Mathematical Biosciences and Engineering

MBE, 19(12): 13829–13850.
DOI: 10.3934/mbe.2022644
Received: 24 August 2022
Revised: 11 September 2022
Accepted: 15 September 2022
Published: 19 September 2022

http://www.aimspress.com/journal/MBE

Research article

iPseU-TWSVM: Identification of RNA pseudouridine sites based on TWSVM

Mingshuai Chen[1,2,†], Xin Zhang[3,†], Ying Ju[4], Qing Liu[5,*] and Yijie Ding[2,*]

[1] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China
[2] Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, Zhejiang, China
[3] Beidahuang Industry Group General Hospital, Harbin, China
[4] School of Informatics, Xiamen University, Xiamen, China
[5] Department of Anesthesiology, Hospital (T.C.M) Affiliated to Southwest Medical University, Luzhou, China

† These Authors contribute equally to this work.

* **Correspondence:** Email: wuxi_dyj@csj.uestc.edu.cn, 1105859368@qq.com.

Abstract: Biological sequence analysis is an important basic research work in the field of bioinformatics. With the explosive growth of data, machine learning methods play an increasingly important role in biological sequence analysis. By constructing a classifier for prediction, the input sequence feature vector is predicted and evaluated, and the knowledge of gene structure, function and evolution is obtained from a large amount of sequence information, which lays a foundation for researchers to carry out in-depth research. At present, many machine learning methods have been applied to biological sequence analysis such as RNA gene recognition and protein secondary structure prediction. As a biological sequence, RNA plays an important biological role in the encoding, decoding, regulation and expression of genes. The analysis of RNA data is currently carried out from the aspects of structure and function, including secondary structure prediction, non-coding RNA identification and functional site prediction. Pseudouridine (У) is the most widespread and rich RNA modification and has been discovered in a variety of RNAs. It is highly essential for the study of related functional mechanisms and disease diagnosis to accurately identify У sites in RNA sequences. At present, several computational approaches have been suggested as an alternative to experimental methods to detect У sites, but there is still potential for improvement in their performance. In this study, we present a model based on twin support vector machine (TWSVM) for У site identification. The model combines a variety of feature representation techniques and uses the max-relevance and min-redundancy methods to obtain the optimum feature subset for training. The independent testing accuracy is improved by 3.4% in comparison to current advanced У site predictors. The outcomes demonstrate that our model has better generalization performance and improves the accuracy of У site identification. iPseU-TWSVM can be a helpful tool to identify У sites.

Keywords: pseudouridine sites; twin support vector machine; max-relevance and min-redundancy; RNA

1. Introduction

Various chemical modifications, including cytosine modification, uridine isomerization, and adenosine methylation, have been found in cellular RNA [1] and have been linked to important biological and physiological functions in cells [2]. У modification is a common posttranscriptional RNA modification known as the fifth base in RNA [3]. It is commonly present in a variety of species, and research has revealed that tRNA and rRNA contain large amounts of it [4]. Numerous biological processes have shown У to be crucial, and distinct У modifications serve different purposes at various places [5–7]. Therefore, the discovery of У sites in RNA sequences is crucial for both fundamental and applied biological research.

Initially, researchers identified У modification sites based on biochemical experiments. At first, researchers used paper chromatography to find У modification sites in the RNA of yeast, which was achieved by using RNA decomposition enzymes to decompose RNA and electrophoresis to separate out column chromatography on the upper layer of paper [3–8]. Later researchers successively used high-performance liquid chromatography and mass spectrometry to detect modification sites [9]. With the growing interest in this field, researchers have proposed a variety of high-throughput sequencing technologies, including Ψ-seq [10,11], PseudoU-seq [12] and CeU-Seq [13], and successfully used them to detect У sites. However, the methods described above are reliant on time-consuming, expensive, and difficult biochemical experiments, which are susceptible to environmental factors, and the sequencing process becomes increasingly difficult as the sequence length increases. Therefore, robust, fast, and inexpensive calculation methods are needed to predict У sites in RNA sequences.

First, Panwar and colleagues proposed a tRNAmod model to predict У sites in tRNA [14]. Then, a web server (PPUS) based on support vector machine (SVM) was proposed by Li et al. to identify У sites in S.cerevisiae and H.sapiens [15]. The frequency composition of nucleotides and the pseudo K-tuple nucleotide composition (PseKNC) were merged for feature representation in the iRNA-PseU model that Chen et al. created [16]. Subsequently, He et al. developed the SVM model (PseUI) to identify У sites in H.sapiens, S.cerevisiae and M.musculus, which combined a variety of feature extraction techniques including position-specific dinucleotide propensity (PSDP) [17]. Later, utilizing convolutional neural networks, Tahir et al. created a predictor (iPseU-CNN) [18]. Extreme gradient boosting (XGboost) was used by Liu et al. to create a new model known as XG-PseU [19]. Lv et al. also proposed a method called RF-PseU, which utilizes the LGBM algorithm for feature selection while combining the random forest algorithm for classification [20]. Saad et al. proposed a convolutional neural network model MU-PseUDeep [21], which combines sequence and secondary structural features to predict У sites. Li et al. built the model Porpoise by utilizing multiple type features and inputting them into the stacked ensemble learning framework [22]. Although the aforementioned techniques have proven successful in correctly identifying У sites in RNA sequences, they might still use more work in comparison to high-performance predictors [23–28].

In this study, we build a У site identification model (iPseU-TWSVM) based on TWSVM, and Figure 1 depicts the model construction process. The model combines multiple feature representation methods, including Kmer, ENAC and EIIP. To obtain the best subset of features, the mRMR approach is utilized. The model is then evaluated using 10-fold cross-validation (10-CV) and independent testing (I-testing). The average I-testing accuracy of the iPseU-TWSVM is 3.4% higher than that of current advanced predictors, demonstrating the better generalization performance of our model. Therefore, iPseU-TWSVM may become an effective tool for У site identification.

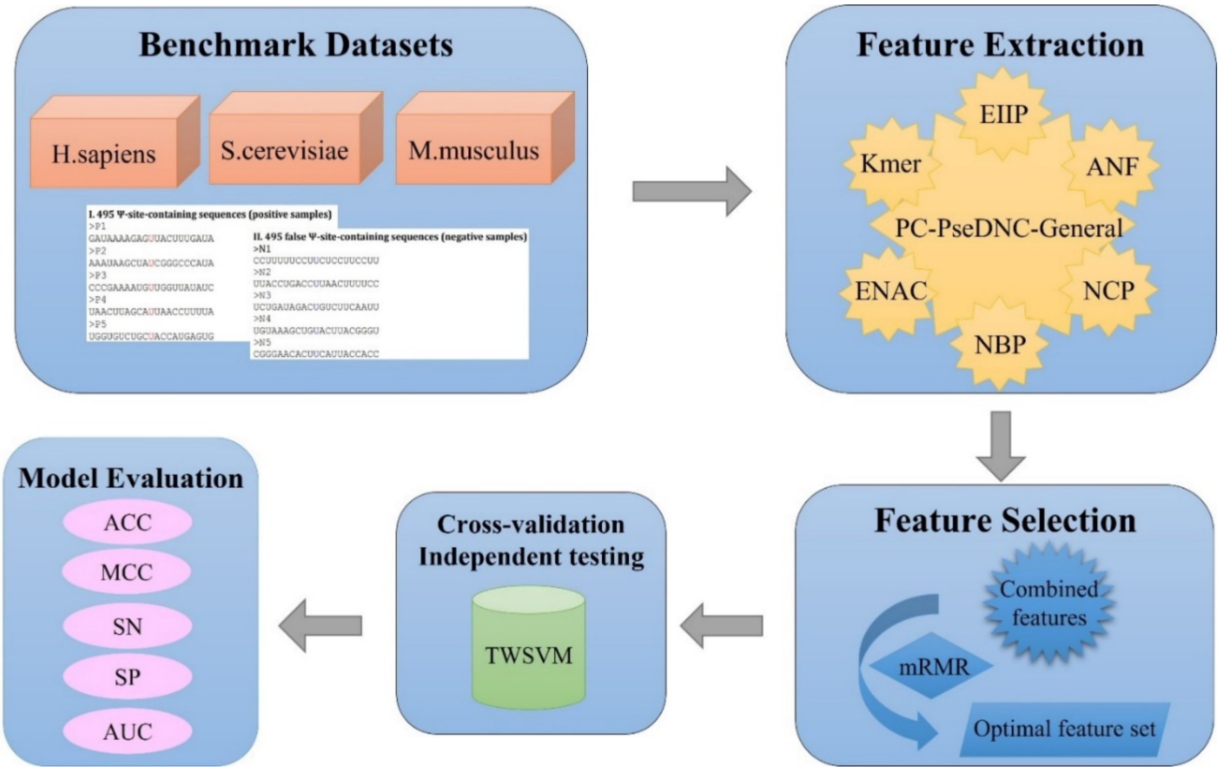


Figure 1. The flowchart of iPseU-TWSVM.

