IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON FEATURE REPRESENTATION AND LEARNING METHODS WITH APPLICATIONS IN LARGE-SCALE BIOLOGICAL SEQUENCE ANALYSIS

Received April 15, 2020, accepted April 27, 2020, date of publication April 30, 2020, date of current version May 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2991477

iRNA-m5C_NB: A Novel Predictor to Identify RNA 5-Methylcytosine Sites Based on the Naive Bayes Classifier

LIJUN DOU[1,2], XIAOLING LI[3], HUI DING[4], LEI XU[5], AND HUAIKUN XIANG[1]

[1]School of Automotive and Transportation Engineering, Shenzhen Polytechnic, Shenzhen 518055, China
[2]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, China
[3]Department of Oncology, Heilongjiang Province Land Reclamation Headquarters General Hospital, Harbin 150001, China
[4]Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China
[5]School of Electronic and Communication Engineering, Shenzhen Polytechnic, Shenzhen 518055, China

Corresponding authors: Lei Xu (csleixu@szpt.edu.cn) and Huaikun Xiang (xianghuaikun@szpt.edu.cn)

ABSTRACT As one of the widespread RNA post-transcriptional modifications (PTCMs), 5-Methylcytosine (m5C) plays vital roles in better understanding of basic biological mechanisms and major disease treatments. In experiments, traditional high-throughput approaches to find m5C sites are usually expensive and laborious. Additionally, facing with a large number of RNA sequences, developing accurate computational methods to distinguish m5C and non-m5C sites is an efficient solution. Here we introduced a novel predictor, called iRNA-m5C_NB, to identify m5C sites in *Home sapiens* using Naive Bayes (NB) algorithm. In this method, unbalanced dataset Met935 is firstly analyzed using efficient hybrid-sampling strategy SMOTEEEN. Then top 57 features are selected by the ANOVA F-value from four kinds of well-performance feature extraction techniques, including Bi-profile Bayes (BPB), enhanced Nucleic Acid Composition (ENAC), electron-ion interaction pseudopotentials (EIIP) and mMGap_1. Based on the jackknife test, the evaluated recall for the unbalanced training dataset Met935 is up to 82.81% with *MCC* of 0.63. And for the independent dataset Test1157, the predictor still shows high recall of 70.06% and *MCC* of 0.34. It is the first m5C predictor constructed using the unbalanced dataset, and the recall scores are increased by 19.82% and 59.23% for jackknife and independent tests compared with the latest tool RNAm5CPred, respectively. We demonstrate that the proposed predictor iRNA-m5C_NB outperforms other state-of-art models, which hopes to be an efficient and reliable method to identify m5C sites.

INDEX TERMS 5-Methylcytosine, bi-profile Bayes, naive Bayes, unbalanced data, feature selection.

I. INTRODUCTION

5-Methylcytosine (m5C) is one of the widely spread post-transcriptional modifications (PTCMs) in rRNA, tRNA and mRNA sequences, which has been found in many organisms [1]–[4]. Specifically, m5C can be formed on carbon atom by the catalysis of RNA methyltransferase (such as NSUN2 and DNMT2), where a methyl group is attached in the 5th position of the cytosine (C) ring [5]. As a research hotpot in recent years, m5C has been discovered in various biological processes, such as tRNA stabilization, rRNA translational fidelity and codon identification [5]–[10]. Meanwhile, it is proved that m5C has important effect on many major human diseases, including breast cancer, autosomal-recessive intellectual disability and Dubowitz syndrome [11]–[16]. Therefore, fast and efficient recognition of m5C is the primary task of further researches on biological mechanisms and valuable applications. Although kinds of biological experiments have been proposed to detect m5C sites (i.e. bisulfite treatment [17], [18], m5C RNA immunoprecipitation (m5C-RIP) [19], 5-azacytidine-mediated RNA immunoprecipitation (Aza-IP) [20] as well as methylation iCLIP (miCLIP) [21], it is believed that corresponding costs of time and money are very high. Meanwhile, the number of RNA sequences shows sharp accumulation with the mature sequencing techniques. Therefore, constructing high-performance computational models to predict m5C becomes a reliable method to resolve this problem.

To our best know, totally eight models have been built to recognize RNA m5C sites [22]–[29]. Except the tool

The associate editor coordinating the review of this manuscript and approving it for publication was Dariusz Mrozek.

PEA-m5C for *Arabidopsis thaliana* [26], the remaining seven [22]–[25], [27]–[29] are all involved in the identification of human sequences (abbreviated as *H. sapiens*). The first computational model, m5C-PseDNC, was proposed by Feng *et al.* using the Support vector machine (SVM). The prediction accuracy is 90.42% over jackknife test, where pseudo dinucleotide composition (PseDNC) features with three physiochemical characteristics (entropy, enthalpy and free energy) were used to encode RNA sequences [22]. Then, RF-based model iRNAm5C-PseDNC was provided by Qiu *et al.* using PseDNC features with ten important properties considered, where jackknife test achieves high accuracy of 92.37% [23]. Later on, Zhang *et al.* developed a novel model m5C-HPCR by heuristic nucleotide physicochemical property reduction algorithm (HPCR), in which *MCC* and *AUC* are up to 0.859 and 0.962 [24]. Sabooh *et al.* developed new method pM^5CS-Comp-mRMR based on the Kmer features ($k = 2 \sim 4$). Particularly, feature selection approach Minimum Redundancy and Maximum Relevance (mRMR) was applied to choose effective features, which finally gives the accuracy value of 93.33% [25]. And compressive and cell-specific predictor RNAm5Cfinder was established by Li *et al.* using binary encoding (BE) features to analyze m5C sites in eight tissues/cell types, corresponding AUC values are both higher than 0.77 and 0.87 [27]. At same time, Lv *et al.* introduced iRNA-m5C model using four integrated features, including Kmer, BE, pseudo k-tuple nucleotide composition (PseKNC) and Natural Vector (NV) [28]. The jackknife accuracy is up to 92.9%. Very recently, Fang *et al.* published a new predictor RNAm5CPred based on combination of three nucleotide compositions, namely Kmer, K-spaced nucleotide pair frequencies (KSNPFs, same as mMKGap in this paper) and PseDNC [29], where the recall and *MCC* are 68.79% and 0.154 over independent test.

The datasets are the most basic and important part for constructing model. For the five tools, namely m5C-PseDNC [22], M5C-HPCR [24], Pm5cs-Comp-Mrmr [25], iRNA-m5C [28] and RNAm5CPred [29], they all used the balanced training dataset Met240 (containing 120 positive and 120 negative instances) collected by Feng *et al.* [22]. And for iRNAm5C-PseDNC using the unbalanced dataset Met1900 (475 positive and 1425 negative samples), there are large amount of redundant sequences with the accuracy and *MCC* achieve 92.37% and 0.79. It means that serious overfitting problem is existed in this model [23]. As for the latest predictor RNAm5CPred [29], kinds of sequences datasets (Balanced: Met240; Unbalanced: Met1900, Met935, Train935, Train839, Test96 and Test1157) were all investigated. However, the model was finally constructed using Met240 by comparison of models results based on Met240 and Met935. Finally, the results over independent dataset Test1157 are unsatisfied ($R_e = 68.79\%$, $S_p = 53.70\%$, $P_{re} = 18.19\%$ and $MCC = 0.154$). Meanwhile, the jackknife performances using unbalanced Met935 are still low ($R_e = 62.99\%$, $S_p = 99.50\%$, $MCC = 0.749$, $P_{re} = 95.24\%$), as well as independent test using Test1157 ($R_e = 10.83\%$, $S_p = 93.00\%$, $MCC = 0.050$, $P_{re} = 19.54\%$). In general, although the high accuracies (more than 93%) were reported using the balanced dataset over jackknife test, it is an urgent need to construct the high-performance model using the unbalanced data based on the fact that the m5C sites is distributed unbalanced. In another hand, the number of Met240 is so small that it lacks statistics characteristics.

In this paper, we focused on the identification of RNA m5C sites in *H. sapiens* using the unbalanced dataset Met935 and Test1157. Figure 1 displays the basic flowchart of this work. Based on the training dataset Met935, several unbalanced strategies are firstly tested using the single BPB features, where four algorithms are also applied simultaneously, including RF, SVM, AdaBoost and NB. After preliminary studies, hybrid-sampling technique SMOTEENN and NB algorithms are selected for the next experiments. Then, we investigate the results of five popular sequence representations, where four features (BPB, ENAC, EIIP and mMGap_1) are finally used. The model is finally constructed using the efficient top 57 features selected by the ANOVA F-value.

II. MATERIALS AND METHODS

A. BENCHMARK DATASETS

In the present work, two unbalanced benchmark datasets Met935 and Test1157 are used for cross validation and independent tests, as well as balanced Met240 for nucleotide distribution analysis. As mentioned above, Met240 is the first benchmark dataset collected by Feng *et al.* [22] to construct m5C sites model. It is obtained from the popular RNA modification database RMBase [1] with 120 positive and 120 negative instances. Met935 is built by Fang *et al.* [29], which includes 127 positive and 808 negative samples. Specifically, positive sequences are also obtained from RMBase [1], and the negative sequences are obtained from 1425 non-m5C samples in Met1900 collected by Qiu et al [23]. Testing dataset Test1157, containing 157 m5C and 808 non-m5C sequences, is used to evaluate the model performances over independent test, which is selected from Gene Expression Omnibus datasets (GEO) website with gse90963 (https://www.ncbi.nlm.nih.gov/geo/) by Feng *et al.* [22]. It is noted that the sequence similarity is less than 70.00% using CD-HIT program for the mentioned three datasets [30]. More details can be found in [22], [23], [29], [31]–[34].

