Ense-i6mA: Identification of DNA N^6-Methyladenine Sites Using XGB-RFE Feature Selection and Ensemble Machine Learning

Q1 Xueqiang Fan, Bing Lin, Jun Hu, and Zhongyi Guo

***Abstract*—DNA N^6-methyladenine (6mA) is an important epigenetic modification that plays a vital role in various cellular processes. Accurate identification of the 6mA sites is fundamental to elucidate the biological functions and mechanisms of modification. However, experimental methods for detecting 6mA sites are high-priced and time-consuming. In this study, we propose a novel computational method, called Ense-i6mA, to predict 6mA sites. Firstly, five encoding schemes, i.e., one-hot encoding, gcContent, Z-Curve, *K*-mer nucleotide frequency, and *K*-mer nucleotide frequency with gap, are employed to extract DNA sequence features. Secondly, eXtreme gradient boosting coupled with recursive feature elimination is applied to remove noisy features for avoiding over-fitting, reducing computing time and complexity. Then, the best subset of features is fed into base-classifiers composed of Extra Trees, eXtreme Gradient Boosting, Light Gradient Boosting Machine, and Support Vector Machine. Finally, to minimize generalization errors, the prediction probabilities of the base-classifiers are aggregated by averaging for inferring the final 6mA sites results. We conduct experiments on two species, i.e., Arabidopsis thaliana and Drosophila melanogaster, to compare the performance of Ense-i6mA against the recent 6mA sites prediction methods. The experimental results demonstrate that the proposed Ense-i6mA achieves area under the receiver operating characteristic curve values of 0.967 and 0.968, accuracies of 91.4% and 92.0%, and Mathew's correlation coefficient values of 0.829 and 0.842 on two benchmark datasets, respectively, and outperforms several existing state-of-the-art methods.**

***Index Terms*—DNA N^6-methyladenine sites, sequence-based encoding, bioinformatics, feature selection, ensemble learning.**

I. Introduction

DNA N^6-methyladenine (6mA) refers to the modification of introducing a methyl (CH3) group to the sixth position of an adenine ring catalyzed by DNA methyltransferases [1], [2], [3], [4]. 6mA is an important epigenetic modification that does not change the DNA segment but could alter the role of DNA molecules. It plays a crucial role in a wide variety of biological processes, such as gene expression regulation, regulating gene transcription, DNA repair and replication, cell division and differentiation, and etc [2], [4], [5], [6], [7]. However, these biological function of 6mA in eukaryotes, especially higher eukaryotes, remain largely unclear due to the distribution patterns of 6mA are rather species-specific which result in diverse functional roles [8]. Locating genomic 6mA distributions is fundamental for the elucidation of potential biological functions of DNA 6mA modification.

Accurate identification of 6mA sites in the genome is the most important step to facilitate the characterization of 6mA distribution patterns and further functional analysis. To this end, a number of experimental methods are applied to detect 6mA sites of DNA, e.g., methylated DNA immunoprecipitation sequencing [9], liquid chromatography coupled with tandem mass spectrometry [10], and single-molecule real-time sequencing [11]. However, these methods are time-consuming and laborious. Due to the important of 6mA and the difficulty in experimentally identifying 6mA sites, together with the fact that a large amount of unannotated DNA sequences have quickly accumulated, the development of computational methods for the fast prediction of 6mA sites solely from DNA sequence has become a hot topic in bioinformatics.

Extracting effective features from DNA sequences is considered the most important step in developing accurate computational methods to predict 6mA sites. During the recent years, a series of computational methods have emerged for predicting 6mA sites. According to feature attributes being extracted from sequence, the features used by the existing identification of 6mA sites methods can be roughly divided into three categories, i.e., physicochemical properties [12], [13], sequence information [14], [15], and evolutionary information [16], [17]. Most current methods, e.g., SpineNet-6mA [18], iDNA6mA (5-step rule) [19], Deep6mA [20], LA6mA [21], AL6mA [21], and I-DNAN6mA, solely utilize one-hot encoding (OHE) to extract sequence information for predicting 6mA sites. Unlike these methods, i6mA-vote [22] introduces one-hot encoding method for dinucleotides (One-hot2) to extract sequence information for the first time. i6mA-DNC [23] uses dinucleotide representation method to extract sequence information. To our

Manuscript received 20 November 2022; accepted 24 June 2024. This work was supported by the National Natural Science Foundation of China under Grant 61775050 and Grant 61902352. *(Corresponding authors: Jun Hu; Zhongyi Guo.)* Q2

Xueqiang Fan, Bing Lin, and Zhongyi Guo are with the School of Computer and Information, Hefei University of Technology, Hefei 230009, China (e-mail: fanxueqiang2022b@163.com; linbing2021s@163.com; guozhongyi@hfut.edu.cn).

Jun Hu is with the College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China (e-mail: hujunum@zjut.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TCBB.2024.3421228, provided by the authors.

Digital Object Identifier 10.1109/TCBB.2024.3421228

knowledge, i6mA-Pred [24] is the first computational method of 6mA sites identification that uses chemical properties with respect to amino/keto bases, strong/weak hydrogen bond, and ring structures (RFHC), and position-specific nucleotide frequencies (PPNF) to obtain physicochemical properties and sequence information in DNA sequence, respectively. Besides RFHC, i6mA-stack [25] also utilizes Dinucleotide Physicochemical Properties (DPCP), Trinucleotide Physicochemical Properties (TPCP), and Electron-Ion-Interaction Pseudo Potentials of Nucleotides (EIIP), and one-hot encoding (OHE) to dig out physicochemical properties and sequence information, respectively. To extract evolutionary information from sequences, MM-6mAPred [16] uses a 1st-order Markov model (MM) that indicates the transition probability between adjacent nucleotides for identifying 6mA sites. In addition to choosing an appropriate feature extraction scheme, another key factor for success of 6mA sites identification is the choice of classification algorithms.

Appropriate classification algorithms can speed up training and efficiently learn the relationship between features and labels. A wide variety of machine learning algorithms are used to predict 6mA sites, such as Support Vector Machine (SVM) [26], eXtreme Gradient Boosting (XGB) [27], Logistic Regression (LR) [28], Bagging [29], Random Forest (RF) [30], Fully-Connected Neural Networks (FCN) [31], Convolutional Neural Networks (CNN) [32], Bidirectional Long Short-Term Memory Recurrent Neural Networks (BiLSTM) [33], and etc. i6mA-Pred [24] combines the SVM classifier with RFHC and PPNF to learn 6mA sites prediction model. It is observed that i6mA-Pred reaches an accuracy of 83.13% in the jackknife test on the rice genome. Unlike i6mA-Pred, i6mA-DNC [23] and iDNA6mA (5-step rule) [19] use CNN and FCN to predict 6mA sites. i6mA-DNC and iDNA6mA (5-step rule) obtain 86.64% and 88.60% of accuracy on the rice genome. In Deep6mA [20], OHE is fed into an ensemble of three neural network units, i.e., CNN, BiLSTM, and FCN, to train the prediction model of 6mA sites and Deep6mA accurately predicts 6mA sites. In LA6mA and AL6mA [21], BiLSTM and self-attention mechanism are used to capture discriminative information from OHE for predicting 6mA sites. The accuracies of LA6mA and AL6mA reach 91.5% and 87.8%, and 90.9% and 88.4% in the Drosophila melanogaster and Arabidopsis thaliana genome, respectively. Nevertheless, despite the efficiency and accuracy achieved, the running speed and performance of 6mA sites prediction methods remain room for further improvements.

(i) The influence of DNA sequence features on 6mA sites prediction is not fully elucidated. It is still improved in 6mA sites prediction by extracting features based on DNA sequences. (ii) By revisiting existing 6mA sites identification methods, it was found that all of them employ fused feature generated in series with multiple single-view features directly as input of the machine learning algorithms. Although the usage of single-view feature or fused multi-view features can fully represent the information contained in the DNA sequence, in most of the cases it introduces redundant or irrelevant information inevitably that will seriously reduce the efficiency of 6mA prediction model. Hence, eliminating noise in the feature is also an important step in the process of 6mA sites identification. (iii) Facing the avalanche of new DNA sequences produced in the post-genomic era, choosing an effective classifier is also a major challenge for researchers.

To address the important issues mentioned above, in this study, we propose a novel 6mA sites prediction method, termed Ense-i6mA. Firstly, two benchmark datasets are collected and each DNA sequence is encoded into OHE, K-mer nucleotide frequency (KNF) [34], gcContent [35], [36], Z-Curve [37], [38], and K-mer nucleotide frequency with gaps (KNFG) [15]. Compared to the single-view feature, the fusion feature can obtain more comprehensive DNA information. Secondly, the XGB coupled with recursive feature elimination (XGB-RFE) is applied to 6mA sites prediction to remove noisy features for avoiding over-fitting, reducing computing time and complexity. Finally, an ensemble classifier consisting of two stages is used as the final classifier. In the first phase, four base-classifiers, i.e., Extra Trees (ET), SVM, XGB, and Light Gradient Boosting Machine (LGBM), are selected from thirteen machine-learning algorithms for the first time. In the second phase, to minimize generalization errors, the prediction probabilities of the base-classifiers are aggregated by averaging for inferring the final 6mA sites results. We conduct experiments on two benchmark datasets to compare the performance of Ense-i6mA against the recent 6mA sites prediction methods. Benchmarking results demonstrate that Ense-i6mA yields substantial performance achieve over previous methods, highlighting its promising potential in solving the 6mA sites prediction problem. Finally, based on the proposed Ense-i6mA, we implement a new standalone-version predictor for predicting 6mA sites, which is freely available at https://github.com/XueQiangFan/Ense-i6mA for academic use.