2. Materials and methods

2.1. Datasets

In this work, we train and evaluate our models using datasets created by Chen et al. [29]. The steps of constructing the benchmark dataset were as follows: 1307 positive samples and 33,280 negative samples were obtained at first, and then the subset-balancing treatment was adopted to reduce the number of negative samples according to Euclidean distance. The obtained distance values were sorted in ascending order, and the first 1307 negative samples were selected to form the negative subset. The training datasets contained data from three species, namely, H.sapiens, S.cerevisiae and M.musculus. The H.sapiens training dataset included 495 positive samples and 495 negative samples; the S.cerevisiae dataset included 314 positive samples and 314 negative samples; and the M.musculus dataset included 944 samples, half of which were positive samples. There were just two species in the I-testing datasets: H.sapiens and S.cerevisiae. Each of them included 200 samples, of which only half were positive and half were negative.

2.2. Feature extraction

Different types of features reflect biological significance from different perspectives, including sequence composition and physicochemical properties. In this work, a variety of types of features are used to comprehensively consider the composition, distribution and physicochemical properties of nucleotides in the sequence from various aspects to further improve the prediction performance of subsequent work.

2.2.1. Kmer

One effective technique for extracting RNA sequence characteristics is Kmer, which reflects the frequency of k adjacent nucleotides in the sequence. The frequencies of the k-neighboring nucleotides are used to generate the feature vector [30]. The method is provided by the web server Pse-in-One2.0 (http://bioinformatics.hitsz.edu.cn/Pse-in-One2.0/) [31].

2.2.2. PC-PseDNC-General

The approach offers 22 different physicochemical properties to create the pseudo-dinucleotide composition [32–34]. It overwrites the local sequence order and the global sequence order information into the feature vector. The relevant features are expressed in this form:

$$Vector = (m_1 m_2 \cdots m_{16} m_{16+1} \cdots m_{16+\lambda})^T \tag{1}$$

with

$$m_i = \begin{cases} \frac{q_i}{\sum_{j=1}^{16} q_j + \alpha \sum_{k=1}^{\lambda} \rho_k} & (1 \le i \le 16) \\ \frac{\alpha \rho_{\sigma-16}}{\sum_{j=1}^{16} q_j + \alpha \sum_{k=1}^{\lambda} \rho_k} & (16 + 1 \le i \le 16 + \lambda) \end{cases} \tag{2}$$

where $q_i (i = 1,2,\cdots,16)$ represents the 16 dinucleotides' normalized frequency of occurrence; $\alpha (0 \le \alpha \le 1)$ is the weight factor; and λ is the highest counted rank. ρ_{k} is the k-tier correlation factor.

ρ_{k} displays the relationship between the sequence orders of all neighboring dinucleotides along a specific RNA sequence, which can be written as

$$\rho_k = \frac{1}{l-k-1} \sum_{j=1}^{l-k-1} C(R_j, R_{j+k}) \quad (k = 1,2,\cdots,\lambda; \lambda < l - 1) \tag{3}$$

where $C(R_j, R_{j+k})$ indicates the correlation function expressed as

$$C(R_j, R_{j+k}) = \frac{1}{\sigma}\sum_{g=1}^{\sigma}[P_g(R_j) - P_g(R_{j+k})]^2 \tag{4}$$

where parameter σ is the number of physicochemical properties studied; $P_g(R_j)$ and $P_g(R_{j+k})$ are the related values of the gth property for the dinucleotides R_j at position j and R_{j+k} at position j+k.

2.2.3. Nucleotide chemical properties (NCP)

The coding method reflects that each nucleotide in the sequence has different chemical structures and binding properties. The ring structures of the four RNA nucleotides (ACGU) differ from one another, hydrogen bond, and functional group. Based on these differences, they may be represented with a 3D coordinate [35].

2.2.4. Accumulated nucleotide frequency (ANF)

The method incorporates data on each nucleotide's distribution in the RNA sequence as well as its frequency [35]. We can calculate the density d_i of an RNA sequence's ith prefix subsequence. It is defined as

$$d_i = \frac{1}{i}\sum_{j=1}^{i} f(x_j),\ where\ f(x_j) = \begin{cases} 1, & if\ x_j = x_i \\ 0, & otherwise \end{cases} \tag{5}$$

where i is the length of the sliding string and x_j represents the nucleotide at the jth position.

2.2.5. Electron-Ion Interaction Pseudopotentials (EIIP)

The EIIP values represent the energy of the delocalized electron in the nucleotide. The nucleotides in the DNA sequence have previously been denoted by the EIIP values of A, G, C and T [36]. In the RF-PseU method [20], each nucleotide in an RNA sequence was also coded by EIIP feature vectors.

2.2.6. Enhanced Nucleic Acid Composition (ENAC)

Using a fixed length window, the approach was used to determine the nucleotide composition [20,35]. Afterward, RNA sequences were converted into equally long feature vectors. Sequence length and sliding window size are two factors that affect the dimension of ENAC coding.

$$E = (b_1, b_2, \cdots, b_n),\ where\ b_i = \frac{N_i}{N}, i \in \{A, C, G, U\} \tag{6}$$

where N is the sliding window size and n is the coding dimension.

2.2.7. Nucleotide binary profiles (NBP)

Binary profiles provide the position specific composition of nucleotides in RNA fragments [35,36]. A four-digit binary vector is used to encode each nucleotide. Dibinary profiles are different from binary profiles in that they are encoded for 16 dinucleotides, i.e., AA is denoted by (0,0,0,0).

2.3. Feature selection

2.3.1. Max-relevance and min-redundancy (mRMR)

mRMR [37] is a commonly used feature selection method for compressing feature vector space. The goal of this technique is to identify a subset of features from the initial feature set that have the lowest correlation between features and the highest correlation with the output result. It considers the connection between features as well as the association between features and labels. The mechanism of feature selection is as follows.

The mutual information is used to find the feature subset *S* containing *m* features first, so that the *m* features found have the maximum correlation with the category *c*. The correlation between the feature subset *S* and the category *c* is defined by the average value of all mutual information between each feature and category as shown in (7).

$$maxD(S,c), D = \frac{1}{|S|}\sum_{x_i \in S} I(x_i;c) \tag{7}$$

where $I(x_i;c)$ is mutual information; S is a subset of features of length *m*; x_i is the ith feature in S and c is category variable.

Then the features selected by the maximum correlation may be redundant, and (8) is used to eliminate the redundancy among *m* features.

$$minR(S), R = \frac{1}{|S|^2}\sum_{x_i, x_j \in S} I(x_i;x_j) \tag{8}$$

The final feature subset *S* is obtained by combining the maximum correlation *D* with the minimum redundancy *R*.

$$mRMR = \max\left[\frac{1}{|S|}\sum_{x_i \in S} I(x_i;c) - \frac{1}{|S|^2}\sum_{x_i, x_j \in S} I(x_i;x_j)\right] \tag{9}$$

Compared with other feature selection methods, the proposed algorithm considers the redundancy among features, further optimizes the feature subset, and solves the problem that the maximum dependency is difficult to achieve. However, only approximate optimal solutions can be obtained in practical applications.

2.4. Classifiers

2.4.1. Twin support vector machine (TWSVM)

Consider the binary classification issue using the training datasets

$$D_{train} = \{(u_1,1),(u_2,1),\cdots,(u_m,1),(u_{m+1},-1),(u_{m+2},-1),\cdots,(u_{m+n},-1)\}, \tag{10}$$

where $u_i \in R^n, i = 1,2,\cdots,m+n$.

Let $T = (u_1,u_2,\cdots,u_m)^T \in R^{m\times n}, F = (u_{m+1},u_{m+2},\cdots,u_{m+n})^T \in R^{n\times n}$ and $l = m+n$.

TWSVM [38] looks for a pair of nonparallel hyperplanes in the linear case.

$$w_+u + b_+ = 0 \text{ and } w_-u + b_- = 0 \tag{11}$$

where $w_+ \in R^n, w_- \in R^n, b_+ \in R, b_- \in R$ by solving the following pair of QPPs:

$$\min_{w_+,b_+,\xi_-} \frac{1}{2}(Tw_+ + e_+b_+)^T(Tw_+ + e_+b_+) + c_1e_-^T\xi_-$$

$$s.t. -(Fw_+ + e_-b_+) + \xi_- \geq e_-,\ \xi_- \geq 0 \tag{12}$$

and

$$\min_{w_-,b_-,\xi_+} \frac{1}{2}(Fw_- + e_-b_-)^T(Fw_- + e_-b_-) + c_2e_+^T\xi_+$$

$$s.t. (Tw_- + e_+b_-) + \xi_+ \geq e_+, \xi_+ \geq 0 \tag{13}$$

where c_1, c_2 are the penalty parameters, e_+, e_- are all 1 vectors ($e_+, e_- = [1 \cdots 1]^T$) whose dimensions are the same as the number of positive and negative samples respectively, and ξ_+, ξ_- are slack vectors of appropriate dimension.