B. RNA FEATURE REPRESENTATION

Efficient RNA feature representation is important to building the machine-learning-based predictors. Various state-of-art feature-extraction platforms have been proposed to conveniently encode RNA segments [35]–[38]. In this paper, six kinds of RNA features are applied to determine whether the nucleotide C can be modified.

1) BI-PROFILE BAYES (BPB)

BPB is a popular sequence-encoding technique, which is widely chosen to solve identification subjects in

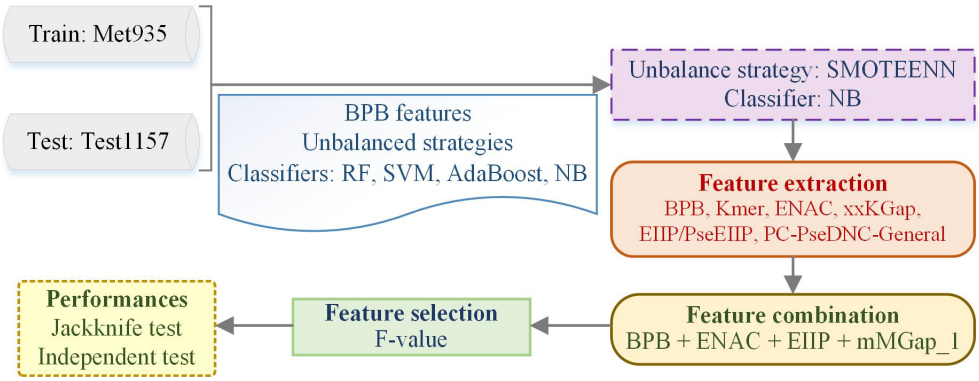

FIGURE 1. Flowchart of iRNA-m5C_NB model to identify RNA m5C sites using Naive Bayes in this work.

bioinformatics [39]–[48]. In this method, nucleotide distribution properties in positive and negative samples are separately used to represent sequences, which well reflects the sequence position-specific information. Considering a l-length RNA sequence $S = R_1R_2R_3 \ldots R_l$, BPB features can be formulated as

$$V_{BPB} = (p_1, p_2, \ldots, p_l, n_{l+1}, n_{l+2}, \ldots, n_{2l})^T \quad (1)$$

where $(p_1, p_2, \ldots, p_l)$ and $(n_{l+1}, n_{l+2}, \ldots, n_{2l})$ give the corresponding nucleotide occurrence probabilities at each location $i(i = 1, 2, ..., l)$ in positive and negative samples, respectively. Considering the nucleotide C always locating in the center of the sequence, we remove two features p_l and p_{2l} for the center C (i.e. p_l and p_{2l} always keep 1.00 for all samples). Therefore, BPB can induces totally $2(l-1)$-dimensional features.

2) KMER

As one well-known vector model, Kmer is simply expressed as the k-tuple or k-neighboring nucleotides composition [26], [28], [35], [49],

$$\mathrm{V}_{Kmer} = [f_1^k, f_2^k, f_3^k, \ldots, f_i^k, \ldots, f_{4^k}^k]^T \quad (2)$$

where f_i^k indicates the calculated frequencies of i-th k-tuple. Obviously, Kmer will induce a 4^k-dimensional vector. In this work, we set $k = 1 \sim 4$ to generate sequence features.

3) ENHANCED NUCLEIC ACID COMPOSITION (ENAC)

In ENAC method, nucleotide frequencies in a length-fixed subsequence are calculated to represent RNA instance, which is usually thought to be an improved version of NAC approach (i.e. Kmer with $k = 1$). Many subsequences will be obtained when the nucleotide window continuously slides from 5' to the 3' terminus over full RNA segment [50]. If we set the subsequence length as m, a $(l - m + 1) \times 4)$-dimensional ENAC feature vector can be obtained. Here we use the default window length 5 to carry out our research.

4) XXKGAP

Similarly, xxKGAP feature is one variation of Kmer method implemented in PyFeat package [36], where the composition of subsequences with k-gaps is used to describe sequences. In this paper, we adapt monoMonoKGap (mMKGap), monoDiKGap (mDKGap) and monoTriKGap (mTKGap) features with $k = 1 \sim 3$ to model.

5) ELECTRON-ION INTERACTION PSEUDOPOTENTIALS (EIIP) AND EIIP OF TRINUCLEOTIDE (PseEIIP)

Based on the reported electron-ion interaction pseudopotentials values of four nucleotides (i.e. $\mathrm{EIIP_A} = 0.1260$, $\mathrm{EIIP_C} = 0.1340$, $\mathrm{EIIP_G} = 0.0806$ and $\mathrm{EIIP_T} = 0.1335$) [51], two effective feature-extraction techniques EIIP and PseEIIP are introduced for prediction researches [43], [52]–[54].

In EIIP scheme, the RNA sequence is directly replaced as the related EIIP values [51]. Furthermore, PseEIIP feature can be expressed using the extended average EIIP value of related trinucleotides,

$$V_{PseEIIP} = [EIIP_{AAA}.f_{AAA}, EIIP_{AAC}.f_{AAC}, \ldots, EIIP_{UUU}.f_{UUU}] \quad (3)$$

Here, $EIIP_{XYZ}$ represents the EIIP value of the i-th trinucleotide XYZ by $EIIP_{XYZ} = EIIP_X + EIIP_Y + EIIP_Z$ (i.e., the sum of three related nucleotides X, Y and Z), and f_{XYZ} *is* the related frequency of XYZ. These two methods EIIP and PseEIIP form l and 64-dimensional numeric vectors, respectively.

6) PC-PseDNC-GENERAL

The PC-PseDNC-General method is a frequently used encoding technique to predict RNA sites [37], [55]–[57], which successfully incorporates sequential information and physicochemical properties of dinucleotides. It induces $(16 + \lambda)$-dimensional features,

$$V_{PC-PseDNC-General} = (d_1 \cdots d_{16} d_{16+1} \cdots d_{16+\lambda})^T \quad (4)$$

where

$$d_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{16} f_i + \omega \sum_{j=1}^{\lambda} \theta_j}, & (1 \le k \le 16) \\ \dfrac{\omega \theta_{k-16}}{\sum_{i=1}^{16} f_i + \omega \sum_{j=1}^{\lambda} \theta_j}, & (16+1 \le k \le 16+\lambda) \end{cases} \quad (5)$$

here f_k is the normalized frequency of the k-th dinucleotide ($k = 1, 2, \ldots, 16$), and λ is the highest counted rank of considered RNA sequence correlations. And ω is the associated weighting factor in the range of $0 \sim 1$; θ_j indicates the j-tier correlation factor, which well demonstrates the sequence-order correlations of the most contiguous dinucleotides. In this study, we change the λ parameter from 2 to 5 to extract different features.

7) XXKGAP

Similarly, xxKGAP feature is one variation of Kmer method implemented in PyFeat package [36], where the composition of subsequences with k-gaps is used to describe sequences. In this paper, we choose monoMonoKGap (mMKGap), monoDiKGap (mDKGap) and monoTriKGap (mTKGap) features with $k = 1 \sim 3$ to construct.

C. MACHINE LEARNING ALGORITHMS

The powerful and efficient machine learning platform based on Python language Scikit-learn package [58] was applied to construct model and analyze features. Here four useful classifiers NB, RF, AdaBoost and SVM were used for prediction task with default parameters.

1) NB

NB is a useful supervised classification algorithm based on Bayes' theorem under the ''naive'' assumption [59]–[61], which can be defined as

$$f_{nb}(F) = \frac{P(c=+)}{P(c=-)} \prod_i^n \frac{P(f_i|c=+)}{P(f_i|c=-)} \quad (6)$$

where $F = (f_1, f_2, \ldots, f_n)$ indicates the object and involved $f_1, f_2, \ldots, f_n$ give the associated features. And c labels the class of samples (positive class: $c = +$; negative class: $c = -$). It has been widely used in bioinformatics researches with good performances [62]–[64]. Here we use the GaussianNB algorithm for classification.

2) RF

RF is a widely used tree-based ensemble estimator in bioinformatics [65]–[78]. In this method, the voting results of a number of decision tree classifiers are finally treated as the output prediction performances [79].

3) SVM

SVM is a useful supervised learning algorithm [80], [81], which has been extensively deployed in bioinformatics [82]–[97]. Low-dimensional feature space can be effectively transformed into high-dimensional Hilbert space to find the best margin for hyperplane using the radial basis kernel function (RBF).

4) ADABOOST

AdaBoost is a widely applied ensemble classifier to improve model performances in bioinformatics [34], [98]–[100]. In this method, various weaker learners are fitted using bicluster-based classifiers, such as small decision trees. The good prediction performances can be finally generated by integrating those classifiers through weighted vote/sum [101]–[103].

D. FEATURE SELECTION AND VISUALIZATION

In order to analyze importance of different features and simply the model, three feature-selection methods are used to rank associated features, namely AdaBoost, F-value and Chi2 implemented in *sklearn* toolkit [58]. F-value and Chi2 are the two traditional univariate feature selection approaches, where the best feature is selected using univariate statistical tests [104]–[108]. Specifically, first one calculates the F-value for the all studied samples. And Chi2 selects important features using the chi-squared stats between each non-negative feature and class. Meanwhile, t-Distributed Stochastic Neighbor Embedding (t-SNE) [109] is applied to visualize distribution by reducing the dimension of original high-dimensional data. Similarities between data points are firstly converted into joint probabilities. Then, Kullback-Leibler divergence between those joint probabilities is optimized to illustrate data distribution.

E. UNBALANCED STRATEGY

In this paper, several unbalanced strategies are used to solve the unbalanced problem of training dataset, including resampling methods and ensemble classifiers [49], [110]–[116]. For the hybrid-sampling method SMOTEENN [117], Synthetic Minority Over-sampling Technique (SMOTE) and undersampling method Edited Nearest Neighbours (ENN) are incorporated to balance the dataset. Specifically, SMOTE is first applied to generate new examples in minority class [118] followed by ENN to remove the mixed samples. More details can be found in [119].