II. Materials and Methods

A. Benchmark Datasets

To evaluate the performance of our proposed I-DNAN6mA, in this study, we chose two well-known datasets contained the DNA 6mA sites data for two species i.e., Arabidopsis thaliana and Drosophila melanogaster, which are previously employed to assess the 6mA sites prediction models in the recently published studies [21], [34] as the benchmark datasets. These raw DNA data of Arabidopsis thaliana and Drosophila melanogaster are collected from the PacBio public database [35]. For each organism, Zhang et al. randomly divides it into the training and independent testing subset at a ratio of 9:1. The number of positive and negative samples is the same for each subset. For more detailed information on the dataset construction, please refer to [21], [34]. All the datasets can be downloaded from https://github.com/XueQiangFan/Ense-i6mA. Finally, the number of samples included in each dataset is shown in Supplemental Table S1.

B. Feature Extraction

Extracting effective features from DNA sequences which contain significant discriminatory information is considered the most important step in developing accurate computational methods to predict 6mA sites. To encode the DNA sequences into vectors recognized by machine-learning, given a DNA sequence with 41-nt, five encoding schemes, i.e., OHE, gcContent [36], [37], Z-Curve [38], [39], KNF [40], and KNFG [15], are used to extract DNA sequence features:

- Every DNA sequence transformed into a 41 × 4 matrix (total 164 elements) after one-hot coding.
- Generally, DNAs with high gcContent scores is more stable than DNA with low gcContent scores. gcContent calculated by:

$$\mathrm{gcContent} = \frac{\sum_i^L C + \sum_i^L G}{\sum_i^L A + \sum_i^L C + \sum_i^L G + \sum_i^L T} \tag{1}$$

- Z-Curve theory is often used in genomic sequence analysis. Each sequence is represented by three elements. It is defined as following:

$$\mathrm{Z-Curve} = [\mathrm{x},\ \mathrm{y},\ \mathrm{z}] \tag{2}$$

$$\begin{cases} x = \left(\sum_i^L A + \sum_i^L G\right) - \left(\sum_i^L C + \sum_i^L T\right) \\ y = \left(\sum_i^L A + \sum_i^L C\right) - \left(\sum_i^L G + \sum_i^L T\right) \\ z = \left(\sum_i^L A + \sum_i^L T\right) - \left(\sum_i^L G + \sum_i^L C\right) \end{cases} \tag{3}$$

- KNF (total 84 elements), which reflects the sequence background differences between the 6mA sites and non-6mA sites, used to calculate the frequencies of adjacent nucleotides in the DNA sequence. In this study, K values are set 1, 2, and 3.
- KNFG (total 720 elements) generated by PyFeat tool [15], a python-based feature generation tool for DNA, RNA and protein sequences.

The detail steps of generating the above descriptors are described in Supplementary Text S1.

C. Feature Selection Using XGB-RFE

A pre-requisite in developing powerful computational models for 6mA sites prediction is to extract sufficient discriminative features to construct accurate models. By visiting existing 6mA sites prediction methods, most of the methods use a variety of coding strategies to generate more DNA features that in most of the cases introduce redundant or irrelevant information inevitably, and at the same time produce feature sparsity problem. Therefore, it will eventually result in over-fitting issue and reducing the generalization capacity of the prediction model. Feature selection which can enhance the performance of the prediction by selecting optimum features, is one of the effective techniques in diverse domains, e.g., pattern recognition, machine learning, and bioinformatics, to remove the noisy information from the actual data.

To find out which features are most suitable to identify 6mA sites, the eXtreme Gradient Boosting (XGB), coupled with recursive feature elimination (RFE) algorithm, is employed to score different meta features and select the optimal meta features to construct the best subset of features (BFS). In this study, XGB and RFE (XGB-RFE) are combined for the first time in the field of 6mA sites identification. Specifically, BFS can be constructed with the following three steps [27], [41]:

Step 1: Sequence-based Feature Encoding

Given a DNA sequence with 41 nucleotides, five encoding schemes, i.e., ONE, gcContent, Z-Curve, KNF, and KNFG, are used to encode 164, 1, 3, 84, and 720-dimensional vectors, respectively. The five types of features are fused to engender a new feature group, which consists of a total of 972-dimensional features for each sequence. The fused feature groups and labels for all sequence constitute a sample dataset ***D***:

$$\boldsymbol{D} = \left\{\left(\xi^1, \eta^1\right), \left(\xi^2, \eta^2\right), \ldots,\ \left(\xi^i, \eta^i\right), \ldots, \left(\xi^n, \eta^n\right)\right\} \tag{4}$$

where *n* is the total number of samples; the element

$$\left(\xi^i, \eta^i\right) = \left[x_1^i, x_2^i, \ldots, x_j^i, \ldots x_{972}^i, y^i\right] \tag{5}$$

means that the *i*-th DNA sequence contains 972 features and a label y^i.

Step 2: Feature Importance Ranking and Elimination of Junk Features

A tree ensemble model, i.e., XGB, uses *M* additive functions to predict the 6mA sites.

$$\tilde{\boldsymbol{y}}^{\boldsymbol{i}} = \sum_{\boldsymbol{m}=\boldsymbol{1}}^{\boldsymbol{M}} \boldsymbol{f}_{\boldsymbol{m}}\left(\boldsymbol{\xi}^{\boldsymbol{i}}\right) \tag{6}$$

where $\boldsymbol{f}_{\boldsymbol{m}}(\boldsymbol{\xi}^{\boldsymbol{i}})$ denotes the importance score of *i*-th feature vector on *m*-th tree. Thus, the objective function can be expressed as:

$$\boldsymbol{O}\left(\emptyset\right) = \sum_i o\left(\tilde{y}^i, y^i\right) + \gamma \tag{7}$$

where $o(\tilde{y}^i, y^i)$ means the loss between the predicted and ground truth values; $\gamma = \sum_m \omega(f_m)$, $\omega(\cdot)$ controls the complexity of the model. Then, the objective function becomes as follows after one iteration generate a tree:

$$\boldsymbol{O}(\emptyset)_{(\boldsymbol{t})} = \sum_i o\left[\left(\tilde{y}_{(t-1)}^i + f_{(t)}\left(\boldsymbol{\xi}^{\boldsymbol{i}}\right)\right), y^i\right] + \gamma \tag{8}$$

where $\tilde{y}_{(t-1)}^i + f_{(t)}(\boldsymbol{\xi}^{\boldsymbol{i}})$ represents the predicted value of *t*-th iteration. Assuming that the *m-1*-th tree weight is known while producing the *m*-th tree.

$$\boldsymbol{O}_{(\boldsymbol{t})} = \sum_{i=1}^{T}\left[o\left(\tilde{y}_{(t-1)}^i, y^i\right) + \delta_i f_{(t)}\left(\boldsymbol{\xi}^{\boldsymbol{i}}\right) + \frac{1}{2}\mu_i f_{(t)}^2\left(\boldsymbol{\xi}^{\boldsymbol{i}}\right)\right] + \gamma \tag{9}$$

where $\boldsymbol{O}_{(t)}$ is the objective function; δ_i and μ_i mean the first- and second-order statistics of the loss function, respectively. Obtaining the importance ranking of features, the lowest scoring features are eliminated using RFE from the current feature space and the remaining features are used as the feature dataset $\boldsymbol{D}^*$ for the next iteration.

Step 3: Iterative Optimization

Repeating step 2, the final BFS contained the 80-dimensional most important features is selected from the fused feature group for each sequence.

D. Architecture of Ense-i6mA

Machine learning, especially the ensemble learning has recently been proven to be a fascinating algorithm and successfully applied in a wide variety of computational bioinformatics domains, such as DNA-binding protein [42], ncRNA-protein interactions [43], protein-protein interactions [44], and etc. Ensemble learning combines multiple classifiers and uses a certain rule to integrate a series of learner results to obtain better results than the single classifier. In this study, an ensemble classifier, termed Ense-i6mA, is established to predict 6mA sites. Framework of Ense-i6mA mainly consists of two-phase, including the first stage base-classifier learning and the second stage integrated predicted probabilities.

Considering the different feature learning spaces and class recognition capabilities of different machine learning algorithms, this study expects to choose excellent base classifiers to train the prediction model for identifying 6mA sites. In the first phase, thirteen machine-learning algorithms, Logistic Regression (LR), K-nearest neighbor (KNN), decision tree (DT), Gaussian NB (NB), Bagging, Random Forest (RF), AdaBoost (AB), Gradient Boosting (GB), Linear Discriminant Analysis (LDA), Extra Trees (ET), eXtreme Gradient Boosting (XGB), Light Gradient Boosting Machine (LGBM), and Support Vector Machine (SVM), are investigated by contrast experiments to select base-classifier. These machine learning-based classifiers are implemented and tuned using the Scikit-learn Python library [45]. By comparing the prediction performance of thirteen machine-learning algorithms on the derived data set BFS of the training data set over five-fold cross-validation tests, SVM, XGB, LGBM, and ET are used as base-classifiers. In the second phase, to minimize generalization errors, the prediction probabilities of the base-classifiers are aggregated by averaging to obtain the final 6mA sites probability. Ense-i6mA can mine the essential discrimination features that characterize DNA 6mA sites through ensemble learning, and its prediction performance is superior to that of the individual classifier. The detailed flow of the Ense-i6mA algorithm is presented in the three steps in Algorithm 1.