Minimizing the objective function means making a hyperplane as close as possible to one type of data, and the constraint requires that the distance between the hyperplane and the other type of data is at least greater than 1. Their corresponding Lagrange dual problems are

$$\max_{\alpha}\ e_-^T\alpha - \frac{1}{2}\alpha^TJ(K^TK)^{-1}J^T\alpha$$

$$s.t. 0 \leq \alpha \leq c_1e_- \tag{14}$$

and

$$\max_{\gamma} e_+^T\gamma - \frac{1}{2}\gamma^TK(J^TJ)^{-1}K^T\gamma$$

$$s.t. 0 \leq \gamma \leq c_2e_+ \tag{15}$$

where

$$K = [T \quad e_+] \in R^{m\times(n+1)},\ J = [F \quad e_-] \in R^{n\times(n+1)}. \tag{16}$$

The solution to the primary problem can be acquired by addressing the dual problem, which can be obtained by

$$(w_+^T, b_+)^T = -(K^TK)^{-1}J^T\alpha, \tag{17}$$

$$(w_-^T, b_-)^T = -(J^TJ)^{-1}K^T\gamma. \tag{18}$$

Therefore, an unknown point $u \in R^n$ is predicted to the Class by

$$Class = \arg\min_{s=-,+}|w_su + b_s|, \tag{19}$$

where $|\cdot|$ is the perpendicular distance of point u from the planes $w_su + b_s = 0, s = -, +$.

This method not only divides a large quadratic programming problem into two small quadratic programming problems, which improves the training speed, but also is not very sensitive to noise.

2.5. Model evaluation metrics and methods

Five indicators were widely used to assess how well the built models performed [39–41], accuracy (ACC), sensitivity (SN), specificity (SP), Matthew correlation coefficient (MCC), and integral area under the receiver operating characteristic curve (auROC), which were calculated using the following equations.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{20}$$

$$SN = \frac{TP}{FN+TP} \tag{21}$$

$$SP = \frac{TN}{TN+FP} \tag{22}$$

$$MCC = \frac{TN\times TP-FN\times FP}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FN)\times(TN+FP)}} \tag{23}$$

where TP, TN, FP, and FN represent true positive, true negative, false positive and false negative, respectively.

We use 10-CV for comparison [42–62]. The training datasets are equally divided into ten subsets. The remaining one subset is tested after the proposed model has been trained using nine subsets. After each subset is tested once, the procedure is repeated ten times, and the average results represent the final performance. Finally, I-testing was used in the testing datasets to evaluate the training model.

3. Results and discussion

3.1. Comparison with different feature combinations

Feature extraction affects the results of subsequent sequence classification. To obtain better performance, this paper studied seven different features, including Kmer, PC-PseDNC-General and ANF, EIIP, ENAC, NCP + NBP from the RF-PseU method [20]. These features were first used in the experiment separately, and then multiple features were selected for different combinations according to the test results to obtain better experimental results.

Table 1 lists the results of feature combination for the H_990 dataset using the TWSVM methods. The first six rows are the performance of single features, with Kmer, EIIP, ENAC and NCP + NBP returning the best results, which are roughly distributed in the range of 0.57–0.59. Since the test results of single features were lower than those of the RF-PseU predictor, we combined several features that perform well and used the mRMR method to select the best feature for model construction. For the H_990 dataset, the combined characteristics listed in Table 1 have four results, including Kmer + ENAC, Kmer + EIIP, Kmer + NCP + NBP and Kmer + PC-PseDNC-General + ANF + EIIP + ENAC + NCP + NBP. Using the TWSVM method, the result of combined features was usually approximately 1–3% higher than that of single features, with a maximum ACC value of 0.65 and the best feature combination being KMER + PC-PseDNC-Generel +ANF + EIIP + ENAC + NCP + NBP. Compared with the RF-PseU predictor, the accuracy was improved by 0.7%.

We chose this feature combination for the 10-CV of the training set H_990 and applied it to the I-testing of the testing set H_200.

Table 1. Results of feature combination for the H_990 dataset using the TWSVM method.

Feature Subset	TWSVM				
	ACC	MCC	SN	SP	AUC
Kmer	0.59	0.181	0.533	0.646	0.618
PC-PseDNC-General	0.534	0.07	0.434	0.635	0.543
ANF	0.526	0.053	0.568	0.485	0.518
EIIP	0.572	0.144	0.525	0.618	0.6
ENAC	0.587	0.178	0.472	0.7	0.59
NCP + NBP	0.584	0.172	0.53	0.639	0.582
Kmer + NCP + NBP	0.59	0.182	0.532	0.648	0.587
Kmer + ENAC	0.603	0.208	0.582	0.624	0.62
Kmer + EIIP	0.606	0.212	0.585	0.626	0.625
Kmer + PC-PseDNC-Generel + ANF + EIIP + ENAC + NCP + NBP	0.65	0.301	0.697	0.602	0.682

Table 2 lists the test results of different feature combinations on the S_628 dataset using the TWSVM method. Similar to the results in Table 1, the test results of Kmer, EIIP, ENAC and NCP + NBP were better. We combined those features that reported better performance. Except for Kmer + NCP + NBP, the results of other combined features were improved compared with the single feature results. Among them, the performance of the feature combination Kmer + PC – PseDNC – Generel + ANF + EIIP + ENAC + NCP + NBP was the best, with test results improved by approximately 6% compared to other combinations. We also used the independent test set S_200 to test this feature combination.

Table 2. Results of feature combination for the S_628 dataset using the TWSVM method.

Feature Subset	TWSVM				
	ACC	MCC	SN	SP	AUC
Kmer	0.627	0.259	0.625	0.63	0.67
PC-PseDNC-General	0.574	0.153	0.666	0.483	0.589
ANF	0.584	0.175	0.716	0.452	0.584
EIIP	0.614	0.238	0.473	0.755	0.632
ENAC	0.634	0.326	0.435	0.836	0.693
NCP + NBP	0.669	0.34	0.659	0.678	0.711
Kmer + NCP + NBP	0.664	0.33	0.653	0.675	0.683
Kmer + ENAC	0.653	0.308	0.631	0.675	0.683
Kmer + EIIP	0.631	0.263	0.628	0.634	0.662
Kmer + PC-PseDNC-Generel+ ANF + EIIP + ENAC + NCP + NBP	0.722	0.45	0.656	0.786	0.758

Table 3 shows the test results for the dataset M_994. The feature combination Kmer + PC-PseDNC-Generel + ANF + EIIP + ENAC + NCP + NBP had the best performance, and all test indicators were higher than the rest of the feature combinations. The combination was increased by approximately 5%, the MCC was increased by approximately 10%, and the AUC was also significantly improved.

Table 3. Results of feature combination for the M_944 dataset using the TWSVM method.

Feature Subset	TWSVM				
	ACC	MCC	SN	SP	AUC
Kmer	0.584	0.17	0.676	0.49	0.614
PC-PseDNC-General	0.541	0.084	0.517	0.566	0.539
ANF	0.553	0.113	0.71	0.396	0.527
EIIP	0.625	0.276	0.826	0.424	0.632
ENAC	0.664	0.329	0.667	0.661	0.7
NCP + NBP	0.662	0.326	0.636	0.688	0.703
Kmer + NCP + NBP	0.677	0.36	0.623	0.731	0.73
Kmer + ENAC	0.662	0.326	0.657	0.667	0.704
Kmer + EIIP	0.636	0.274	0.697	0.574	0.667
Kmer + PC-PseDNC-Generel + ANF + EIIP + ENAC + NCP + NBP	0.728	0.462	0.795	0.661	0.775

3.2. Comparison of different feature selection methods

In this study, we contrast the mRMR approach and the LightGBM method [64] since feature selection is a crucial component of model construction. Figure 2 shows their accuracy on the training datasets of the three species. The findings demonstrate that the performance of the mRMR technique is superior, which further enhances the classification accuracy of the model. The accuracy of the mRMR method on the three species is greater than that of the LightGBM approach.