F. CRITERIA FOR PERFORMANCES EVALUATION

Although performances of existing tools are finally evaluated over jackknife test, we firstly used 10-fold cross validation (10-fold CV) for preliminary experiments. Then, jackknife and independent tests are used to give objective results. For the 10-fold CV, the training dataset are randomly split into 10 subsets on average. Later, the model is trained using 9 subsets and tested using the remaining one. Repeat this process 10 times until each subset is used once as testing set. The average performance of related 10 folds is used as the final scores of models. As a special case, jackknife test is a special case of k-fold CV, where k is equal to the total number of samples.

Based on the 10-fold CV and independent tests, six metrics associated with the confusion matrix, namely recall

or sensitivity (R_e, S_n), specificity (S_p), accuracy (A_{CC}) and Matthew's correlation coefficient (MCC), precision (P_{re}) and $F1$ score are used to check performances of classification models, which are defined as bellow (7), as shown at the bottom of this page.

Here, TP and TN give the number of predicted true positive and negative instances, whereas FP and FN indicate the number of false positive and negative sequences, respectively. Furthermore, the AUC value (area under ROC curve) is also applied to represent the prediction results, which is no sensitive to the thresholds of predicted probability [120]–[127].

III. RESULTS AND DISCUSSION

A. ANALYSIS OF NUCLEOTIDE DISTRIBUTION

First of all, the nucleotide distribution characteristics are displayed in Figure 2 for the unbalanced dataset Met935 (Left) and balanced Met240 (Right). The enriched and depleted nucleotides are calculated by the differences of nucleotide frequencies between positive and negative samples (i.e. $p_i - n_{l+i}$ at position i, see details in Sec. II). Obviously, there are big differences existed between two datasets. For Met935, corresponding distribution is quietly different in individual place, which can be clearly seen for the nucleotides near the center. For the nucleotide at upstream position -1, C and A are separately enriched in positive and negative samples, respectively. On the contrary, G and U are obviously located in downstream positions 1∼4 and 5∼7, respectively (positive instances), while A, C, U enriched in positions 1∼3 as well as A, C in 4∼6 (negative instances). As for Met240, the distribution is basically uniform and simple. Specifically, nucleotides C and G are widely distributed in positive samples, whereas A and U in negative sequences, except for upstream position -20. Generally, there is obvious differences existed between the unbalanced and balanced datasets, where former is more complex and weaker.

As a supplement, corresponding visualization of these two datasets, i.e. Met935 (Left) and Met240 (Right), is also plotted in Figure 3. Here BPB features are finally transferred into a 2-dimensional vector to conveniently display. For the unbalanced dataset Met935, positive samples are basically placed in the entire feature space, only a few gathers at the bottom right. However, it is simpler and clearer for balanced Met240, where almost positive samples are clustered in the upper right. Considering the fact that m5C and non-m5C sequences are unbalanced, we can demonstrate that the unbalanced-dataset-based predictor is more reasonable and accurate, but also more difficult than the model using balanced dataset to diagnose m5C sites.

B. PRELIMINARY RESULTS OF DIFFERENT UNBALANCED STRATEGIES WITH SEVERAL ALGORITHMS

There are many unbalanced strategies overcoming the unbalanced problems and various algorithms constructing models. Using BPB features, we perform several preliminary experiments to investigate different unbalanced approaches using benchmark dataset Met935, including resampling and ensemble techniques. At the same time, four kinds of algorithms, namely NB, RF, SVM and AdaBoost, are separately applied to select the efficient algorithm. There are totally seven metrics (R_e *or* S_n, S_p, A_{CC}, MCC, AUC, P_{re} and $F1$) are used to evaluate performances. The best algorithm is mainly decided by combing the recall and specificity scores of training and testing datasets, especially for the recall over independent test. Among those experiments, we found that the combination of SMOTEENN and NB showed best results. Results of several unbalanced techniques and classifiers are listed as following to demonstrate the optimizing process.

Table1 summarizes the prediction performances of six unbalanced strategies using NB method, where the best results are obtained using SMOTEENN approach labeled as superscript **a**. For the training dataset Met935, it can be seen that all methods show good performances (R_e and S_p achieve about 80.00%), except for the under-sampling technique ENN with low R_e of 52.76%. However, the results are generally unsatisfactory for the testing dataset. Particularly, five models are almost focused on the prediction of negative samples, which ignored the prediction of positive samples with low R_e scores, including SMOTE, ADASYN, ENN, SMOTEENN and SMOTETomek. Setting popular over-sampling technique SMOTE as an example, 10-fold CV experiment shows high recall for positive results (R_e = 80.69%), however, independent test gives bad score (R_e = 45.86%). As for the SMOTEENN, unified and better performances can be found (Met935: R_e = 88.59%, S_p = 85.83%,

$$\begin{cases} R_e, S_n = \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\ S_p = \dfrac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \\ Acc = \dfrac{TP+TN}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \\ MCC = \dfrac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}} \\ P_{re} = \dfrac{TP}{\mathrm{TP}+\mathrm{FP}} \\ F1 = \dfrac{2\times P_{re}\times R_e}{P_{re}+R_e} \end{cases} \tag{7}$$

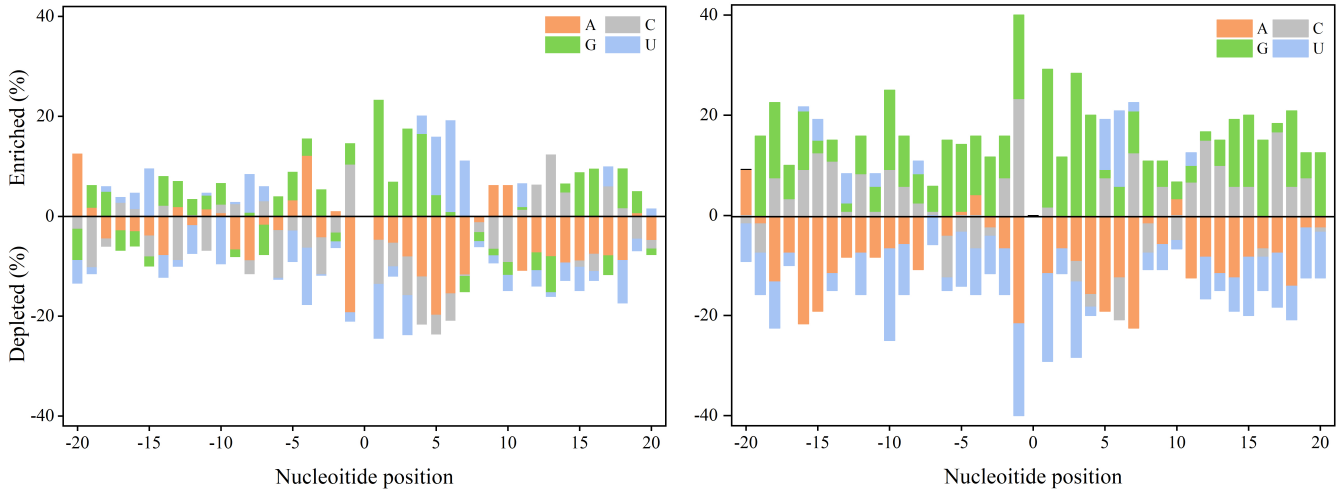

FIGURE 2. Nucleotide distribution characteristics between positive and negative sequences for the unbalanced dataset Met935 (Left) and balanced dataset Met240 (Right).

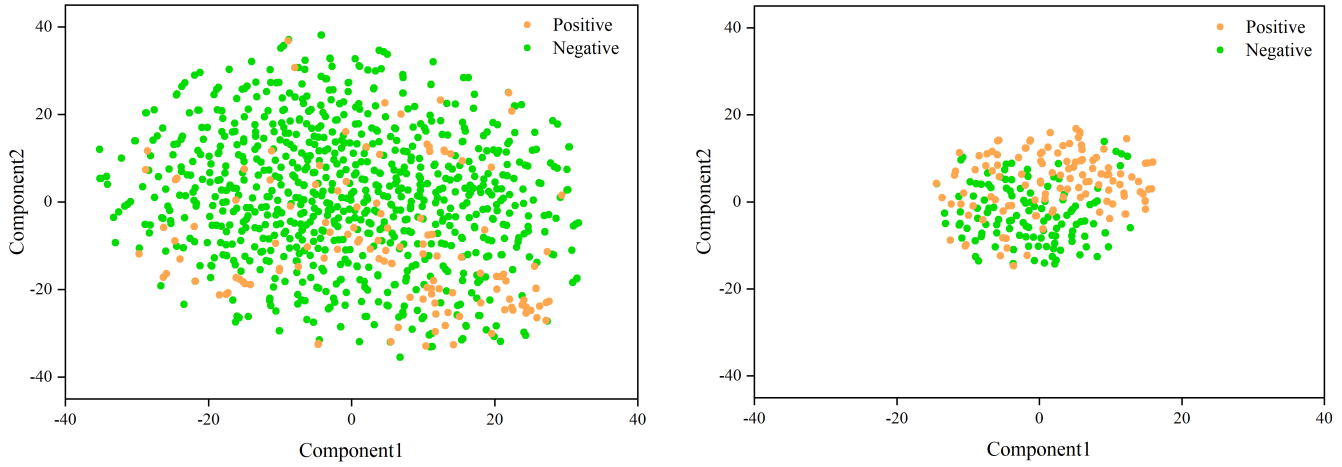

FIGURE 3. Visualization of unbalanced Met935 (Left) and balanced Met240 (Right) by t-SNE method using 80 BPB features.

TABLE 1. NB-based prediction performances of different unbalanced strategies using BPB features.