E. Model Construction

In this study, a novel method is proposed, called Ense-i6mA, for identifying 6mA sites. The flow chart is shown in Fig. 1. All experiments are performed on Windows Server 10 Inter Core i7-9750H CPU @2.60 Hz, 16.0 GB of RAM, and Python 3.7 programming. The detailed steps of Ense-i6mA are described as follows:

1) Collecting two benchmark 6mA sites datasets from previous literatures.
2) Five encoding schemes, i.e., OHE, KNF, Z-Curve, gcContent, and KNFG, are applied to extract DNA feature for given DNA sequence with-41nts. Experimental results show that the fused feature could extract complementary and representative information compared with the single feature.
3) XGB-RFE is utilized to remove noisy features for avoiding over-fitting, speed up training, reducing computing time and complexity. The performance of XGB-RFE with other feature selection methods, i.e., Principal Component Analysis (PCA) [46], SVM-RFE [47], LR-RFE [48], RF-RFE [49], and AB-RFE, is also evaluated by Sn, Sp, ACC, MCC, and auROC.
4) SVM, XGB, LGBM, and ET, algorithms are stacked to build up base-classifiers. The BFS generated by steps (3) are fed into the base-classifiers and the output the 6mA

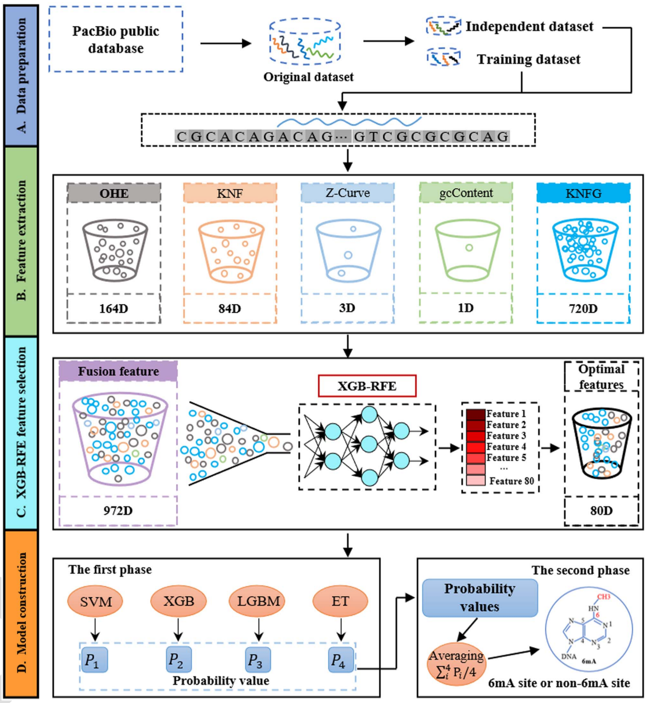


Fig. 1. The overall flow for identifying 6mA sites by Ense-i6mA. (A) Data preparation. (B) Feature extraction. (C) XGB-RFE feature selection. (D) Model construction.

Algorithm 1: Ense-i6mA Algorithm.

```
Input: Dataset D = {$(X_1, Y_1), (X_2, Y_2), \ldots,(X_n, Y_n)$};
       Feature selection FS = XGB-RFE;
       Base-classifiers $c_1$=SVM, $c_2$=XGB, $c_3$=LGBM,
        $c_4$=ET.
Output: ensemble classifier C
    $D^* = \varnothing$;
    Step 1: construct the best subset of features
    for i = 1, 2, 3,. . ., n do
    $X_i^{'}$= FS($X_i$, $Y_i$);
    end for
    $D^*$= {$(X_1^{'}, Y_1), (X_2^{'}, Y_2), \ldots,(X_n^{'}, Y_n)$};
    Step 2: train the base-classifiers
    for t = 1, 2, 3, 4 do
    $h_t$=$c_t(D^*)$;
    end for
    H = {$h_1(x), h_2(x), h_3(x), h_4(x)$};
    Step 3: aggregate results by averaging
    C = $\sum_t^4 h_t/4$;
    return C
```

site probabilities of the base-classifiers are aggregated by averaging for concluding the final results.

5) The effectiveness of Ense-i6mA is validated on two benchmark datasets. The performance of Ense-i6mA with other compared methods, i.e., SVM, XGB, LGBM, ET, GB, DeepM6A, i6mA-DNC, iDNA6mA, 3-mer-LR, LA6mA, and AL6mA, is assessed on the independent testing datasets using Sn, Sp, ACC, MCC, and auROC.

F. Evaluation Metrics

In this study, the performance of the proposed method is assessed by using the following four classical evaluation indexes of binary classification, namely sensitivity (Sn), specificity (Sp), accuracy (ACC) and Mathew's correlation coefficient (MCC), which are respectively expressed as follows:

$$\mathrm{Sn} = 1 - \frac{\alpha^{+}_{-}}{\alpha^{+}} \tag{10}$$

$$\mathrm{Sp} = 1 - \frac{\alpha^{-}_{+}}{\alpha^{-}} \tag{11}$$

$$\mathrm{ACC} = 1 - \frac{\alpha^{+}_{-} + \alpha^{-}_{+}}{\alpha^{+} + \alpha^{-}} \tag{12}$$

$$\mathrm{MCC} = \frac{1 - \frac{\alpha^{+}_{-}+\alpha^{-}_{+}}{\alpha^{+}+\alpha^{-}}}{\sqrt{\left(1 + \frac{\alpha^{-}_{+}-\alpha^{+}_{-}}{\alpha^{+}}\right)\left(1 + \frac{\alpha^{+}_{-}-\alpha^{-}_{+}}{\alpha^{+}}\right)}} \tag{13}$$

where α^{+} (i.e., true positive) is the total number of 6mA sites, α^{+}_{-} is the number of 6mA sites incorrectly predicted as non-6mA sites, α^{-} is the total number of non-6mA sites, α^{-}_{+} is the number of non-6mA sites incorrectly predicted as 6mA sites. MCC measures the correlation between the expected class and the predicted class. The MCC measure ranges from -1 to 1, and the other three evaluation measures range between 0 to 1. Furthermore, this study also uses the receiver operating characteristic (ROC) curve evaluate the performance of the proposed method. The area under the ROC curve (auROC) is a comprehensive indicator of the performance quality of a binary classifier. The value 0.5 of auROC is equivalent to random prediction, while 1 of auROC means a perfect one.

III. Results and Discussions

A Performance Comparison of Different Features

In this section, the discriminative performances of the five sequence-based features and one combination feature of them, i.e., OHE, KNF, Z-Curve, gcContent, KNFG, and the fusion feature, are investigated. Three commonly individual machine learning algorithms, i.e., Logistic Regression (LR), K-nearest neighbor (KNN), and Random Forest (RF), are used to assess each feature by performing five-fold cross-validation tests on the training datasets of Arabidopsis thaliana and Drosophila melanogaster, respectively. Among them, the number of LR iterations is 500, the neighbors of the KNN method are set as 7, and RF sets the number of base decision trees to 500 and the maximum learning depth to 10. Table I summarizes the discriminative average performance results of these features. Supplemental Figs. S1 and S2 demonstrate ROC curves of LR, KNN, and RF algorithms with different features on A.thaliana and D.melanogaster, respectively.

From Table I and Figs. S1 and S2, we can easily find that the fusion feature consistently outperforms other five individual features, i.e., OHE, KNF, Z-Curve, gcContent, and KNFG concerning the five evaluation indexes. Taking the results of the LR algorithm on training dataset A.thaliana as example, the Sn, Sp, ACC, MCC, and auROC of the fusion feature are 0.871, 0.876, 0.873, 0.746, and 0.939, respectively, which are 2.60%, 4.16%, 3.31%, 7.96%, and 3.00% higher than those of the second-best feature, i.e., OHE, respectively. Furthermore, Table I also provides performance comparison of different features in terms of Sn under the fixed Sp (i.e., 0.8 and 0.9). It can be also observed that the fusion feature performed best under fixed Sp in most cases, followed by OHE. These experimental results demonstrate that the five single-view features contain complementary information.

B. Performance Comparison of Different Feature Selection Methods

Choosing one appropriate feature selection method can remove the noise while reducing the feature dimension and selecting the optimal features. In this study, the discriminative performances of six feature selection methods, i.e., PCA, SVM-RFE, LR-RFE, RF-RFE, AB-RFE, and XGB-FRE, are investigated by observing the performances of LR, KNN, and RF algorithms again on training datasets over five-fold cross-validation tests. The optimal features of these feature selection methods with default parameters are set to 100. The prediction results are shown in Table II. Supplemental Figs. S3 and S4 illustrate ROC curves of LR, KNN, and RF algorithms with different feature selection methods on the training datasets over five-fold cross-validation tests, respectively.

