Mathematical Biosciences and Engineering

MBE, 18 (4): 3348–3363.
DOI: 10.3934/mbe.2021167
Received: 11 February 2021
Accepted: 25 March 2021
Published: 15 April 2021

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

Research article

Computational identification of N4-methylcytosine sites in the mouse genome with machine-learning method

Hasan Zulfiqar[1], Rida Sarwar Khan[1], Farwa Hassan[1], Kyle Hippe[2], Cassandra Hunt[2], Hui Ding[1,*], Xiao-Ming Song[1,3,*] and Renzhi Cao[2,*]

[1] School of Life Science and Technology and Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China.
[2] Department of Computer Science, Pacific Lutheran University, Tacoma 98447, USA.
[3] School of Life Sciences, North China University of Science and Technology, Tangshan, Hebei 063210, China.

* **Correspondence:** Email: hding@uestc.edu.cn, songxiaoming116@163.com, caora@plu.edu.

Abstract: N4-methylcytosine (4mC) is a kind of DNA modification which could regulate multiple biological processes. Correctly identifying 4mC sites in genomic sequences can provide precise knowledge about their genetic roles. This study aimed to develop an ensemble model to predict 4mC sites in the mouse genome. In the proposed model, DNA sequences were encoded by *k-mer*, enhanced nucleic acid composition and composition of *k*-spaced nucleic acid pairs. Subsequently, these features were optimized by using minimum redundancy maximum relevance (mRMR) with incremental feature selection (IFS) and five-fold cross-validation. The obtained optimal features were inputted into random forest classifier for discriminating 4mC from non-4mC sites in mouse. On the independent dataset, our model could yield the overall accuracy of 85.41%, which was approximately 3.8% - 6.3% higher than the two existing models, i4mC-Mouse and 4mCpred-EL respectively. The data and source code of the model can be freely download from https://github.com/linDing-groups/model_4mc.

Keywords: DNA modification; feature extraction; feature selection; N4-methylcytosine; random forest

1. Introduction

DNA modifications, such as demethylation and methylation, play important roles in the regulation of gene expression [1]. At the site of (*5′*-C-phosphate-G-*3′*), the methylation of cytosine is an important

epigenetic trait, which is closely related to cell proliferation and chromosomal stability protection [2,3]. 5-methylcytosine (5mC), 4-methylcytosine (4mC), and 3-methylcytosine are the most common methylations of cytosine in eukaryotic and prokaryotic genomes [4,5]. 5mC is the common kind of methylation of cytosine and, relates to many cancerous and neural diseases [6,7]. 4mC is also an effective modification that guards its own genetic information from deterioration through restriction enzymes [8–10]. Accurate recognition of 4mC could provide key clues for understanding its regulation roles. Currently, several experimental methodologies, including mass spectrometry, reduced-representation bisulfite sequencing, and single-molecule real-time sequencing, have been developed to identify 4mC sites [11–13]. Although these methodologies are helpful in the identification of 4mC sites, they are highly expensive when implemented on extensively large sequencing data. Thus, a bioinformatics tool to identify 4mC sites is urgently needed. At present, some computational methods have been presented to identify 4mC sites. In 2017, an innovative prediction model based on the confirmed 4mC dataset was constructed to predict 4mC sites in several species [14]. Afterwards, an iterative feature representative algorithm was designed based on the benchmark dataset of Chen et al. [15], which helped to learn and train the features from numerous progressive models to predict 4mC sites. iEC4mC-SVM [16] was developed to predict the 4mC in the Escherichia coli by using light gradient boosting machine feature selection technology. DNA4mc-LIP [17], a linear integration tool, was developed by combining existing prediction methods to identify 4-methyl cytosine sites in multiple species. Then, Meta-4mCpred [18] was developed to predict 4mC sites in the genomes of six species. However, to date, only two predictors, i4mC-Mouse and 4mCpred-EL are available for recognizing 4mC sites in mice [19,20]. These two methods employed various features and machine learning algorithms on the sequence data of mice derived from the Meth-SMRT database [21]. Although both i4mC-Mouse and 4mCpred-EL can produce good outcomes, there is still room for further improvement by extracting more feature information.

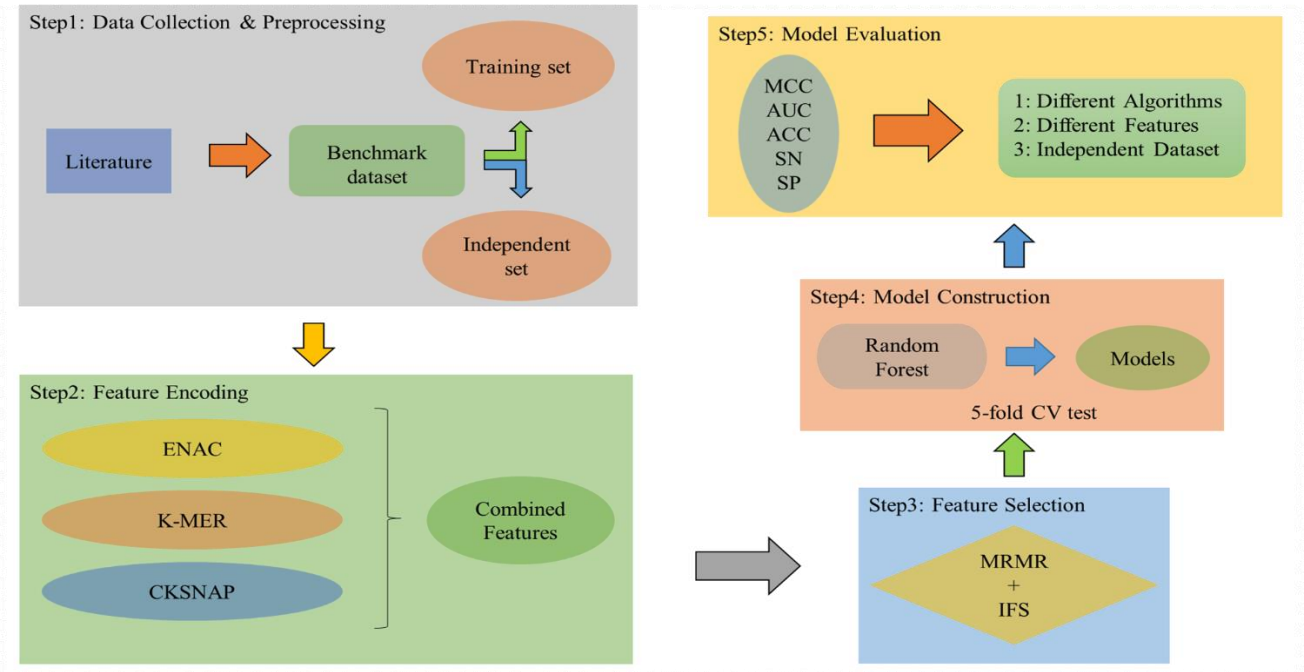

Figure 1. The workflow of the prediction of 4mC sites in mouse genome.