Unbalanced strategy	Met935							Test1157						
	R_e (S_n, %)	S_p (%)	A_{cc} (%)	*MCC*	*AUC*	P_{re} (%)	*F1*	R_e (S_n, %)	S_p (%)	A_{cc} (%)	*MCC*	*AUC*	P_{re} (%)	*F1*
SMOTE	80.32	80.69	80.51	0.61	0.89	80.62	0.80	45.86	84.50	79.26	0.26	0.80	31.72	0.37
ADASYN	79.30	77.97	78.64	0.57	0.86	78.63	0.79	48.41	81.40	76.92	0.24	0.80	29.01	0.36
ClusterCentroids	100.00	89.76	94.88	0.90	0.98	90.71	0.95	100.00	0.10	13.66	0.01	0.80	13.58	0.24
ENN	52.76	89.13	82.83	0.41	0.79	50.38	0.52	52.87	81.50	77.61	0.28	0.80	30.97	0.39
SMOTEENN[a]	88.59	85.83	87.95	0.69	0.94	95.45	0.92	65.61	71.60	70.79	0.27	0.80	26.61	0.38
SMOTETomek	80.32	80.69	80.51	0.61	0.89	80.62	0.80	45.86	84.50	79.26	0.26	0.80	31.72	0.37

[a]Unbalanced strategy with best performances.

AUC = 0.94; Test1157: R_e = 65.61%, S_p = 71.60%, AUC = 0.80). After comprehensive comparison, we finally choose SMOTEENN to deal with the unbalanced dataset in the next discussion.

Similarly, Table 2 lists the results of different algorithms based on SMOTEEENN strategy, including NB, RF, SVM and AdaBoost. It can be seen that the calculated results for four classifiers over 10-fold CV are exactly high, while bad

TABLE 2. BPB results of different algorithms using the unbalanced strategy SMOTEENN.

Algorithm	Met935							Test1157						
	R_e (S_n, %)	S_p (%)	A_{cc} (%)	MCC	AUC	P_{re} (%)	$F1$	R_e (S_n, %)	S_p (%)	A_{cc} (%)	MCC	AUC	P_{re} (%)	$F1$
NB[a]	88.59	85.83	87.95	0.69	0.94	95.45	0.92	65.61	71.60	70.79	0.27	0.80	26.61	0.38
RF	98.64	95.83	97.99	0.94	0.99	98.76	0.99	74.52	52.10	55.14	0.18	0.82	19.63	0.31
SVM	100.00	92.08	98.18	0.95	1.00	97.70	0.99	59.24	67.40	66.29	0.19	0.77	22.20	0.32
AdaBoost	99.50	95.42	98.57	0.96	1.00	98.65	0.99	96.82	12.30	23.77	0.10	0.83	14.77	0.26

[a]Algorithm with best performances.

TABLE 3. NB results of several kinds of features using the unbalanced strategy SMOTEENN.

Features	Fea_num	Met935							Test1157						
		R_e (S_n, %)	S_p (%)	A_{cc} (%)	MCC	AUC	P_{re} (%)	$F1$	R_e (S_n, %)	S_p (%)	A_{cc} (%)	MCC	AUC	P_{re} (%)	$F1$
Kmer_1	4	86.29	76.54	81.95	0.63	0.91	82.09	0.84	58.60	54.10	54.71	0.09	0.69	16.70	0.26
Kmer_2	16	91.47	89.93	90.84	0.81	0.97	92.87	0.92	45.22	65.00	62.32	0.07	0.70	16.86	0.25
Kmer_3	256	94.64	94.85	94.72	0.89	0.99	96.93	0.96	41.40	70.40	66.46	0.09	0.71	18.01	0.25
Kmer_4	1024	96.28	89.10	94.79	0.84	0.97	97.12	0.97	43.95	63.80	61.11	0.05	0.70	16.01	0.23
ENAC	148	90.41	87.10	89.71	0.72	0.97	96.29	0.93	65.61	72.10	71.22	0.27	0.84	26.96	0.38
mMGap_1	16	91.08	80.91	87.02	0.73	0.95	87.77	0.89	50.32	68.80	66.29	0.14	0.69	20.20	0.29
mMGap_2	16	91.42	83.49	88.28	0.75	0.95	89.42	0.90	49.68	72.30	69.23	0.16	0.70	21.97	0.30
mMGap_3	16	91.27	81.89	87.54	0.74	0.95	88.41	0.90	47.77	72.50	69.14	0.15	0.70	21.43	0.30
EIIP	41	84.56	72.97	81.44	0.55	0.87	89.46	0.87	76.43	50.70	54.19	0.19	0.81	19.58	0.31
PseEIIP	64	95.52	95.31	95.44	0.90	0.98	97.34	0.96	36.94	70.00	65.51	0.05	0.71	16.20	0.23
PC-PseDNC-Gneral	18	92.38	87.90	90.53	0.80	0.97	91.57	0.92	49.68	64.30	62.32	0.10	0.70	17.93	0.26

results over independent tests. For example, SVM model obtains the high R_e of 100.00% on training dataset, however only score of 59.24% on testing dataset. Combined the performances of Met935 and Test1157, we believe that NB is the best candidate to construct prediction model labeled as superscript **a**. Related R_e and S_p achieve 88.59%, 85.83% and 65.61%, 71.60% for two datasets, respectively.

C. EVALUATED RESULTS OF SINGLE FEATURES USING NB ALGORITHM

Considering the various sequence features, here we further investigate five feature extraction techniques (see Table 3), including Kmer, ENAC, mMGap, EIIP/PseEIIP and PC-PseDNC-General. The second column ''Fea_num'' indicates the feature dimension. For Kmer results, it can be found that associated results for different k values do not differ much, especially for the values of R_e for Test1157, which are generally in 41.40%∼58.60%. However, 148-dimensional ENAC features give the exciting results, where R_e and S_p reach 90.41%, 87.10%, 65.61% and 72.10% over 10-fold CV and independent tests. As for the xxKGap features, due to the performances are average, here we only list three mMKGap results with $k = 1 \sim 3$, which associated with the dinucleotides frequencies of X_X, X_ _X and X_ _ _X. It can be seen that the R_e and S_p are basically in high scores (80.91%∼91.42%) for training dataset, where the independent results often fail, especially for the R_e results (47.77%∼50.32%). Among these three features, mMGap_1 shows the better results. As for the two electron-ion interaction associated features EIIP and PseEIIP, simple EIIP give the relatively better results, where the recall results for negative samples are up to 84.56% and 76.43% for two experiments. As for the PC-PseDNC-General features with parameters λ from 2 to 5, the results of independent test are still disappointed, where only the results with default parameters ($\lambda = 2, \omega = 0.1$) are listed in this table. In summary, we selected three of listed features combing with BPB features to construct the comprehensive model, including ENAC, mMGap_1, EIIP features.

D. MODEL OPTIMIZATION AND COMPARISON WITH EXISTING TOOLS

We incorporate four well-performance features, including BPB, ENAC, EIIP and mMGap_1, to build prediction tool. There are totally 285 features concluded with the prediction results ($R_e = 89.10\%$, $S_p = 92.02\%$; Test1157: $R_e = 44.59\%$, $S_p = 89.50\%$). It can be found that the prediction performances for negative samples in Testing dataset are not very well. Thus, we applied three useful feature-selection methods F-value, Chi2 and AdaBoost to analyze the feature importance and remove redundant features. Finally, we find that top 57 features based on F-value show the best performances. Especially, the evaluated results between training and testing experiments are relatively unified and higher

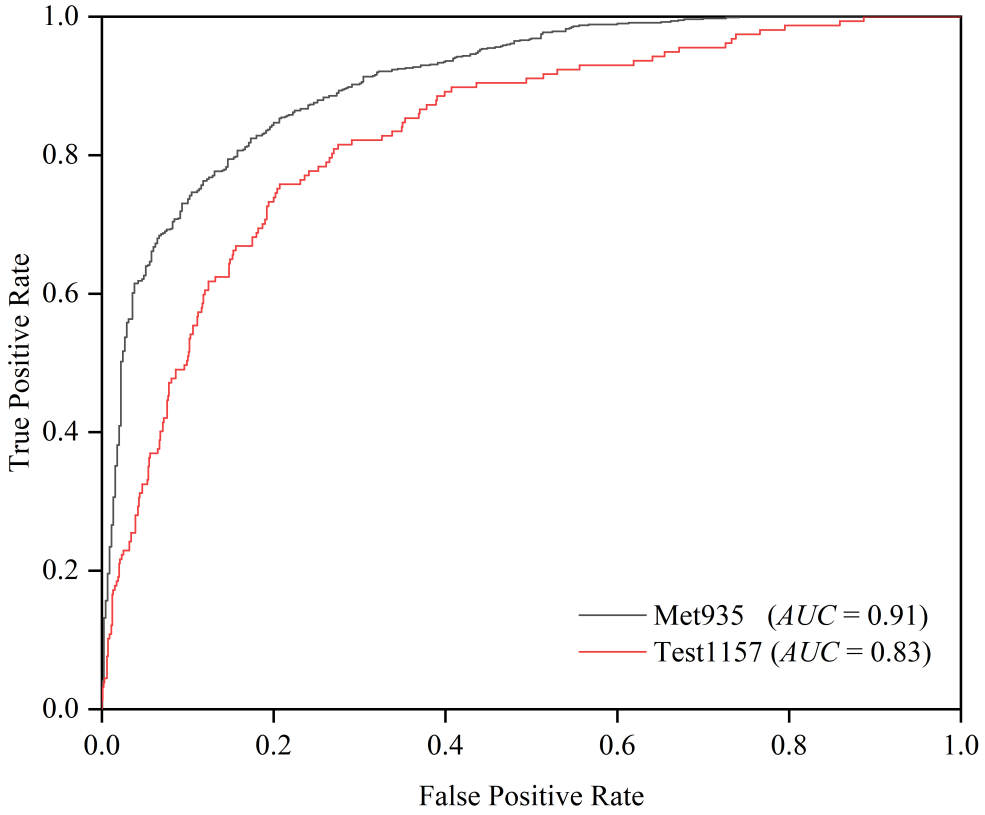

FIGURE 4. ***ROC*** **curves of training dataset Met935 (black line) and testing dataset Test1157 (red line) using the efficient top 57 features selected by ANOVA F-value method.**

TABLE 4. Comparison of our model and three reported prediction tools, where performances are obtained using efficient top 57 features by feature selection method ANOVA F-value.