Table II shows that the performance of XGB-RFE is superior to that of the other five feature selection methods. Specifically, XGB-RFE with LR, KNN, and RF gains the highest MCC and auROC values, which are two overall measurements of the quality of the binary classification, among all feature selection methods on each training dataset. Taking the results of XGB-RFE with KNN on the training dataset of A.thaliana as an example, XGB-RFE achieves 127.96% and 32.11%, 7.14% and 1.96%, 8.85% and 1.85%, 7.91% and 2.40%, and 15.21% and 4.45% average enhancements of MCC and auROC values, respectively, compared to the other five feature selection methods, i.e., PCA, SVM-RFE, LR-RFE, RF-RFE, and AB-RFE. In addition, XGB-RFE shares the highest Sn, Sp, ACC, Sn (Sp = 0.8), and Sn (Sp = 0.9) values. The numerous experimental results shown in Table II and Figs. S3 and S4 indicate that the performance is indeed enhanced after applying feature selection.

C. Selection of Base Classifiers

To determine the most suitable base classifiers, we evaluate the performance of 13 machine learning classifiers (i.e., LR,

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT FEATURES ON THE TRAINING DATASETS OVER FIVE-FOLD CROSS-VALIDATION TESTS USING THE LR, KNN, AND RF ALGORITHMS

Training dataset	Method	Feature	Sn [a]	Sp [a]	ACC [a]	MCC [a]	auROC [a]	Sn [b]	Sn [c]
Arabidopsis thaliana	LR	OHE	0.849	0.841	0.845	0.691	0.912	0.781	0.873
		KNF	0.676	0.750	0.713	0.428	0.785	0.447	0.613
		Z-Curve	0.644	0.641	0.643	0.286	0.703	0.313	0.469
		gcContent	0.659	0.634	0.646	0.294	0.695	0.287	0.464
		KNFG	0.731	0.788	0.759	0.519	0.838	0.566	0.717
		Fusion	0.871	0.876	0.873	0.746	0.939	0.913	0.845
	KNN	OHE	0.871	0.693	0.783	0.574	0.854	0.630	0.747
		KNF	0.652	0.716	0.684	0.370	0.745	0.395	0.556
		Z-Curve	0.594	0.652	0.623	0.247	0.659	0.244	0.412
		gcContent	0.661	0.554	0.608	0.216	0.647	0.222	0.398
		KNFG	0.662	0.755	0.708	0.419	0.755	0.446	0.602
		Fusion	0.808	0.743	0.776	0.552	0.849	0.743	0.582
	RF	OHE	0.826	0.824	0.825	0.650	0.882	0.679	0.835
		KNF	0.729	0.723	0.726	0.452	0.794	0.461	0.641
		Z-Curve	0.656	0.641	0.648	0.297	0.707	0.312	0.482
		gcContent	0.659	0.634	0.646	0.293	0.695	0.288	0.464
		KNFG	0.745	0.748	0.747	0.493	0.819	0.530	0.682
		Fusion	0.824	0.869	0.846	0.694	0.922	0.869	0.792
Drosophila melanogaster	LR	OHE	0.836	0.871	0.853	0.708	0.923	0.794	0.898
		KNF	0.669	0.715	0.692	0.384	0.754	0.382	0.559
		Z-Curve	0.600	0.636	0.617	0.236	0.664	0.240	0.382
		gcContent	0.591	0.518	0.555	0.109	0.587	0.164	0.288
		KNFG	0.735	0.773	0.754	0.509	0.835	0.520	0.699
		Fusion	0.880	0.891	0.885	0.771	0.948	0.873	0.929
	KNN	OHE	0.798	0.800	0.799	0.699	0.829	0.400	0.798
		KNF	0.701	0.599	0.651	0.302	0.698	0.307	0.463
		Z-Curve	0.573	0.593	0.583	0.166	0.613	0.167	0.317
		gcContent	0.394	0.624	0.507	0.018	0.533	0.124	0.228
		KNFG	0.749	0.589	0.670	0.343	0.737	0.361	0.517
		Fusion	0.886	0.587	0.738	0.500	0.836	0.541	0.712
	RF	OHE	0.677	0.937	0.805	0.634	0.868	0.705	0.783
		KNF	0.682	0.713	0.698	0.400	0.773	0.436	0.593
		Z-Curve	0.624	0.615	0.619	0.239	0.662	0.233	0.392
		gcContent	0.588	0.527	0.558	0.116	0.586	0.154	0.301
		KNFG	0.721	0.730	0.725	0.451	0.806	0.495	0.637
		Fusion	0.826	0.925	0.876	0.755	0.942	0.847	0.917

[a] *Results computed with prediction cutoff threshold value set as 0.5.*
[b] *Results computed with the fixed specificity at 0.9.*
[c] *Results computed with the fixed specificity at 0.8.*

KNN, RF, Decision Tree (DT), Gaussian NB (NB), Bagging, AdaBoost (AB), Gradient Boosting (GB), Linear Discriminant Analysis (LDA), Extra Trees (ET), eXtreme Gradient Boosting (XGB), Light Gradient Boosting Machine (LGBM), and Support Vector Machine (SVM)) on the training datasets over five-fold cross-validation tests. The parameter of 13 machine learning classifiers are as follows, i.e., the number of iterations of LR, XGB, and LGBM is 500; the closest neighbor of KNN is 5; ET and RF set the number of base decision trees to 500 and the maximum learning depth to 10; SVM uses the RBF kernel function; the 'n_estimators' of AB, Bagging, AB, GB, XGB, and LGBM are all set as 500; DT, LDA, and NB use default parameters. These classifiers are implemented using the Scikit-learn Python library [45]. Table III demonstrates the prediction results of 13 classifiers on the training datasets over five-fold cross-validation tests. The ROC curves can be seen in Fig. 2.

According to the MCC and auROC values listed in Table III and the ROC curves presented in Fig. 2, we can find that the five top-ranked classifiers are ET, XGB, LGBM, SVM, and GB, respectively. Concretely, the ET acts as the best performer followed by XGB, LGBM, SVM, and GB. ET is the only classifier to obtain MCC > 0.78 and auROC > 0.95 on both training datasets. It is noted that LGBM gains comparable performance to XGB in terms of MCC and auROC values. The MCC and auROC values of XGB and LGBM classifiers both exceed 0.76 and 0.947, respectively. Furthermore, we observe that the MCC and auROC values of SVM are 1.94%

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT FEATURE SELECTION METHODS ON THE TRAINING DATASETS OVER FIVE-FOLD CROSS-VALIDATION TESTS USING THE LR, KNN, AND RF ALGORITHM

Training dataset	Method	Feature selection method	Sn [a]	Sp [a]	ACC [a]	MCC [a]	auROC [a]	Sn [b]	Sn [c]
Arabidopsis thaliana	LR	-	0.871	0.876	0.873	0.746	0.939	0.913	0.845
		PCA	0.673	0.683	0.678	0.357	0.743	0.381	0.549
		SVM-RFE	0.849	0.869	0.859	0.719	0.928	0.827	0.896
		LR-RFE	0.851	0.872	0.861	0.723	0.923	0.823	0.891
		RF-RFE	0.843	0.856	0.850	0.700	0.921	0.814	0.878
		AB-RFE	0.854	0.866	0.860	0.721	0.927	0.817	0.892
		XGB-RFE	0.854	0.870	0.862	0.736	0.930	0.824	0.893
	KNN	-	0.808	0.743	0.776	0.552	0.849	0.743	0.582
		PCA	0.498	0.814	0.655	0.329	0.710	0.356	0.513
		SVM-RFE	0.896	0.800	0.848	0.700	0.920	0.806	0.896
		LR-RFE	0.903	0.778	0.842	0.689	0.921	0.808	0.891
		RF-RFE	0.869	0.825	0.847	0.695	0.916	0.805	0.879
		AB-RFE	0.861	0.788	0.825	0.651	0.898	0.740	0.849
		XGB-RFE	0.911	0.832	0.872	0.750	0.938	0.832	0.911
	RF	-	0.824	0.869	0.846	0.694	0.922	0.869	0.792
		PCA	0.703	0.713	0.708	0.416	0.775	0.433	0.599
		SVM-RFE	0.834	0.910	0.872	0.746	0.938	0.847	0.904
		LR-RFE	0.837	0.909	0.871	0.748	0.938	0.844	0.900
		RF-RFE	0.823	0.906	0.864	0.732	0.934	0.827	0.889
		AB-RFE	0.823	0.910	0.871	0.744	0.937	0.843	0.904
		XGB-RFE	0.834	0.927	0.881	0.764	0.944	0.859	0.912
Drosophila melanogaster	LR	-	0.880	0.891	0.885	0.771	0.948	0.873	0.929
		PCA	0.624	0.590	0.606	0.212	0.655	0.240	0.398
		SVM-RFE	0.871	0.879	0.875	0.751	0.944	0.852	0.922
		LR-RFE	0.878	0.893	0.885	0.771	0.946	0.869	0.926
		RF-RFE	0.870	0.885	0.877	0.755	0.943	0.856	0.920
		AB-RFE	0.862	0.869	0.865	0.732	0.935	0.837	0.905
		XGB-RFE	0.893	0.900	0.889	0.777	0.950	0.872	0.933
	KNN	-	0.886	0.587	0.738	0.500	0.836	0.541	0.712
		PCA	0.460	0.741	0.598	0.208	0.646	0.227	0.377
		SVM-RFE	0.911	0.794	0.853	0.711	0.926	0.829	0.908
		LR-RFE	0.920	0.777	0.850	0.706	0.911	0.817	0.902
		RF-RFE	0.912	0.784	0.849	0.703	0.928	0.826	0.904
		AB-RFE	0.903	0.664	0.785	0.585	0.859	0.651	0.781
		XGB-RFE	0.932	0.773	0.853	0.716	0.932	0.850	0.932
	RF	-	0.826	0.925	0.876	0.755	0.942	0.847	0.917
		PCA	0.602	0.662	0.631	0.263	0.692	0.275	0.454
		SVM-RFE	0.852	0.927	0.889	0.780	0.951	0.882	0.934
		LR-RFE	0.844	0.910	0.877	0.756	0.942	0.855	0.910
		RF-RFE	0.853	0.933	0.893	0.788	0.950	0.888	0.933
		AB-RFE	0.838	0.893	0.865	0.732	0.936	0.832	0.905
		XGB-RFE	0.861	0.941	0.900	0.804	0.957	0.903	0.941