To address the aforementioned issues, an ensemble model was established to predict 4mC sites in mice. Figure 1 shows the workflow of the proposed model. First, three types of feature descriptors, k-mer, enhanced nucleic acid composition and composition of k-spaced nucleic acid pairs, were used as features to input into a random forest classifier [22] for identifying 4mC sites. After this, the mRMR [23] with IFS [24,25] technique was utilized to get optimal feature vectors. Finally, the best model was examined on an independent dataset. The outcomes on independent-samples indicated that the proposed model outpaced the two existed predictors, i4mC-Mouse and 4mCpred-EL.

2. Materials and methods

A reliable and accurate dataset is necessary to establish a prediction model. Therefore, we obtained the benchmark dataset from Hasan et al. work [20], and Manavalan et. al. [19]. In their study, they excluded similar sequences using 70% as cutoff of sequence identity [26]. After this elimination procedure, they finally obtained the benchmark dataset of 906 positive and 906 negative sequences with length of 41bp. Subsequently, the benchmark data were separated into 80% training data and 20% independent data to objectively estimate the efficiencies and performances of predictors, as shown in Table 1.

Table 1. The distribution of sample numbers in benchmark dataset.

Attribute	Training Data	Independent Data	Total
Positive	746	160	906
Negative	746	160	906
Total	1492	320	1812

2.1. Feature descriptors

Selecting the feature-encodings that are instructive and autonomous is an important stage in creating machine learning based models, such as BioSeq-Analysis2.0 [27], IDP-Seq2Seq [28], ACPred [29], iBitter-SCM [30], iTTCA-Hybrid [31], Meta-iAVP [32], PseKRAAC [33], iBLP [34] and so on [35,36]. Expressing the DNA sequences with a mathematical manifestation is very important in functional element identification. Zhang et al. obtained optimal nonamer composition to represent the sequences of mRNA [37]. Dao et al. used three types of feature encodings physiochemical properties, binary encodings and nucleotide chemical properties [38]. Yang et al. identify recombination site based on k-mer composition [39]. Dou et al. used k-mer nucleotide composition, nucleotide chemical properties and pseudo dinucleotide composition to identify RNA modification site [40]. Wei et al. identified circRNA-disease associations based on matrix factorization [41]. Zheng et al. developed reduced amino acid clusters [42]. Lv et al. applied k-tuple nucleotide frequency component, nucleotide pair spectrum encoding and natural vector in 3D genome [43]. Here, three types of feature-encoding approaches were presented to describe the DNA sequences.

2.1.1. k-mer nucleotide compositions (k-mer NC)

k-mer NC can reflect short-range nucleotide interaction of sequences [44–46]. The (N-k+1) nucleotide residues can be obtained via a sliding window method by setting the window size of *k bp*

with step size of 1 *bp* to examine a sequence with *N bp*. An arbitrary sample *M* with the sequence length of *N* (here *N* is 41bp) can be characterized as

$$M = R_1 R_2 R_3 \ldots\ldots R_i \ldots\ldots R_{(N-1)} R_N \quad (1)$$

where R_i signifies the nucleotide (A, T, C, and G) at the *i*-th position. The sequences can be transformed into the 4^k-D vector using *k-mer* nucleotide composition as follows

$$M_k = [f_1^{k-tuple} f_2^{k-tuple} \ldots\ldots f_i^{k-tuple} \ldots\ldots f_{4^k}^{k-tuple}]^T \quad (2)$$

where *T* denotes the transposition of the vector, and $f_1^{k\text{-}tuple}$ symbolizes the occurrence of the *i*-th *k-mer* nucleotide composition in the sequence. When k =1, a DNA sample can be deciphered into a 4-D vector $M_1 = [f(A), f(G), f(C), f(T)]^T$. When k = 2, the DNA sample can be described by a 16-dimension vector. In this study, the value of k was set as (1, 2, … 6). Therefore, a sequence sample can be transformed into a 5460 ($4^1 + 4^2 + 4^3 + 4^4 + 4^5 + 4^6$) dimension vectors formulated as follows

$$M = M_1 \cup M_2 \cup M_3 \cup M_4 \cup M_5 \cup M_6 \quad (3)$$

2.1.2. Enhanced nucleic acid composition (ENAC)

The ENAC calculates the nucleic acid composition based on the sequence window. It can be used to formulate the sequence with equal length. The enhanced nucleic acid composition can be calculated as

$$Q = [\frac{N_{A,win1}}{k} \frac{N_{G,win1}}{k}, \frac{N_{C,win1}}{k}, \frac{N_{T,win1}}{k}, \frac{N_{A,win2}}{k} \ldots \frac{N_{G,winL-k+1}}{k}, \frac{N_{T,win\,l-k+1}}{k}] \quad (4)$$

In Equation (4), k characterizes the size of the sliding window, $N_{A,win}$ denotes the number of nucleotide A in the sliding window *p, T* ∈ [G, C, A, T], and (p = 1, 2, ..., L-k+1). In this study, the sliding window was set to 5. Then the feature dimension is 148.

2.1.3. Composition of k-spaced nucleic acid pairs (CKSNAP)

The CKSNAP embodies the incidence of nucleotide pairs disconnected by any k nucleotide (k = 0, 1, 2, 3, 4 5). The composition of k-spaced nucleic acid pairs feature comprises 16 nucleotide pairs [AA, AG, ... TG, TT]. By taking k = 1 as an instance, composition of k-spaced nucleic acid pairs can be specified as follows:

$$Q = [\frac{N_{A*A}}{N_{Total}}, \frac{N_{A*G}}{N_{Total}}, \ldots \frac{N_{T*G}}{N_{Total}}, \frac{N_{T*T}}{N_{Total}}]_{16} \quad (5)$$

where * signifies (A, G, C, and T), N_{Y*Z} signifies the number of nucleotides Y*Z pairs in the sequence, and N_{Total} embodies the total number of single-spaced nucleotide pairs in the sequence. If the nucleic acid pair AA appears j times in the nucleotide sequence, the composition of the nucleic acid pair AA can be equal to j divided by the total number of 0-spaced nucleic acid pairs N_{Total} in the nucleotide sequence. For k = 0, 1, 2, 3, 4 and 5, the value of N_{Total} is $P-1$, $P-2$, $P-3$, $P-4$, $P-5$ and $P-6$ for a nucleotide sequence of length P, respectively. In this study, k = 2 and the dimension of the composition of k-spaced nucleic acid pairs feature was 48.

