Journal of Computer Science and Frontier Technologies
ISSN Print: 3104-4204
ISSN Online: 3104-4212
CD ISSN: 3104-4220
https://doi.org/10.63313/JCSFT.9056

A Method for the Identification and Intensity Prediction of DNA Enhancers Based on Feature Extraction Algorithms

XiFeng Li
College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou 730070, PR China
Email: 2023222396@nwnu.edu.cn

How to cite this paper: Li, X. F. (2026). A method for the identification and intensity prediction of DNA enhancers based on feature extraction algorithms. *Journal of Computer Science and Frontier Technologies*, 3(1), 9-20. ISSN Print: 3104-4204; ISSN Online: 3104-4212.
https://doi.org/10.63313/JCSFT.9056
Published: 2026-03-20

Copyright © 2026 by author(s) and Erytis Publishing Limited.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/



Abstract

DNA enhancers are important non-coding cis-regulatory elements that control gene transcription. To reduce reliance on high-cost experimental methods such as ChIP-seq and DNase-seq, this paper constructs a classification framework based on sequence features and supervised learning for two sequence-level tasks: enhancer identification (enhancer/non-enhancer) and enhancer strength prediction (strong/weak enhancer).Methodologically, k-mer frequency features and CKSNAP-based dinucleotide interval features were extracted separately, and these were concatenated to form a fused representation. Under a unified hierarchical 5-fold cross-validation framework, the performance of Support Vector Machines (SVM) and XGBoost models was compared. The results indicate that, for both tasks, XGBoost performed slightly better or was comparable to SVM under the same feature set;under the same classifier, the fusion of k-mer and CKSNAP features outperforms either feature alone. Specifically, XGBoost based on the fusion features achieved accuracy rates of 0.796 and 0.681 for enhancer identification and strength classification, respectively, with overall superior ROC/AUC performance. This indicates that local k-mer composition information and short-range distance-dependent information are complementary, thereby enhancing the model's discriminative power and stability.

Keywords

DNA Enhancer; Enhancer Identification; Strength Prediction; K-Mer; CKSNAP; SVM; Xgboost

1. Introduction

DNA enhancers are important non-coding cis-regulatory elements in the genome that play a key role in precisely regulating gene expression levels, maintaining cell type specificity, and governing developmental programs.Enhancers are typically 50–1,500 bp in length and can be located upstream, downstream, or within the intron regions of target genes; they can even span considerable linear distances[1].

Through chromatin 3D folding and DNA looping, they establish spatial contact with promoter regions, thereby facilitating the recruitment of transcription factors and the transcriptional machinery, increasing transcription initiation frequency, and enhancing gene expression[2].Extensive research indicates that sequence variations within enhancer regions are closely associated with various complex diseases, including cancer, neurological disorders, and immune-related diseases, suggesting that regulatory abnormalities at the enhancer level serve as a crucial link between genotype and phenotype.Based on their ability to regulate target gene expression, enhancers are typically classified into two categories: strong enhancers and weak enhancers, with strong enhancers contributing more significantly to transcriptional activation. Therefore, accurately identifying enhancers at the DNA sequence level and further distinguishing between strong and weak enhancers is of great significance for gaining a deeper understanding of gene expression regulatory networks and elucidating disease-related regulatory mechanisms[3,4].

With the advancement of high-throughput sequencing technologies and functional genomics techniques (such as ChIP-seq, ATAC-seq, and STARR-seq), a vast amount of omics data related to enhancer activity has been continuously accumulated, providing a rich data foundation for enhancer identification and functional strength assessment[5].Early studies on enhancers largely relied on sequence alignment and conservation analysis, or directly utilized experimental signal thresholds for screening, which yielded satisfactory results under specific data conditions. However, such methods often heavily depend on expensive and condition-sensitive experimental approaches, making them difficult to scale up or apply in more complex scenarios.To reduce reliance on experimental data, an increasing number of studies in recent years have attempted to construct enhancer identification and strength classification models based solely on DNA sequences themselves, proposing various feature extraction methods and modeling strategies, such as pseudo-k-mer nucleotide composition (PseKNC), multi-scale k-mer frequency, physicochemical property encoding, and deep learning embedding representations.Overall, these methods have advanced the development of enhancer prediction at the sequence level, but some shortcomings remain: on the one hand, certain deep learning methods rely on complex network architectures, which demand large sample sizes and significant computational resources, and suffer from relatively poor interpretability;on the other hand, many studies still focus primarily on a single type of feature in their design, with relatively limited exploration of systematic comparisons and complementary integration of different types of sequence features. Particularly for more refined classification tasks such as distinguishing between strong and weak enhancers, how to fully exploit the multidimensional information within sequences while ensuring computational efficiency and interpretability remains worthy of in-depth research[6].