- means results computed without using feature selection methods.
[a] Results computed with prediction cutoff threshold value set as 0.5.
[b] Results computed with the fixed specificity at 0.9.
[c] Results computed with the fixed specificity at 0.8.

and 0.63%, 2.21% and 0.58% average lower than, respectively, the corresponding values achieved by XGB and LGBM on both training datasets. By revisiting Table III, it is apparent that the Sp values reached by these five classifiers are largely more than the Sn values they achieve. The reason for this is that they predict too many false negatives. Thus, ET, XGB, LGBM, SVM, and GB are provisionally selected as the base classifiers.

To further analyze the combined performance of these 13 machine learning classifiers, we rank these methods using the sum of Z-scores of all evaluation indexes. Fig. 3(a) and (b) show the comprehensive performance of all methods in the A.thaliana and D.melanogaster genome, respectively. It can be found that the comprehensive performance of ET is the best among all methods in both A.thaliana and D.melanogaster genomes, followed by XGB, LGBM, SVM, and GB.

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT MACHINE LEARNING ALGORITHMS ON THE TRAINING DATASETS OVER FIVE-FOLD CROSS-VALIDATION TESTS

Training dataset	Method	Sn [a]	Sp [a]	ACC [a]	MCC [a]	auROC [a]	Sn [b]	Sn [c]
Arabidopsis thaliana	LR	0.848	0.863	0.855	0.710	0.923	0.818	0.891
	KNN	0.896	0.801	0.849	0.701	0.906	0.791	0.896
	DT	0.836	0.825	0.831	0.661	0.831	0.478	0.841
	NB	0.787	0.853	0.820	0.642	0.892	0.728	0.833
	Bagging	0.849	0.875	0.862	0.725	0.923	0.819	0.891
	RF	0.816	0.870	0.843	0.687	0.911	0.774	0.864
	AB	0.856	0.846	0.851	0.703	0.922	0.804	0.885
	LDA	0.848	0.862	0.855	0.711	0.922	0.817	0.886
	SVM	0.849	0.908	0.878	0.758	0.941	0.858	0.911
	XGB	0.853	0.913	0.883	0.767	0.947	0.865	0.921
	LGBM	0.855	0.918	0.886	0.774	0.947	0.872	0.919
	ET	0.853	0.928	0.890	0.783	0.951	0.882	0.926
	GB	0.841	0.883	0.862	0.724	0.928	0.827	0.879
Drosophila melanogaster	LR	0.875	0.887	0.881	0.762	0.946	0.861	0.928
	KNN	0.924	0.794	0.859	0.725	0.916	0.823	0.918
	DT	0.850	0.834	0.842	0.685	0.842	0.512	0.855
	NB	0.810	0.896	0.853	0.709	0.920	0.801	0.881
	Bagging	0.876	0.895	0.886	0.772	0.937	0.870	0.916
	RF	0.822	0.880	0.851	0.703	0.920	0.796	0.879
	AB	0.877	0.872	0.874	0.750	0.943	0.848	0.918
	LDA	0.868	0.885	0.876	0.754	0.946	0.859	0.927
	SVM	0.867	0.925	0.895	0.791	0.955	0.886	0.937
	XGB	0.883	0.929	0.906	0.813	0.961	0.906	0.946
	LGBM	0.877	0.932	0.904	0.810	0.960	0.902	0.945
	ET	0.868	0.952	0.910	0.822	0.961	0.909	0.944
	GB	0.873	0.907	0.890	0.781	0.949	0.879	0.922

[a] *Results computed with prediction cutoff threshold value set as 0.5.*
[b] *Results computed with the fixed specificity at 0.9.*
[c] *Results computed with the fixed specificity at 0.8.*

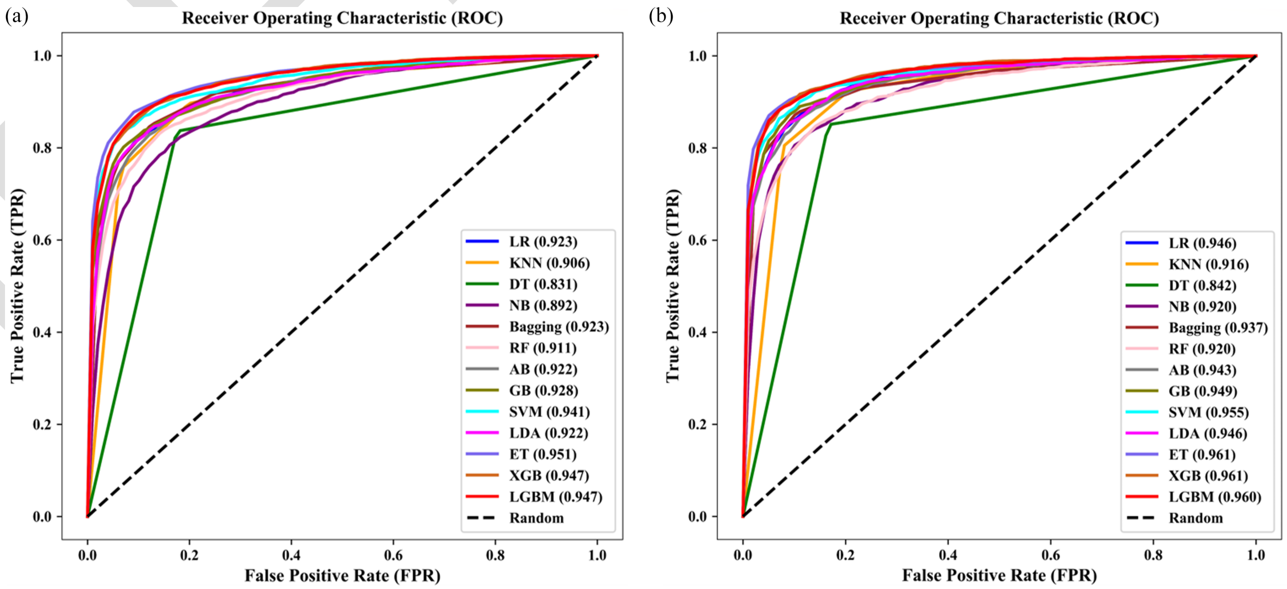


Fig. 2. ROC curves of different machine learning classifiers on the training datasets over five-fold cross-validation tests: (a) A.thaliana and (b) D.melanogaster.

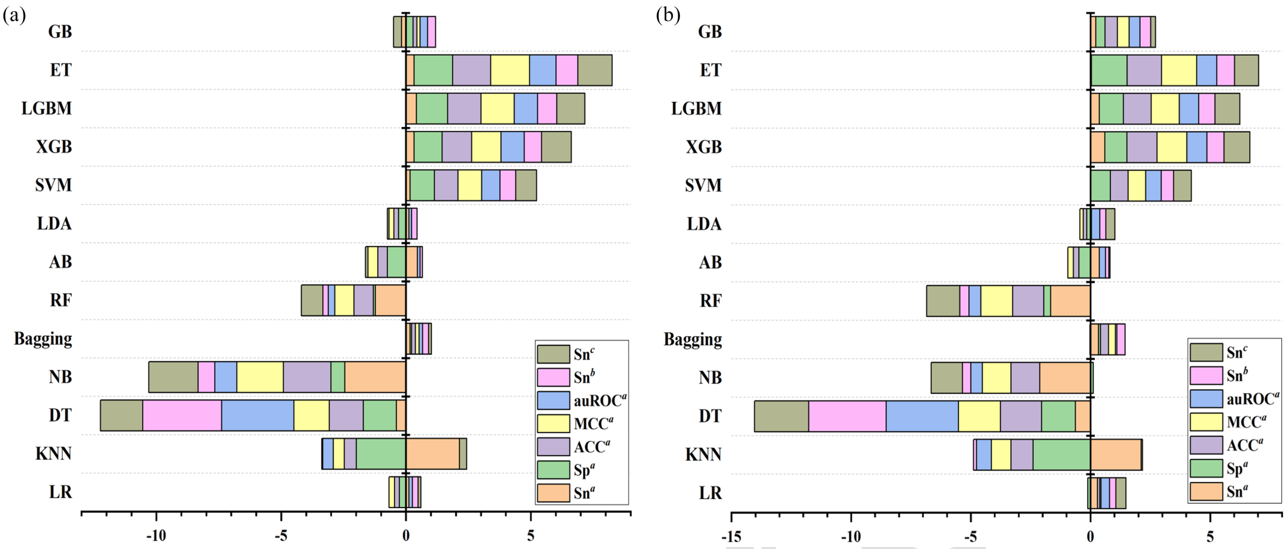


Fig. 3. Ranking of various classifiers in the global performance evaluation. (a) and (b) are ranked according to the sum of the Z-scores of all the evaluation indexes on the A.thaliana and D.melanogaster, respectively.

