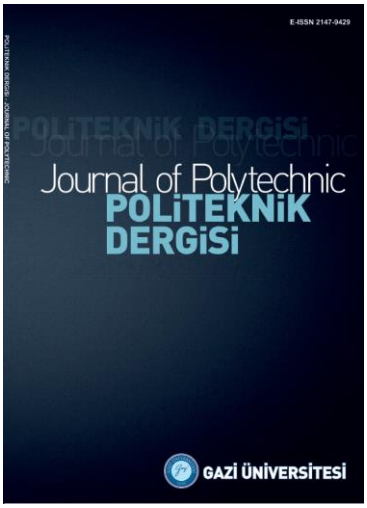

POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

Improving long non-coding RNA prediction through recursive feature elimination and XGBoost

Tekrarlayan özellik eliminasyonu ve XGBoost ile uzun kodlamayan RNA tahmininin iyileştirilmesi

***Yazar(lar) (Author(s))**: Freshta Alizada[1], Volkan ALTUNTAŞ[2]*

ORCID[1]: 0009-0009-7632-0274

ORCID[2]: 0000-0003-3144-8724

To cite to this article: Alizada F, and Altuntaş V., "Improving Long Non-Coding RNA Prediction through Recursive Feature Elimination and XGBoost ", ***Journal of Polytechnic***, *(*): *, (*).

Bu makaleye şu şekilde atıfta bulunabilirsiniz: Alizada F. ve Altuntaş V., "Tekrarlayan Özellik Eliminasyonu ve XGBoost ile Uzun Kodlamayan RNA Tahmininin İyileştirilmesi", ***Politeknik Dergisi***, *(*): *, (*).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: 10.2339/politeknik.1627668

Improving Long Non-Coding RNA Prediction through Recursive Feature Elimination and XGBoost

Highlights

- *Presenting a robust machine learning pipeline for distinguishing lncRNA sequences from protein-coding RNAs (messenger RNAs).*
- *Employing recursive feature elimination as the feature selection algorithm to address the dimensionality issue of the feature dataset.*
- *Accessing higher predictive performance in the context of lncRNA prediction compared to three established lncRNA prediction tools in the literature.*

Graphical Abstract

We followed several steps and employed some machine learning algorithms as it is seen in bellow Figure to predict lncRNA sequences.

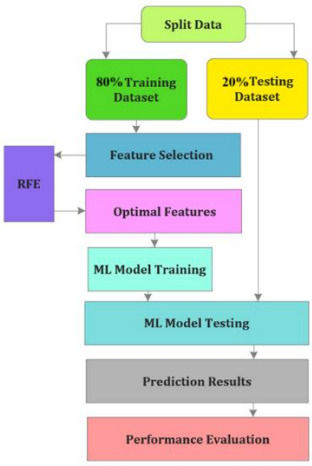

Figure. Proposed pipeline for lncRNAs prediction.

Aim

This study aims to distinguish lncRNAs from mRNAs by designing a robust machine learning pipeline.

Design & Methodology

As lncRNAs feature dataset, We utilized the lncRNA feature dataset from an existing research paper in the literature. To address the dimensionality issue, we applied the Recursive Feature Elimination (RFE) algorithm. Subsequently, we employed various machine learning classification algorithms to predict lncRNAs.

Originality

While most of studies in the literature often employ deep learning or design novel decomposition models to reduce dimensionality and achieve high accuracy in predicting lncRNAs, our study proposes a robust pipeline using existing machine learning algorithms, which demonstrates higher accuracy rates than most of them.

Findings

Recursive Feature Elimination (RFE) with XGBoost Classifier achieved the highest accuracy rate (92.57%) compared to combinations of RFE with other classifiers used in this study.

Conclusion

This study amied to distinguish lncRNAs from protein-coding RNAs by empolying RFE algorithm as features selection algorithm and SVM, NN, REPTree, LR and RF as classifiers. Concequently, combination of RFE and RF has gained the higherst accuracy rate of 93.42%.

Declaration of Ethical Standards

The authors of this research paper state that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

Improving Long Non-Coding RNA Prediction through Recursive Feature Elimination and XGBoost

Araştırma Makalesi / Research Article

Freshta Alizada[1*], Volkan ALTUNTAŞ[2]

[1] Faculty of Engineering and Natural Sciences, Department of Computer Engineering, Bursa Technical University, Turkey
[2] Faculty of Engineering and Natural Sciences, Department of Computer Engineering, Bursa Technical University, Turkey

(Geliş/Received : 27.01.2025 ; Kabul/Accepted : 18.05.2025 ; Erken Görünüm/Early View : 22.05.2025)

ABSTRACT

In recent years, advancements in high-throughput technologies have uncovered numerous concealed layers known as Non-Coding Ribonucleic Acids (ncRNAs), shifting the protein-centric view of genomes. NcRNAs, previously considered insignificant segments of the genome, are now recognized as essential functional components in prokaryotic and eukaryotic organisms. Long non-coding RNAs (lncRNAs) are a unique category of ncRNAs with 200 nucleotides length, which are instrumental in key biological functions, including cellular differentiation, regulatory mechanisms, and epigenetic modifications. Despite the similarities between lncRNAs and messenger RNAs (mRNAs), there is a fundamental difference: mRNAs encode proteins, whereas lncRNAs do not. This study aims to distinguish these two RNA classes from each other by designing a robust machine learning (ML) pipeline employing Recursive Feature Elimination (RFE) for dimensionality reduction of dataset and XGBoost (XGB) classification model. Whereas previous studies trained and tested machine learning models using the complete set of dataset features, we employ the RFE technique to reduce the number of features, thereby we achieve a more optimal dataset with relevant features. To evaluate the predictive performance of our pipeline, we used error rate, accuracy, precision, recall, and F1-score. Compared to three existing lncRNA identification tools in the literature, our pipeline demonstrated superior prediction accuracy and precision at 93.42% and 94.19% respectively.