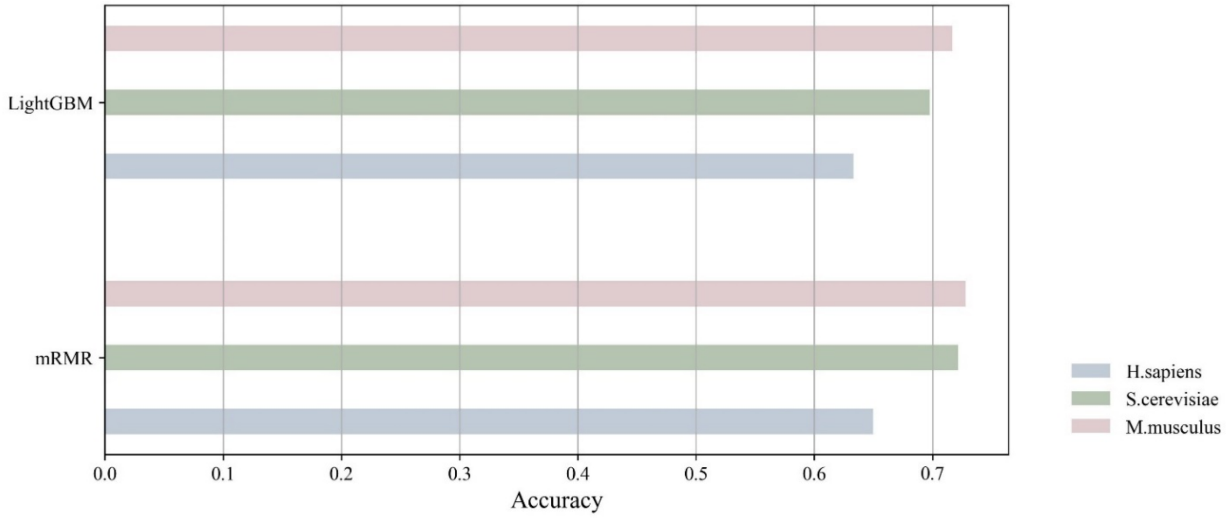
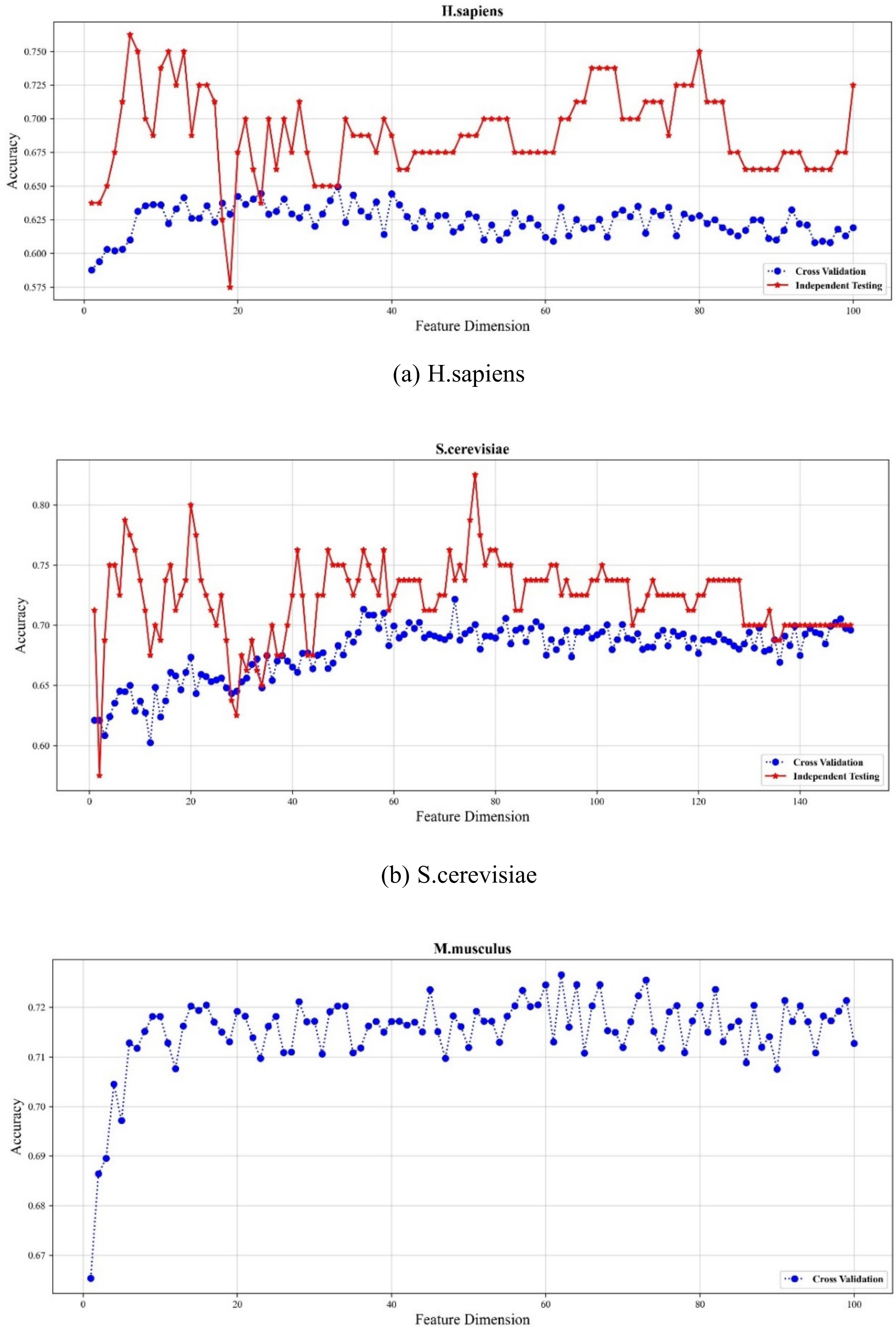


Figure 2. Contrasting various approaches to feature picking.

3.3. Optimization with different feature dimensions



(a) H.sapiens

(b) S.cerevisiae

(c) M.musculus

Figure 3. Accuracy of the twin support vector machine predictor varied with feature dimension for all three species: (a) H.sapiens; (b) S.cerevisiae; (c) M.musculus.

The accuracy of classification results may be successfully increased by feature selection. We initially utilized the mRMR technique to pick feature subsets with high correlation with class labels and low feature redundancy to obtain the optimum feature dimension. To further obtain the feature dimension with the best precision, the incremental feature selection approach was applied. After many experiments, we found that the accuracy of I-testing and 10-CV fluctuates as the number of characteristics rises and the highest accuracy mostly appeared within 100 or 150 dimensions, as illustrated in Figure 3. The accuracy of each species initially increases rapidly as the feature dimension increases, and then fluctuates continuously. For H.sapiens species, the highest 10-CV accuracy of 0.65 was obtained when the feature dimension reached 33, while the highest independent test accuracy of 0.763 was obtained at relatively low dimensions. The highest 10-CV accuracy of 0.722 and independent test accuracy of 0.825 for S. cerevisiae species were between 60–80 dimensions, obtained in 72 and 76 dimensions, respectively. M. musculus species only showed 10-CV results with the highest value at a feature dimension of 62.

Table 4. Results of feature subsets selection of H.sapiens species.

Feature	The original dimension	The dimension after feature selection
NCP	63	14
EIIP	21	21
NBP	84	15
ENAC	105	41
ANF	21	8
Kmer	64	1
PC-PseDNC-general	22	0

Table 5. Results of feature subsets selection of S.cerevisiae species.

Feature	The Original Dimension	The dimension after feature selection
NCP	93	22
EIIP	31	31
NBP	124	25
ENAC	155	53
ANF	31	19
Kmer	256	0
PC-PseDNC-general	18	0

Table 6. Results of feature subsets selection of M.musculus species.