Tools	Met935							Test1157						
	R_e (S_n, %)	S_p (%)	A_{cc} (%)	MCC	AUC	P_{re} (%)	$F1$	R_e (S_n, %)	S_p (%)	A_{cc} (%)	MCC	AUC	P_{re} (%)	$F1$
M5C-HPCR								62.42	51.10	52.64	0.09		16.70	
iRNA-m5C								43.95	49.20	48.49	-0.05		11.96	
RNAm5CPred								68.79	53.70	55.75	0.15		18.91	
RNAm5CPred[a]	62.99	99.50	94.55	0.75		95.24		10.83	93.00	81.85	0.05		19.54	
This work	82.81	81.11	82.20	0.63	0.91	88.59	0.86	70.06	75.60	74.85	0.34	0.83	31.07	0.43

[a] Although this model is finally constructed using Met240, the results for unbalanced datasets are still listed in Ref. [29].

(Met935: $R_e = 82.81\%$, $S_p = 82.00\%$, $A_{cc} = 82.52\%$, $MCC = 0.63$, $AUC = 0.91$, $P_{re} = 88.59\%$ and $F1 = 0.86$; Test1157: $R_e = 82.81\%$, $S_p = 81.11\%$, $A_{cc} = 82.20\%$, $MCC = 0.63$, $AUC = 0.91$, $P_{re} = 88.59\%$ and $F1 = 0.86$).

In order to conviniently compare performances of different models, we further perform jackknife test for taining dataset Met935. Table 4 lists our results and simultaneously compare with three proposed models, including M5C-HPCR, iRNA-m5C and RNAm5CPred [24], [28], [29]. It is noted that the results of M5C-HPCR [24] and iRNA-m5C [28] are excerpted from Fang *et al.*'s paper, where the unbalanced dataset Test1157 is apllied to test the related model efficency. Because these two are both constructed using the balanced dataset Met240, the performces of independent tests seems to be low, where $\boldsymbol{R_e}$, $\boldsymbol{S_p}$, $\boldsymbol{MCC}$ are only 62.42%, 51.10%, 0.09 and 43.95%, 49.20%, -0.05 for two experiments, respectively. As for the lateset model RNAm5CPred, although many datasets are investigated, the final model is still construted using Met240, where coresponding independent results for Test1157 are still needed to be improved ($\boldsymbol{R_e} = 68.79\%$, $\boldsymbol{S_p} = 53.70\%$, $\boldsymbol{A_{cc}} = 55.75\%$, $\boldsymbol{MCC} = 0.15$, $\boldsymbol{P_{re}} = 18.91\%$). As for the model based on the Met935 (labelled as RNAm5CPred$^{\boldsymbol{a}}$ in 4$^{\text{th}}$ row), associated $\boldsymbol{R_e}$ and $\boldsymbol{S_p}$ are 62.99% and 99.50% for training dataset, 10.83% and 93.00% for testing dataset. Although our $\boldsymbol{S_p}$ is 18.40% lower than RNAm5CPred$^{\boldsymbol{a}}$, there are totally 19.82% improvement for

R_e, which well reflects the high sensetivity of our model for positive samples. As for the testing dataset Test1157, our R_e and S_p are up to 70.06% and 75.05%. Corresponding ROC curves are also plotted in Figure 4, where black and red lines indicate Met935 and Test1157 with the AUC values of 0.91 and 0.83, respectively. Compared to the RNAm5CPred[a] results, our testing R_e is largely increased by 59.23%, where MCC is reached from 0.05 to 0.34. It can be concluded that our model is a more accurate predictor to identify m5C modifications.

IV. CONSLUSION

As one important epigenetic modification, 5-Methylcytosine (m5C) plays vital roles in researching various biological mechanisms and major diseases. In this work, we constructed an efficient NB-based model iRNA-m5C_NB to distinguish RNA m5C and non-m5C sites in *H. sapiens*. Unbalanced strategy SMOTEENN and classification method NB is firstly selected during series of preliminary experiments using BPB features. Then, top 57 features are selected from a 285-diemnsional combined feature vector "BPB + ENAC + EIIP + mMGap_1" using ANOVA F-value and applied to construct the prediction model. Jackknife test on training dataset Met935 shows well results ($R_e = 82.81\%$, $S_p = 81.11\%$, $A_{cc} = 82.20\%$, $MCC = 0.63$, $AUC = 0.91$, $P_{re} = 88.59\%$ and $F1 = 0.86$), as well as independent test on Test1157 ($R_e = 82.81\%$, $S_p = 81.11\%$, $A_{cc} = 82.20\%$, $MCC = 0.63$, $AUC = 0.91$, $P_{re} = 88.59\%$ and $F1 = 0.86$). Although the specificity is about 18.39% and 17.40% lower than RNAm5CPred for two datasets, the recall/sensitivity are surprisingly increased by 19.82% and 59.24%, respectively. Our new model reports *MCC* of 0.34 compared with original value of 0.05 for RNAm5CPred tool. It can be obviously demonstrated that this model outperforms other predictors. We believe that iRNA-m5C_NB model has great potential in predicting m5C modification sites in RNA sequences.

ACKNOWLEDGMENT

(Lijun Dou and Xiaoling Li contributed equally to this work.)

REFERENCES

[1] W.-J. Sun, J.-H. Li, S. Liu, J. Wu, H. Zhou, L.-H. Qu, and J.-H. Yang, "RMBase: A resource for decoding the landscape of RNA modifications from high-throughput sequencing data," *Nucleic Acids Res.*, vol. 44, no. 1, pp. 259–265, Jan. 2016.

[2] P. Boccaletto, M. A. Machnicka, E. Purta, P. Piątkowski, B. Bagiński, T. K. Wirecki, V. de Crécy-Lagard, R. Ross, P. A. Limbach, A. Kotter, M. Helm, and J. M. Bujnicki, "MODOMICS: A database of RNA modification pathways. 2017 update," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D303–D307, Jan. 2018.

[3] L. Zhang, Y. He, H. Wang, H. Liu, Y. Huang, X. Wang, and J. Meng, "Clustering count-based RNA methylation data using a nonparametric generative model," *Current Bioinf.*, vol. 14, no. 1, pp. 11–23, Dec. 2018.

[4] Q. Zou, P. Xing, L. Wei, and B. Liu, "Gene2vec: Gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA," *RNA*, vol. 25, no. 2, pp. 205–218, Feb. 2019.

[5] B. Brzezicha, M. Schmidt, I. Makałowska, A. Jarmołowski, J. Pieńkowska, and Z. Szweykowska-Kulińska, "Identification of human tRNA:m^5C methyltransferase catalysing intron-dependent m^5C formation in the first position of the anticodon of the pre-$tRNA^{Leu}$(CAA)," *Nucleic Acids Res.*, vol. 34, no. 20, pp. 6034–6043, Nov. 2006.

[6] M. Goll, "Methylation of tRNA(AsP) by the DNA methyltransferase homolog Dnmt2," *Science*, vol. 311, pp. 388–395, Feb. 2006.

[7] M. Schaefer, T. Pollex, K. Hanna, F. Tuorto, M. Meusburger, M. Helm, and F. Lyko, "RNA methylation by $Dnmt^2$ protects transfer RNAs against stress-induced cleavage," *Genes Develop.*, vol. 24, no. 15, pp. 1590–1595, Aug. 2010.

[8] J. E. Squires and T. Preiss, "Function and detection of 5-methylcytosine in eukaryotic RNA," *Epigenomics*, vol. 2, no. 5, pp. 709–715, Oct. 2010.

[9] S. Blanco, A. Kurowski, J. Nichols, F. M. Watt, S. A. Benitah, and M. Frye, "The RNA–methyltransferase misu ($NSun_2$) poises epidermal stem cells to differentiate," *PLoS Genet.*, vol. 7, no. 12, 2011, Art. no. e1002403.

[10] X. Yang, "5-methylcytosine promotes mRNA export—NSUN2 as the methyltransferase and ALYREF as an m5C reader," *Cell Res.*, vol. 27, no. 5, pp. 606–625, May 2017.

[11] M. Frye, I. Dragoni, S.-F. Chin, I. Spiteri, A. Kurowski, E. Provenzano, A. Green, I. O. Ellis, D. Grimmer, A. Teschendorff, C. C. Zouboulis, C. Caldas, and F. M. Watt, "Genomic gain of 5p15 leads to over-expression of misu (NSUN2) in breast cancer," *Cancer Lett.*, vol. 289, no. 1, pp. 71–80, Mar. 2010.

[12] L. Abbasi-Moheb, "Mutations in NSUN2 cause autosomal-recessive intellectual disability," *Amer. J. Hum. Genet.*, vol. 90, no. 5, pp. 847–855, 2012.

[13] M. P. Guy, M. Shaw, C. L. Weiner, L. Hobson, Z. Stark, K. Rose, V. M. Kalscheuer, J. Gecz, and E. M. Phizicky, "Defects in tRNA anticodon loop 2'-o-methylation are implicated in nonsyndromic X-Linked intellectual disability due to mutations in FTSJ1," *Hum. Mutation*, vol. 36, no. 12, pp. 1176–1187, Dec. 2015.