Keywords: Recursive Feature Elimination, XGBoost, lncRNAs, Bioinformatics, Machine Learning.

Tekrarlayan Özellik Eliminasyonu ve XGBoost ile Uzun Kodlamayan RNA Tahmininin İyileştirilmesi

ÖZ

Son yıllarda, yüksek verimli teknolojilerdeki ilerlemeler, kodlamayan Ribonükleik Asitler (ncRNA'lar) olarak bilinen çok sayıda gizli katmanı ortaya çıkararak genomların protein merkezli görüşünü değiştirdi. Daha önce genomun önemsiz bölümleri olarak kabul edilen NcRNA'lar, artık prokaryotik ve ökaryotik organizmalarda temel işlevsel bileşenler olarak kabul ediliyor. Uzun kodlamayan RNA'lar (lncRNA'lar), hücresel farklılaşma, düzenleyici mekanizmalar ve epigenetik modifikasyonlar dahil olmak üzere temel biyolojik işlevlerde etkili olan 200 nükleotid uzunluğundaki benzersiz bir ncRNA kategorisidir. LncRNA'lar ve haberci RNA'lar (mRNA'lar) arasındaki benzerliklere rağmen, temel bir fark vardır: mRNA'lar protein kodlar, oysa lncRNA'lar kodlamaz. Bu çalışma, veri kümesinin boyutsallığını azaltmak için Tekrarlayan Özellik Eliminasyonu (RFE) ve XGBoost (XGB) sınıflandırma modelini kullanan sağlam bir makine öğrenimi (ML) boru hattı tasarlayarak bu iki RNA sınıfını birbirinden ayırmayı amaçlamaktadır. Önceki çalışmalar, veri kümesi özelliklerinin tamamını kullanarak makine öğrenimi modellerini eğitmiş ve test etmişken, biz özellik sayısını azaltmak için RFE tekniğini kullanıyoruz, böylece ilgili özelliklere sahip daha optimum bir veri kümesi elde ediyoruz. Boru hattımızın tahmin performansını değerlendirmek için hata oranı, doğruluk, kesinlik, geri çağırma ve F1 puanını kullandık. Literatürdeki üç mevcut lncRNA tanımlama aracıyla karşılaştırıldığında, boru hattımız sırasıyla %93,42 ve %94,19'da üstün tahmin doğruluğu ve kesinlik gösterdi.

Anahtar Kelimeler : Özyinelemeli Özellik Giderme, XGBoost, lncRNA'lar, Biyoinformatik, Makine Öğrenm.

1. INTRODUCTION

Over the past few years, there has been considerable progress in handling and interpreting biological data [1][2]. Advancements in high-throughput technologies have uncovered numerous concealed layers known as Non-Coding Ribonucleic Acids (ncRNAs), situated between transcription and translation processes. These RNAs are not translated into proteins [3]. NcRNAs, once considered as transcriptional noise [4] because of their poor conservation [5] and insignificant segments of genomes, in recent years, have emerged as essential functional components in both prokaryotic and eukaryotic organisms [6]. NcRNAs have become one of the stars of modern biology [7]. NcRNAs are commonly categorized into two primary groups based on transcript length. Small Non-Coding RNAs (sncRNAs) refer to shorter sequences [6], while Long Non-Coding RNAs

*Sorumlu Yazar (Corresponding Author)
e-posta : Freshta.Alizada579@gmail.com

(lncRNAs) are designated for transcripts longer than 200 nucleotides [6-11] and have no protein potential [3, 4], [8], [12, 13]. The initial set of lncRNAs was identified about two decades ago [14]. As research has advanced, lncRNAs, previously dismissed as dark matter or insignificant, have gradually come to light. In 2007, Rinn et al. from Stanford University initiated formal lncRNA research with an article published in Cell, marking a significant starting point [15]. Recent studies have underscored the pivotal role of lncRNAs in crucial biological processes such as cellular differentiation, epigenetics [11], and regulation [1, 3], [6]. Additionally, lncRNAs have been implicated in gene expression [3], [6, 7], [12], [16], [17], translation, transcription [18], and the pathogenesis of complex diseases [16]. Differentiating between protein-coding transcripts (i.e., messenger RNAs) and lncRNAs proves to be a surprisingly challenging task in practice [19] because lncRNAs and mRNAs share similarities in their sequence lengths, poly (A) tails, and splicing structures. Additionally, lncRNAs occasionally tend to encode long open reading frame (ORF) [12] and the primary distinction between lncRNAs and mRNAs is the absence of discernible coding potential in lncRNAs [20]. Consequently, distinguishing between lncRNAs and mRNAs remains a challenge. A variety of computational algorithms have been proposed in recent years to distinguish between lncRNAs and mRNAs [12].

More information on existing methods for differentiating them and the current study will be given in the related works section.

1.1 Related Works

Approximately 2% of the human genome is dedicated to encoding proteins, while the remainder consists of ncRNAs [21, 22]. Consequently, distinguishing this large part of the genome from proteins (i.e., mRNAs) helps scientists delve deeply into them and find solutions for the types of diseases mentioned in articles [16] and [23], which claim thousands of human lives. According to [1], machine learning (ML) algorithms are utilized to integrate biological and biomedical datasets to identify lncRNAs, but numerous challenges lie in their way because these datasets have intrinsic complexity behind their huge size, which causes the existence of high-dimensionality, incompleteness, bias, heterogeneity, dynamism, and noise. Conversely, numerous bioinformatics applications utilize ML algorithms for analyzing sequence data. Since most ML algorithms accept numerical data, sequences have to be transformed into numbers. To avoid long sequences of numbers, the most efficient approach is choosing relevant features from the sequences.

