EnsemPseU: Identifying pseudouridine sites with an ensemble approach

Yue Bi, Dong Jin, Cangzhi Jia[*]

School of Science, Dalian Maritime University, Dalian 116026, China

Corresponding author: Cangzhi Jia (cangzhijia@dlmu.edu.cn)

This work was supported by Fundamental Research Funds for the Central Universities (3132019323, 3132020170), the National Natural Science Foundation of Liaoning Province (20180550307, 20170540092), and the Natural Science Foundation of China (61803065).

ABSTRACT Pseudouridine (Ψ) is the most prevalent RNA modification, which is formed from uridine through an isomerization reaction. With the increasing availability of genomic and proteomic samples, computer-aided pseudouridine-synthase-specific Ψ site recognition is becoming possible. In this paper, we propose an ensemble approach to identify pseudouridine sites, named EnsemPseU. First, five sequence-encoding strategies, namely, kmer, binary encoding, enhanced nucleic acid composition (ENAC), nucleotide chemical property (NCP), and nucleotide density (ND), were applied to extract sequence information. Then, chi-square feature selection was used to reduce the feature dimensionality and remove redundant information. Finally, an ensemble algorithm integrating support vector machine (SVM), extreme gradient boosting (XGBoost), na ïve Bayes (NB), k-nearest neighbor (KNN), and random forest (RF) was used to build our prediction model. Upon testing, the results showed that the accuracy improved 5.3% for *H. sapiens*, 6.09% for *S. cerevisiae*, and 5.55% for *M. musculus* after chi-square feature selection. Moreover, upon evaluation via 10-fold cross-validation and an independent test, our proposed model EnsemPseU outperformed the other best existing model. The source code and data sets are available at https://github.com/biyue1026/EnsemPseU.

INDEX TERMS Machine learning, ensemble learning, pseudouridine site prediction, feature selection.

I. INTRODUCTION

Pseudouridine (Ψ) is the most prevalent RNA modification, which is formed from uridine through an isomerization reaction. It was discovered as early as the 1950s and was initially studied in noncoding RNA types such as ribosomal RNA (rRNA) and transfer RNA (tRNA) [1, 2]. However, recent reports have shown that Ψ is also present in messenger RNA (mRNA). It has been detected in transcripts obtained from humans, yeast, and the unicellular eukaryotic parasite *Toxoplasma gondii* [3-5]. The production of Ψ, an isomer of uridine, is catalyzed by highly conserved pseudouridine synthase, which detaches the uridine residue's base from its sugar, followed by "rotating" it 180° along the N3–C6 axis, and subsequently reattaches the base's 5-carbon to the 1′-carbon of the sugar. Ψ can be considered to play important roles in structure, function, and metabolism [6]. Therefore, the identification of Ψ sites is crucial for revealing the biological principle concerned.

Experimental verification plays an important role in identifying Ψ sites, but is costly and time-consuming. Instead, some computational methods have been developed to identify Ψ sites, which have the advantages of rapidity, low cost, and efficiency. In 2015, Li *et al*. built the first computational model called PPUS to predict the pseudouridine-synthase-specific Ψ sites in *H. sapiens* and *S. cerevisiae* [7]. They used the nucleotides around Ψ as features and employed SVM as the classifier. The following year, Chen *et al*. developed another model called iRNA-PseU to identify Ψ sites in *H. sapiens*, *S. cerevisiae*, and *M. musculus*, and also employed SVM as the classifier [8]. But they considered the combination of the occurrence frequency density distributions of the nucleotides and their chemical properties into the general form of pseudo k-tuple nucleotide composition (PseKNC) as feature vectors. Based on the same SVM classifier, He *et al*. developed PseUI using hybrid features including nucleotide composition (NC), dinucleotide composition (DC), pseudo dinucleotide composition (PseDNC), position-specific nucleotide propensity (PSNP), and position-specific dinucleotide propensity (PSDP) in 2018[9]. In 2019, two new models, iPse-CNN and XG-PseU, were built to identify Ψ sites. iPse-CNN proposed by Tahir *et al.* is a convolutional neural network-based method using one-hot features [10],while XG-PseU proposed by Liu *et al.* is an XGBoost-based method with optimal features from six types of features, namely, nucleotide composition (NC), dinucleotide composition (DNC), trinucleotide composition (TNC), nucleotide chemical property (NCP), nucleotide density (ND), and one-hot encoding [11]. It is worthy of note that Liu *et al.* built the newest benchmark datasets according to the latest RMBase v2.0 database [12]. Despite these efforts, the performance of computational model still needs further improvement.