2.1.4. Feature selection with mRMR and IFS

The insertion of noisy features might result in the unsatisfactory performance of a model. Dao et al. proposed a two-step feature selection strategy to exclude noise [47]. Feng et al. used a mRMR technique to reduced noise [48]. Shao et al. performed three ranking algorithms to exclude irrelevant features [49]. Cheng et al. used MetaMap to reduced noisy features [50]. Other computational works did the similar works [51–53]. Therefore, the selection of features is an obligatory phase to remove the less important features and increase the productivity of a model [54]. Many feature selection and ranking techniques are available, such as f-score , mRMR [23], MRMD [55], chi-square [56]. In this study, mRMR with IFS [24,57] was applied to obtain the optimal feature subset. mRMR is a filter-based selection technique [58] to achieve an optimal model. Compactness functions are described as *y* and *z*, and *P (y)* and *P (z)* are the two corresponding probabilities. *P (y, z)* is the possibility of compactness, and the common information between the two functions can be demarcated as

$$I(y;z) = \iint P\,(y,z)\, log\, \frac{P(y,z)}{P(y)P(z)}\, dydz \tag{6}$$

In shared information, searching a subset S with m optimum features helps to determine the feature transmission, which majorly depends on the target $\{yi\}$ class q.

$$max\ d(S,q), d = \frac{1}{|S|}\sum_{y_i \in S} I(y_i, q) \qquad (i = 1,2,3 \ldots m) \tag{7}$$

Minimum redundancy can be defined as

$$min\ r(S,q), r = \frac{1}{|S|^2}\sum_{y_i, y_{j \in S}} I\left(y_i, y_j\right) \tag{8}$$

Final selection criteria can be articulated as:

$$max\emptyset\ (d,r), \emptyset = d - r \tag{9}$$

The principle of the mRMR technique is to use a typical redundancy and relevance to rank features to acquire the best subset. Mostly, if a model was built on a high-dimensional feature subset, it can produce overfitting and informational redundancy problems. Therefore, mRMR (minimum redundancy maximum relevance) with the IFS (Incremental Feature Selection) [24,59] technique and five-fold cross-validation method was applied to examine the optimal feature subset with the maximum accuracy. We ranked all features according to the $\emptyset$ -values and obtained new feature vectors, which is given in the below equation 10.

$$I^* = [h_1, h_2, h_3 \ldots h_n]^T \tag{10}$$

The first feature subset comprises the feature with the highest $\emptyset$ -value $I^* = [h_1]^T$. By adding the second highest $\emptyset$-value to the first subset, the second feature subset $I^* = [h_1, h_2]^T$ is formed and by adding the third highest $\emptyset$-value to the second feature subset, the third feature subset $I^* = [h_1, h_2, h_3]^T$ is formed [47]. The process was repeated until all the features were considered.

2.1.5. Machine learning classifier

Support vector machine is very famous and has been used in many bioinformatics tools [44-46].

It performs binary classification on data in supervised learning. We have used a free available package LibSVM version 3.21, which can be easily downloaded from https://www.csie.ntu.edu.tw/~cjlin/libsvm/ to train and test the model. We have used rbf kernel function due to its efficiency in non-linear classification. We have optimized cost and gamma parameters of RBF kernel function by using grid search with searching space $[2^{-5},2^{5}]$ for cost and $[2^{-12},2^{1}]$ for gamma. Naïve Bayes classifier has been widely used in bioinformatics due to its simplicity and better performance [60]. It is a classification technique and totally depends on Bayes theorem. Ada boost classifier is also very famous and has been widely used in bioinformatics [61]. It is an ensemble technique and combines various classifiers to enhance the accuracy. The main idea of this is to set the classifiers weights and trained the data in each iteration. We implemented these classifiers in Weka (version 3.8.4) [62]. Random forest is a combined knowledge technique extensively applied in bioinformatics [63,64]. The underlying principle is to combine several weak classifiers. The outcome is attained by the voting process therefore, the outcome of the model has higher exactness and simplification. The model was constructed using a random forest algorithm [22] and the complete procedure is clearly described in [65]. Scikit - learn package (v - 0.22.1) [66,67] was used to execute the random forest classifiers. Firstly, we used randomized search CV and then grid search CV to tune hyperparameter. The best tuned parameters of the proposed model are given in Table 3.

2.1.6. Evaluation metrics

Matthews correlation coefficient (MCC), accuracy (Acc), sensitivity (Sn) and specificity (Sp) were used in this study to check the overall efficiency of the model defined as Equation 11.

$$\begin{cases} Sn = \frac{TP}{TP + FN} \\ Sp = \frac{TN}{TN + FP} \\ Acc = \frac{TP + TN}{TP + FP + TN + FN} \\ MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TN + FN) \times (TP + FP) \times (TN + FP)}} \end{cases} \tag{11}$$

where *TP* represents the correctly identified 4mC sequences in benchmark data and *FP* signifies the 4mC sequences false-classified as non-4mC. Likewise, *TN* represents the correctly recognized non-4mC sequences in the data and *FN* signifies the non-4mC sequences, which were false-classified as 4mC. Consequently, the receiver operating characteristic (ROC) curve was used to illustrate the efficiency of the model graphically. The ROC curvature could assess the projecting ability of the proposed model on the whole assortment of resultant values. The area under the curve was premeditated to check the efficiency of the model. A good classifier gave $AUC = 1$, and the arbitrary performance gave $AUC = 0.5$.

3. Results and discussion

3.1. Composition analysis of sequences

The sequence pattern around the modification site is an operative stage to predict and interpret the genetic meanings of variations [68,69]. In this study, Two Sample Logo [70] (http://www.twosamplelogo.org/cgi-bin/tsl/tsl.cgi) was used to examine the distribution of nucleotides around 4mC. Figure 2 shows that nucleotide distribution among positive and negative sequences are different in regions flanking the nucleotide C. Both T and C nucleotides were individually abundant at the upstream and downstream of the positive sequences, whereas A and G were correspondingly enriched at the upstream and downstream of the negative samples. Some nucleotides tend to act continuously along the sequences.

For example, five sequential C nucleotides (6–10, 13–17 and 35–39) were found in positive sequences, while three successive A nucleotides (1–3), (8–10) and six repeated A nucleotides (36–41) were observed in negative sequences. Figure 2 also shows that there was significant difference between 4mC samples and non-4mC samples (*t*-test, *P*-value < *0.05*). Above results suggested that the nucleotides distribution in different positions are helpful for the accurate classification of 4mC and non-4mC samples.