Relevant features selected in various studies include: Guanine and Cytosine (GC) content, sequence length, k-mer (k= 1 up to 6), and open reading frame (ORF) [1]; k-mer (k= 1 up to 5), sequence-order, and correlation coefficient factors [12]; sequence features such as k-mer (2-15), CG content, and structured features including binary and Quadri nary representations, as well as minimum free energy [3]; k-mer frequency features and spectrum features [24]; weighted k-mer (k=1 up to 3), pseudo nucleotide composition, hexamer usage bias, Fickett score, ORF, UTR regions, and HMMER score [21].

During recent years, many different computational algorithms have been developed for differentiating lncRNAs. These include: Decomposing model for feature selection in lncRNAs [1], IDlncRNA using ML algorithms [3], LNCRI [21], NCYPred [6], LncDLSM [24], ncRFP [16], LPGNMF [17], FexRNA [25], IPCARF [26], RFLDA [27]. These tools applied machine learning algorithms across multiple species, particularly plants, humans, and animals. This approach has provided enhanced insights into lncRNAs. According to [1], some studies have explored the utilization of multiple features to extract meaningful information from lncRNAs, resulting in the creation of high-dimensional feature vectors. The presence of a high-dimensional feature dataset contributes to the curse of dimensionality problem, which can decrease the performance of ML algorithms. As stated in [25], the process of feature selection (FS) is a crucial step aimed at identifying the most pertinent features to enhance the performance of ML algorithms. The fundamental objective of FS is to eliminate noise from the data, thereby mitigating the risk of overfitting and improving the predictive performance of a model. Consequently, FS involves selecting a smaller subset of features that exhibit superior or comparable predictive performance, particularly in the context of predictive modeling (supervised models). By utilizing fewer features, ML algorithms improve their ability to generalize across heterogeneous datasets, reduce computational costs, and simplify the model's complexity.

To distinguish lncRNAs from mRNAs, the existence of high-dimensional feature datasets has led us to pursue efficient FS techniques. In academic literature, there are multiple approaches using different FS techniques to reduce the dimensionality of feature datasets for biological data classification. For instance, [1] (reduced from 5468 to 10 lncRNAs) reports experiments using Metaheuristics and a decomposition model containing rounds and a voting scheme, [24] (reduced from 1355 to 8 lncRNAs) describes the use of the hierarchical neural network (HINN) algorithm for FS, and [23] (reduced from 3435 to 234 lncRNAs) reported using Tree-based, L1-based, and Variance threshold algorithms for FS.[25] used 7 different FS algorithms (e.g., for the Filter-based FS category, chi-squared is used. For wrapper-based, RFE using random forest (RF), RFE using logistic regression, RFE with k-fold cross-validation using LR and RF models are used. Finally, for the embedded FS category, embedded based on LR and RF classifiers are used) and applied them on 17 extracted features of lncRNAs. However, we observed a lack of studies focused on FS using RFE for high-dimensional datasets of lncRNAs.

RFE has been employed in numerous studies as a feature selection (FS) technique and has consistently yielded favorable outcomes. As reported in [25], in CPC2, a list of 23 features is compiled. Subsequently, the RFE feature selection technique, in conjunction with 10-fold cross-validation, is applied to this set of candidate features. The outcome of this process is the selection of 4 features deemed as the most significant for estimating the coding potential of a transcript. Similarly, in the case of CPC2, [25] applies RFE to its extracted 17 features to identify the most relevant features providing sufficient information about lncRNAs. Therefore, based on the successful results of RFE in previous works with small numbers of features, we propose using RFE on high-dimensional datasets of lncRNAs to identify the best features. In our approach, we employed RFE on a dataset with a dimensionality of 5467 features. Furthermore, we examined the impact of the features chosen by the decomposition model on the predictive performance of three ML algorithms: J48, REPTree, and Random Forest, in the classification of lncRNAs task. According to [1], we selected these machine learning algorithms because they generate interpretable predictive models, enabling a deeper understanding of the internal decision-making mechanisms. As a result, domain experts can verify the knowledge employed by the models for classifying new sequences. The main contributions of our work are:

- Presenting a robust machine learning pipeline for distinguishing lncRNA sequences from protein-coding RNAs (messenger RNAs).
- Employing recursive feature elimination as the feature selection algorithm to address the dimensionality issue of the feature dataset.
- Accessing higher predictive performance in the context of lncRNA prediction compared to two established lncRNA prediction tools in the literature.

The remainder of this article presents our materials and methods in section 2, experiments, results and discussion in section 3, and conclusions in the last part.

2. MATERIAL AND METHOD

In this part of our research paper, we outline our methodological approach crafted to attain the intended aims. To obtain our objective we integrated machine learning methods. Figure 1 summarizes our methodology in this study.

2.1 Dataset

The dataset utilized in this study, sourced from recent literature [1], comprises 18,040 RNA sequences (i.e., 9020 lncRNAs and 9020 mRNAs) from five species: *Arabidopsis thaliana*, *Cucumis sativus*, *Glycine max*, *Oryza sativa*, and *Populus trichocarpa*. Each species contributes 1,804 lncRNA sequences and 1,804 mRNA sequences. Within this dataset, two classes are identified: the positive class, consisting of lncRNAs, and the negative class, consisting of protein-coding genes (mRNAs). The lncRNA data were obtained from two public databases, PLNlncRbase and GreenNC (as described in [1]), while the mRNA sequences were extracted from Phytozome (further explanation is provided in [1]).