To develop a more effective model for identifying Ψ sites, we propose an ensemble model called EnsemPseU (https://github.com/biyue1026/EnsemPseU) that integrates support vector machine (SVM), extreme gradient boosting (XGBoost), na ïve Bayes (NB), k-nearest neighbor (KNN), and random forest (RF) based on a majority voting strategy. We collected and applied the latest datasets used in XG-PseU. Then five encoding methods are employed to extract features, namely, kmer, binary, enhanced nucleic acid composition (ENAC), nucleotide chemical property (NCP), and nucleotide density (ND). Besides, chi-square feature selection is used to reduce the feature dimensionality and redundant information. Furthermore, 10-fold cross-validation, jackknife and independent tests were used to evaluate the model performance. Figure 1 displays the specific framework of EnsemPseU.

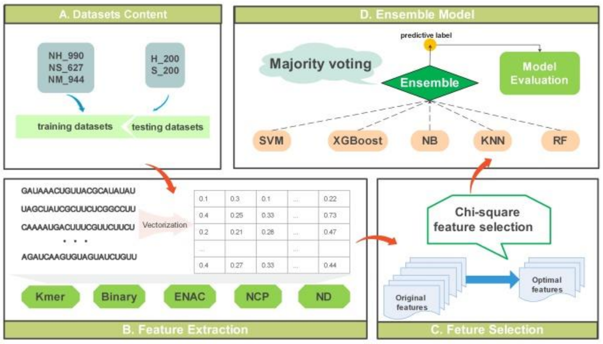

FIGURE 1. **The framework of EnsemPseU.**

II. MATERIALS AND METHODS

A. BENCHMARK DATASETS

Based on RMBase [13]. Chen *et al*. built the first benchmark datasets named H_990, S_628, and M_944, which included three species: *H. sapiens*, *S. cerevisiae*, and *M. musculus*, respectively [8]. With the updated version of RMBase (*i.e.*, RMBase v2.0) [12], Liu *et al.* established the second benchmark datasets named NH_990, NS_627, and NM_944, which also collected from *H. sapiens*, *S. cerevisiae*, and *M. musculus*, respectively [11]. In our study, we used NH_990, NS_627, and NM_944 to implement experiments. We also adopted the independent datasets H_200 and S_200 to evaluate model performance, which were built by Chen *et al.* [8]. A detailed statistical summary about these datasets is listed in Supplementary Table S1.

B. FEATURE EXTRACTION

1) KMER

Kmer is a widely used feature extraction approach in bioinformatics, which has been used in various fields, such as the identification of microRNA precursors [14], enhanced regulatory sequence prediction [15], and the phenotypic classification of metagenomic colon cancer reads [16]. For a RNA sequence, kmer can be defined as the frequencies of occurrence of k neighboring nucleic acids. The kmer (*k*=2) can be denoted as:

$$\mathcal{F} = \frac{N(t)}{L}, \quad t \in \{AA, AC, AG, \dots, UU\} \tag{1}$$

where *N(t)* is the number of kmer type *t* and *L* is the length of a RNA sequence. In this study, *k* ranges from 1 to 5.

2) BINARY

Through binary encoding, each nucleotide is represented by a four-dimensional binary vector, that is, nucleotide A is encoded by (1, 0, 0, 0), nucleotide C is encoded by (0, 1, 0, 0), nucleotide G is encoded by (0, 0, 1, 0), and nucleotide U is encoded by (0, 0, 0, 1).

3) ENAC

ENAC encoding was proposed for the first time by Chen [17]. It calculates the nucleic acid composition based on the fixed-length window that continuously slides from the 5′ to the 3′ terminus of each nucleotide sequence. It is usually applied to encode the nucleotide sequence with an equal length. The description of ENAC is as follows:

$$V = [\frac{N_{A,win_1}}{S}, \frac{N_{C,win_1}}{S}, \frac{N_{G,win_1}}{S}, \frac{N_{U,win_1}}{S}, \frac{N_{A,win_2}}{S}, \dots, \frac{N_{G,win_{L-S+1}}}{S}, \frac{N_{U,win_{L-S+1}}}{S}] \tag{2}$$

where S is the size of the sliding window, N_{t,win_r} is the number of nucleic acids *t* in the sliding window *r*, $t \in \{A, C, G, U\}$, and $r = 1, 2, \dots, L - S + 1$. In this work, the sliding window's length *S* is fixed as 5.

4) NCP

It has been demonstrated that considering the nucleotides' chemical properties may help to improve the model's predictive performance. We used the same strategy as XG-PseU [11] to encode each nucleotide with a three-dimensional vector, that is, A (1,1,1), C (0,1,0), G (1,0,0), and U (0,0,1).