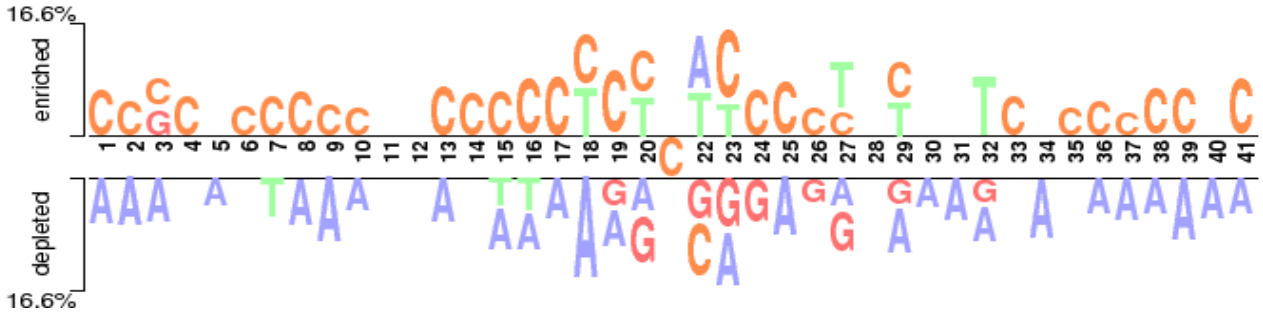

Figure 2. Compositional preferences of sequence between 4mC and non-4mC sites.

3.2. Performance evaluation

Based on sequence feature, we constructed a model to identify 4mC site. First, the training data were converted into feature vectors using feature descriptors (*k*-mer, composition of *k*-spaced nucleic acid pairs, enhanced nucleic acid composition, and feature fusion). Subsequently, the feature vectors of each encoding model were evaluated by random forest classifier using a five-fold CV test. mRMR with IFS method was used to pick out the best feature subset for the sake of better prediction accuracy. Figure 3 shows the IFS curve for searching optimal features. Table 2 recorded that performances of the three single-encoding models and the feature fusion model. The AUCs of single-encoding models (k-mer, CKSNAP, and ENAC) were 0.88, 0.80, and 0.79, respectively. The AUC of k-mer was around 1%–4% higher than those of the other encodings. Fusion feature-based model could produce the best results. In this optimal model, the Acc, MCC, Sn, Sp, and AUC were 79.91%, 0.598, 81.88%, 78.12% and 0.908, respectively. Figure 4 also shows the AUC of random forest based fusion model on training dataset and independent dataset by using five-fold cross validation. The best parameters were shown in Table 3.

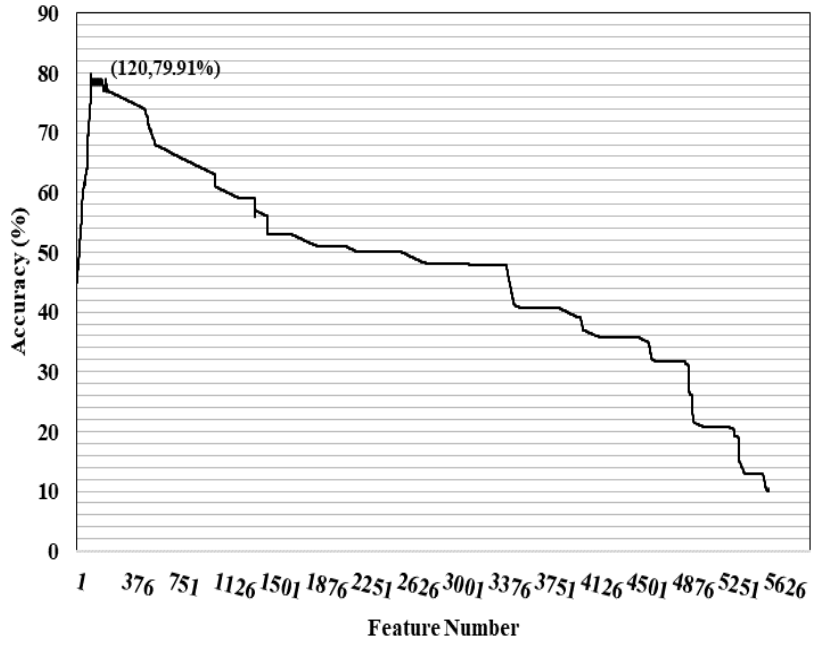

Figure 3. IFS curve of the optimal features.

Table 2. Performance on the basis of single encoding model using random forest.

Method	*k*	FS	Dimension	Ac (%)	MCC	Sn (%)	Sp (%)	AUC
CKSNAP	2	No	48	72.28	0.448	71.09	70.00	0.787
		Yes	7	72.54	0.450	72.00	71.00	0.800
ENAC	5	No	148	70.02	0.418	75.00	68.82	0.77.6
		Yes	13	70.98	0.425	77.00	67.00	0.790
k-mer	6	No	5460	76.92	0.557	77.20	78.34	0.873
		Yes	4088	75.66	0.539	76.80	77.34	0.863
		Yes	2426	77.32	0.563	79.20	77.64	0.878
		Yes	1221	78.12	0.568	80.20	78.14	0.883
		Yes	100	78.57	0.571	80.77	77.18	0.887
Fusion		No	5656	77.95	0.567	80.20	78.10	0.881
model		Yes	4020	77.80	0.561	78.45	**79.20**	0.881
		Yes	3105	78.30	0.581	80.25	79.10	0.893
		Yes	2088	77.90	0.578	78.55	78.04	0.886
		Yes	1023	79.54	0.596	81.32	78.40	0.903
		Yes	**120**	**79.91**	**0.598**	**81.69**	78.12	**0.908**