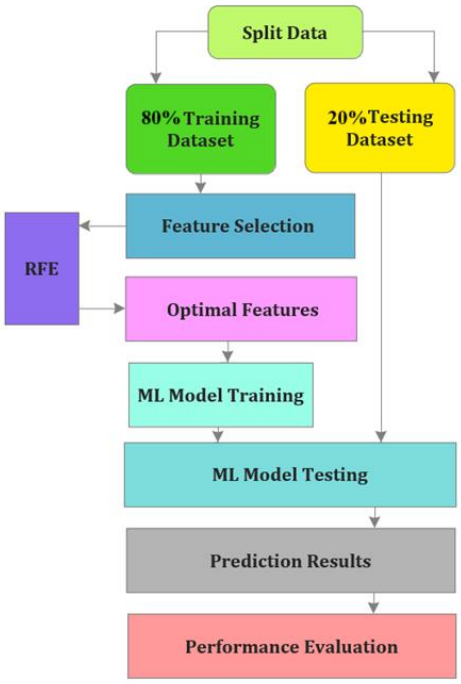

Figure 1. Our methodology diagram

As part of the data preprocessing, sequence redundancy was eliminated at 80% identity using CD-HIT-EST (more description in [1], [21]), and only sequences longer than 200 nucleotides were retained. To construct the feature vector and extract the most relevant features, four indicators—Guanine and Cytosine content or GC content, k-mer (k= 1 up to 6), sequence length, and ORF—were considered [1]. As a result, four feature vectors were derived for each sequence in the dataset, including GC content with 1 feature, k-mer frequencies with 5460 features, sequence length with 1 feature, and ORF with 5 features.

According to [28], combining multiple feature sets into a single, joined feature vector preserves the unique discriminating information from each original set while reducing redundancy caused by correlations between different sets. This approach enhances the robustness and predictive performance of models. Therefore, all four feature vectors were concatenated, resulting in a dataset with 18,040 rows and 5,467 columns. In previous studies [5] and [21], training and testing datasets were split using 70:30 and 80:20 ratios, respectively. Following the approach in [21], we used an 80:20 split, where 80% of the dataset was used for training and 20% for testing, as illustrated in Figure 1.

2.2 Feature Selection

Feature selection involves identifying relevant subsets of features within a dataset [29][30]. As the more features in a dataset means the more expensive and time-consuming computational complexity and our used dataset [1] is with dimension 18040 x 5467, Choosing the

most relevant features through feature selection (FS) is crucial for enhancing the performance of any Ml algorithms [25]. Therefore, we decided to apply the wrapper method (i.e., Recursive Feature Elimination) for FS to reduce the high-dimensionality problem of dataset.

2.2.1 Recursive feature elimination

RFE is a backward feature selection algorithm categorized under Wrapper methods, wherein it recursively reduces the feature set size using a specific underlying algorithm to choose features [31]. The RFE selection method involves a repetitive process where features are ranked based on their importance. During each iteration, the importance of each feature is assessed, and the least relevant feature is removed. This process continues until reaching the desired number of features which is declared by programmer [32].

For instance, if there is a dataset which has N features such as$f_1, f_2, f_3,\ldots f_n$, therefore, the importance of each feature is gained according to equation (1)

$$FI(f_i) = \frac{\sum_{i=1}^{T} \Delta G_t(f_i)}{T} \quad (1)$$

Where FI refers to feature importance, $\Delta G_t(f_i)$ is the reduction in impurity for feature f_i in tree t, and T is the total number of trees in the random forest.

Then features are ranked and according to equation (2) the feature with the least importance is deleted.

$$F = F - \{f_{least}\} \quad (2)$$

Where F is the set of all the features and f_{least} is the least important feature.

2.3 Evaluation Metrics

In our study, the success percentages of ML algorithms were assessed by five measures: recall (RE) [23], accuracy (ACC), precision (PR), F1-score (F1) [1], and error rate (ER). These measures were utilized to assess the ML model's predictive performance on testing dataset. Four quantitative variables—true positive (TP), true negative (TN), false positive (FP), and false negative (FN)—were computed for the testing dataset [33] where, TP represents accurately predicted lncRNAs, TN denotes correctly predicted mRNAs, FP indicates false positives where negative entities are incorrectly predicted as lncRNAs, and FN signifies false negatives where positive entities are incorrectly predicted as mRNAs.

$$Accuracy = \frac{TP + PN}{TP + TN + FP + FN} \quad (3)$$

$$F1\ score = \frac{2TP}{2TP + FP + FN} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{P} \quad (6)$$

$$Error\ Rate = 1 - Accuracy \quad (7)$$

2.4 McNemar's Test

There are different statistical methods such as Wilcoxon signed-rank test, R test, J test [34], and McNemar's test [35] to assess and determine the statistical significance of predictive performance between binary classifiers. Therefore, the best binary classification models can be identified from a statistical standpoint.In this study, the same of literature [35] we employed McNemar's test to evaluate the predictive performance between all the classification models used to classify lncRNAs and mRNAs. According to this literature, McNemar's test is a non-parametric method which is employed to assess the error rates of two binary classifiers applied on the same dataset. It determines whether the difference in predictive performance between two binary classifiers is statistically significant. McNemar's test formula is given in equation (8).

$$\chi^2 = \frac{(|n_{10} - n_{01}| - 1)^2}{n_{10} + n_{01}} \quad (8)$$

If we consider A and B as two binary classification models, in equation (8), χ^2 represents McNemar's test result, n_{10} represents the number of instances classified as positive by A but negative by B, and n_{01} represents the number of instances classified as negative by A but positive by B.

In order to determine whether the differences between the classification models are statistically significant, this study computed the p-value for each pairwise McNemar's test result.

3. RESULTS AND DISCUSSIONS

3.1 Experimental Setups

For all experiments reported in this study, we used Python Version 3.11.4) and PyCharm (Version 2022.2.2 Edu) as the development environment. To achieve optimal performance in training and prediction from the machine learning models, we employed the grid search method for hyper-parameter tuning. The hyper-parameters located in Table 1 were determined through grid search for our ML models.