5) ND

The nucleotide density (ND) encoding integrates the nucleotide frequency information and the distribution of each nucleotide reflected in the RNA sequence [11]. The density d_i of any nucleotide N_j at position i in the RNA sequence is defined by the following formula:

$$d_i = \frac{1}{\|S_i\|} \sum_{j=1}^{l} f(N_j) \tag{3}$$

$$f(N_j) = \begin{cases} 1 & if\ N_j\ is\ the\ nucleotide\ concerned \\ 0 & otherwise \end{cases} \tag{4}$$

where $\|S_i\|$ is the length of the sliding substring concerned, while *l* is the corresponding locator's sequence position. Take the sequence "UUACGGCCUA" as an example. The density of "U" is 1 (1/1), 1 (2/2), and 0.333 (3/9) at positions 1, 2, and 9, respectively. The density of "A" is 0.333 (1/3) and 0.5 (2/10) at positions 3 and 10, respectively. The density of "C" is 0.25 (1/4), 0.286 (2/7), and 0.375 (3/8) at positions 4, 7, and 8, respectively. The density of "G" is 0.2 (1/5) and 0.333 (2/6) at positions 5 and 6, respectively.

C. CHI-SQUARE FEATURE SELECTION

The chi-square (χ2) test is often applied to determine the independence of two events in statistics [18]. It can also act as a feature selection method that can evaluate the degree of association of features with respect to the class labels [19]. The features with higher scores in the chi-square test may be deemed as priority features for classification. The basic formula of the chi-square test is as follows:

$$\chi^2 = \sum \frac{(A-E)^2}{E} \tag{5}$$

where *A* is the observed value and *E* is the expected value. For discrete data, the chi-square test can directly test whether two features are related, while the range of continuous value features needs to be discretized into intervals.

D. ENSEMBLE MODEL CONSTRUCTION

For the identification of pseudouridine sites, the common practice is to find a suitable classifier that can precisely recognize the special sites. Nevertheless, a single classifier often fails to achieve satisfactory predictive performance. As demonstrated by a series of previous studies, such as on protein fold pattern recognition [20], enzyme functional classification [21], protein subcellular location prediction [22], protein–protein binding site identification [23], multiple lysine PTM site identification in proteins [24], recombination spot identification [25], the enhancer identification [26] and so on, the ensemble predictor based on fusing of an array of individual predictors via a voting system or other strategy can yield much better predictive performance [25].

Here, five popular classification algorithms (SVM, XGBoost, NB, KNN, and RF) were integrated into an ensemble model based on a majority voting strategy. The prediction label of a test sample is determined by the formula:

$$Y = mode\{P_{SVM}, P_{XGBoost}, P_{NB}, P_{KNN}, P_{RF}\} \tag{6}$$

Assuming that we get the results of five classifiers as follows: P_{SVM}=1, $P_{XGBoost}$=1, P_{NB}=0, P_{KNN}=0 and P_{RF}=*1*. We would classify the test sample as class 1 according to $Y = mode\{1,1,0,0,1\}$ for the number of 1 is larger than that of 0.

SVM is a commonly used machine learning algorithm in bioinformatics, which makes the linear indivisible samples in original space become separable by mapping the original sample features in low-dimensional space to high-dimensional space [27, 28]. Kernel functions are used to map low-dimensional space to high-dimensional space, and have been widely developed for different classification scenarios, including Gaussian radial basis function (RBF) and linear/polynomial/sigmoid kernel.

XGBoost is a machine learning algorithm based on gradient tree boosting [29]. It can find the overall optimal solution by performing the second-order Taylor expansion to the loss function, and adds regularization items to the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2989469, IEEE Access

objective function. In theory, XGBoost is an improvement on the gradient boosting decision tree (GBDT) algorithm and is more efficient for avoiding over-fitting.

NB is a classification method since it is based on the simple assumption that attributes are conditionally independent of each other when the target value is given [30]. In this paper, we used Bernoulli NB as one of the base classifiers in our study. The Bernoulli NB classifier works under the assumptions that data features are independent and have Bernoulli distributions.

KNN algorithm is a commonly employed unsupervised algorithm that clusters samples by calculating their similarities/distances [31]. The key idea of the KNN algorithm is that, the sample would also belong to a category if the most of the *k* nearest samples of this sample belong to a certain category.

RF classifier is a widely employed ensemble classifier that produces multiple decision trees, using a randomly selected subset of training samples and variables [32]. When applying RF, the number of decision trees is an important parameter and should be tested exhaustively based on the specific application or biological question, for optimal predictive performance [17].