D. Integrated Classifiers With Averaging Strategy

To minimize the generalization error and enhance the performance of 6mA prediction, on the basis of subsection ‘Selection of base classifiers’, we empirically examine the predictive performance of single and ensemble classifiers on both training datasets over five-fold cross-validation tests. In the present subsection, two ensemble learning schemes, i.e., averaging and voting strategies, are considered to combine five base classifiers, i.e., ET, XGB, LGBM, SVM, and GB. Note that, the five individual classifiers should be first combined according to the priority of their overall performance, then each combiner is integrated by averaging or voting strategies. Hence, here, the performance of six integrated classifiers, i.e., ET+XGB+Averaging, ET+XGB+LGBM+Voting, ET+XGB+LGBM+Averaging, ET+XGB+LGBM+SVM+Averaging, ET+XGB+LGBM+SVM+GB+Voting, and ET+XGB+LGBM+SVM+GB+Averaging, are researched. For ease of description, these six integrated classifiers mentioned above are named Averaging2, Voting3, Averaging3, Averaging4, Voting5, and Averaging5, respectively. Table IV summarizes the compared results and Supplemental Fig. S5 displays the ROC curves of different classifiers.

From Table IV and Fig. S5, it is clear that the performance of Averaging4 is superior to that of the other single and integrated classifiers. In detail, by observing Table IV, we can easily find that, out of four averaging strategy-based classifiers, Averaging4 acts as the best performer followed by Averaging2, Averaging3, and Averaging5. For example, compared with Averaging2, the second-best classifier from the viewpoint of ACC, MCC, and auROC values, Averaging4 achieves average 0.28%, 0.56%, and 0.52% improvements in ACC, MCC, and auROC values on both training datasets. In addition, among two voting strategy-based classifiers, i.e., Voting3 and Voting5, the classifier Voting3 shows excellent prediction performance. For the classifier Voting3, the prediction accuracy value is 0.894, MCC value is 0.793, and auROC values is 0.954. Although the overall prediction performance of the Voting5 is slightly lower than that of Voting3, Voting5 achieves a better Sn value on the training dataset Drosophila melanogaster. It has not escaped from our notice that the performance of the averaging strategy-based classifiers is consistently higher than that of the voting strategy-based classifiers. Meanwhile, Averaging4 achieves the highest MCC and auROC values. However, when base-classifier GB is added to Averaging4, the overall prediction performance of 6mA sites (i.e., Voting5 and Averaging5) drops. We also rank the methods by using the sum of the Z-scores of global metrics to analyze the comprehensive performance of various 6mA sites prediction methods. It can be found that Averaging4 has the best comprehensive performance in both A.thaliana Fig. 4(a) and D. melanogaster Fig. 4(b) genomes. Therefore, Averaging4, i.e., ET+XGB+LGBM+SVM+Averaging, is employed as the final model of Ense-i6mA.

E. Comparison With Existing 6mA Sites Identification Methods

The purpose of this section is to experimentally demonstrate the efficacy of the proposed Ense-i6mA by comparing it with other recently state-of-the-art 6mA sites prediction methods on both independent testing datasets, including DeepM6A [34], i6mA-DNC [23], iDNA6mA (5-step rule) [19], 3-mer-LR [21], LA6mA, and AL6mA [21]. For an objective and fair comparison, all the methods use the same training datasets and independent testing datasets. The attributes of the feature used by the existing methods mentioned in the introduction section can be generally categorized into three major groups, i.e., physicochemical properties, sequence information, and evolutionary information. Here, DeepM6A, iDNA6mA (5-step rule), LA6mA,

TABLE IV
PERFORMANCE COMPARISON OF DIFFERENT METHODS ON THE TRAINING DATASETS OVER FIVE-FOLD CROSS-VALIDATION TESTS

Training dataset	Method	Sn [a]	Sp [a]	ACC [a]	MCC [a]	auROC [a]	Sn [b]	Sn [c]
Arabidopsis thaliana	ET	0.853	0.928	0.890	0.783	0.951	0.882	0.926
	XGB	0.853	0.913	0.883	0.767	0.947	0.865	0.921
	LGBM	0.855	0.918	0.886	0.774	0.947	0.872	0.919
	SVM	0.849	0.908	0.878	0.758	0.941	0.858	0.911
	GB	0.841	0.883	0.862	0.724	0.928	0.827	0.879
	Averaging2 [d]	0.866	0.924	0.894	0.793	0.954	0.886	0.933
	Voting3 [e]	0.859	0.918	0.888	0.778	0.949	0.873	0.923
	Averaging3 [f]	0.865	0.917	0.891	0.785	0.956	0.887	0.927
	Averaging4 [g]	0.870	0.925	0.897	0.796	0.961	0.890	0.935
	Voting5 [h]	0.858	0.911	0.884	0.769	0.946	0.866	0.919
	Averaging5 [i]	0.858	0.912	0.885	0.772	0.951	0.875	0.925
Drosophila melanogaster	ET	0.868	0.952	0.910	0.822	0.961	0.909	0.944
	XGB	0.883	0.929	0.906	0.813	0.961	0.906	0.946
	LGBM	0.877	0.932	0.904	0.810	0.960	0.902	0.945
	SVM	0.867	0.925	0.895	0.791	0.955	0.886	0.937
	GB	0.873	0.907	0.890	0.781	0.949	0.879	0.922
	Averaging2 [d]	0.887	0.938	0.913	0.826	0.963	0.918	0.949
	Voting3 [e]	0.879	0.935	0.907	0.816	0.961	0.911	0.948
	Averaging3 [f]	0.885	0.939	0.912	0.825	0.963	0.918	0.948
	Averaging4 [g]	0.889	0.943	0.915	0.832	0.966	0.917	0.950
	Voting5 [h]	0.882	0.929	0.906	0.813	0.960	0.907	0.944
	Averaging5 [i]	0.885	0.934	0.909	0.821	0.962	0.911	0.949

[a] *Results computed with prediction cutoff threshold value set as 0.5.*
[b] *Results computed with the fixed specificity at 0.9.*
[c] *Results computed with the fixed specificity at 0.8.*
[d] *Results computed by integrating ET and XGB with averaging strategy.*
[e] *Results computed by integrating ET, XGB, and LGMB with averaging strategy.*
[f] *Results computed integrating ET, XGB, and LGMB with voting strategy.*
[g] *Results computed integrating ET, XGB, LGMB, and SVM with averaging strategy.*
[h] *Results computed integrating ET, XGB, LGMB, SVM, and GB with voting strategy.*
[i] *Results computed integrating ET, XGB, LGMB, SVM, and GB with averaging strategy.*

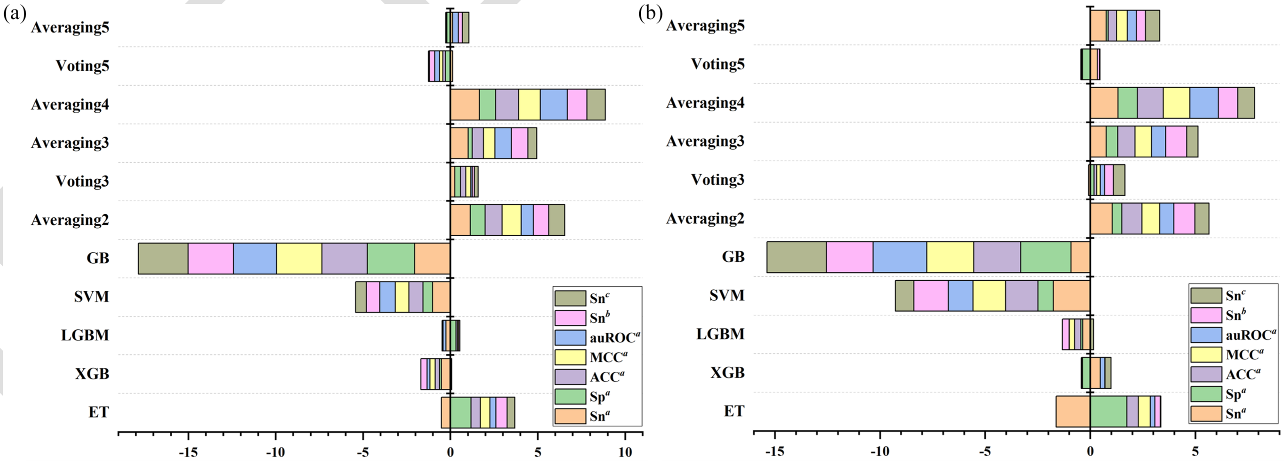


Fig. 4. Ranking of the methods in the global performance evaluation. (a) and (b) are ranked according to the sum of the Z-scores of all the evaluation metrics on the A.thaliana and D.melanogaster, respectively.

and AL6mA use OHE to identify 6mA sites; i6mA-DNC and 3-mer-LR predict 6mA sites in the DNA sequences based on dinucleotide components and 3-mer nucleotide frequency, respectively. Unlike these methods, Ense-i6mA incorporates OHE, KNF, Z-Curve, gcContent, KNFG, and XGB-RFE feature selection method for identifying 6mA sites. Table V and Fig. S6 summarize the performance compared results of the seven 6mA sites prediction methods on both independent testing datasets.