3.2 Feature Selection Process

In our study, we employed RFE with Random Forest as the underlying estimator to train the model with all features. We applied RFE to our training dataset with the purpose of reducing the dimension of it . The same of literature [1], the desired number of features which should be found by RFE in our case was set to 10 as well. Table 2 presents the 10 optimal features which were extracted by RFE and the best decomposing model of literature [1]. According to Table 2, it is significant to highlight both algorithms has extracted *score* and *cdsSizes* features jointly in their optimal feature datasets.

3.3 Training and Testing Phase

In order to assess the efficiency of the optimal features extracted by the RFE FS algorithm, we had to choose an optimal classifier. Therefore, we chose six state-of-the-art classification algorithms which work perfect in binary classification task such as, REPTree [1][10], Random Forest [36],[37],[38][39] Support Vector Machine (SVM)[19] [36],[40],[41] Logistic Regression [36],[42], and Artificial Neural Network (ANN) [43][44][45], and XGBoost (XGB) [21]. We trained them using the RFE's optimal features dataset. Then, to evaluate the prediction performance of each ML model, we used testing dataset to assess the prediction performance of each ML model and recorded the results.

Table 1. Illustrates the hyper-parameters for utilized classifiers in this article.

Model	Parameter	Value
Random Forest	bootstrap	true
	max-depth	10
	mean-samples-leaf	4
	mean-samples-split	2
	n-estimators	100
Logestic Regression	c	0.001
	max-iter	100
	penalty	i2
	solver	newton-cg
Support Vector Machine	c	0.1
	gamma	1
	kernel	linear
Artificial Neural Network	activation	logistic
	alpha	0.001
	hidden-layer-sizes	10
	learning-rate	constant
	solver	adam
	max-iter	200
REPTree	criterion	entropy
	max-depth	5
	min-samples-leaf	1
	min-samples-split	2
XGBoost	colsample-bytree	0.8
	learning-rate	0.01
	max-depth	3
	subsample	0.8
	n-estimators	50

Table 2. Features extracted by RFE and M1-GA

FRE	M1-GA
AC	GTCCCC
GG	CCGGCA
TA	CGCCTC
TT	CGGAGT
GGA	CGTTAG
tamanho	CTAGGT
cdsStop	GGGGGG
score	**score**
cdsSizes	**cdsSizes**
cdsPercent	TCACGG

As shown in Table 3, SVM demonstrates the lowest performance among the models,

with an ACC of 92.07% and PR of 86.45%. Following SVM, the performances of ANN, LR, and REPTree are also relatively lower, with accuracies of 92.12%, 92.37%, and 93.37%, respectively, and corresponding precision values of 92.48%, 87.45%, and 94.12%. The highest performance is owned by both RF and XGB models, with the same accuracy (ACC) of 93.42%. However, XGB was chosen as the best algorithm due to its precision (PR) of 94.19%, slightly outperforming RF, which has a precision of 94.17%.

In addition to computing the evaluation metrics, we conducted McNemar's test to statistically assess the classification models and identify the most effective one to be used in conjunction with the RFE feature selection algorithm. Table 4 contains the results of McNemar's test for all the pairs of classifiers and the values which are statistically significant are shown in bold. According to the received results from Table 4, RF, REPTree and XGBoost Classification models' predictive performances are not so different because the p-value

between them is not smaller than 0.05 and RF and XGBoost classification models' predictive performances are almost the same because the result is infinite.

Consequently, although according to McNemar's test RF and XGBoost predictive performances are the same, we introduced RFE feature selection combined with XGB as the optimal model for our study because of the superior performance that it had in precision and outperformed the RF model.

Table 3. Depicting the prediction performance of ML models trained with RFE's optimal features, on test dataset. In this table the highest and the lowest values are bold for more visibility

Features Reduction Algorithm	ML Prediction Algorithm	ER	ACC	PR	Recall	F1 score
RFE	Random Forest	**6.57%**	**93.42%**	94.17%	93.542%	**93.40%**
	REPTree	6.62%	93.37%	94.12%	93.37%	93.34%
	SVM	7.92%	**92.07%**	**86.45%**	**99.77%**	92.64%
	ANN	7.62%	92.37%	87.45%	98.94%	92.84%
	LR	7.87%	92.12%	92.48%	**92.12%**	**92.09%**
	XGB	**6.57%**	**93.42%**	**94.19%**	93.542%	**93.40%**