III. RESULTS AND DISCUSSION

To measure the performance of our model, we used four metrics, sensitivity (Sn), specificity (Sp), accuracy (Acc), and the Matthew's correlation coefficient (MCC), which have been used in a series of studies to evaluate the effectiveness of predictors [33-35]. These measurements are defined as follows:

$$Sn = 1 - \frac{N_-^+}{N^+} \tag{7}$$

$$Sp = 1 - \frac{N_+^-}{N^-} \tag{8}$$

$$Acc = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} \tag{9}$$

$$MCC = \frac{1 - \frac{N_-^+ + N_+^-}{N^+ + N^-}}{\sqrt{\left(1 + \frac{N_+^- - N_-^+}{N^+}\right)\left(1 + \frac{N_-^+ - N_+^-}{N^-}\right)}} \tag{10}$$

where N^+ represents the total number of positive samples (Ψ); N^- represents the total number of negative samples $(non\ \Psi)$; N_-^+ represents the number of positive samples incorrectly predicted as negative samples; and N_+^- represents the number of negative samples incorrectly predicted as positive samples. K-fold (k=10) cross-validation, jackknife and independent tests were also used to determine the model's generalizability.

A. THE RESULTS OF FEATURE SELECTION

In our study, we chose the kmer, binary, ENAC, NCP, and ND encoding schemes to convert nucleotide sequences to numeral characteristics. The total number of features for each nucleotide sequence exceeds 1700 in this case. As has been investigated, highly dimensional feature vectors may contain some noisy information and influence the model's performance. Therefore, it is essential to apply a feature selection approach to reduce the irrelevant features from a large number of original features. The feature selection process calculates the score of each probable feature based on a specific feature selection technique and then selects the best *k* features [36].

The bottleneck in this approach is how to decide the number *k* of optimal features. Here, we combine the selection of *k* value with the parameter optimization of the ensemble model. Specifically, we took the values of *k* ranging in [100, 1000] with the step of 100, and parameter optimization process was implemented at each *k* value based on 10-fold cross-validation. We optimize the parameter γ of SVM, boosting the learning rate (r) of XGBoost, k neighbors (k) of KNN, and the number of decision trees (n) of RF. In addition to the parameters mentioned above, the remaining parameters follow the default of the scikit-learn (https://github.com/scikit-learn/scikit-learn). Based on our artificial experience, we created 6×5×5×5=750 combinations by setting $\gamma \in \{0.01, 0.02, 0.04, 0.06, 0.08, 0.1\}$, $r \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, $k \in \{1, 3, 5, 7, 9\}$ and $n \in \{100, 200, 300, 400, 500\}$.

For each *k* value, 10-fold cross-validation test was implemented according to different parameter combinations running 750 times. The optimization results can be seen in the Supplementary file. To estimate the performance of each *k* value objectively, we calculated the average accuracy of 750 results of 10-fold cross-validation. For better observation, we drew the change of accuracy along with the change of *k* in Figure 2.

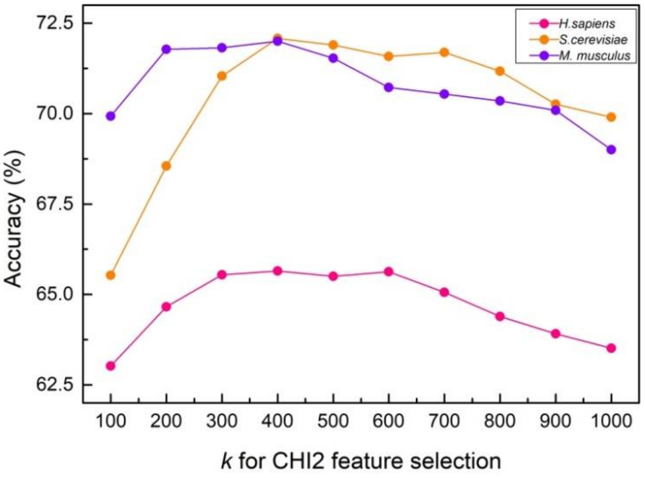

FIGURE 2 **The accuracy comparison of different *k* values for three species.**

For *H. sapiens*, the accuracy reached 60.35% under 1600 dimensions of original features. When *k* = 400, the accuracy peaked at 65.65%, which was an improvement of 5.3% in comparison with the accuracy of original features. For *S. cerevisiae*, the accuracy was 65.99% under 1720 dimensions of original features. At *k* = 400, the accuracy increased to 72.08%, which was an improvement of 6.09% in comparison with the accuracy of original features. For *M. musculus*, the accuracy was 66.45% under 1600 dimensions of original features. When *k* = 400, the accuracy reached up to 72%, which was an improvement of 5.55% in comparison with the accuracy of original features. The improvements of accuracy indicate that χ^2 feature selection can eliminate redundant information and retain useful information. Therefore, we determined 400 as the number of optimal features for above three species. After that, we selected the parameter with the highest Acc to build the model with *k*=400 for each species. See Supplementary Table S2 for detailed corresponding settings.