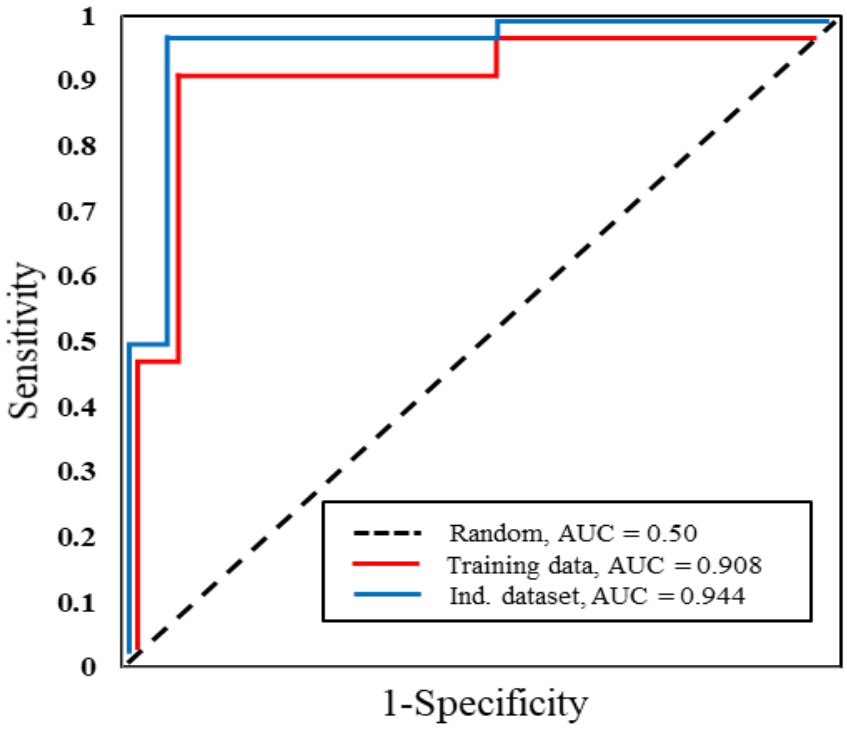

Figure 4. The ROC curve was evaluated on the training and independent dataset by a 5-fold cross validation test.

Table 3. Best parameters of the proposed model by 5-fold CV test.

Best Parameters	
'Bootstrap'	True
'Max-depth'	30
'Max-features'	2
'Min-samples-leaf'	1
'Min-samples-split'	8
'n-estimators'	40

3.3. Performance evaluation of different ML algorithms

k-mer, CKSNAP, ENAC and their fusion were inputted into three machine learning classifiers, namely Adaboost, SVM, and Naive Bayes algorithm, for comparing with random forest classifier-based models [71]. Cross-validation is a statistical analysis method and has been widely used in machine learning to train and test model. A five-fold CV test was used to elevate their corresponding machine learning constraints on individual encoding classifiers. In five-fold CV, the benchmark dataset was arbitrarily separated into five groups of about equal size. Each group was individualistically tested by the model which trained with the remaining four groups. Therefore, the five-fold CV method was performed five times, and the average of the results was the final result. Finally, an ideal model was achieved for each classifier. The results are shown in Table 4. We noticed that fused feature did produce high accuracy except Adaboost (69.57%). Then, comparison between feature fusion-based models with single-encoding based models indicates that the multiple information was effective to achieve

better results. As shown in Figure 5, based on fused features, random forest model exhibits higher accuracy compare with other three machine learning models. Particularly, the AUC of the feature fusion model based on random forest classifier was 1%–10% higher than that of the other models, indicating that the random forest model was the best for 4mC identification.

Table 4. Performances of all the models using different machine learning approaches.

Classifier	Method	Acc (%)	MCC	Sn (%)	Sp (%)	AUC
RF	CKSNAP	72.54	0.450	72.00	71.00	0.800
	ENAC	70.98	0.425	77.00	67.00	0.790
	k-mer	78.57	0.571	80.00	77.00	0.880
	Fusion	**79.91**	**0.598**	**81.88**	78.12	**0.908**
AB	CKSNAP	69.03	0.381	69.00	69.00	0.746
	ENAC	67.02	0.342	72.40	65.40	0.736
	k-mer	70.30	0.406	70.60	70.20	0.772
	Fusion	69.57	0.391	69.20	69.70	0.766
SVM	CKSNAP	66.75	0.335	65.50	67.20	0.668
	ENAC	49.93	-0.01	59.90	49.90	0.499
	k-mer	76.74	0.536	73.60	**78.50**	0.767
	Fusion	77.56	0.571	77.25	77.10	0.862
NB	CKSNAP	67.09	0.342	65.70	67.60	0.744
	ENAC	68.83	0.377	68.60	68.90	0.755
	k-mer	77.61	0.554	81.80	75.50	0.854
	Fusion	78.75	0.576	81.60	77.20	0.863

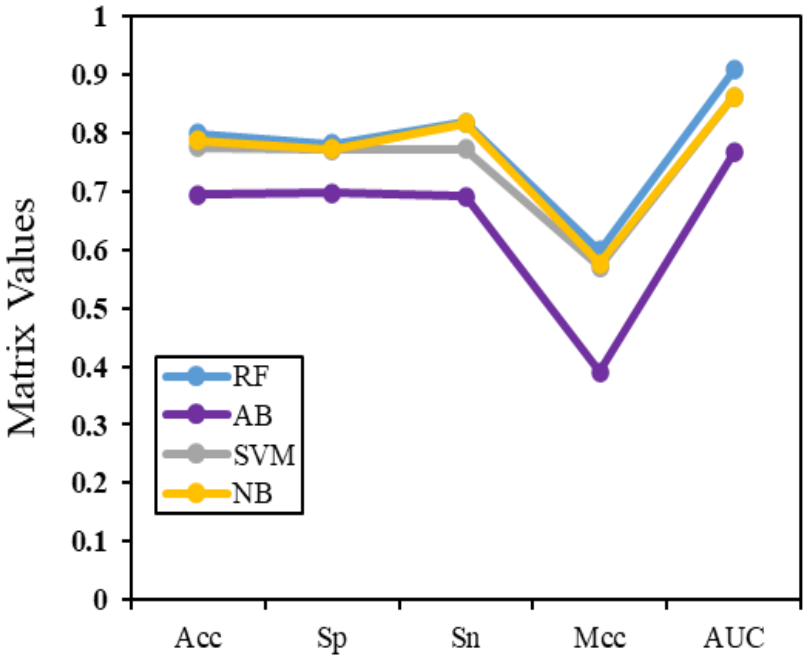

Figure 5. Matrix values of feature fusion models on four different ML algorithms. Performances were evaluated on the training dataset by 5-fold cross-validation test.

3.4. Comparison with existing models on an independent dataset

Independent dataset test was used to examine and compare the anticipated model with already published models. Two existing models, i4mC-Mouse and 4mCpred-EL could provide 4mC identification in mouse. Therefore, the efficiency of the proposed model was assessed against that of the aforementioned two existed models on the same independent dataset (160 4mC, and 160 non-4mC), as shown in Table 5. The MCC, Sn, Sp, Acc, and AUC of the i4mC-Mouse were 0.633, 80.71%, 82.52%, 81.61%, and 0.920, respectively. The MCC, Sn, Sp, Acc, and AUC of the 4mCpred-EL were 0.584, 75.72%, 82.51%, 79.10% and 0.881, respectively. The Feature Fusion model could produce 0.711, 82.00%, 89.13%, 85.41%, and 0.944, respectively for MCC, Sn, Sp, Acc, and AUC. Obviously, our proposed model outpaced both existing models by 2.4% and 6.3% in AUC which is shown in Figure 6. The good performance of the proposed model was due to the use of different and accurate encoding schemes and the selection of suitable classifiers.