[14] K. Bohnsack, C. Höbartner, and M. Bohnsack, "Eukaryotic 5-methylcytosine (m5C) RNA methyltransferases: Mechanisms, cellular functions, and links to disease," *Genes*, vol. 10, no. 2, p. 102, 2019.

[15] X. Chen, Y.-Z. Sun, H. Liu, L. Zhang, J.-Q. Li, and J. Meng, "RNA methylation and diseases: Experimental results, databases, Web servers and computational models," *Briefings Bioinf.*, vol. 20, no. 3, pp. 896–917, May 2019.

[16] Y.-M. Feng, "Gene therapy on the road," *Current Gene Therapy*, vol. 19, no. 1, p. 6, May 2019.

[17] M. Frommer, L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy, and C. L. Paul, "A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands.," *Proc. Nat. Acad. Sci. USA*, vol. 89, no. 5, pp. 1827–1831, Mar. 1992.

[18] M. Schaefer, T. Pollex, K. Hanna, and F. Lyko, "RNA cytosine methylation analysis by bisulfite sequencing," *Nucleic Acids Res.*, vol. 37, no. 2, p. e12, 2008.

[19] I. Masiello and M. Biggiogera, "Ultrastructural localization of 5-methylcytosine on DNA and RNA," *Cellular Mol. Life Sci.*, vol. 74, no. 16, pp. 3057–3064, Aug. 2017.

[20] V. Khoddami and B. R. Cairns, "Identification of direct targets and modified bases of RNA cytosine methyltransferases," *Nature Biotechnol.*, vol. 31, no. 5, pp. 458–464, May 2013.

[21] S. Hussain, A. A. Sajini, S. Blanco, S. Dietmann, P. Lombard, Y. Sugimoto, M. Paramor, J. G. Gleeson, D. T. Odom, J. Ule, and M. Frye, "NSun2-mediated Cytosine-5 methylation of vault noncoding RNA determines its processing into regulatory small RNAs," *Cell Rep.*, vol. 4, no. 2, pp. 255–261, Jul. 2013.

[22] P. Feng, H. Ding, W. Chen, and H. Lin, "Identifying RNA 5-methylcytosine sites via pseudo nucleotide compositions," *Mol. BioSyst.*, vol. 12, no. 11, pp. 3307–3311, 2016.

[23] W.-R. Qiu, S.-Y. Jiang, Z.-C. Xu, X. Xiao, and K.-C. Chou, "IRNAm5C-PseDNC: Identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition," *Oncotarget*, vol. 8, no. 25, pp. 41178–41188, Jun. 2017.

[24] M. Zhang, Y. Xu, L. Li, Z. Liu, X. Yang, and D.-J. Yu, "Accurate RNA 5-methylcytosine site prediction based on heuristic physical-chemical properties reduction and classifier ensemble," *Anal. Biochem.*, vol. 550, pp. 41–48, Jun. 2018.

[25] M. F. Sabooh, N. Iqbal, M. Khan, M. Khan, and H. F. Maqbool, "Identifying 5-methylcytosine sites in RNA sequence using composite encoding feature into Chou's PseKNC," *J. Theor. Biol.*, vol. 452, pp. 1–9, Sep. 2018.

[26] J. Song, J. Zhai, E. Bian, Y. Song, J. Yu, and C. Ma, "Transcriptome-wide annotation of m5C RNA modifications using machine learning," *Frontiers Plant Sci.*, vol. 9, p. 519, Apr. 2018.

[27] J. Li, Y. Huang, X. Yang, Y. Zhou, and Y. Zhou, "RNAm5Cfinder: A Web-server for predicting RNA 5-methylcytosine (m5C) sites based on random forest," *Sci. Rep.*, vol. 8, no. 1, Dec. 2018, Art. no. 17299.

[28] H. Lv, Z. M. Zhang, S. H. Li, J. X. Tan, W. Chen, and H. Lin, "Evaluation of different computational methods on 5-methylcytosine sites identification," *Briefings Bioinf.*, vol. 2, Jun. 2019, Art. no. bbz048.

[29] T. Fang, Z. Zhang, R. Sun, L. Zhu, J. He, B. Huang, Y. Xiong, and X. Zhu, "RNAm5CPred: Prediction of RNA 5-Methylcytosine sites based on three different kinds of nucleotide composition," *Mol. Therapy Nucleic Acids*, vol. 18, pp. 739–747, Dec. 2019.

[30] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: Accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, Dec. 2012.

[31] Q. Zou, G. Lin, X. Jiang, X. Liu, and X. Zeng, "Sequence clustering in bioinformatics: An empirical study," *Briefings Bioinf.*, vol. 21, no. 1, pp. 1–10, 2020.

[32] J. Fu, Y. Luo, M. Mou, H. Zhang, J. Tang, Y. Wang, and F. Zhu, "Advances in current diabetes proteomics: From the perspectives of Label- free quantification and biomarker selection," *Current Drug Targets*, vol. 21, no. 1, pp. 34–54, Dec. 2019.

[33] F. Li, Y. Zhou, X. Zhang, J. Tang, Q. Yang, Y. Zhang, Y. Luo, J. Hu, W. Xue, Y. Qiu, Q. He, B. Yang, and F. Zhu, "SSizer: Determining the sample sufficiency for comparative biological study," *J. Mol. Biol.*, Feb. 2020, doi: 10.1016/j.jmb.2020.01.027.

[34] B. Manavalan, S. Basith, T. H. Shin, L. Wei, and G. Lee, "MAHTPred: A sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation," *Bioinformatics*, vol. 35, no. 16, pp. 2757–2765, Aug. 2019.

[35] B. Liu, X. Gao, and 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.*, vol. 47, no. 20, p. e127, Nov. 2019.

[36] R. Muhammod, S. Ahmed, D. Md Farid, S. Shatabda, A. Sharma, and A. Dehzangi, "PyFeat: A Python-based effective feature generation tool for DNA, RNA and protein sequences," *Bioinformatics*, vol. 35, no. 19, pp. 3831–3833, Oct. 2019.

[37] B. Liu, H. Wu, and K.-C. Chou, "Pse-in-One 2.0: An improved package of Web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Natural Sci.*, vol. 09, no. 04, pp. 67–91, 2017.

[38] Z. Chen, P. Zhao, F. Li, T. T. Marquez-Lago, A. Leier, J. Revote, Y. Zhu, D. R. Powell, T. Akutsu, G. I. Webb, K.-C. Chou, A. I. Smith, R. J. Daly, J. Li, and J. Song, "ILearn: An integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data," *Briefings Bioinf.*, Apr. 2019, doi: 10.1093/bib/bbz041.

[39] D. Mrozek, B. Malysiak, and S. Kozielski, "An optimal alignment of proteins energy characteristics with crisp and fuzzy similarity awards," in *Proc. IEEE Int. Fuzzy Syst. Conf.*, Jul. 2007, pp. 1–6.

[40] J. Shao, D. Xu, S.-N. Tsai, Y. Wang, and S.-M. Ngai, "Computational identification of protein methylation sites through bi-profile Bayes feature extraction," *PLoS ONE*, vol. 4, no. 3, 2009, Art. no. e4920.

[41] B. Malysiak-Mrozek and D. Mrozek, "An improved method for protein similarity searching by alignment of fuzzy energy signatures," *Int. J. Comput. Intell. Syst.*, vol. 4, no. 1, pp. 75–88, Feb. 2011.

[42] X. Zhao, J. Zhang, Q. Ning, P. Sun, Z. Ma, and M. Yin, "Identification of protein pupylation sites using bi-profile bayes feature extraction and ensemble learning," *Math. Problems Eng.*, vol. 2013, Oct. 2013, Art. no. 283129.

[43] W. He, C. Jia, Y. Duan, and Q. Zou, "70ProPred: A predictor for discovering sigma70 promoters based on combining multiple features," *BMC Syst. Biol.*, vol. 12, no. S4, p. 44, Apr. 2018.

[44] Z. Ju and S.-Y. Wang, "Predicting lysine lipoylation sites using bi-profile bayes feature extraction and fuzzy support vector machine algorithm," *Anal. Biochem.*, vols. 561–562, pp. 11–17, Nov. 2018.

[45] L. Dou, X. Li, H. Ding, L. Xu, and H. Xiang, "Is there any sequence feature in the RNA pseudouridine modification prediction problem?" *Mol. Therapy Nucleic Acids*, vol. 19, pp. 293–303, Mar. 2020.

[46] T. Li, R. Song, Q. Yin, M. Gao, and Y. Chen, "Identification of S-nitrosylation sites based on multiple features combination," *Sci. Rep.*, vol. 9, Feb. 2019, Art. no. 3098.

[47] M. Zhang, F. Li, T. T. Marquez-Lago, A. Leier, C. Fan, C. K. Kwoh, K.-C. Chou, J. Song, and C. Jia, "MULTiPly: A novel multi-layer predictor for discovering general and specific types of promoters," *Bioinformatics*, vol. 35, no. 17, pp. 2957–2965, Sep. 2019.

[48] X. Zhao, Q. Jiao, H. Li, Y. Wu, H. Wang, S. Huang, and G. Wang, "ECFS-DEA: An ensemble classifier-based feature selection for differential expression analysis on expression profiles," *BMC Bioinf.*, vol. 21, no. 1, p. 43, Dec. 2020.

[49] J. Yin, W. Sun, F. Li, J. Hong, X. Li, Y. Zhou, Y. Lu, M. Liu, X. Zhang, N. Chen, X. Jin, J. Xue, S. Zeng, L. Yu, and F. Zhu, "VARIDT 1.0: Variability of drug transporter database," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D1042–D1050, Jan. 2020.

[50] Z. Chen, P. Zhao, F. Li, A. Leier, T. T. Marquez-Lago, Y. Wang, G. I. Webb, A. I. Smith, R. J. Daly, K.-C. Chou, and J. Song, "IFeature: A Python package and Web server for features extraction and selection from protein and peptide sequences," *Bioinformatics*, vol. 34, no. 14, pp. 2499–2502, Jul. 2018.