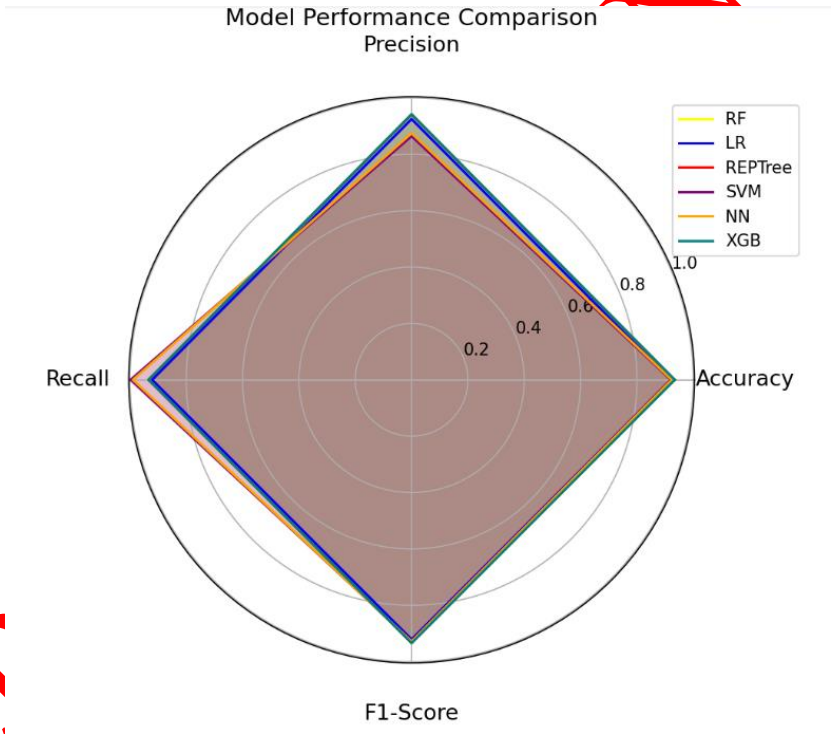

Figure 2. Radar chart showing the performance of ML models based on ACC, PR, Recall and F1-score

Table 4. McNemar's results between all the pairs of binary classification models' predictive performances. **p-value < 0.05**

Classifiers	RF	LR	REPTree	SVM	ANN	XGBoost
RF		**0.0001**	0.4795	**0.0001**	**0.0001**	infinite
LR			**0.0001**	0.8802	**0.0077**	**0.0001**
REPTree				**0.0001**	**0.0001**	0.4795
SVM					0.1183	**0.0001**
ANN						**0.0001**
XGBoost						

3.4 Evaluation with Other Classifier Tools

To evaluate the performance of our best model (i.e., RFE-XGB), like article [16], we compared it with two tools in the literature: M1-GA [1] and IPCARF [26]. As mutual dataset between our proposed method and the two other methods mentioned before, we utilized our dataset which is detailed in Section 2. For each tool, we utilized the FS classification models as described in their respective methodologies. To ensure a fair comparison, we used the same 80:20 ratio for training and testing datasets as in our proposed method, RFE-XGB. The results of this comparison are presented in Table 3.

As shown in Table 3, RFE-XGB achieved the highest values for ACC, PR, Recall, and F1-score, as well as the lowest ER, demonstrating the success of our model in this study. Figure 3 illustrates the ROC curve, depicting the prediction performance of our proposed method alongside the two other tools on the testing dataset. Notably, the performance of M1-GA and RFE-XGB is very close, while IPCARF exhibits the lowest prediction performance on the testing dataset.

Table 3. Comparison of RFE-XGB algorithm performance with two literature tools. The highest and lowest values are highlighted in bold for clarity.

Tools	ER	ACC	PR	Recall	F1 score
RFE-XGB	**6.57%**	**93.42%**	**94.19%**	**93.42%**	**93.40%**
M1-GA	6.65%	93.34%	94.1%	93.34%	93.31%
IPCARF	**11.77%**	**88.22%**	**88.68%**	**88.22%**	**88.19%**

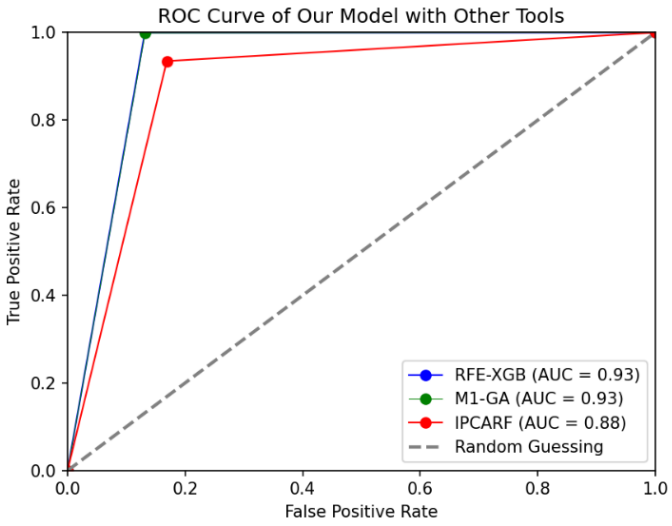

Figure 3. ROC curve of comparison between RFE-XGB and two other lncRNAs identification tools.

4. CONCLUSION

In this study, we developed a validated and robust machine learning pipeline to reduce the dimensionality of features in lncRNAs and mRNAs and accurately differentiate between them. We utilized a dataset from the literature [1], which consists of 5,467 features of lncRNAs and mRNAs from five plant species. To identify the most relevant features and reduce dimensionality, we employed the RFE algorithm, extracting the 10 most relevant features. For classification, we evaluated six machine learning classifiers: RF, REPTree, ANN, SVM, LR and XGBoost. Among these, XGBoost demonstrated the best prediction performance on the testing dataset and was selected as the optimal classifier to pair with the RFE algorithm.

To assess the effectiveness of our pipeline, we compared its performance with two other lncRNA identification tools from the literature. As shown in Table 3, our method outperformed these existing tools, achieving higher percentages in ACC, PR, recall and F1-score.

One limitation of this work is the exclusive use of machine learning classification algorithms. As future work, we plan to integrate RFE with deep learning classifiers and evaluate their performance. The findings of this study have the potential to be further utilized in advancing lncRNA prediction methodologies.

REFERENCES

[1] Bonidia R. P.,Machida J.S.,Negri T.C.,Alves W.A.L.,Kashiwabara A.Y.,Domingues D.S., Charvalho A.D., Paschoal A.R. and Shanches D.S., "A Novel Decomposing Model with Evolutionary Algorithms for Feature Selection in Long Non-Coding RNAs", ***IEEE Access***, 8: 181683–181697, (2020).

[2] Nuray, B. and Altuntaş, V., "RNA m6A Modifikasyon Bölgelerinin Sınıflandırılması için Öznitelik Çıkarma ve Boyut Azaltma Yöntemlerinin Karşılaştırılması". ***Politeknik Dergisi,*** pp.1-1. (2024).