In order to find an appropriate feature selection method, we have also carried out experiments using other two kinds of feature ranking methods including mRMR, F-score, and compared them on 10-fold cross-validation. The comparison results shown in Figure 3 indicated that the prediction results of chi-square are the best for all of three species. Therefore, we adopted the chi-square feature selection method to construct our predictor.

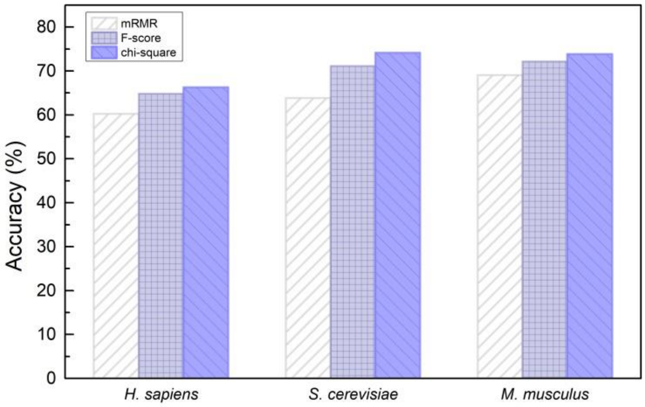

FIGURE 3. Comparison of mRMR, F-score and chi-square feature selection.

B. EFFECTIVENESS OF ENSEMBLE MODEL

During our experiments, we chose five different machine learning methods SVM, XGBoost, NB, KNN, and RF as basic classifiers and integrated them by a majority voting strategy. The parameters for each classifier were determined as described in the above section on feature selection. Then we trained the model for each species. To verify the ensemble's effectiveness, we compared ensemble model with individual classifiers on 10-fold cross-validation. The detailed results can be seen in Supplementary Table S3. For more convenient observation, we plotted accuracy of ensemble model and individual models for three species as show in Figure 4-6.

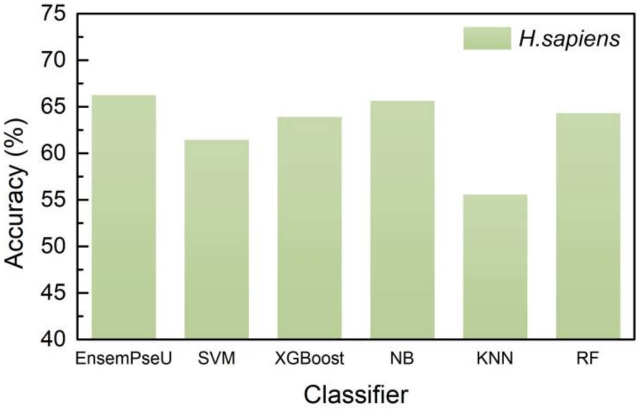

FIGURE 4 The comparison of individual and ensemble classifiers for *H. sapiens*.

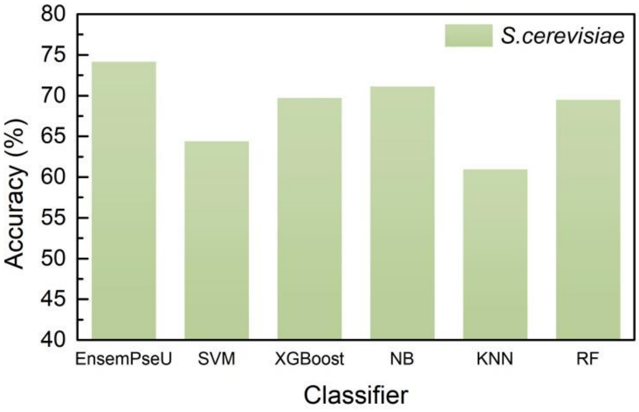

FIGURE 5 The comparison of individual and ensemble classifiers for *S. cerevisiae*.

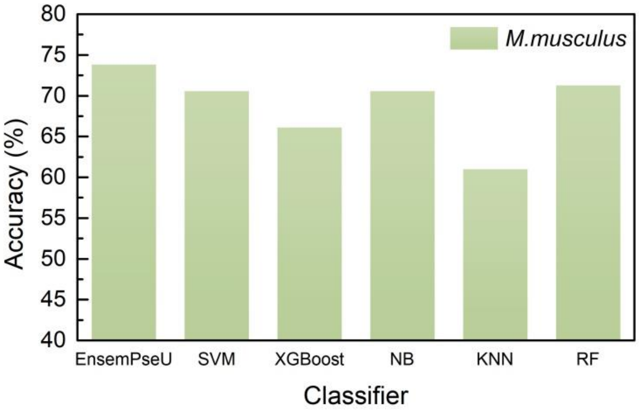

FIGURE 6 The comparison of individual and ensemble classifiers for *M. musculus*.