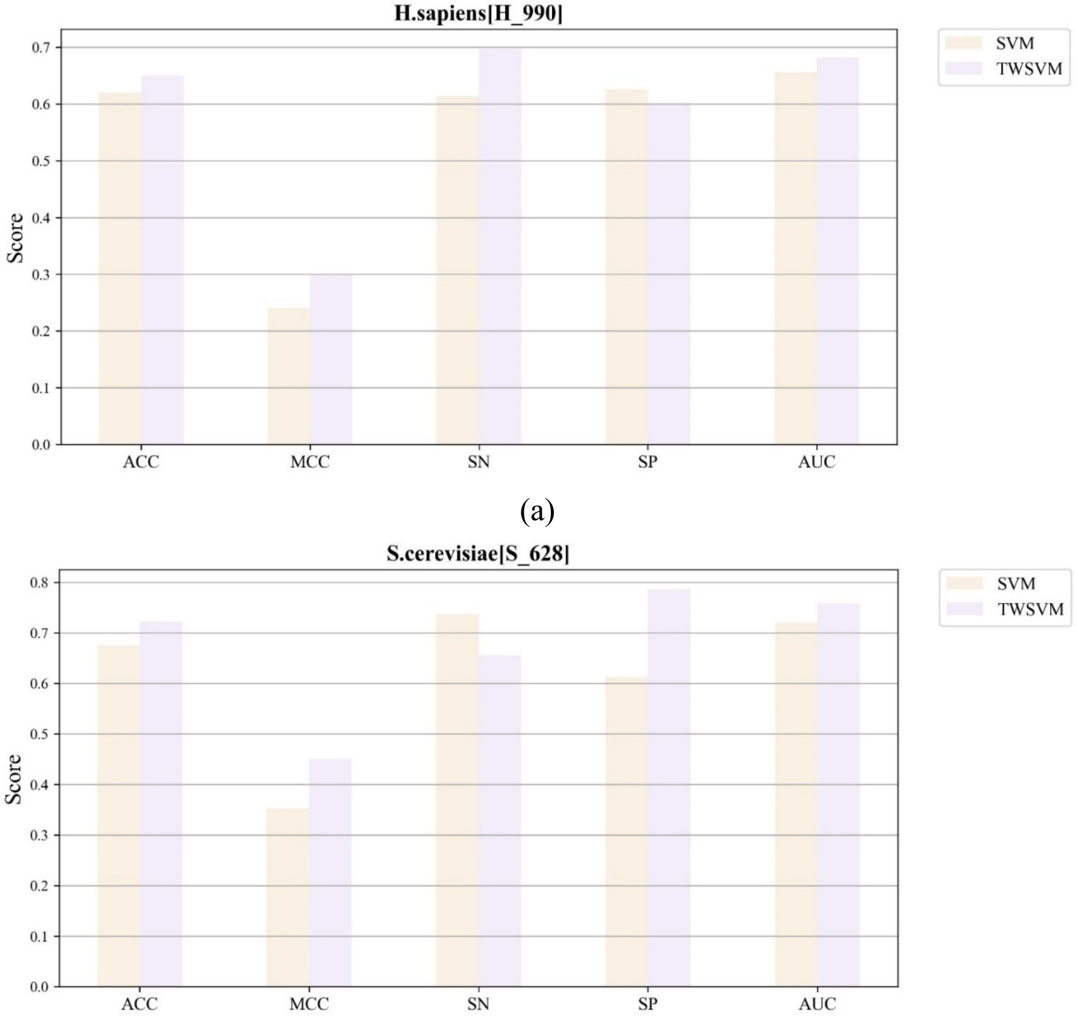
Feature	The original dimension	The dimension after feature selection
NCP	63	11
EIIP	21	21
NBP	84	17
ENAC	105	43
ANF	21	7
Kmer	64	1
PC-PseDNC-general	22	0

Tables 4–6 show the changes of feature dimensions after feature selection and the distribution of feature subsets after optimization for the three species. It can be found that ENAC and EIIP occupy a large number in the optimized feature subset of the three species, followed by NCP and NBP, ANF and Kmer occupy a small number, and there is no PC-PseDNC-General in the optimized

feature subset. It indicates that each feature has different contributions in the model, and ENAC and EIIP play an important role in the model.

3.4. Comparison with SVM classifier

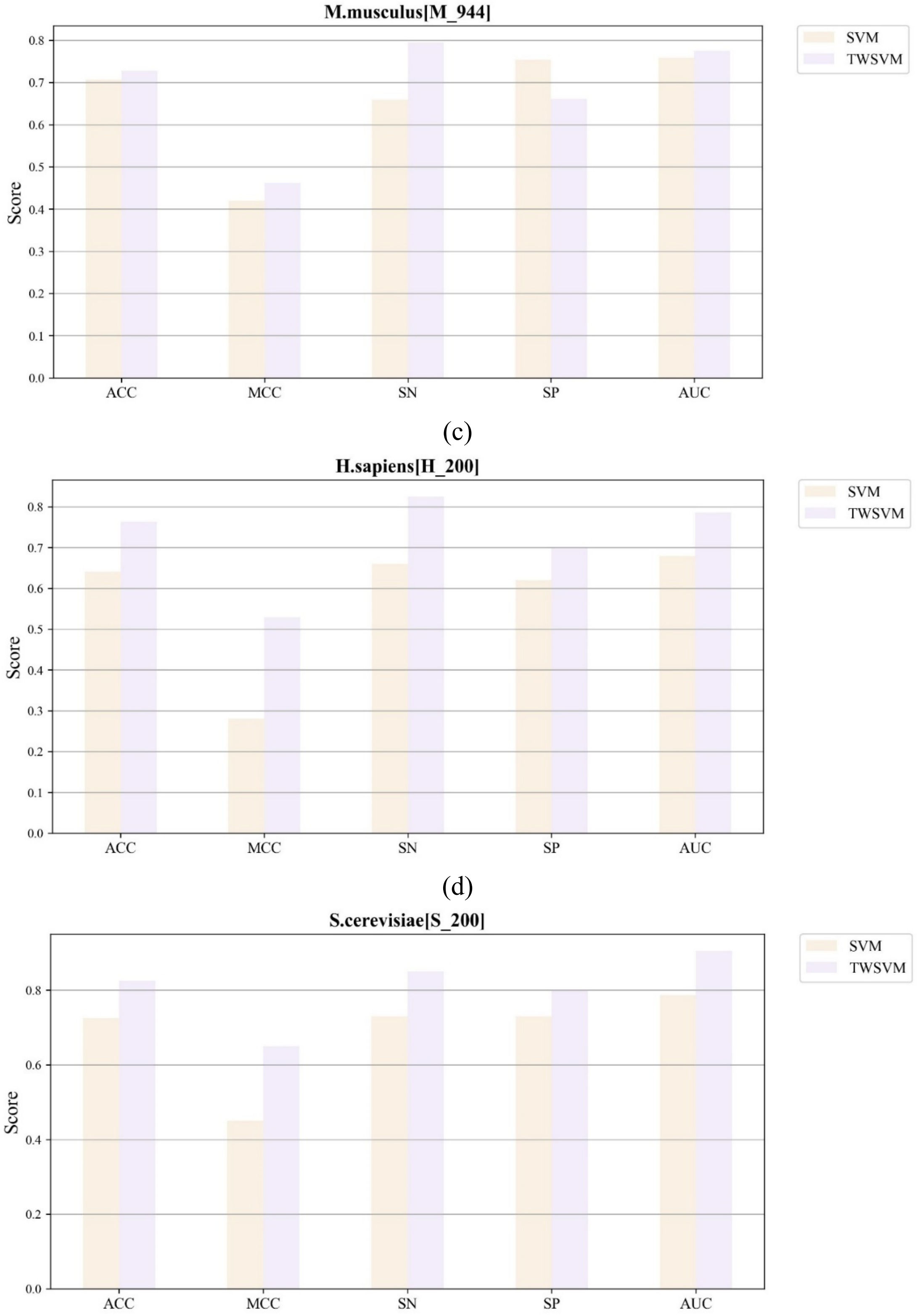
Since many previous researchers built У sites recognition models based on support vector machines, we employed SVM [65] as a classifier in the same feature space to compare the performance of TWSVM with that of SVM. Figure 4 displays how it performed. The ACC, MCC, and AUC based on the TWSVM model were found to be larger than those based on the SVM model for the 10-CV results of the three species, while the independent test results may have more clearly indicated the difference between the two. All of the evaluation metrics outperformed the SVM model. As a result, we concluded that the TWSVM model performs much better than the SVM model, suggesting that it may be better suited for identifying У sites in RNA sequences.



(a)

(b)

Continued on next page



(c)

(d)

(e)

Figure 4. Comparison of 10-CV scores and I-testing scores for SVM and TWSVM. (a) is for the training datasets of H.sapiens; (b) is for the training datasets of S.cerevisiae; (c) is for the training datasets of f M.musculus; (d) is for the testing datasets of H.sapiens; (e) is for the testing datasets of S.cerevisiae.

3.5. Comparison with previous methods

Table 7. Comparison of cross-validation scores of current advanced У sites predictors and iPseU-TWSVM.

Species	Classifier	Cross-validation				
		ACC	MCC	SN	SP	AUC
H.sapiens	iRNA-PseU	0.604	0.21	0.61	0.598	0.64
	PseUI	0.642	0.28	0.649	0.636	0.68
	iRNA-CNN	0.667	0.34	0.65	0.688	/
	XG-PseU	0.661	0.32	0.635	0.687	0.7
	RF-PseU	0.643	0.29	0.661	0.626	0.7
	iPseU-TWSVM	0.65	0.301	0.697	0.602	0.682
S.cerevisiae	iRNA-PseU	0.645	0.29	0.647	0.643	0.81
	PseUI	0.641	0.3	0.647	0.675	0.69
	iRNA-CNN	0.682	0.37	0.664	0.705	/
	XG-PseU	0.682	0.37	0.668	0.695	0.77
	RF-PseU	0.748	0.49	0.772	0.724	0.81
	iPseU-TWSVM	0.722	0.45	0.656	0.786	0.758
M.musculus	iRNA-PseU	0.691	0.38	0.733	0.648	0.75
	PseUI	0.704	0.41	0.799	0.703	0.71
	iRNA-CNN	0.718	0.44	0.748	0.691	/
	XG-PseU	0.72	0.45	0.765	0.676	0.74
	RF-PseU	0.748	0.5	0.731	0.765	0.796
	iPseU-TWSVM	0.728	0.462	0.795	0.661	0.775

Table 8. Comparison of I-testing scores of current advanced У sites predictors and iPseU-TWSVM.