Based on the above background and needs, this study focuses on the problem of

enhancer identification and strength prediction “relying solely on DNA sequence information.” It systematically investigates sequence feature extraction and machine learning classification modeling methods, aiming to construct a lightweight, efficient, and easily reproducible enhancer prediction framework.On the one hand, by predicting the presence and strength of enhancers based solely on genomic sequences without relying on additional omics experimental data, this work provides a scalable computational tool for enhancer functional annotation and related biological research. On the other hand, through sequence-level feature analysis, it helps reveal the differential patterns between strong and weak enhancers in terms of base composition and local dependencies, offering a new perspective for understanding enhancer sequence characteristics and transcriptional regulatory mechanisms.

Overall, this study builds upon existing research to construct an enhancer classification framework based on sequence features and supervised learning, and conducts systematic experiments and comparisons on two tasks: “enhancer identification” and “distinguishing strong from weak enhancers.” Through a unified evaluation process, the performance of different features and model configurations is objectively analyzed, providing a methodological foundation and performance baseline for future work that integrates richer sequence features, introduces more complex model structures, and extends to larger-scale data and multi-species scenarios.

2. Data and Methods

2.1. Dataset

This study utilized the benchmark dataset employed in the paper by Liu et al[7]. (2019), titled "iEnhancer-5Step: Identifying enhancers using hidden information of DNA sequences via Chou's 5-step rule and word embedding," published in Analytical Biochemistry. This dataset has become the standard dataset in the field of enhancer classification research.The dataset was constructed based on information from nine different cell lines, and all DNA sequences were standardized into 200-bp fragments. To ensure data quality, the researchers used the CD-HIT tool to perform deduplication, removing samples with sequence similarity exceeding 20%.The complete dataset contains 1,484 enhancer samples and 1,484 non-enhancer samples. Since this study focuses on the classification of strong and weak enhancers, we used only the 1,484 enhancer samples, which were further categorized into 742 strong enhancers and 742 weak enhancers.

2.2. k-mer

In this study, DNA sequences are treated as strings composed of {A, C, G, T}. The concept of k-mer features involves sliding a window of length k sliding a window of

length from left to right along the sequence, treating each resulting contiguous segment as a "word," and then counting the frequency of each "word" across the entire sequence. The advantage of this approach is that it explicitly represents local base composition preferences in the sequence without requiring any additional annotation information[8].For regulatory elements such as enhancers, many function-related differences often manifest as the enrichment or depletion of local segments, and k-mer frequencies provide a direct quantification of these differences. Let a DNA sequence be $S = s_1 s_2 \cdots s_L$, with a length of L . When K is fixed, the window can generate $L-K+1$ k-mers across the sequence. For any specific k-mer (denoted as u), its occurrence count can be written as

$$c(u) = \sum_{i=1}^{L-k+1} I\left(s_i s_{i+1} \cdots s_{i+k-1} = u\right) \quad (1)$$

where $I(\cdot)$ is the indicator function, which takes 1 if the condition holds and 0 otherwise. To eliminate the influence of length on the count, relative frequency is typically used as a feature component:

$$f(u) = \frac{c(u)}{L-k+1} \quad (2)$$

When the alphabet size is 4, the total number of possible k-mers is 4^k . By arranging these k-mers in a fixed lexicographical order (e.g., from AA...A to TT...T), the sequence can be mapped to a fixed-dimensional vector $x \in \mathbb{R}^{4^k}$, where the j th dimension corresponds to the frequency $f\left(u_j\right)$ of the j th k-mer. In this study, k=3 is used in the experiments; therefore, each 200-base sequence corresponds to a 64-dimensional 3-mer frequency vector.This setting strikes a better balance between expressive power and dimensionality: too small a value of k leads to insufficient representation, while too large a value of k tends to exacerbate sparsity and increase dimensionality.

2.3. CKSNAP

The distinction between CKSNAP (Composition of K-Spaced Nucleotide Pairs) and k-mers is that CKSNAP is not limited to analyzing contiguous segments formed by adjacent bases; instead, it focuses on "base pairs separated by a certain number of positions." In DNA sequences, many patterns do not necessarily appear in strictly contiguous form; particularly in fixed-length segments, positional relationships separated by a few bases may still carry discriminative information.By introducing the spacing parameter k , CKSNAP captures this short-range dependency structure at the "dinucleotide pair" level[9-11].