For all three species, the ensemble model achieved the best performance, while KNN achieved the lowest Acc. With respect to a basic classifier, NB classifier reached the best performance with Acc of 65.66% for *H. sapiens*, 71.14% for *S. cerevisiae*, and 70.58% for *M. musculus*.

C. COMPARISON WITH EXISTING METHOD

Previous models for identifying Ψ sites, iRNA-PseU, PseUI, and iPseU-CNN, all used the old training sets, H_990, S_628, and M_944 for *H. sapiens*, *S. cerevisiae*, and *M. musculus*, respectively, except for XG-PseU. Since Liu *et al*. updated the datasets, we compared our model EnsemPseU with the latest predictor XG-PseU to evaluate the ability to identify Ψ sites. The results obtained by EnsemPseU and XG-PseU on 10-fold cross-validation and independent tests are listed in Table 1 and Table 2, respectively.

TABLE 1 Comparison of EnsemPseU with XG-PseU on 10-fold cross-validation test.

Species	Methods	Sn (%)	Sp (%)	Acc (%)	MCC
H. sapiens	XG-PseU	63.45	68.65	66.05	0.32
	EnsemPseU	**63.46**	**69.09**	**66.28**	**0.33**
S. cerevisiae	XG-PseU	**77.35**	69.48	73.42	0.47
	EnsemPseU	73.88	**74.45**	**74.16**	**0.49**
M. musculus	XG-PseU	65.92	**76.30**	71.10	0.43
	EnsemPseU	**75.43**	72.25	**73.84**	**0.48**

TABLE 2 Comparisons of EnsemPseU with XG-PseU on independent tests.

Species	Methods	Sn (%)	Sp (%)	Acc (%)	MCC
H. sapiens	XG-PseU	-	-	67.5	-
	EnsemPseU	73.00	66.00	**69.50**	0.39
S. cerevisiae	XG-PseU	-	-	71.0	-
	EnsemPseU	85.00	65.00	**75.00**	0.51

Note: '-' denotes that the metric value is not given in XG-PseU.

For *H. sapiens*, EnsemPseU demonstrated the effectiveness and advantages when compared to XG-PseU on both 10-fold cross-validation test and independent test with respect to Sn, Sp, Acc, and MCC. For *S. cerevisiae*, the results of 10-fold cross-validation tests of EnsemPseU are higher than those of XG-PseU with respect to Sp, Acc, and MCC, except for Sn. Moreover, the results of independent test of EnsemPseU are higher than that of XG-PseU. For *M. musculus*, the results of 10-fold cross-validation are higher than those of XG-PseU with respect to Sn, Acc, and MCC, except for Sp. In addition, the ROC curves on 10-fold cross-validation were also plotted as shown in Figure 7. EnsemPseU and XG-PseU attained the same AUC (the area under ROC) 0.700 for *H. sapiens*. Nevertheless, EnsemPseU obtained the AUC of 0.786 and 0.775 which is better than XG-PseU with AUC of 0.74 and 0.77 for *S. cerevisiae* and *M. musculus*, respectively.

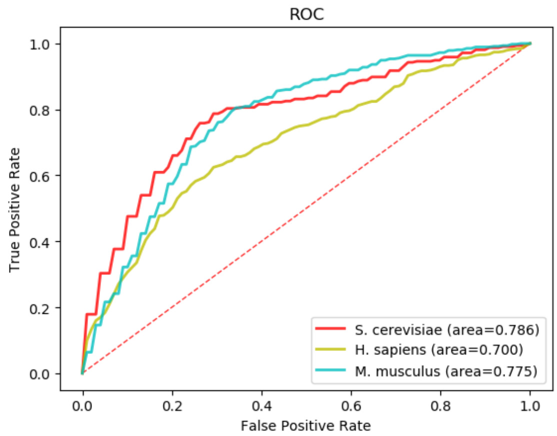

FIGURE 7 The ROC curves of EnsemPseU for three species on 10-fold cross-validation.

Since jackknife test is usually regarded as the most objective test method [8], and so, we provide the jackknife test results for *H. sapiens, S. cerevisiae, and M. musculus* in Table 3, respectively.

TABLE 3 Results of EnsemPseU on jackknife test.

Species	Sn (%)	Sp (%)	Acc (%)	MCC
H. sapiens	63.03	70.91	66.97	0.34
S. cerevisiae	72.61	72.84	72.73	0.45
M. musculus	74.36	71.40	72.88	0.46

In conclusion, EnsemPseU showed better performance regarding most indicators, especially for *S. cerevisiae*. However, we also noticed that both EnsemPseU and XG-PseU achieved the best MCC for *S. cerevisiae*, but the lowest one for *H. sapiens*. That is to say, the conservatism of sequences varies among different species.