Table 5. Comparison between proposed model and existing methods.

Method	Acc (%)	MCC	Sn (%)	Sp (%)	References	AUC
4mCpred-EL	79.10	0.584	75.72	82.51	[19]	0.881
i4mC-Mouse	81.61	0.633	80.71	82.52	[20]	0.920
model_4mc	**85.41**	**0.711**	**82.00**	**89.13**	**Our Work**	**0.944**

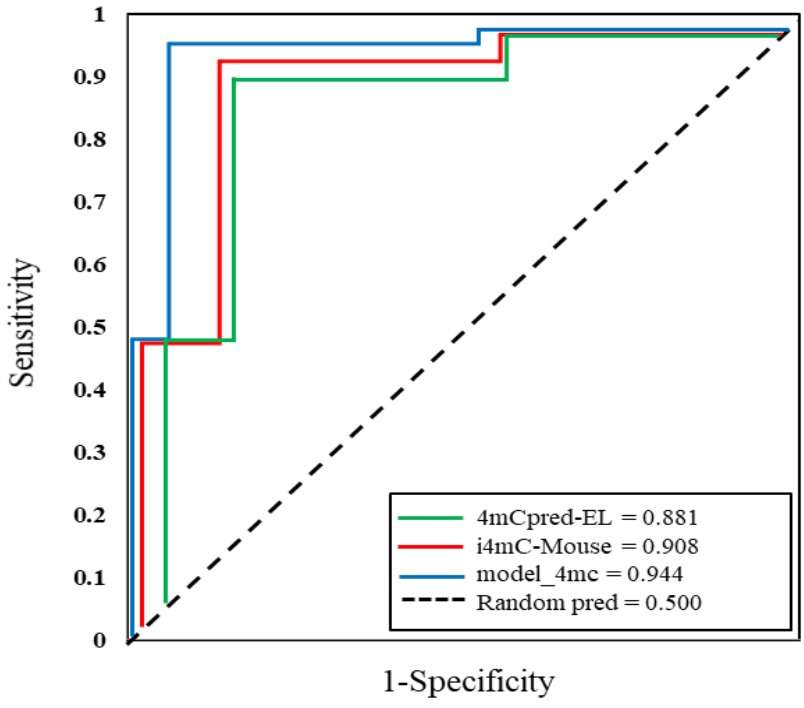

Figure 6. AUC of proposed model and two existing tools.

4. Conclusions

4mC is a DNA modification with a series of significant genetic progressions such as regulation of

gene expression and cell differentiation. The identification of 4mC sites in the whole genome is vital for understanding their genetic roles. To date, numerous predictors have been established to classify 4mC sites in diverse species [14,17,18,72–74], but only two methods 4mCpred-EL [19] and i4mC-Mouse [20] exist for mice. In this study, an advanced ensemble model was established to identify 4mC sites in the mouse genome. In the proposed model, DNA sequences were encoded using *k*-mer, CKSNAP and ENAC. Then, these encoding-features were optimized by using mRMR with IFS. On the basis of the top feature subset, the finest 4mC sorting model was achieved by the random forest classifier using a five-fold CV test. The estimated outcomes on independent data showed that the proposed model provided outstanding generalization capability. Further studies will aim to create a user-friendly web server for the projected model. Also, additional feature selection methods and algorithms will be implemented to further improve the efficiency to classify 4mC sites.

Acknowledgments

This work has been supported by the China Postdoctoral Science Foundation (2020M673188).

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. D. Liu, G. Li, Y. Zuo, Function determinants of TET proteins: The arrangements of sequence motifs with specific codes, *Brief Bioinform.*, **20** (2019), 1826–1835.
2. A. Jeltsch, R. Z. Jurkowska, New concepts in DNA methylation, *Trends Biochem. Sci.*, **39** (2014), 310–318.
3. D. Schübeler, Function and information content of DNA methylation, *Nature*, **517** (2015), 321–326.
4. B. M. Davis, M. C. Chao, M. K. Waldor, Entering the era of bacterial epigenomics with single molecule real time DNA sequencing, *Curr. Opin. Microbiol.*, **16** (2013), 192–198.
5. T. P. Meakin, N. Pillay, S. Beck, 3-methylcytosine in cancer: an underappreciated methyl lesion? *Epigenomics*, **8** (2016), 451–454.
6. K. D. Robertson, DNA methylation and human disease, *Nat. Rev. Genet.*, **6** (2005), 597–610.
7. M. M. Suzuki, A. Bird, DNA methylation landscapes: provocative insights from epigenomics, *Nat. Rev. Genet.*, **9** (2008), 465–476.
8. H. P. Schweizer, Bacterial genetics: Past achievements, present state of the field, and future challenges, *Biotechniques*, **44** (2008), 633–641.
9. L. M. Iyer, S. Abhiman, L. Aravind, Natural history of eukaryotic DNA methylation systems, *Prog. Mol. Biol. Transl. Sci.*, **101** (2011), 25–104.
10. W. He, C. Jia, Q. Zou, 4mCPred: Machine learning methods for DNA N4-methylcytosine sites prediction, *Bioinformatics*, **35** (2019), 593–601.
11. B. A. Flusberg, D. R. Webster, J. H. Lee, K. J. Travers, E. C. Olivares, T. A. Clark, et al., Direct detection of DNA methylation during single-molecule, real-time sequencing, *Nat. Methods*, **7** (2010), 461–465.
12. R. Doherty, C. Couldrey, Exploring genome wide bisulfite sequencing for DNA methylation analysis in livestock: a technical assessment, *Front. Genet.*, **5** (2014), 126.