Species	Classifier	Independent testing				
		ACC	MCC	SN	SP	AUC
H.sapiens	iRNA-PseU	0.65	0.3	0.6	0.7	/
	PseUI	0.655	0.31	0.63	0.7	/
	iRNA-CNN	0.69	0.4	0.777	0.68	/
	XG-PseU	0.675	/	/	0.608	/
	RF-PseU	0.75	0.5	0.78	0.72	0.8
	iPseU-TWSVM	0.763	0.529	0.825	0.7	0.786
S.cerevisiae	iRNA-PseU	0.6	0.2	0.63	0.57	/
	PseUI	0.685	0.37	0.65	0.72	/
	iRNA-CNN	0.735	0.47	0.688	0.778	/
	XG-PseU	0.71	/	/	/	/
	RF-PseU	0.77	0.54	0.75	0.79	0.838
	iPseU-TWSVM	0.825	0.65	0.85	0.8	0.905

Table 9. Average accuracy comparison between iPseU-TWSVM and current advanced У sites predictors.

Scores type	iPseU-TWSVM	RF-PseU	XG-PseU	iRNA-CNN	PseUI	iRNA-PseU
Cross-validation	0.7	0.713	0.687	0.689	0.662	0.647
Independent testing	0.794	0.76	0.693	0.713	0.7	0.625

The effectiveness of iPseU-TWSVM was also evaluated in comparison to other advanced predictors, such as iRNA-PseU [16], PseUI [17], iPseU-CNN [18], XG-PseU [19] and RF-PseU [20]. The 10-CV and I-testing results of the advanced У site predictors using iPseU-TWSVM are contrasted in Tables 7 and 8, respectively. The 10-CV results reveal that the accuracy of iPseU-TWSVM on H.sapiens is 1.7% less accurate than that of the best predictor iPseU-CNN on this species, and the accuracy on S.cerevisiae and M.musculus is 0.722 and 0.728, respectively, which is 2.6 and 2.0% less accurate than that of the best predictor RF-PseU. Although iPseU-TWSVM does not perform optimally on the training set, iPseU-TWSVM has higher accuracy than other predictors on all species in terms of I-testing. H.sapiens and S.cerevisiae are 1.3 and 5.5% more accurate in I-testing than the best predictor RF-PseU, with corresponding accuracy values of 0.763 and 0.825, respectively. We also calculated the average accuracy of several species so that we could compare the predictors' performance in depth. As shown in Table 9, the 10-CV accuracy of iPseU-TWSVM is 1.3% less than that of RF-PseU. In terms of I-testing, iPseU-TWSVM is significantly improved by 3.4% compared with RF-PseU. iPseU-TWSVM performs much better overall than the other predictors. The findings demonstrate that iPseU-TWSVM, a very practical technique, has greater generalization performance and is more appropriate for recognizing У sites in RNA sequences.

4. Conclusions

This work proposes the use of a novel model called iPseU-TWSVM to identify RNA У sites across various species. We have used an efficient feature selection method to obtain the best feature subset and selected TWSVM as the classifier to increase recognition accuracy. Finally, we compared advanced predictors and found that iPseU-TWSVM significantly improved the independent test accuracy by 3.4%, while the accuracy of cross validation was lower by 1.3%. Through comprehensive analysis, it was concluded that the relatively poor performance of the training datasets was due to the following two reasons. One is that the features used by the best predictor are different, and the other is that the classifier of the model is different. The above results indicate that iPseU-TWSVM had better generalization performance and could more accurately identify У sites from RNA sequences. It is anticipated that iPseU-TWSVM will be effective in identifying RNA У sites.

The contribution of this work has the following three aspects: (i) the model uses TWSVM as a classifier, which improves the accuracy of the model and improves the training speed; (ii) the model has good generalization performance and can be applied to the prediction of other sites in the sequence; (iii) further accurate identification of У sites in the sequence lays the foundation for disease control and related drug development. At the same time, this work also has the following shortcomings: (i) in the feature selection part, only two algorithms are compared, and subsequent research can try other algorithms to further improve the feature subset; (ii) the model uses TWSVM as a classifier. In the original problem of TWSVM, only empirical risk is minimized, but structural risk is not minimized. Moreover, the algorithm can only obtain approximate solutions. Subsequent research can consider improving TWSVM or try other classification algorithms as the classifier of the model to improve the prediction performance. Future work will study emerging methods [66–81] to further improve the accuracy of the model.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China (NSFC 62172076, 62072385), and the Municipal Government of Quzhou under Grant Number 2020D003 and 2021D004.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. P. Boccaletto, M. A. Machnicka, E. Purta, P. Piatkowski, B. Baginski, T. K. Wirecki, et al., MODOMICS: A database of RNA modification pathways. 2017 update, *Nucleic Acids Res.*, **46** (2018), D303–D307. https://doi.org/10.1093/nar/gkx1030
2. J. Song, C. Yi, Chemical modifications to RNA: A new layer of gene expression regulation, *ACS Chem. Biol.*, **12** (2017), 316–325. https://doi.org/10.1021/acschembio.6b00960
3. F. F. Davis, F. W. Allen, Ribonucleic acids from yeast which contain a fifth nucleotide, *J. Biol. Chem.*, **227** (1957), 907–915. https://doi.org/10.1016/s0021-9258(18)70770-9
4. W. E. Cohn, Pseudouridine, a carbon-carbon linked ribonucleoside in ribonucleic acids: Isolation, structure, and chemical characteristics, *J. Biol. Chem.*, **235** (1960), 1488–1498. https://doi.org/10.1002/jbmte.390020410
5. T. Fujiwara, H. Harigae, Molecular pathophysiology and genetic mutations in congenital sideroblastic anemia, *Free Radical Biol. Med.*, **133** (2019), 179–185. https://doi.org/10.1016/j.freeradbiomed.2018.08.008
6. N. Guzzi, M. Ciesla, P. C. T. Ngoc, S. Lang, S. Arora, M. Dimitriou, et al., Pseudouridylation of tRNA-derived fragments steers translational control in stem cells, *Cell*, **173** (2018), 1204–1216. https://doi.org/10.1016/j.cell.2018.03.008
7. J. Karijolich, Y. T. Yu, Converting nonsense codons into sense codons by targeted pseudouridylation, *Nature*, **474** (2011), 395–398. https://doi.org/10.1038/nature10165
8. R. W. Holley, G. A. Everett, J. T. Madison, A. Zamir, Nucleotide sequences in the yeast alanine transfer ribonucleic acid, *J. Biol. Chem.*, **240** (1965), 2122–2128. https://doi.org/10.1016/s0021-9258(18)97435-1
9. C. Y. Gradeen, D. M.Billay, S. C. Chan, Analysis of bumetanide in human urine by high-performance liquid chromatography with fluorescence detection and gas chromatographyl/mass spectrometry, *J. Anal. Toxicol.*, **14** (1990), 123–126. https://doi.org/10.1093/jat/14.2.123
10. A. Basak, C. C. Query, A pseudouridine residue in the spliceosome core is part of the filamentous growth program in yeast, *Cell Rep.*, **8** (2014), 966–973. https://doi.org/10.1016/j.celrep.2014.07.004
11. T. M. Carlile, M. F. Rojas-Duran, B. Zinshteyn, H. Shin, K. M. Bartoli, W. V. Gilbert, Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells, *Nature*, **515** (2014), 143–146. https://doi.org/10.1038/nature13802