IV. CONCLUSION

In this paper, we proposed an ensemble approach through integrating five classifiers: SVM, XGBoost, NB, KNN, and RF. The chi-square test was employed to remove redundant information. Testing revealed that the accuracy improved 5. 3% for *H. sapiens*, 6.09% for *S. cerevisiae*, and 5.55% for *M. musculus* after chi-square feature selection. Moreover, upon evaluation by 10-fold cross-validation and an independent test, our proposed model EnsemPseU outperformed the other best existing model.

REFERENCES

[1] X. Y. Li, S. Q. Ma, and C. Q. Yi, "Pseudouridine: the fifth RNA nucleotide with renewed interests," *Current Opinion in Chemical Biology*, Review vol. 33, pp. 108-116, Aug 2016.

[2] H. Adachi, M. D. De Zoysa, and Y.-T. Yu, "Post-transcriptional pseudouridylation in mRNA as well as in some major types of noncoding RNAs," *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, vol. 1862, no. 3, pp. 230-239, 2019.

[3] A. F. Lovejoy, D. P. Riordan, and P. O. Brown, "Transcriptome-wide mapping of pseudouridines:

pseudouridine synthases modify specific mRNAs in S. cerevisiae," *PloS one*, vol. 9, no. 10, p. e110799, 2014.

[4] S. Schwartz *et al.*, "Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA," *Cell*, vol. 159, no. 1, pp. 148-162, 2014.

[5] M. A. Nakamoto, A. F. Lovejoy, A. M. Cygan, and J. C. Boothroyd, "mRNA pseudouridylation affects RNA metabolism in the parasite Toxoplasma gondii," *RNA*, vol. 23, no. 12, pp. 1834-1849, 2017.

[6] M. Charette and M. W. Gray, "Pseudouridine in RNA: what, where, how, and why," *IUBMB life*, vol. 49, no. 5, pp. 341-351, 2000.

[7] Y. H. Li, G. G. Zhang, and Q. H. Cui, "PPUS: a web server to predict PUS-specific pseudouridine sites," *Bioinformatics*, Article vol. 31, no. 20, pp. 3362-3364, Oct 2015.

[8] W. Chen, H. Tang, J. Ye, H. Lin, and K. C. Chou, "iRNA-PseU: Identifying RNA pseudouridine sites," *Molecular Therapy-Nucleic Acids*, Article vol. 5, p. 9, Jul 2016, Art. no. e332.

[9] J. J. He, T. Fang, Z. Z. Zhang, B. Huang, X. L. Zhu, and Y. Xiong, "PseUI: Pseudouridine sites identification based on RNA sequence information," *Bmc Bioinformatics*, Article vol. 19, p. 11, Aug 2018, Art. no. 306.

[10] M. Tahir, H. Tayara, and K. T. Chong, "iPseU-CNN: Identifying RNA Pseudouridine Sites Using Convolutional Neural Networks," *Molecular Therapy-Nucleic Acids*, Article vol. 16, pp. 463-470, Jun 2019.

[11] K. Liu, W. Chen, and H. Lin, "XG-PseU: an eXtreme Gradient Boosting based method for identifying pseudouridine sites," *Molecular genetics and genomics : MGG*, 2019 Aug 07 (Epub 2019 Aug 2019.

[12] J. J. Xuan *et al.*, "RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data," *Nucleic Acids Research*, Article vol. 46, no. D1, pp. D327-D334, Jan 2018.

[13] W. J. Sun *et al.*, "RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data," *Nucleic Acids Research*, Article vol. 44, no. D1, pp. D259-D265, Jan 2016.

[14] B. Liu, L. Y. Fang, S. Y. Wang, X. L. Wang, H. T. Li, and K. C. Chou, "Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy," *Journal of Theoretical Biology*, Article vol. 385, pp. 153-159, Nov 2015.

[15] M. Ghandi, D. Lee, M. Mohammad-Noori, and M. A. Beer, "Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features," *Plos Computational Biology*, Article vol. 10, no. 7, p. 15, Jul 2014, Art. no. e1003711.

[16] A. Kishk *et al.*, "A Hybrid Machine Learning Approach for the Phenotypic Classification of Metagenomic Colon Cancer Reads Based on Kmer Frequency and Biomarker Profiling," in *2018 9th Cairo International Biomedical Engineering Conference*, A. Eldeib, T. Basha, and I. Yassine, Eds. (Cairo International Biomedical Engineering Conference, New York: Ieee, 2018, pp. 118-121.