[3] Sreeshma C. M., Manu M. and Gopakumar G., "Identification of Long Non-Coding RNA From Inherent Features Using Machine Learning Techniques",

International Conference on Bioinformatics and Systems Biology", BSB 2018: 97–102, (2018).

[4] Chen M., Peng Y., Li A., Deng Y., Deng Y. and Li Z., "A Novel lncRNA-Disease Association Prediction Model Using Laplacian Regularized Least Squares and Space Projection-Federated Method", ***IEEE Access***, 8: 111614–111625, (2020).

[5] Zampetaki A., Albrecht A. and Steinhofel K., "Long Non-Coding RNA Structure and Function: Is There A Link?", ***Frontiers in Physiology***, 9(AUG): 1–8, (2018).

[6] Lima D. D. S., Amichi L. J. A., Fernandez M. A., Constantino A. A. and Seixas F. A. V., "NCYPred: A Bidirectional LSTM Network with Attention for Y RNA and Short Non-Coding RNA Classification", **IEEE/ACM Transactions on Computational Biology and Bioinformatics**,20(1): 557–565,(2023).

[7] Alessio E., Bonadio R. S., Buson L., Chemello F. and Cagnin S., "A Single Cell But Many Different Transcripts: A journey into the world of long non-coding RNAs**",** ***International Journal of Molecular Sciences***,21(1), (2020).

[8] Wang W., Min L.,Qiu X., Wu X., Liu C., Ma J., Zhang D. and Zhu L., "Biological Function of Long Non-Coding RNA (LncRNA) Xist", ***Frontiers in Cell and Developmental Biology***, 9(Jue): 1–27, (2021).

[9] Xuan P., Zhao Y., Cui H., Zhan L., Jin Q., Zhang T. and Nakaguchi T., "Semantic Meta-Path Enhanced Global and Local Topology Learning for lncRNA-Disease Association Prediction", ***IEEE-ACM Transactions on Computational Biology and Bioinformatics***, 20(2): 1480–1491, (2023).

[10] Wang B. and Zhang J., "Logistic Regression Analysis for LncRNA-Disease Association Prediction Based on Random Forest and Clinical Stage Data", ***IEEE Access***, 8: 35004–35017, (2020).

[11] Xie G., Jiang J. and Sun Y., "LDA-LNSUBRW: LncRNA-Disease Association Prediction Based on Linear Neighborhood Similarity and Unbalanced Bi-Random Walk", ***IEEE/ACM Transactions on Computational Biology and Bioinformatics***,19(2): 989–997, (2022).

[12] Hu J. and Andrews B., "Distinguishing Long Non-Coding RNAs From mRNAs Using A Two-layer Structured Classifier", ***IEEE International Conference on Computational Advances in Bio and Medical Sciences, ICCABS***,2017-Octob: 1–5, (2017).

[13] Shen C., Mao D., Tang J., Liao Z. and Chen S., "Prediction of LncRNA-Protein Interactions Based on Kernel Combinations and Graph Convolutional Networks", ***IEEE Journal of Biomedical and Health Informatics,*** 28(4): 1937–1948, (2024).

[14] Shen C., Ding Y., Tang J., Jiang L. and Guo F., "LPI-KTASLP: Prediction of LncRNA-Protein Interaction by Semi-Supervised Link Learning with Multivariate Information", ***IEEE Access***, 7: 13486–13496, (2019).

[15] Liu X. Q., Li B. X., Zeng G. R., Liu Q. Y. and Ai D. M., "Prediction of Long Non-Coding RNAs Based on Deep Learning", ***Genes***, 10(4): (2019).

[16] Wang L., Zheng S.,Zhang H.,Qiu Z.,Zhong X.,Liu H.and Liu Y., "NcRFP: A Novel end-To-end Method for Non-Coding RNAs Family Prediction Based on Deep Learning", ***IEEE/ACM Transactions on Computational Biology and Bioinformatics***, 18(2): 784–789, (2021).

[17] Zhang T., Wang M., Xi J. and Li A., "LPGNMF: Predicting Long Non-Coding RNA and Protein Interaction Using Graph Regularized Nonnegative Matrix Factorization", ***IEEE/ACM Transactions on Computational Biology and Bioinformatics***, 17(1): 189–197, (2020).

[18] Zhang T., Tang Q., Nie F., Zhao Q. and Chen W., "DeepLncPro: an Interpretable Convolutional Neural Network Model for Identifying Long Non-Coding RNA Promoters", ***Briefings in Bioinformatics,*** 23(6): 1–9, (2022).

[19] Schneider H. W., Raiol T., Brigido M. M., Walter M. E. M. T. and Stadler P. F., "A Support Vector Machine based method to distinguish long non-coding RNAs from protein coding transcripts", ***BMC Genomics***, vol. 18(1): 1–14, (2017).

[20] Budak H., Kaya S. B. and Cagirici H. B., "Long Non-Coding RNA in Plants in the Era of Reference Sequences", ***Frontiers in Plant Science***.,11(March): 1–10, (2020).

[21] Musleh S., Islam M. T. and Alam T., "LNCRI: Long Non-Coding RNA Identifier in Multiple Species", ***IEEE Access***, 9: 167219–167228, (2021).

[22] Ping P., Wang L., Kuang L., Ye S., Iqbal M. F. B. and Pei T., "A Novel Method for LncRNA-Disease Association Prediction Based on an lncRNA-Disease Association Network", ***IEEE-ACM Transactions on Computational Biology and Bioinformatics***, 16(2): 688–693, (2019).