For a given interval k , base pairs of the form s_i, $s_{\{i+k+1\}}$ are extracted from the sequence, where $i = 1, 2, \ldots, L-k-1$. There are 16 types of dinucleotide pairs, denoted by the set

$$\mathcal{P} = AA, AC, \ldots, TT \quad (3)$$

For any class of base pairs $p \in \mathcal{P}$, their count is

$$c_k(p) = \sum_{i=1}^{L-k-1} I\left(s_i s_{i+k+1} = p\right) \tag{4}$$

and normalize it to a frequency

$$f_k(p) = \frac{c_k(p)}{L-k-1} \tag{5}$$

Thus, under a fixed interval k , a 16-dimensional vector can be obtained

$$k = [f_k(AA), f_k(AC), \ldots, f_k(TT)] \tag{6}$$

Compared to k-mer, CKSNAP places greater emphasis on "pairwise relationships" and their distribution differences across different intervals; the two differ in their focus on information, and thus possess complementary potential in subsequent modeling.

2.4. SVM

The core objective of SVM is to learn a classification hyperplane in the feature space that maximizes the margin between the two classes of samples[12]. Let the training set be $(x_i, y_i)_{i=1}^n$, where$x_i \in \mathbb{R}^d$, and the class labels are $y_i \in (-1, +1)$. In the ideal case of linear separability, SVM aims to find w and b such that

$$y_i(w^\top x_i + b) \geq 1, \quad i = 1, 2, \ldots, n \tag{7}$$

and maximizes the geometric margin. The equivalent convex optimization formulation is to minimize $\frac{1}{2}|w|^2$. Since real-world data often contains noise or overlap, SVM introduces slack variables to obtain a soft margin model:

$$\min_{w,b,\xi} \frac{1}{2}|w|^2 + C\sum_{i=1}^{n}\xi_i \text{ s.t. } y_i(w^\top x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, 2, \ldots, n \tag{8}$$

where (C > 0) balances the maximization of the margin with the penalty for misclassification. When samples are linearly inseparable in the original space, SVM replaces the inner product with similarity under the kernel mapping via a kernel function, thereby achieving linear separation in high-dimensional space. The kernel function is written as

$$K(x, z) = \phi(x)^\top \phi(z) \tag{9}$$

This study adopts the RBF kernel (Gaussian kernel):

$$K(x, z) = \exp\left(-\gamma |x - z|^2\right) \tag{10}$$

where controls the range of the kernel function, influencing the complexity of the decision boundary. Since k-mers and CKSNAP are frequency-based features, their scales may vary across dimensions. Therefore, this study performs standardization before SVM training to make features in each dimension more comparable in

numerical scale, thereby preventing certain dimensions from gaining an unreasonable advantage in distance calculations.

2.5. XGBoost

XGBoost is an efficient implementation of Gradient Boosting Trees (GBDT). It constructs the final prediction by summing the outputs of multiple weak learners through the iterative addition of regression trees[13]. Let the tree model added in the i-th round be $f_t(\cdot)$, and the prediction for sample x_i be $\hat{y}_i = \sum t = 1^T f_t(x_i)$.

In binary classification tasks, the log loss is commonly used, and the probability of the positive class is obtained by applying a sigmoid transformation to $\hat{y}_i$. The objective function for XGBoost optimization consists of a loss term and a regularization term:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}i) + \sum t = 1^T \Omega(f_t) \tag{11}$$

where the regularization term is used to control the complexity of the tree model, and a common form is

$$\Omega(f) = \alpha m + \frac{\lambda}{2} \sum_{j=1}^{m} w_j^2 \tag{12}$$

Here, `$m$ ` is the number of leaf nodes, `$w_j$ ` is the output weight of the `$j$ `th leaf node, and `$\alpha$ ` and `$\lambda$ ` are the complexity penalty coefficients.During training, an additive modeling approach is adopted: in the t th round, the outputs of the first $t-1$ trees are fixed, and the objective function of the current round is approximated using first- and second-order gradients, thereby efficiently selecting splits and computing leaf weights. Compared to SVM, which relies on kernel distance metrics, XGBoost automatically learns nonlinear separations and feature interactions through its tree structure, typically offering strong expressive power and robustness for frequency-based concatenated features.