[17] Z. Chen *et al.*, "iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data," *Briefings in bioinformatics*, 2019 Apr 24 (Epub 2019 Apr 2019.

[18] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *In International Conference on Machine Learning (ICML)*, vol. 97, p. 35.

[19] O. Rehman, H. Q. Zhuang, A. M. Ali, A. Ibrahim, and Z. W. Li, "Validation of miRNAs as Breast Cancer Biomarkers with a Machine Learning Approach," *Cancers*, Article vol. 11, no. 3, p. 10, Mar 2019, Art. no. 431.

[20] H. B. Shen and K. C. Chou, "Ensemble classifier for protein fold pattern recognition," *Bioinformatics*, Article vol. 22, no. 14, pp. 1717-1722, Jul 2006.

[21] H. B. Shen and K. C. Chou, "EzyPred: A top-down approach for predicting enzyme functional classes and subclasses," *Biochemical and Biophysical Research Communications*, Article vol. 364, no. 1, pp. 53-59, Dec 2007.

[22] K. C. Chou and H. B. Shen, "Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms," *Nature Protocols*, Article vol. 3, no. 2, pp. 153-162, 2008.

[23] J. H. Jia, Z. Liu, X. Xiao, B. X. Liu, and K. C. Chou, "Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition," *Journal of Biomolecular Structure & Dynamics*, Article vol. 34, no. 9, pp. 1946-1961, 2016.

[24] W. R. Qiu, B. Q. Sun, X. Xiao, Z. C. Xu, and K. C. Chou, "iPTM-mLys: identifying multiple lysine PTM sites and their different types," *Bioinformatics*, Article vol. 32, no. 20, pp. 3116-3123, Oct 2016.

[25] B. Liu, S. Y. Wang, R. Long, and K. C. Chou, "iRSpot-EL: identify recombination spots with an ensemble learning approach," *Bioinformatics*, Article vol. 33, no. 1, pp. 35-41, Jan 2017.

[26] B. Liu, K. Li, D. S. Huang, and K. C. Chou, "iEnhancer-EL: identifying enhancers and their strength with ensemble learning approach," *Bioinformatics*, Article vol. 34, no. 22, pp. 3835-3842, Nov 2018.

[27] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988-999, 1999.

[28] V. Vapnik, *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995.

[29] T. Q. Chen, C. Guestrin, and M. Assoc Comp, *XGBoost: A Scalable Tree Boosting System* (Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining). New York: Assoc Computing Machinery, 2016, pp. 785-794.

[30] T. M. Mitchell, "Machine learning. 1997," *Burr Ridge, IL: McGraw Hill*, vol. 45, no. 37, pp. 870-877, 1997.

[31] Y. D. Cai, T. Huang, L. L. Hu, X. H. Shi, L. Xie, and Y. X. Li, "Prediction of lysine ubiquitination with mRMR feature selection and analysis," *Amino Acids*, Article vol. 42, no. 4, pp. 1387-1395, Apr 2012.

[32] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001.

[33] L. Y. Wei, S. X. Wan, J. S. Guo, and K. K. L. Wong, "A novel hierarchical selective ensemble classifier with bioinformatics application," *Artificial Intelligence in Medicine*, Article vol. 83, pp. 82-90, Nov 2017.

[34] L. Y. Wei, P. W. Xing, G. T. Shi, Z. L. Ji, and Q. Zou, "Fast Prediction of Protein Methylation Sites Using a Sequence-Based Feature Selection Technique," *Ieee-Acm Transactions on Computational Biology and Bioinformatics*, Article; Proceedings Paper vol. 16, no. 4, pp. 1264-1273, Jul-Aug 2019.

[35] M. Zhang *et al.*, "MULTiPly: a novel multi-layer predictor for discovering general and specific types of promoters," *Bioinformatics*, Article vol. 35, no. 17, pp. 2957-2965, Sep 2019.

[36] I. S. Thaseen and C. A. Kumar, "Intrusion detection model using fusion of chi-square feature selection and multi class SVM," *Journal of King Saud University-Computer and Information Sciences*, vol. 29, no. 4, pp. 462-472, 2017.

Yue Bi received a B.S. degree from Liaoning Normal University, China. She is currently working toward a Master's degree in the School of Science, Dalian Maritime University, China. Her research interests are bioinformatics, machine learning, and deep learning.

Dong Jin received a B.S. degree from Ludong University, China. He is currently working toward a Master's degree in the School of Science, Dalian Maritime University, China. His research interests include bioinformatics, computational biology, and machine learning.

Cangzhi Jia received a Ph.D. from the School of Mathematical Sciences, Dalian University of Technology, in 2007. She is an Associate Professor in the School of Science, Dalian Maritime University, China. Her research interests include mathematical modeling in bioinformatics and machine learning.