As described in Table V, we can see that DeepM6A has better prediction results for the 6mA sites in DNA for the existing prediction methods. The Sn, Sp, ACC, MCC, and auROC values are 0.894, 0.931,0.826, and 0.966, 0.901, 0.939,

TABLE V
PERFORMANCE COMPARISON BETWEEN THE PROPOSED ENSE-I6MA AND OTHER EXISTING METHODS FOR IDENTIFYING 6MA SITES ON THE INDEPENDENT TESTING DATASETS

Testing dataset	Method	Sn [a]	Sp [a]	ACC [a]	MCC [a]	auROC [a]	Sn [b]	Sn [c]
Arabidopsis thaliana	DeepM6A [*]	0.894	0.931	0.913	0.826	0.966	0.920	0.956
	i6mA-DNC [*]	0.846	0.909	0.878	0.757	0.944	0.853	0.912
	iDNA6mA [*,#]	0.843	0.889	0.866	0.733	0.932	0.833	0.902
	3-mer-LR [*]	0.669	0.728	0.699	0.397	0.773	0.411	0.577
	LA6mA [*]	0.899	0.917	0.909	0.817	0.962	0.912	0.948
	AL6mA [*]	0.862	0.905	0.884	0.768	0.945	0.867	0.927
	Ense-i6mA	0.899	0.930	0.914	0.829	0.967	0.919	0.951
Drosophila melanogaster	DeepM6A [*]	0.901	0.939	0.920	0.841	0.969	0.930	0.959
	i6mA-DNC [*]	0.869	0.917	0.893	0.787	0.947	0.878	0.916
	iDNA6mA [*,#]	0.883	0.843	0.863	0.727	0.937	0.846	0.904
	3-mer-LR [*]	0.68	0.702	0.691	0.383	0.753	0.347	0.558
	LA6mA [*]	0.909	0.915	0.912	0.824	0.966	0.921	0.955
	AL6mA [*]	0.84	0.916	0.878	0.758	0.941	0.848	0.92
	Ense-i6mA	0.902	0.940	0.920	0.842	0.968	0.921	0.949

[a] *Results computed with prediction cutoff threshold value set as 0.5.*
[b] *Results computed with the fixed specificity at 0.9.*
[c] *Results computed with the fixed specificity at 0.8.*
[*] *Results excerpted from [21].*
[#] *iDNA6mA stands for iDNA6mA (5-step rule).*

0.920,0.841, and 0.969, respectively, on the independent testing datasets Arabidopsis thaliana and Drosophila melanogaster. As expected, the 3-mer-LR, which is developed based on individual classifier LR algorithm, gained the lowest prediction performance in terms of five evaluation indexes. However, the novel method Ense-i6mA proposed in this study achieves comparable recognition performance as DeepM6A, and even superior to DeepM6A in certain evaluation indices. Taking the results of the proposed Ense-i6mA methods on the independent testing dataset Arabidopsis thaliana as an example, Ense-i6mA achieves the highest Sp, ACC, MCC, auROC values except Sn. Especially, the MCC and auROC, which are two most important indexes to assess the overall performance of the 6mA prediction methods, of Ense-i6mA are 0.829 and 0.967, which are 0.36% and 0.10%, 9.51% and 2.44%, 13.10% and 3.76%, 108.82% and 25.10%, 1.47% and 0.52%, and 7.94% and 2.33% higher than DeepM6A, i6mA-DNC, iDNA6mA (5-step rule), 3-mer-LR, LA6mA, and AL6mA, respectively. Furthermore, Table V also provides performance comparison of different methods in terms of Sn under the fixed Sp (i.e., 0.8 and 0.9). For two independent testing datasets, it is easy to find that DeepM6A performs best under fixed Sp followed by Ense-i6mA.

By revisiting Table V, it is noteworthy that although five deep learning-based methods, i.e., DeepM6A, i6mA-DNC, iDNA6mA (5-step rule), LA6mA, and AL6mA, obtain good performance, the proposed Ense-i6mA is the solely ensemble learning-based approach that achieves Sn>0.899, ACC>0.914, MCC>0.829 and auROC>0.967 on both model organisms. In addition, we also observe that DeepM6A, iDNA6mA (5-step rule), LA6mA and AL6mA, and i6mA-DNC use 164 = (41 × 4) and 640 = (40 × 16) meta-features, respectively, whereas the proposed Ense-i6mA only utilizes 80 meta-features (48.78% of DeepM6A, iDNA6mA (5-step rule), LA6mA and AL6mA, and 12.5% of i6mA-DNC). This may portend that Ense-i6mA can achieve performance comparable to or even higher than DeepM6A with less computation time and complexity. In summary, these results further validate the effectiveness and robustness of Ense-i6mA, indicating that Ense-i6mA is a powerful prediction method.

IV. CONCLUSION

Accurate identification of 6mA sites in DNA is crucial to elucidate the function of 6mA epigenetic modification. In this study, a new calculational method, called Ense-i6mA, is implemented for predicting 6mA sites in DNA. Experimental results have demonstrated that Ense-i6mA outperforms other existing state-of-the-art prediction methods, i.e., DeepM6A, i6mA-DNC, iDNA6mA (5-step rule), 3-mer-LR, LA6mA, and AL6mA. The superior performance of the proposed Ense-i6mA is primarily due to the following three aspects. Firstly, five discriminative feature sources, i.e., OHE, KNF, Z-Curve, gc-Content, and KNFG, are employed to extract more discriminative information from the data sets. Secondly, XGB-RFE is employed to remove noisy features while reducing computing time and complexity. Finally, the proposed Ense-i6mA leverages ensemble learning to further improve predictive performance of 6mA sites.

Despite its good performance, the proposed Ense-i6mA still has potential disadvantages and room for improvement. For instance, the feature representations used in this study should hardly adequately represent the identifiability of the 6mA sites data. Our further research work comprises the following four directions to further enhance the prediction efficacy of 6mA sites: (1) designing high discriminative feature source; (2) developing an excellent feature selection tool; (3) designing a more accurate method by combining Ense-i6mA and other state-of-the-art 6mA sites prediction methods; (4) establishing a user-friendly web-server to help potential researchers and end-users of Ense-i6mA. Finally, we believe that Ense-i6mA will be exploited as a useful tool to accelerate the progress of DNA function detection and understanding.

References

[1] G.-Z. Luo et al., "Characterization of eukaryotic DNA N6-methyladenine by a highly sensitive restriction enzyme-assisted sequencing," *Nat. Commun.*, vol. 7, no. 1, pp. 1–6, 2016.

[2] G.-Z. Luo, M. A. Blanco, E. L. Greer, C. He, and Y. Shi, "DNA N6-methyladenine: A new epigenetic mark in eukaryotes?," *Nat. Rev. Mol. Cell Biol.*, vol. 16, no. 12, pp. 705–710, 2015.

[3] T. P. Wu et al., "DNA methylation on N6-adenine in mammalian embryonic stem cells," *Nature*, vol. 532, no. 7599, pp. 329–333, 2016, doi: 10.1038/nature17640.

[4] G. Zhang et al., "N6-methyladenine DNA modification in Drosophila," *Cell*, vol. 161, no. 4, pp. 893–906, 2015.

[5] K.-J. Wu, "The epigenetic roles of DNA N6-methyladenine (6mA) modification in eukaryotes," *Cancer Lett.*, vol. 494, pp. 40–46, 2020.

[6] K. Vasu and V. Nagaraja, "Diverse functions of restriction-modification systems in addition to cellular defense," *Microbiol. Mol. Biol. Rev.*, vol. 77, no. 1, pp. 53–72, 2013.

[7] Z. Liang et al., "DNA N6-adenine methylation in Arabidopsis thaliana," *Dev. Cell*, vol. 45, no. 3, pp. 406–416.e3, 2018.

[8] J. Liu et al., "Abundant DNA 6mA methylation during early embryogenesis of zebrafish and pig," *Nat. Commun.*, vol. 7, no. 1, pp. 1–7, 2016.

[9] K. R. Pomraning, K. M. Smith, and M. Freitag, "Genome-wide high throughput analysis of DNA methylation in eukaryotes," *Methods*, vol. 47, no. 3, pp. 142–150, 2009.

[10] S. Frelon, T. Douki, J.-L. Ravanat, J.-P. Pouget, C. Tornabene, and J. Cadet, "High-performance liquid chromatography− tandem mass spectrometry measurement of radiation-induced base damage to isolated and cellular DNA," *Chem. Res. Toxicol.*, vol. 13, no. 10, pp. 1002–1010, 2000.

[11] B. A. Flusberg et al., "Direct detection of DNA methylation during single-molecule, real-time sequencing," *Nat. Methods*, vol. 7, no. 6, pp. 461–465, 2010.

[12] B. Liu, F. Liu, L. Fang, X. Wang, and K.-C. Chou, "repDNA: A Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects," *Bioinformatics*, vol. 31, no. 8, pp. 1307–1309, 2015.

[13] P. Feng, W. Chen, and H. Lin, "Prediction of CpG island methylation status by integrating DNA physicochemical properties," *Genomics*, vol. 104, no. 4, pp. 229–233, 2014.