12. S. Schwartz, D. A. Bernstein, M. R. Mumbach, M. Jovanovic, R. H. Herbst, B. X. Leon-Ricardo, et al., Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA, *Cell*, **159** (2014), 148–162. https://doi.org/10.1016/j.cell.2014.08.028
13. X. Li, P. Zhu, S. Ma, J. Song, J. Bai, F. Sun, et al., Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome, *Nat. Chem. Biol.*, **11** (2015), 592–597. https://doi.org/10.1038/nchembio.1836
14. B. Panwar, G. P. Raghava, Prediction of uridine modifications in tRNA sequences, *BMC Bioinf.*, **15** (2014), 326. https://doi.org/10.1186/1471-2105-15-326
15. Y. H. Li, G. Zhang, Q. Cui, PPUS: A web server to predict PUS-specific pseudouridine sites, *Bioinformatics*, **31** (2015), 3362–3364. https://doi.org/10.1093/bioinformatics/btv366
16. W. Chen, H. Tang, J. Ye, H. Lin, K. C. Chou, iRNA-PseU: Identifying RNA pseudouridine sites, *Mol. Ther. Nucleic Acids*, **5** (2016), e332. https://doi.org/10.1038/mtna.2016.37
17. J. He, T. Fang, Z. Zhang, B. Huang, X. Zhu, Y. Xiong, PseUI: Pseudouridine sites identification based on RNA sequence information, *BMC Bioinf.*, **19** (2018), 306. https://doi.org/10.1186/s12859-018-2321-0
18. M. Tahir, H. Tayara, K. T. Chong, iPseU-CNN: Identifying RNA pseudouridine sites using convolutional neural networks, *Mol. Ther. Nucleic Acids*, **16** (2019), 463–470. https://doi.org/10.1016/j.omtn.2019.03.010
19. K. Liu, W. Chen, H. Lin, XG-PseU: An eXtreme Gradient Boosting based method for identifying pseudouridine sites, *Mol. Genet. Genomics*, **295** (2020), 13–21. https://doi.org/10.1007/s00438-019-01600-9
20. Z. Lv, J. Zhang, H. Ding, Q. Zou, RF-PseU: A random forest predictor for RNA pseudouridine sites, *Front. Bioeng. Biotechnol.*, **8** (2020), 134. https://doi.org/10.3389/fbioe.2020.00134
21. S. M. Khan, F. He, D. Wang, Y. Chen, D. Xu, Mu-pseudeep: A deep learning method for prediction of pseudouridine sites, *Comput. Struct. Biotechnol. J.*, **18** (2020), 1877–1883. https://doi.org/10.1016/j.csbj.2020.07.010
22. F. Li, X. Guo, P. Jin, J. Chen, D. Xiang, J. Song, Porpoise: A new approach for accurate prediction of RNA pseudouridine sites, *Briefings Bioinf.*, **22** (2021), bbab245. https://doi.org/10.1093/bib/bbab245
23. Y. Q. Qian, H. Meng, W. Z. Lu, Z. J. Liao, Y. J. Ding, H. J. Wu, Identification of DNA-binding proteins via hypergraph based laplacian support vector machine, *Curr. Bioinf.*, **17** (2022), 108–117. https://doi.org/10.2174/1574893616666210806091922
24. S. Naseer, W. Hussain, Y. D. Khan, N. Rasool, NPalmitoylDeep-PseAAC: A predictor of N-palmitoylation sites in proteins using deep representations of proteins and PseAAC via modified 5-Steps rule, *Curr. Bioinf.*, **16** (2021), 294–305. https://doi.org/10.2174/1574893615999200605142828
25. S. W. Sun, L. Xu, Q. Zou, G. H. Wang, BP4RNAseq: A babysitter package for retrospective and newly generated RNA-seq data analyses using both alignment-based and alignment-free quantification methods, *Bioinformatics*, **37** (2021), 1319–1321. https://doi.org/10.1093/bioinformatics/btaa832
26. L. Zhang, Z. Huang, L. Kong, CSBPI_Site: Multi-information sources of features to RNA binding sites prediction, *Curr. Bioinf.*, **16** (2021), 691–699. https://doi.org/10.2174/1574893615666210108093950

27. Z. Zhang, F. Cui, W. Su, L. Dou, A. Xu, C. Cao, Q. Zou, webSCST: An interactive web application for single-cell RNA-sequencing data and spatial transcriptomic data integration, *Bioinformatics*, **38** (2022), 3488–3489. https://doi.org/ 10.1093/bioinformatics/btac350
28. X. Wang, S. Wang, H. Fu, X. Ruan, X. Tang, DeepFusion-RBP: Using Deep Learning to Fuse Multiple Features to Identify RNA-binding Protein Sequences, *Curr. Bioinf.*, **16** (2021), 1089–1100. https://doi.org/ 10.2174/1574893616666210618145121
29. W. Chen, H. Ding, X. Zhou, H. Lin, K. C. Chou, iRNA(m6A)-PseDNC: Identifying N(6)-methyladenosine sites using pseudo dinucleotide composition, *Anal. Biochem.*, **561** (2018), 59–65. https://doi.org/10.1016/j.ab.2018.09.002
30. L. Wei, M. Liao, Y. Gao, R. Ji, Z. He, Q. Zou, Improved and promising identification of human microRNAs by incorporating a high-quality negative set, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **11** (2014), 192–201. https://doi.org/10.1109/TCBB.2013.146
31. B. Liu, X. Gao, H. Zhang, BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches, *Nucleic Acids Res.*, **47** (2019), e127. https://doi.org/10.1093/nar/gkz740
32. W. Chen, X. Zhang, J. Brooker, H. Lin, L. Zhang, K. C. Chou, PseKNC-General: A cross-platform package for generating various modes of pseudo nucleotide compositions, *Bioinformatics*, **31** (2015), 119–120. https://doi.org/10.1093/bioinformatics/btu602
33. H. Yang, H. Lv, H. Ding, W. Chen, H. Lin, iRNA-2OM: A sequence-based predictor for identifying 2'-O-Methylation sites in Homo sapiens, *J. Comput. Biol.*, **25** (2018), 1266–1277.https://doi.org/10.1089/cmb.2018.0004
34. B. Liu, BioSeq-Analysis: A platform for DNA, RNA and protein sequence analysis based on machine learning approaches, *Briefings Bioinf.*, **20** (2019), 1280–1294. https://doi.org/10.1093/bib/bbx165
35. Y. Hu, T. Zhao, N. Zhang, Y. Zhang, L. Cheng, A review of recent advances and research on drug target identification methods, *Curr. Drug Metab.*, **20** (2019), 209–216. https://doi.org/10.2174/1389200219666180925091851
36. A. S. Nair, S. P. Sreenadhan, A coding measure scheme employing electron-ion interaction pseudopotential (EIIP), *Bioinformation*, **1** (2006), 197–202.
37. H. Peng, F. Long, C. Ding, Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy, *IEEE Trans. Pattern Anal. Mach. Intell.*, **27** (2005), 1226–1238. https://doi.org/10.1109/TPAMI.2005.159
38. Y. Tian, Z. Qi, Review on: Twin support vector machines, *Ann. Data Sci.*, **1** (2014), 253–277. https://doi.org/10.1007/s40745-014-0018-4
39. L. Cheng, J. Sun, W. Xu, L. Dong, Y. Hu, M. Zhou, OAHG: An integrated resource for annotating human genes with multi-level ontologies, *Sci. Rep.*, **6** (2016), 1–9. https://doi.org/10.1038/srep34820
40. L. Y. Wei, S. X. Wan, J. S. Guo, K. K. L. Wong, A novel hierarchical selective ensemble classifier with bioinformatics application, *Artif. Intell. Med.*, **83** (2017), 82–90. https://doi.org/10.1016/j.artmed.2017.02.005
41. B. Liu, C. C. Li, K. Yan, DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks, *Briefings Bioinf.*, **21** (2020), 1733–1741. https://doi.org/10.1093/bib/bbz098