13. J. Boch, U. Bonas, Xanthomonas AvrBs3 family-type III effectors: discovery and function, *Annu. Rev. Phytopathol.*, **48** (2010), 419–436.
14. W. Chen, H. Yang, P. Feng, H. Ding, H. Lin, iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties, *Bioinformatics,* **33** (2017), 3518–3523.
15. 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.
16. Z. Lv, D. Wang, H. Ding, B. Zhong, L. Xu, Escherichia coli DNA N-4-methycytosine site prediction accuracy improved by light gradient boosting machine feature selection technology, *IEEE Access*, **8** (2020), 14851–14859.
17. Q. Tang, J. Kang, J. Yuan, H. Tang, X. Li, H. Lin, et al., DNA4mC-LIP: A linear integration method to identify N4-methylcytosine site in multiple species, *Bioinformatics*, **36** (2020), 3327–3335.
18. 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.
19. B. Manavalan, S. Basith, T. H. Shin, D. Y. Lee, L. Wei, G. Lee, 4mCpred-EL: An ensemble learning framework for identification of DNA N4-methylcytosine sites in the mouse genome, *Cells*, **8** (2019), 1332.
20. M. M. Hasan, B. Manavalan, W. Shoombuatong, M. S. Khatun, H. Kurata, i4mC-Mouse: Improved identification of DNA N4-methylcytosine sites in the mouse genome using multiple encoding schemes, *Comput. Struct. Biotechnol. J.*, **18** (2020), 906–912.
21. P. Ye, Y. Luan, K. Chen, Y. Liu, C. Xiao, Z. Xie, MethSMRT: An integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing, *Nucleic Acids Res.*, (2016), DOI: 10.1093/nar/gkw950.
22. A. Liaw, M. Wiener, Classification and regression by random forest, *R. News*, **2** (2002), 18–22.
23. N. D. Jay, S. P. Cavanagh, C. Olsen, N. E. Hachem, G. Bontempi, B. H. Kains, mRMRe: An R package for parallelized mRMR ensemble feature selection, *Bioinformatics*, **29** (2013), 2365–2368.
24. W. Yang, X. J. Zhu, J. Huang, H. Ding, H. Lin, A brief survey of machine learning methods in protein sub-golgi localization, *Curr. Bioinform.*, **14** (2019), 234–240.
25. K. Liu, W. Chen, iMRM: A platform for simultaneously identifying multiple kinds of RNA modifications, *Bioinformatics*, **36** (2020), 3336–3342.
26. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: Accelerated for clustering the next-generation sequencing data, *Bioinformatics*, **28** (2012), 3150.
27. 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.
28. Y. J. Tang, Y. H. Pang, B. Liu, IDP-Seq2Seq: Identification of intrinsically disordered Regions based on sequence to sequence learning, *Bioinformaitcs*, (2020), DOI: 10.1093/bioinformatics/btaa667.
29. N. Schaduangrat, C. Nantasenamat, V. Prachayasittikul, W. Shoombuatong, ACPred: A computational tool for the prediction and analysis of anticancer peptides, *Molecules*, **24** (2019), 1973.
30. P. Charoenkwan, J. Yana, N. Schaduangrat, C. Nantasenamat, M. M. Hasan, W. Shoombuatong, iBitter-SCM: Identification and characterization of bitter peptides using a scoring card method with propensity scores of dipeptides, *Genomics*, **112** (2020), 2813–2822.
31. P. Charoenkwan, C. Nantasenamat, M. M. Hasan, W. Shoombuatong, iTTCA-Hybrid: Improved

and robust identification of tumor T cell antigens by utilizing hybrid feature representation, *Anal. Biochem.*, **599** (2020), 113747.
32. N. Schaduangrat, C. Nantasenamat, V. Prachayasittikul, W. Shoombuatong, Meta-iAVP: A sequence-based meta-predictor for improving the prediction of antiviral peptides using effective feature representation, *Int. J. Mol. Sci.*, **20** (2019), 5743.
33. P. Charoenkwan, C. Nantasenamat, M. M. Hasan, W. Shoombuatong, Meta-iPVP: A sequence-based meta-predictor for improving the prediction of phage virion proteins using effective feature representation, *J. Comput. Aided Mol. Des.*, **34** (2020), 1105–1116.
34. V. Laengsri, C. Nantasenamat, N. Schaduangrat, P. Nuchnoi, V. Prachayasittikul, W. Shoombuatong, TargetAntiAngio: A sequence-based tool for the prediction and analysis of anti-angiogenic peptides, *Int. J. Mol. Sci.*, **20** (2019), 2950.
35. Y. Zuo, Y. Li, Y. Chen, G. Li, Z. Yan, L. Yang, PseKRAAC: A flexible web server for generating pseudo k-tuple reduced amino acids composition, *Bioinformatics.*, **33** (2017), 122–124.
36. D. Zhang, H. D. Chen, H. Zulfiqar, S. S. Yuan, Q. L. Huang, Z. Y. Zhang, et al., iBLP: An xgboost-based predictor for identifying bioluminescent proteins, *Comput. Math. Methods Med.*, **2021** (2021), 15.
37. Z. Y. Zhang, Y. H. Yang, H. Ding, D. Wang, W. Chen, H. Lin, Design powerful predictor for mRNA subcellular location prediction in homo sapiens, *Brief Bioinform.*, **22** (2020), 526–535.
38. F. Y. Dao, H. Lv, Y. H. Yang, H. Zulfiqar, H. Gao, H. Lin, Computational identification of N6-methyladenosine sites in multiple tissues of mammals, *Comput. Struct. Biotechnol. J.*, **18** (2020), 1084–1091.
39. H. Yang, W. Yang, F. Y. Dao, H. Lv, H. Ding, W. Chen, et al., A comparison and assessment of computational method for identifying recombination hotspots in Saccharomyces cerevisiae, *Brief Bioinform.*, **21** (2020), 1568–1580.
40. L. J. Dou, X. Li, H. Ding, L. Xu, H. Xiang, Is there any sequence feature in the RNA pseudouridine modification prediction problem? *Mol. Ther. Nucleic Acids.*, **19** (2020), 293–303.
41. H. Wei, B. Liu, iCircDA-MF: Identification of circRNA-disease associations based on matrix factorization, *Brief Bioinform.*, **21** (2020), 1356–1367.
42. L. Zheng, D. Liu, W. Yang, L. Yang, Y. Zuo, RaacLogo: a new sequence logo generator by using reduced amino acid clusters, *Brief Bioinform.*, (2020), DOI: 10.1093/bib/bbaa096.
43. H. Lv, F. Y. Dao, H. Zulfiqar, W. Su, H. Ding, L. Liu, et al., A sequence-based deep learning approach to predict CTCF-mediated chromatin loop, *Brief Bioinform.*, (2021), DOI: 10.1093/bib/bbab031.
44. F. Y. Dao, H. Lv, H. Zulfiqar, H. Yang, W. Su, H. Gao, et al., A computational platform to identify origins of replication sites in eukaryotes, *Brief Bioinform.*, **22** (2020), 1940–1950.
45. B. Liu, BioSeq-Analysis: A platform for DNA, RNA, and protein sequence analysis based on machine learning approaches, *Brief Bioinform.*, **20** (2019), 1280–1294.
46. L. Zheng, S. Huang, N. Mu, H. Zhang, J. Zhang, Y. Chang, et al., RAACBook: A web server of reduced amino acid alphabet for sequence-dependent inference by using Chou's five-step rule, *Database-Oxford*., **2019** (2019), baz131.
47. F. Y. Dao, H. Lv, F. Wang, C. Q. Feng, H. Ding, W. Chen, et al., Identify origin of replication in saccharomyces cerevisiae using two-step feature selection technique, *Bioinformatics*, **35** (2019), 2075–2083.
48. C. Q. Feng, Z. Y. Zhang, X. J. Zhu, Y. Lin, W. Chen, H. Tang, et al., iTerm-PseKNC: A sequence-based