[23] Ganapaneni M. D., Paruchuru K. H., Ambati J. H., Valavala M. and Sobin C. C., "Detecting Long Non-Coding RNAs Responsible for Cancer Development", ***Proceedings - 2022 OITS International Conference on Information Technology, OCIT 2022***, 164–169,(2022).

[24] Wang Y., Zhao P., Du H., Cao Y., Peng Q. and Fu L., "LncDLSM: Identification of Long Non-Coding RNAs with Deep Learning-Based Sequence Model", ***IEEE Journal of Biomedical and Health Informatics***,27(4): 2117–2127,(2023).

[25] Amin N., McGrath A., and Chen Y. P. P., "FexRNA: Exploratory Data Analysis and Feature Selection of Non-Coding RNA", ***IEEE/ACM Transactions on Computational Biology and Bioinformatics***, 18(6): 2795–2801,(2021).

[26] Zhu R., Wang Y., Liu J. X. and Dai L. Y., "IPCARF: Improving Lncrna-Disease Association Prediction Using Incremental Principal Component Analysis Feature Selection and A Random Rorest Classifier", ***BMC Bioinformatics***, 22(1): 1–17,(2021).

[27] Yao D., Zhan X., Zhan X., Kwoh C. K., Li P. and Wang J., "A Random Forest Based Computational Model for Predicting Novel LncRNA-Disease Associations", ***BMC Bioinformatics***, 21(1): 1–18,(2020).

[28] Fan X.-N. and Zhang S.-W., "lncRNA-MFDL: Identification of Human Long Non-Coding RNAs by Fusing Multiple Features and Using Deep Learning", ***Molecular BioSystems***., 11(3): 892–897,(2015).

[29] Ventola G. M. M., Noviello T. M. R., D'Aniello S., Spagnuolo A., Ceccarelli M. and Cerulo L., "Identification of Long Non-Coding Transcripts with Feature Selection: A Comparative Study", BMC ***Bioinformatics***,18(1): 187, (2017).

[30] Hatipoğlu, A. and Altuntaş, V.," DeepTFBS: Transkripsiyon Faktörü Bağlanma Bölgeleri Tahmini İçin Derin Öğrenme Yöntemleri Kullanan Hibrit Bir Model". ***Politeknik Dergisi***, pp.1-1. (2024).

[31] Zhang B., Zhang Y. and Jiang X., "Feature selection for global tropospheric ozone prediction based on the BO-XGBoost-RFE algorithm", ***Scientific Reports***, 12(1): 1–10, (2022).

[32] Granitto P. M., Furlanello C., Biasioli F. and Gasperi F., "Recursive Feature Elimination with Random Forest for PTR-MS Analysis of Agro-industrial Products", ***Chemometrics and Intelligent Laboratory Systems***, 83(2): 83–90, (2006).

[33] Tripathi R., Patel S., Kumari V., Chakraborty P. and Varadwaj P. K., "DeepLNC, a Long Non-Coding RNA Prediction Tool Using Deep Neural Network", ***Network Modeling Analysis in Health Informatics and Bioinformatics***, 5(1): 21: (2016).

[34] Gamgam H. and Altunkaynak B., "Test Statistic for Ordered Alternatives based on Wilcoxon Signed Rank." **[Online].** Available: http://dergipark.gov.tr/gujs

[35] Atilkan. Y. et al., "Advancing Crayfish Disease Detection: A Comparative Study of Deep Learning and Canonical Machine Learning Techniques," ***Applied Sciences (Switzerland)***, 14(14): (2024).

[36] Zhang C., Liu C., Zhang X., and Almpanidis G., "An Up-to-Date Comparison of State-of-the-Art Classification Algorithms", ***Expert Systems with Applications***,82: 128–150, (2017).

[37] Ahmad I., Basheri M., Iqbal M. J., and Rahim A., "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection", ***IEEE Access***, 6,33789–33795:(2018).

[38] Nacar E. N. and Erdebilli B., "Makine Öğrenmesi Algoritmaları ile Satış Tahmini," ***Journal of Industrial Engineering***", 2(2): 307–320,(2021).

[39] Camalan. M. and Çavur. M., "Using Random Forest Tree Classification for Evaluating Vertical Cross-Sections in Epoxy Blocks to Get Unbiased Estimates for 3D Mineral Map," ***Politeknik Dergisi***, 24(1), 113–120, (2021).

[40] Jakkula V., "Tutorial on Support Vector Machine (SVM)", ***School of EECS***, Washington State University,1–13:(2011).

[41] Bekçioğulları M. F, Dikici B., Açıkgöz H., and Keçecioğlu Ö. F., "Güneş Enerjisinin Kısa-Dönem Tahmininde Farklı Makine Öğrenme Yöntemlerinin Karşılaştırılması Comparison of Different Machine Learning Methods in Short-Term Forecasting of Solar Energy", ***EMO Bilimsel Dergi***, 11(22):37–45, (2021).

[42] Long W. J., Griffith J. L., Selker H. P., and D'Agostino R. B., "A Comparison of Logistic Regression to Decision-Tree Induction in A Medical Domain", ***Computers and Biomedical Research***, 26(1): 74–97, (1993).

[43] Mitrea C. A., Lee C. K. M., and Wu Z., "A Comparison Between Neural Networks and Traditional Forecasting Methods: A Case Study", International Journal of Engineering Business Management, 1(2):19–24, (2009).

[44] Çalışan. M. and Talu M. F., "Comparison of Methods for Determining Activity from Physical Movements," ***Politeknik Dergisi***, 24(1),17–23,(2021).

[45] Ekincioğlu. G., Akbay. D., and Keser. S., "Estimating Uniaxial Compressive Strength of Sedimentary Rocks with Leeb hardness Using SVM Regression Analysis and Artificial Neural Networks," ***Journal of Polytechnic***, 1–1, (2024).