[51] A. S. Nair and S. P. Sreenadhan, "A coding measure scheme employing electron-ion interaction pseudopotential (EIIP)," *Bioinformation*, vol. 1, no. 6, pp. 197–202, 2006.

[52] T. Bojiá, V. R. Perovi, M. Senłanski, and S. Gliłiá, "Identification of candidate allosteric modulators of the m1 muscarinic acetylcholine receptor which may improve vagus nerve stimulation in chronic tinnitus," *Frontiers Neurosci.*, vol. 11, p. 636, Nov. 2017.

[53] S. Han, Y. Liang, Q. Ma, Y. Xu, Y. Zhang, W. Du, C. Wang, and Y. Li, "LncFinder: An integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property," *Briefings Bioinf.*, vol. 20, no. 6, pp. 2009–2027, Nov. 2019.

[54] W. He, C. Jia, and Q. Zou, "4mCPred: Machine learning methods for DNA N4-methylcytosine sites prediction," *Bioinformatics*, vol. 35, no. 4, pp. 593–601, Feb. 2019.

[55] B. Liu, "BioSeq-analysis: A platform for DNA, RNA and protein sequence analysis based on machine learning approaches," *Briefings Bioinf.*, vol. 20, no. 4, pp. 1280–1294, Jul. 2019.

[56] F. Golabi, M. Shamsi, M. H. Sedaaghi, A. Barzegar, and M. S. Hejazi, "Development of a new oligonucleotide block location-based feature extraction (BLBFE) method for the classification of riboswitches," *Mol. Genet. Genomics*, vol. 295, no. 2, pp. 525–534, Mar. 2020.

[57] M. Zhang, J.-W. Sun, Z. Liu, M.-W. Ren, H.-B. Shen, and D.-J. Yu, "Improving N6-methyladenosine site prediction with heuristic selection of nucleotide physical–chemical properties," *Anal. Biochemistry*, vol. 508, pp. 104–113, Sep. 2016.

[58] F. Pedregosa, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[59] H. Zhang, Y.-L. Kang, Y.-Y. Zhu, K.-X. Zhao, J.-Y. Liang, L. Ding, T.-G. Zhang, and J. Zhang, "Novel naïve bayes classification models for predicting the chemical ames mutagenicity," *Toxicology Vitro*, vol. 41, pp. 56–63, Jun. 2017.

[60] H. Zhang, Z.-X. Cao, M. Li, Y.-Z. Li, and C. Peng, "Novel naïve bayes classification models for predicting the carcinogenicity of chemicals," *Food Chem. Toxicol.*, vol. 97, pp. 141–149, Nov. 2016.

[61] N. Kosylo, J. Smith, M. Conover, L. Chan, H. Zhang, H. Mei, and R. Cao, "Artificial intelligence on job-hopping forecasting: AI on job-hopping," in *Proc. Portland Int. Conf. Manage. Eng. Technol. (PICMET)*, Aug. 2018, pp. 1–5.

[62] P.-M. Feng, H. Ding, W. Chen, and H. Lin, "Naïve bayes classifier with feature selection to identify phage virion proteins," *Comput. Math. Methods Med.*, vol. 2013, Mar. 2013, Art. no. 530696.

[63] B. K. Sarkar, "Hybrid model for prediction of heart disease," *Soft Comput.*, vol. 24, no. 3, pp. 1903–1925, Feb. 2020.

[64] J. Yu, Z. Xuan, X. Feng, Q. Zou, and L. Wang, "A novel collaborative filtering model for LncRNA-disease association prediction based on the Naïve Bayesian classifier," *BMC Bioinf.*, vol. 20, no. 1, p. 396, Dec. 2019.

[65] Y. Ding, J. Tang, and F. Guo, "Identification of protein–protein interactions via a novel matrix-based sequence representation model with amino acid contact information," *Int. J. Mol. Sci.*, vol. 17, no. 10, p. 1623, 2016.

[66] Y. Hu, M. Zhou, H. Shi, H. Ju, Q. Jiang, and L. Cheng, "Measuring disease similarity and predicting disease-related ncRNAs by a novel method," *BMC Med. Genomics*, vol. 10, no. S5, p. 71, Dec. 2017.

[67] J. Pirgazi, A. R. Khanteymoori, and M. Jalilkhani, "GENIRF: An algorithm for gene regulatory network inference using rotation forest," *Current Bioinf.*, vol. 13, no. 4, pp. 407–419, Jul. 2018.

[68] L. Cheng, Y. Jiang, H. Ju, J. Sun, J. Peng, M. Zhou, and Y. Hu, "InfAcrOnt: Calculating cross-ontology term similarities using information flow by a random walk," *BMC Genomics*, vol. 19, no. S1, p. 919, Jan. 2018.