tool for predicting bacterial transcriptional terminators, *Bioinformatics*, **35** (2019), 1469–1477.
49. J. Shao, K. Yan, B. Liu, FoldRec-C2C: protein fold recognition by combining cluster-to-cluster model and protein similarity network, *Brief Bioinform.*, (2020), DOI: 10.1093/bib/bbaa144.
50. L. Cheng, Computational and biological methods for gene therapy, *Curr. Gene Ther.*, **19** (2019), 210–210.
51. L. Cheng, H. Zhao, P. Wang, W. Zhou, M. Luo, T. Li, et al., Computational methods for identifying similar diseases, *Mol. Ther. Nucleic Acids.*, **18** (2019), 590–604.
52. L. Cheng, C. Qi, H. Zhuang, T. Fu, X. Zhang, gutMDisorder: A comprehensive database for dysbiosis of the gut microbiota in disorders and interventions, *Nucleic Acids Res.*, **48** (2020), D554–D560.
53. H. Zulfiqar, M. S. Masoud, H. Yang, S. G. Han, C. Y. Wu, H. Lin, Screening of prospective plant compounds as H1R and CL1R inhibitors and its antiallergic efficacy through molecular docking approach, *Comput. Math. Methods Med.*, **2021** (2021), 9.
54. X. J. Zhu, C. Q. Feng, H. Y. Lai, W. Chen, L. Hao, Predicting protein structural classes for low-similarity sequences by evaluating different features, *Knowl. Based Syst.*, **163** (2019), 787–793.
55. Q. Zou, J. Zeng, L. Cao, R. Ji, A novel features ranking metric with application to scalable visual and bioinformatics data classification, *Neurocomputing*, **173** (2016), 346–354.
56. N. Rachburee, W. Punlumjeak, A comparison of feature selection approach between greedy, ig-ratio, chi-square, and mRMR in educational mining, in 2015 *7th International Conference on Information Technology and Electrical Engineering (ICITEE)*, IEEE, (2015), 420–424.
57. Z. M. Zhang, J. S. Wang, H. Zulfiqar, H. Lv, F. Y. Dao, H. Lin, Early diagnosis of pancreatic ductal adenocarcinoma by combining relative expression orderings with machine learning method, *Front. Cell Dev. Biol.*, **8** (2020), 1076.
58. 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.
59. J. X. Tan, S. H. Li, Z. M. Zhang, C. X. Chen, W. Chen, H. Tang, et al., Identification of hormone binding proteins based on machine learning methods, *Math. Biosci. Eng.*, **16** (2019), 2466–2480.
60. H. Lv, Z. M. Zhang, S. H. Li, J. X. Tan, W. Chen, H. Lin, Evaluation of different computational methods on 5-methylcytosine sites identification, *Brief Bioinform.*, **21** (2020), 982–995.
61. X. Li, L. Wang, E. Sung, AdaBoost with svm-based component classifiers, *Eng. Appl. Artif. Intell.*, **21** (2008), 785–795.
62. E. Frank, M. Hall, L. Trigg, G. Holmes, I. H. Witten, Data mining in bioinformatics using weka, *Bioinformatics.*, **20** (2004), 2479–2481.
63. X. Ru, L. Li, Q. Zou, Incorporating distance-based top-n-gram and random forest to identify electron transport proteins, *J. Proteom. Res.*, **18** (2019), 2931–2939.
64. Z. Lv, J. Zhang, H. Ding, Q. Zou, RF-PseU: A random forest predictor for RNA pseudouridine sites, *Front. Bioeng. Biotechnol.*, **8** (2020), 134.
65. L. Breiman, Random forests, *Mach Learn.,* **45** (2001), 5–32.
66. A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, et al., Machine learning for neuroimaging with scikit-learn, *Front. Neuroinform.*, **8** (2014),14.
67. P. Liang, W. Yang, X. Chen, C. Long, L. Zheng, H. Li, et al., Machine learning of single-cell transcriptome highly identifies mRNA signature by comparing f-score selection with DGE analysis, *Mol. Ther. Nucleic Acids.*, **20** (2020), 155–163.

68. Z. D. Smith, A. Meissner, DNA methylation: Roles in mammalian development, *Nat. Rev. Genet.*, **14** (2013), 204–220.
69. K. Liu, W. Chen, H. Lin, XG-PseU: An extreme gradient boosting based method for identifying pseudouridine sites, *Mol. Genet. Genom.*, **295** (2020), 13–21.
70. V. Vacic, L. M. Iakoucheva, P. Radivojac, Two sample logo: a graphical representation of the differences between two sets of sequence alignments, *Bioinformatics.*, **22** (2006), 1536–1537.
71. Y. Zhang, Y. Li, R. Wang, J. Lu, X. Ma, M. Qiu, PSAC: Proactive sequence-aware content caching via deep learning at the network edge, *IEEE Trans. Netw. Sci. Eng.*, **7** (2020), 2145–2154.
72. H. Lv, F. Y. Dao, D. Zhang, Z. X. Guan, H. Yang, W. Su, et al., iDNA-MS: An integrated computational tool for detecting DNA modification sites in multiple genomes, *iScience*, **23** (2020), 100991.
73. H. Xu, P. Jia, Z. Zhao, Deep4mC: Systematic assessment and computational prediction for DNA N4-methylcytosine sites by deep learning, *Brief Bioinform.*, (2020), DOI: 10.1093/bib/bbaa099.
74. Q. Liu, J. Chen, Y. Wang, S. Li, C. Jia, J. song, et al., DeepTorrent: A deep learning-based approach for predicting DNA N4-methylcytosine sites, *Brief Bioinform.*, (2020), DOI:10.1093/bib/bbaa124.

AIMS Press