42. D. Mrozek, P. Gosk, B. Małysiak-Mrozek, Scaling Ab initio predictions of 3D protein structures in microsoft azure cloud, *J. Grid Comput.*, **13** (2015), 561–585. https://doi.org/10.1007/s10723-015-9353-8
43. R. Cao, J. Cheng, Protein single-model quality assessment by feature-based probability density functions, *Sci. Rep.*, **6** (2016), 23990. https://doi.org/10.1038/srep23990
44. W. Chen, H. Yang, P. Feng, H. Ding, H. Lin, iDNA4mC: Identifying DNA N-4-methylcytosine sites based on nucleotide chemical properties, *Bioinformatics*, **33** (2017), 3518–3523. https://doi.org/ 10.1093/bioinformatics/btx479
45. W. Chen, H. Lv, F. Nie, H. Lin, i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome, *Bioinformatics*, **35** (2019), 2796–2800. https://doi.org/10.1093/bioinformatics/btz015
46. G. Pan, J. Tang, F. Guo, Analysis of co-associated transcription factors via ordered adjacency differences on motif distribution, *Sci. Rep.*, **7** (2017), 43597. https://doi.org/10.1038/srep43597
47. W. He, C. Jia, Q. Zou, 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction, *Bioinformatics*, **35** (2019), 593–601. https://doi.org/10.1093/bioinformatics/bty668
48. L. Jiang, Y. Ding, J. Tang, F. Guo, MDA-SKF: Similarity kernel fusion for accurately discovering miRNA-Disease association, *Front. Genet.*, **9** (2018), 618. https://doi.org/10.3389/fgene.2018.00618
49. Y. Xiong, Q. Wang, J. Yang, X. Zhu, D. Q. Wei, PredT4SE-Stack: Prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method, *Front. Microbiol.*, **9** (2018), 2571. https://doi.org/10.3389/fmicb.2018.02571
50. L. Yu, J. Zhao, L. Gao, Predicting potential drugs for breast cancer based on miRNA and tissue specificity, *Int. J. Biol. Sci.*, **14** (2018), 971–982. https://doi.org/10.7150/ijbs.23350
51. M. Zhang, Y. Xu, L. Li, Z. Liu, X. Yang, D. J. Yu, Accurate RNA 5-methylcytosine site prediction based on heuristic physical-chemical properties reduction and classifier ensemble, *Anal. Biochem.*, **550** (2018), 41–48. https://doi.org/10.1016/j.ab.2018.03.027
52. Y. Ding, J. Tang, F. Guo, Identification of drug-side effect association via multiple information integration with centered kernel alignment, *Neurocomputing*, **325** (2019), 211–224. https://doi.org/10.1016/j.neucom.2018.10.028
53. B. Manavalan, S. Basith, T. H. Shin, L. Wei, G. Lee, Meta-4mCpred: A sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation, *Mol. Ther. Nucleic Acids*, **16** (2019), 733–744. https://doi.org/10.1016/j.omtn.2019.04.019
54. P. Feng, H. Yang, H. Ding, H. Lin, W. Chen, K. C. Chou, iDNA6mA-PseKNC: Identifying DNA N(6)-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC, *Genomics*, **111** (2019), 96–102. https://doi.org/10.1016/j.ygeno.2018.01.005
55. L. Kong, L. Zhang, i6mA-DNCP: Computational identification of DNA N(6)-methyladenine sites in the rice genome using optimized dinucleotide-based features, *Genes*, **10** (2019), 828. https://doi.org/10.3390/genes10100828
56. C. C. Li, B. Liu, MotifCNN-fold: protein fold recognition based on fold-specific features extracted by motif-based convolutional neural networks, *Briefings Bioinf.*, **21** (2020), 2133–2141. https://doi.org/10.1093/bib/bbz133
57. X. Shan, X. Wang, C. D. Li, Y. Chu, Y. Zhang, Y. Xiong, et al., Prediction of CYP450 enzyme-substrate selectivity based on the network-based label space division method, *J. Chem. Inf. Model.*, **59** (2019), 4577–4586. https://doi.org/10.1021/acs.jcim.9b00749

58. X. Wang, X. Zhu, M. Ye, Y. Wang, C. D. Li, Y. Xiong, et al., STS-NLSP: A network-based label space partition method for predicting the specificity of membrane transporter substrates using a hybrid feature of structural and semantic similarity, *Front. Bioeng. Biotechnol.*, **7** (2019), 306. https://doi.org/10.3389/fbioe.2019.00306
59. L. Wei, S. Luan, L. A. E. Nagai, R. Su, Q. Zou, Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species, *Bioinformatics*, **35** (2019), 1326–1333. https://doi.org/10.1093/bioinformatics/bty824
60. L. Wei, R. Su, S. Luan, Z. Liao, B. Manavalan, Q. Zou, et al., Iterative feature representations improve N4-methylcytosine site prediction, *Bioinformatics*, **35** (2019), 4930–4937. https://doi.org/10.1093/bioinformatics/btz408
61. L. Xu, G. Liang, C. Liao, G. D. Chen, C. C. Chang, k-Skip-n-Gram-RF: A random forest based method for Alzheimer's disease protein identification, *Front. Genet.*, **10** (2019), 33. https://doi.org/10.3389/fgene.2019.00033
62. L. H. Roland, C. T. Wannige, A deep learning model for predicting DNA N6-methyladenine (6mA) sites in eukaryotes, *IEEE Access*, **8** (2020), 175535–175545. https://doi.org/10.1109/access.2020.3025990
63. Z. Chen, P. Zhao, F. Li, T. T. Marquez-Lago, A. Leier, J. Revote, 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 Bioinf.*, **21** (2020), 1047–1057. https://doi.org/10.1093/bib/bbz041
64. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, et al., LightGBM: A highly efficient gradient boosting decision tree, in *Advances in Neural Information Processing Systems 30 (NIP 2017)*, **30** (2017), 1–9.
65. C. Cortes, V. Vapnik, Support-vector networks, *Mach. Learn.*, **20** (1995), 273–297. https://doi.org/ 10.1007/BF00994018
66. H. Zhou, H. Wang, Y. Ding, J. Tang, Multivariate information fusion for identifying antifungal peptides with Hilbert-Schmidt independence criterion, *Curr. Bioinf.*, **17** (2022), 89–100. https://doi.org/10.2174/1574893616666210727161003
67. C. Wang, Y. Ju, Q. Zou, C. Lin, DeepAc4C: A convolutional neural network model with hybrid features composed of physico-chemical patterns and distributed representation information for identification of N4 acetylcytidine in mRNA, *Bioinformatics*, **38** (2022), 52–57. https://doi.org/10.1093/bioinformatics/btab611
68. X. Guo, W. Zhou, B. Shi, X. Wang, A. Du, Y. Ding, et al., An efficient multiple kernel support vector regression model for assessing dry weight of hemodialysis patients, *Curr. Bioinf.*, **16** (2021), 284–293. https://doi.org/ 10.2174/1574893615999200614172536
69. E. Scornet, Random forests and kernel methods, *IEEE Trans. Inf. Theory*, **62** (2016), 1485–1500. https://doi.org/10.1109/tit.2016.2514489
70. S. Zhao, Y. Ju, X. Ye, J. Zhang, S. Han, Bioluminescent proteins prediction with voting strategy, *Curr. Bioinf.*, **16** (2021), 240–251. https://doi.org/ 10.2174/1574893615999200601122328
71. M. Niu, Q. Zou, C. Wang, GMNN2CD: Identification of circRNA–disease associations based on variational inference and graph Markov neural networks, *Bioinformatics*, **38** (2022), 2246–2253. https://doi.org/ 10.1093/bioinformatics/btac079

72. A. K. Sharma, R. Srivastava, Protein secondary structure prediction using character Bi-gram embedding and Bi-LSTM, *Curr. Bioinf.*, **16** (2021), 333–338. https://doi.org/10.2174/1574893615999200601122840
73. C. Wang, C. Han, Q. Zhao, X. Chen, Circular RNAs and complex diseases: from experimental results to computational models, *Briefings Bioinf.*, **22** (2021), bbab286. https://doi.org/10.1093/bib/bbac357
74. A. Alim, A. Rafay, I. Naseem, PoGB-pred: Prediction of antifreeze proteins sequences using amino acid composition with feature selection followed by a sequential-based ensemble approach, *Curr. Bioinf.*, **16** (2021), 446–456. https://doi.org/10.2174/1574893615999200707141926
75. Y. Tian, X. Ju, Z. Qi, Y. Shi, Improved twin support vector machine, *Sci. China Math.*, **57** (2013), 417–432. https://doi.org/10.1007/s11425-013-4718-6
76. Y. Zou, H. Wu, X. Guo, L. Peng, Y. Ding, J. Tang, et al., MK-FSVM-SVDD: A multiple kernel-based fuzzy SVM model for predicting DNA-binding proteins via support vector data description, *Curr. Bioinf.*, **16** (2021), 274–283. https://doi.org/10.2174/1574893615999200607173829
77. Q. Tang, F. Nie, Q. Zhao, W. Chen, A merged molecular representation deep learning method for blood–brain barrier permeability prediction, *Briefings Bioinf.*, **2022** (2022), bbac357. https://doi.org/10.1093/bib/bbac357
78. F. Li, X. Guo, D. Xiang, M. E. Pitt, A. Bainomugisa, L. J. M. Coin, Computational analysis and prediction of PE_PGRS proteins using machine learning, *Comput. Struct. Biotechnol. J.*, **20** (2022), 662–674. https://doi.org/ 10.1016/j.csbj.2022.01.0192001-0370
79. F. Sun, J. Sun, Q. Zhao, A deep learning method for predicting metabolite-disease associations via graph neural network, *Briefings Bioinf.*, **23** (2022), bbac266. https://doi.org/10.1093/bib/bbac266
80. F. Li, S. Dong, A. Leier, M. Han, X. Guo, J. Xu, et al., Positive-unlabeled learning in bioinformatics and computational biology: A brief review, *Briefings Bioinf.*, **23** (2021), bbab461. https://doi.org/10.1093/bib/bbab461
81. W. Liu, Y. Jiang, L. Peng, X. Sun, W. Gan, Q. Zhao, et al., Inferring gene regulatory networks using the improved Markov blanket discovery algorithm, *Interdiscip. Sci. Comput. Life Sci.*, **14** (2022), 168–181. https://doi.org/10.1007/s12539-021-00478-9