[69] L. Xu, G. Liang, C. Liao, G.-D. Chen, and C.-C. Chang, "K-Skip-n-Gram-RF: A random forest based method for Alzheimer's disease protein identification," *Frontiers Genet.*, vol. 10, p. 33, Feb. 2019.
[70] P. J. Moore, T. J. Lyons, and J. Gallacher, "Random forest prediction of Alzheimer's disease using pairwise selection from time series data," *PLoS ONE*, vol. 14, no. 2, 2019, Art. no. e0211558.
[71] D. Yao, X. Zhan, and C.-K. Kwoh, "An improved random forest-based computational model for predicting novel miRNA-disease associations," *BMC Bioinf.*, vol. 20, no. 1, p. 624, Dec. 2019.
[72] Q. Liu, W. Shi, and Z. Chen, "Rubber fatigue life prediction using a random forest method and nonlinear cumulative fatigue damage model," *J. Appl. Polym. Sci.*, vol. 137, no. 14, p. 48519, Apr. 2020.
[73] B. Ma, Y. Geng, F. Meng, G. Yan, and F. Song, "Identification of a sixteen-gene prognostic biomarker for lung adenocarcinoma using a machine learning method," *J. Cancer*, vol. 11, no. 5, pp. 1288–1298, 2020.
[74] Y. Zhou, Q. Cui, and Y. Zhou, "NmSEER V2.0: A prediction tool for 2'-O-methylation sites based on random forest and multi-encoding combination," *BMC Bioinf.*, vol. 20, no. S25, p. 690, Dec. 2019.
[75] Z. Lv, S. Jin, H. Ding, and Q. Zou, "A random forest sub-Golgi protein classifier optimized via dipeptide and amino acid composition features," *Frontiers Bioeng. Biotechnol.*, vol. 7, p. 215, 2019.
[76] X. Ru, L. Li, and Q. Zou, "Incorporating distance-based Top-n-gram and random forest to identify electron transport proteins," *J. Proteome Res.*, vol. 18, no. 7, pp. 2931–2939, Jul. 2019.
[77] Q. Yang, "Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data," *Brief Bioinform.*, Jun. 2019, doi: 10.1093/bib/bbz049.
[78] Y. Ding, J. Tang, and F. Guo, "Predicting protein-protein interactions via multivariate mutual information of protein sequences," *BMC Bioinf.*, vol. 17, no. 1, p. 398, Dec. 2016.
[79] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
[80] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[81] J. S.-T. N. Cristianini, *An Introduction of Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000, pp. 93–124.
[82] Q. Jiang, G. Wang, S. Jin, Y. Li, and Y. Wang, "Predicting human microRNA-disease associations based on support vector machine," *Int. J. Data Mining Bioinf.*, vol. 8, no. 3, p. 282–293, 2013.
[83] Y. Ding, J. Tang, and F. Guo, "Identification of drug-target interactions via multiple information integration," *Inf. Sci.*, vols. 418–419, pp. 546–560, Dec. 2017.
[84] L. Xu, G. Liang, S. Shi, and C. Liao, "SeqSVM: A sequence-based support vector machine method for identifying antioxidant proteins," *Int. J. Mol. Sci.*, vol. 19, no. 6, p. 1773, 2018.
[85] X. Xiao, Z.-C. Xu, W.-R. Qiu, P. Wang, H.-T. Ge, and K.-C. Chou, "IPSW(2L)-PseKNC: A two-layer predictor for identifying promoters and their strength by hybrid features via pseudo K-tuple nucleotide composition," *Genomics*, vol. 111, no. 6, pp. 1785–1793, Dec. 2019.
[86] Z. Liao, D. Li, X. Wang, L. Li, and Q. Zou, "Cancer diagnosis through IsomiR expression with machine learning method," *Current Bioinf.*, vol. 13, no. 1, pp. 57–63, Feb. 2018.
[87] L. Wei, R. Su, B. Wang, X. Li, Q. Zou, and X. Gao, "Integration of deep feature representations and handcrafted features to improve the prediction of N6-methyladenosine sites," *Neurocomputing*, vol. 324, pp. 3–9, Jan. 2019.
[88] N. Zhang, Y. Sa, Y. Guo, W. Lin, P. Wang, and Y. Feng, "Discriminating ramos and jurkat cells with image textures from diffraction imaging flow cytometry based on a support vector machine," *Current Bioinf.*, vol. 13, no. 1, pp. 50–56, Feb. 2018.
[89] B. Liu and K. Li, "IPromoter-2L2.0: Identifying promoters and their types by combining smoothing cutting window algorithm and sequence-based features," *Mol. Therapy Nucleic Acids*, vol. 18, pp. 80–87, Dec. 2019.
[90] S. C.-C. Chen, C.-M. Lo, S.-H. Wang, and E. C.-Y. Su, "RNA editing-based classification of diffuse gliomas: Predicting isocitrate dehydrogenase mutation and chromosome 1p/19q codeletion," *BMC Bioinf.*, vol. 20, no. S19, p. 659, Dec. 2019.
[91] S. Basith, B. Manavalan, T. H. Shin, and G. Lee, "SDM6A: A Web-based integrative machine-learning framework for predicting 6 mA sites in the rice genome," *Mol. Therapy Nucleic Acids*, vol. 18, pp. 131–141, Dec. 2019.
[92] Z.-Y. Zhang, Y.-H. Yang, H. Ding, D. Wang, W. Chen, and H. Lin, "Design powerful predictor for mRNA subcellular location prediction in homo sapiens," *Briefings Bioinf.*, Jan. 2020, doi: 10.1093/bib/bbz177.
[93] L. Wei, C. Zhou, H. Chen, J. Song, and R. Su, "ACPred-FL: A sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides," *Bioinformatics*, vol. 34, no. 23, pp. 4007–4016, 2018.
[94] L. Chao, L. Wei, and Q. Zou, "SecProMTB: A SVM-based classifier for secretory proteins of mycobacterium tuberculosis with imbalanced data Set," *Proteomics*, vol. 19, Oct. 2019, Art. no. e1900007.
[95] J. Tang, J. Fu, Y. Wang, B. Li, Y. Li, Q. Yang, X. Cui, J. Hong, X. Li, Y. Chen, W. Xue, and F. Zhu, "ANPELA: Analysis and performance assessment of the label-free quantification workflow for metaproteomic studies," *Briefings Bioinf.*, vol. 21, no. 2, pp. 621–636, Mar. 2020.
[96] R. Su, H. Wu, B. Xu, X. Liu, and L. Wei, "Developing a multi-dose computational model for drug-induced hepatotoxicity prediction based on toxicogenomics data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 4, pp. 1231–1239, Jul. 2019.
[97] Y.-W. Chu, K.-P. Chang, C.-W. Chen, Y.-T. Liang, Z. T. Soh, and L. Hsieh, "MiRgo: Integrating various off-the-shelf tools for identification of micro RNA–target interactions by heterogeneous features and a novel evaluation indicator," *Sci. Rep.*, vol. 10, no. 1, Dec. 2020, Art. no. 1466.
[98] Q. Huang, Y. Chen, L. Liu, D. Tao, and X. Li, "On combining biclustering mining and AdaBoost for breast tumor classification," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 4, pp. 728–738, Apr. 2020.
[99] X. Yang, S. Yang, Q. Li, S. Wuchty, and Z. Zhang, "Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method," *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 153–161, 2020.
[100] Z. Wang, W. He, J. Tang, and F. Guo, "Identification of highest-affinity binding sites of yeast transcription factor families," *J. Chem. Inf. Model.*, vol. 60, no. 3, pp. 1876–1883, Mar. 2020.
[101] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.
[102] E. Keogh and A. Mueen, "Curse of dimensionality," in *Encyclopedia of Machine Learning*, C. Sammut G. I. Webb, Eds. Boston, MA, USA: Springer, 2010, pp. 257–258.
[103] P. Ramzi, F. Samadzadegan, and P. Reinartz, "Classification of hyperspectral data using an AdaBoostSVM technique applied on band clusters," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2066–2079, Jun. 2014.
[104] H. Liu, M. Zhou, and Q. Liu, "An embedded feature selection method for imbalanced data classification," *IEEE/CAA J. Automatica Sinica*, vol. 6, no. 3, pp. 703–715, May 2019.
[105] L. Ali, C. Zhu, N. A. Golilarz, A. Javeed, M. Zhou, and Y. Liu, "Reliable Parkinson's disease detection by analyzing handwritten drawings: Construction of an unbiased cascaded learning system based on feature selection and adaptive boosting model," *IEEE Access*, vol. 7, pp. 116480–116489, 2019.
[106] R. A. Alexander and D. M. Govern, "A new and simpler approximation for ANOVA under variance heterogeneity," *J. Educ. Statist.*, vol. 19, no. 2, pp. 91–101, 1994.
[107] G. Wang, Y. Wang, W. Feng, X. Wang, J. Y. Yang, Y. Zhao, Y. Wang, and Y. Liu, "Transcription factor and microRNA regulation in androgen-dependent and -independent prostate cancer cells," *BMC Genomics*, vol. 9, no. 2, p. S22, 2008.
[108] Y. Zhao, F. Wang, S. Chen, J. Wan, and G. Wang, "Methods of MicroRNA promoter prediction and transcription factor mediated regulatory network," *BioMed Res. Int.*, vol. 2017, Mar. 2017, Art. no. 7049406
[109] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[110] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," *ACM Comput. Surv.*, vol. 49, no. 2, pp. 1–50, Nov. 2016.
[111] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," *ACM Comput. Surv.*, vol. 52, no. 4, pp. 1–36, Sep. 2019.
[112] X. Ru, P. Cao, L. Li, and Q. Zou, "Selecting essential MicroRNAs using a novel voting method," *Mol. Therapy Nucleic Acids*, vol. 18, pp. 16–23, Dec. 2019.
[113] B. Małysiak-Mrozek, T. Baron, and D. Mrozek, "Spark-IDPP: high-throughput and scalable prediction of intrinsically disordered protein regions with spark clusters on the cloud," *Cluster Comput.*, vol. 22, no. 2, pp. 487–508, Jun. 2019.

[114] B. Wang, K. Lu, X. Zheng, B. Su, Y. Zhou, P. Chen, and J. Zhang, "Early stage identification of Alzheimer's disease using a two-stage ensemble classifier," *Current Bioinf.*, vol. 13, no. 5, pp. 529–535, Sep. 2018.
[115] B. Li, J. Tang, Q. Yang, S. Li, X. Cui, Y. Li, Y. Chen, W. Xue, X. Li, and F. Zhu, "NOREVA: Normalization and evaluation of MS-based metabolomics data," *Nucleic Acids Res.*, vol. 45, no. W1, pp. W162–W170, Jul. 2017.
[116] Y. Wang, S. Zhang, F. Li, Y. Zhou, Y. Zhang, Z. Wang, R. Zhang, J. Zhu, Y. Ren, Y. Tan, C. Qin, Y. Li, X. Li, Y. Chen, and F. Zhu, "Therapeutic target database 2020: Enriched resource for facilitating research and early development of targeted therapeutics," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D1031–D1041, Jan. 2020.
[117] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 20–29, Jun. 2004.
[118] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[119] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Trans. Syst., Man, Cybern.*, vols. SMC–2, no. 3, pp. 408–421, Jul. 1972.
[120] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*. New York, NY, USA: Association for Computing Machinery, 2006, doi: 10.1145/1143844.1143874.
[121] G. Wang, Y. Wang, M. Teng, D. Zhang, L. Li, and Y. Liu, "Signal transducers and activators of Transcription-1 (STAT1) regulates microRNA transcription in interferon μ-Stimulated HeLa cells," *PLoS ONE*, vol. 5, no. 7, 2010, Art. no. e11794.
[122] L. Wei, J. Tang, and Q. Zou, "Local-DPP: An improved DNA-binding protein prediction method by exploring local evolutionary information," *Inf. Sci.*, vol. 384, pp. 135–144, Apr. 2017.
[123] X. Qiang, C. Zhou, X. Ye, P.-F. Du, R. Su, and L. Wei, "CPPred-FL: A sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learning," *Briefings Bioinf.*, Sep. 2018.
[124] J. Fu, J. Tang, Y. Wang, X. Cui, Q. Yang, J. Hong, X. Li, S. Li, Y. Chen, W. Xue, and F. Zhu, "Discovery of the consistently well-performed analysis chain for SWATH-MS based pharmacoproteomic quantification," *Frontiers Pharmacol.*, vol. 9, p. 681, Jun. 2018, doi: 10.3389/fphar.2018.00681.
[125] L. Wei, P. Xing, J. Zeng, J. Chen, R. Su, and F. Guo, "Improved prediction of protein–protein interactions using novel negative samples, features, and an ensemble classifier," *Artif. Intell. Med.*, vol. 83, pp. 67–74, Nov. 2017.
[126] Y. Ding, J. Tang, and F. Guo, "Identification of drug-side effect association via multiple information integration with centered kernel alignment," *Neurocomputing*, vol. 325, pp. 211–224, Jan. 2019.
[127] L. Cheng, P. Wang, R. Tian, S. Wang, Q. Guo, M. Luo, W. Zhou, G. Liu, H. Jiang, and Q. Jiang, "LncRNA2Target v2.0: A comprehensive database for target genes of lncRNAs in human and mouse," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D140–D144, Jan. 2019.

LIJUN DOU received the Ph.D. degree in atomic and molecular physics from the Institute of Modern Physics, Chinese Academy of Sciences. She is currently a Postdoctoral Researcher with Shenzhen Polytechnic and the University of Electronic Science and Technology of China. Her research interest includes bioinformatics.

XIAOLING LI received the degree from Harbin Medical University. She is currently the Chief Physician with the Heilongjiang Province Land Reclamation Headquarters General Hospital. Her research interest includes the treatment of malignant tumors.

HUI DING received the Ph.D. degree in science and engineering from Inner Mongolia University, in 2012. She is currently an Associate Professor with the Center for Informational Biology, UESTC.

LEI XU received the B.Sc. and M.Sc. degrees from the School of Computer Science and Technology, Harbin Institute of Technology, in 2006 and 2008, respectively, and the Ph.D. degree from the Department of Computing, The Hong Kong Polytechnic University, in October 2013. She is currently an Assistant Professor with the School of Electronic and Communication Engineering, Shenzhen Polytechnic. Her research interests include bioinformatics and pattern recognition.

HUAIKUN XIANG is currently an Associate Professor with the Department of Automotive and Transportation, Shenzhen Polytechnic. His research interests include analysis of resident travel characteristics based on the big data, research on the evaluation of safe driving behavior based on the coupling of driver and vehicle, and vulnerability analysis for urban congested road networks.