulos, E. A. Weretilnyk, and G. B. Golding, "Prediction of plant lncRNA by ensemble machine learning classifiers," *BMC Genomics*, vol. 19, no. 1, pp. 316–411, 2018.

[21] S. Han, Y. Liang, Q. Ma et al., "LncFinder: an integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property," *Briefings in Bioinformatics*, vol. 20, no. 6, pp. 2009–2027, 2019.

[22] T. d. C. Negri, W. A. L. Alves, P. H. Bugatti, P. T. M. Saito, D. S. Domingues, and A. R. Paschoal, "Pattern recognition analysis on long noncoding RNAs: a tool for prediction in plants," *Briefings in Bioinformatics*, vol. 20, no. 2, pp. 682–689, 2019.

[23] H. B. Cagirici, S. Galvez, T. Z. Sen, and H. Budak, "LncMachine: a machine learning algorithm for long non-coding RNA annotation in plants," *Functional and Integrative Genomics*, vol. 21, no. 2, pp. 195–204, 2021.

[24] S. W. Yin, X. Tian, J. J. Zhang, P. Sun, and G. L. Li, "PCirc: random forest-based plant circRNA identification software," *BMC Bioinformatics*, vol. 22, no. 1, pp. 1–14, 2021.

[25] M. W. Szcześniak, O. Bryzghalov, J. Ciomborowska-Basheer, and I. Makałowska, "CANTATAdb 2.0: expanding the collection of plant long noncoding RNAs," *Methods in Molecular Biology*, vol. 1933, pp. 415–429, 2019.

RETRACTED

[26] C. Yang, L. Yang, M. Zhou et al., "LncADeep: anab initiolncRNA identification and functional annotation tool based on deep learning," *Bioinformatics*, vol. 34, no. 22, pp. 3825–3834, 2018.

[27] J. D. Li, K. W. Cheng, S. H. Wang et al., "Feature selection: a data perspective," *ACM Computing Surveys*, vol. 50, no. 6, pp. 1–45, 2017.

[28] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 2, pp. 185–205, 2005.

[29] A. W. Whitney, "A direct method of nonparametric measurement selection," *IEEE Transactions on Computers*, vol. C-20, no. 9, pp. 1100–1103, 1971.

[30] H. Vafaie and K. D. Jong, "Genetic algorithms as a tool for feature selection in machine learning," in *Proceedings of the Fourth International Conference on Tools with Artificial Intelligence TAI'92*, pp. 200–203, Arlington, VA, USA, November 1992.

[31] R. Sharkawy, K. Ibrahim, M. M. A. Salama, and R. Bartnikas, "Particle swarm optimization feature selection for the classification of conducting particles in transformer oil," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 18, no. 6, pp. 1897–1907, 2011.

[32] R. N. Khushaba, A. Al-Ani, and A. Al-Jumaily, "Differential evolution based feature subset selection," in *Proceedings of the 2008 19th International Conference on Pattern Recognition*, pp. 1–4, IEEE, Tampa, FL, USA, December 2008.

[33] D. Rodrigues, L. A. M. Pereira, T. N. S. Almeida et al., "BCS: a binary cuckoo search algorithm for feature selection," in *Proceedings of the 2013 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 465–468, IEEE, Beijing, China, May 2013.

[34] H. Banati and M. Bajaj, "Fire fly based feature selection approach," *International Journal of Computer Science Issues*, vol. 8, no. 2, pp. 473–481, 2011.

[35] R. Y. M. Nakamura, L. A. M. Pereira, K. A. Costa, D. Rodrigues, J. P. Papa, and X.-S. Yang, "BBA: a binary bat algorithm for feature selection," in *Proceedings of the 25th SIBGRAPI Conference on Graphics, Patterns and Images*, pp. 291–297, IEEE, Ouro Preto, Brazil, August 2012.

[36] D. Rodrigues, X.-S. Yang, A. N. de Souza, and J. P. Papa, "Binary flower pollination algorithm and its application to feature selection," *Recent Advances in Swarm Intelligence and Evolutionary Computation*, Springer, Berlin, Germany, 2015.

[37] E. Emary, H. M. Zawbaa, and A. E. Hassanien, "Binary grey wolf optimization approaches for feature selection," *Neurocomputing*, vol. 172, pp. 371–381, 2016.

[38] A. I. Hafez, H. M. Zawbaa, E. Emary, and A. E. Hassanien, "Sine cosine optimization algorithm for feature selection," in *Proceedings of the 2016 International Symposium on Innovations in Intelligent Systems and Applications*, pp. 1–5, IEEE, Sinaia, Romania, August 2016.

[39] M. Sharawi, H. M. Zawbaa, and E. Emary, "Feature selection approach based on whale optimization algorithm," in *Proceedings of the 2017 Ninth International Conference on Advanced Computational Intelligence*, pp. 163–168, IEEE, Doha, Qatar, February 2017.

[40] H. T. Ibrahim, W. J. Mazher, O. N. Ucan, and O. Bayat, "Feature selection using salp swarm algorithm for real biomedical datasets," *International Journal of Computer Science and Network Security*, vol. 17, no. 12, pp. 13–20, 2017.

[41] J. Too, A. R. Abdullah, and N. M. Mohd Saad, "A new quadratic binary Harris hawk optimization for feature selection," *Electronics*, vol. 8, no. 10, pp. 1130–1156, 2019.

[42] T. Q. Chen and C. Guestrin, "Xgboost: a scalable tree boosting system," in *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, San Francisco, CA, USA, August 2016.

RETRACTED