[14] W. Chen, X. Zhang, J. Brooker, H. Lin, L. Zhang, and K.-C. Chou, "PseKNC-General: A cross-platform package for generating various modes of pseudo nucleotide compositions," *Bioinformatics*, vol. 31, no. 1, pp. 119–120, 2015.

[15] 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, 2019.

[16] C. Pian, G. Zhang, F. Li, and X. Fan, "MM-6mAPred: Identifying DNA N6-methyladenine sites based on Markov model," *Bioinformatics*, vol. 36, no. 2, pp. 388–392, Jan. 15, 2020, doi: 10.1093/bioinformatics/btz556.

[17] D. A. Filatov, "ProSeq: A software for preparation and evolutionary analysis of DNA sequence data sets," *Mol. Ecol. Notes*, vol. 2, no. 4, pp. 621–624, 2002.

[18] Z. Abbas, H. Tayara, and K. to Chong, "SpineNet-6mA: A novel deep learning tool for predicting DNA N6-methyladenine sites in genomes," *IEEE Access*, vol. 8, pp. 201450–201457, 2020.

[19] M. Tahir, H. Tayara, and K. T. Chong, "iDNA6mA (5-step rule): Identification of DNA N6-methyladenine sites in the rice genome by intelligent computational model via Chou's 5-step rule," *Chemometrics Intell. Lab. Syst.*, vol. 189, pp. 96–101, 2019.

[20] Z. Li et al., "Deep6mA: A deep learning framework for exploring similar patterns in DNA N6-methyladenine sites across different species," *PLoS Comput. Biol.*, vol. 17, no. 2, 2021, Art. no. e1008767.

[21] Y. Zhang et al., "Leveraging the attention mechanism to improve the identification of DNA N6-methyladenine sites," *Brief. Bioinf.*, vol. 22, 2021, Art. no. bbab351.

[22] Z. Teng et al., "i6mA-Vote: Cross-species identification of DNA N6-methyladenine sites in plant genomes based on ensemble learning with voting," *Front. Plant Sci.*, vol. 13, pp. 845835–845835, 2022.

[23] S. Park, A. Wahab, I. Nazari, J. H. Ryu, and K. T. Chong, "i6mA-DNC: Prediction of DNA N6-Methyladenosine sites in rice genome based on dinucleotide representation using deep learning," *Chemometrics Intell. Lab. Syst.*, vol. 204, 2020, Art. no. 104102.

[24] W. Chen, H. Lv, F. Nie, and H. Lin, "i6mA-Pred: Identifying DNA N6-methyladenine sites in the rice genome," *Bioinformatics*, vol. 35, no. 16, pp. 2796–2800, 2019.

[25] J. Khanal, D. Y. Lim, H. Tayara, and K. T. Chong, "i6mA-stack: A stacking ensemble-based computational prediction of DNA N6-methyladenine (6mA) sites in the rosaceae genome," *Genomics*, vol. 113, no. 1, pp. 582–592, 2021.

[26] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Their Appl.*, vol. 13, no. 4, pp. 18–28, Jul./Aug. 1998.

[27] T. Chen et al., "Xgboost: Extreme gradient boosting," *R Package Version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[28] D. W. Hosmer, T. Hosmer, S. L. Cessie, and S. Lemeshow, "A comparison of goodness-of-fit tests for the logistic regression model," *Statist. Med.*, vol. 16, no. 9, pp. 965–980, 1997.

[29] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[30] Y. Qi, "Random forest for bioinformatics," in *Ensemble Machine Learning*. Berlin, Germany: Springer, 2012, pp. 307–323.

[31] M. Iliadis, L. Spinoulas, and A. K. Katsaggelos, "Deep fully-connected networks for video compressive sensing," *Digit. Signal Process*, vol. 72, pp. 9–18, 2018.

[32] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *Proc. Int. Conf. Eng. Technol.*, 2017, pp. 1–6.

[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 15, 1997, doi: 10.1162/neco.1997.9.8.1735.

[34] F. Tan et al., "Elucidation of DNA methylation on N6-adenine with deep learning," *Nat. Mach. Intell.*, vol. 2, no. 8, pp. 466–475, 2020.

[35] K. E. Kim et al., "Long-read, whole-genome shotgun sequence data for five model organisms," *Sci. Data*, vol. 1, 2014, Art. no. 140045, doi: 10.1038/sdata.2014.45.

[36] Y. Benjamini and T. P. Speed, "Summarizing and correcting the GC content bias in high-throughput sequencing," *Nucleic Acids Res.*, vol. 40, no. 10, pp. e72–e72, 2012.

[37] F. Hildebrand, A. Meyer, and A. Eyre-Walker, "Evidence of selection upon genomic GC-content in bacteria," *PLos Genet*, vol. 6, no. 9, 2010, Art. no. e1001107.

[38] R. Zhang and C.-T. Zhang, "A brief review: The z-curve theory and its application in genome analysis," *Curr. Genomic.*, vol. 15, no. 2, pp. 78–94, 2014.

[39] C.-T. Zhang, R. Zhang, and H.-Y. Ou, "The Z curve database: A graphic representation of genome sequences," *Bioinformatics*, vol. 19, no. 5, pp. 593–599, 2003.

[40] Y. Zhang et al., "StackRAM: A cross-species method for identifying RNA N6-methyladenosine sites based on stacked ensemble," *Chemometrics Intell. Lab. Syst.*, vol. 222, 2022, Art. no. 104495.

[41] Q. Zhang, P. Liu, X. Wang, Y. Zhang, Y. Han, and B. Yu, "StackPDB: Predicting DNA-binding proteins based on XGB-RFE feature optimization and stacked ensemble classifier," *Appl. Soft Comput.*, vol. 99, 2021, Art. no. 106921.

[42] J. Hu, Y.-S. Bai, L.-L. Zheng, N.-X. Jia, D.-J. Yu, and G. Zhang, "Protein-DNA binding residue prediction via bagging strategy and sequence-based cube-format feature," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 6, pp. 3635–3645, Nov./Dec. 2022.

[43] H.-C. Yi, Z.-H. You, M.-N. Wang, Z.-H. Guo, Y.-B. Wang, and J.-R. Zhou, "RPI-SE: A stacking ensemble learning framework for ncRNA-protein interactions prediction using sequence information," *BMC Bioinf.*, vol. 21, no. 1, pp. 1–10, 2020.

[44] Z.-H. You, W.-Z. Huang, S. Zhang, Y.-A. Huang, C.-Q. Yu, and L.-P. Li, "An efficient ensemble learning approach for predicting protein-protein interactions by integrating protein primary sequence and evolutionary information," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 3, pp. 809–817, May/Jun. 2018.

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

[46] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscip. Rev. Comput. Statist.*, vol. 2, no. 4, pp. 433–459, 2010.

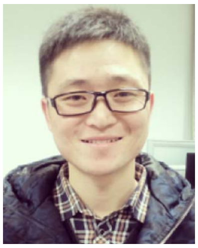
[47] H. Sanz, C. Valim, E. Vegas, J. M. Oller, and F. Reverter, "SVM-RFE: Selection and visualization of the most relevant features through non-linear kernels," *BMC Bioinf.*, vol. 19, no. 1, pp. 1–18, 2018.

[48] J. Augustine and A. Jereesh, “Blood-based DNA methylation marker identification for Parkinson’s Disease prediction,” in *Proc. Int. Conf. Innov. Comput. Commun.*, 2022, pp. 777–784.

[49] Q. Chen, Z. Meng, X. Liu, Q. Jin, and R. Su, “Decision variants for the automatic determination of optimal feature subset in RF-RFE,” *Genes*, vol. 9, no. 6, 2018, Art. no. 301.

Xueqiang Fan is currently working toward the PhD degree with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China. His current research interests include pattern recognition and polarization imaging.

Bing Lin received the BE degree in communication engineering from the Hefei University of Technology, Hefei, China, in 2021. She is currently working toward the master’s degree with the Advanced Electromagnetic Function Laboratory (AEMFLab), Hefei University of Technology. Her research interests include polarization imaging and deep learning.



Jun Hu received the BS degree in computer science from Anhui Normal University, in 2011, and the PhD degree from the School of Computer Science and Engineering, Nanjing University of Science and Technology, in 2018, and a member of Pattern Recognition and Bioinformatics Group, led by professor Dong-Jun Yu. From 2016 to 2017, he acted as a visiting student with the University of Michigan (Ann Arbor) in USA. He is currently a teacher with the College of Information Engineering, Zhejiang University of Technology. His current interests include pattern recognition, data mining and bioinformatics.

Zhongyi Guo received the bachelor’s degree from the Department of Physics, Harbin Institute of Technology, in 2003, and the master’s and doctoral degrees from the Harbin Institute of Technology, in 2005 and 2008, respectively. From 2008 to 2009, he worked as an assistant professor with the Department of Physics, Harbin Institute of Technology. Then he moved to Hanyang University (Korea) and worked as a postdoctor for 2 years. In 2011, he continued to move to HongKong Polytechnic University as a postdoctor for 6 months. In the end of 2011, he joined in and worked as a full professor with the School of Computer and Information, Hefei University of Technology. Now, he focus his research interests include fields of polarization information processing, advanced optical communication, OAM antenna, manipulation of optical fields, and nanophotonics.