2.6. Fusion of k-mer and XGBoost Features

k-mer frequency features primarily characterize the compositional distribution of contiguous base segments in DNA sequences, reflecting the enrichment of local short-sequence patterns; CKSNAP features, on the other hand, introduce a spacing parameter based on dinucleotide pair statistics. By describing the occurrence frequency of base pairs at different intervals, they supplement the short-range dependency information that adjacency statistics cannot cover.Since these two types of features focus on different levels of sequence information—k-mers emphasize “contiguous local composition” while CKSNAP emphasizes “spaced pair relationships”—they are complementary in their representational capabilities. To leverage both types of information simultaneously, this study employs a

feature-level fusion strategy, directly concatenating the k-mer vector and the CKSNAP vector of the same sequence along the feature dimension to construct a unified input representation.

Let the k-mer feature vector corresponding to a given sequence be K , and the CKSNAP feature vector be S . The fused feature vector is denoted as F , and its construction can be expressed as

$$F = [K, S] \tag{13}$$

In the specific implementation, the k-mer frequency vector is first generated according to a fixed dictionary order, followed by the CKSNAP frequency vector in a fixed order; finally, the two vectors are concatenated end-to-end to form a new feature vector. The fused features maintain the semantic clarity of both types of features while jointly capturing information regarding the continuous composition of the sequence and short-range interval dependencies, thereby providing richer discriminative information for subsequent classification models.

2.7. Evaluation Metrics

Both tasks in this study are binary classification problems.In the confusion matrix, TP denotes the number of samples that are truly positive and predicted as positive; TN denotes those that are truly negative and predicted as negative; FP denotes those that are truly negative but predicted as positive; and FN denotes those that are truly positive but predicted as negative. Based on this, unified expressions for sensitivity (SN), specificity (SP), accuracy (ACC), and the Matthews correlation coefficient (MCC) can be derived:

$$\begin{cases} \mathrm{SN} = \dfrac{TP}{TP+FN} \\ \mathrm{SP} = \dfrac{TN}{TN+FP} \\ \mathrm{ACC} = \dfrac{TP+TN}{TP+TN+FP+FN} \\ \mathrm{MCC} = \dfrac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \end{cases} \tag{14}$$

Here, SN reflects the model's ability to detect positive samples, SP characterizes its ability to identify negative samples, ACC measures the overall classification accuracy, and MCC provides a more comprehensive measure of association by considering TP, TN, FP, and FN.For classifiers that output probabilities, a ROC curve can be plotted by varying the decision threshold; the area under the curve (AUC) is used to measure the model's overall discriminative ability across all thresholds, with values ranging from [0,1]. A value closer to 1 indicates stronger discriminative ability[14,15].

3. Results and Discussion

3.1. Analysis of Experimental Results

The results of this study were obtained under a stratified 5-fold cross-validation (Stratified 5-fold CV) framework, in which the data was stratified into five folds based on class proportions. In each iteration, one fold was used as the validation set and the remaining four folds as the training set[16]. After five iterations, the results from each fold were averaged.This setup reduces the randomness associated with a single random split while ensuring consistent positive-to-negative sample ratios across all folds, thereby enabling a fairer and more stable comparison of different feature and model combinations.

Table 1 shows the average accuracy of three feature configurations (k-mer, CKSNAP, and k-mer+CKSNAP fusion) across two model types (SVM and XGBoost). Overall, the accuracy for the enhancer identification task is higher than that for the strong/weak enhancer classification task, indicating that "strong/weak" represents a finer-grained functional distinction, with weaker sequence differences and greater classification difficulty.

In the enhancer identification task, XGBoost (0.790) outperformed SVM (0.775) when using k-mer features;while the two models performed nearly equally with CKSNAP features (SVM 0.777, XGBoost 0.776). When using fused features, both models showed improvements, with SVM reaching 0.785 and XGBoost reaching 0.796—the highest values for both tasks—indicating that fused features provide a stable gain in terms of the five-fold average.

In the strong-weak enhancer classification task, accuracy decreased overall, and the differences became more pronounced. With k-mer features, SVM and XGBoost achieved 0.645 and 0.652, respectively; with CKSNAP features, SVM reached 0.631, while XGBoost improved to 0.666, demonstrating that CKSNAP is more advantageous when used with XGBoost;with both models further improving to 0.657 (SVM) and 0.681 (XGBoost) when using the fused features, achieving the best results for this task as shown in Table 1. A comprehensive review of Table 1 reveals that the five-fold average results support two points: first, XGBoost is slightly superior to SVM under most settings; second, the fused features yield the best results in both tasks, with a more pronounced advantage in the strong-weak classification task.

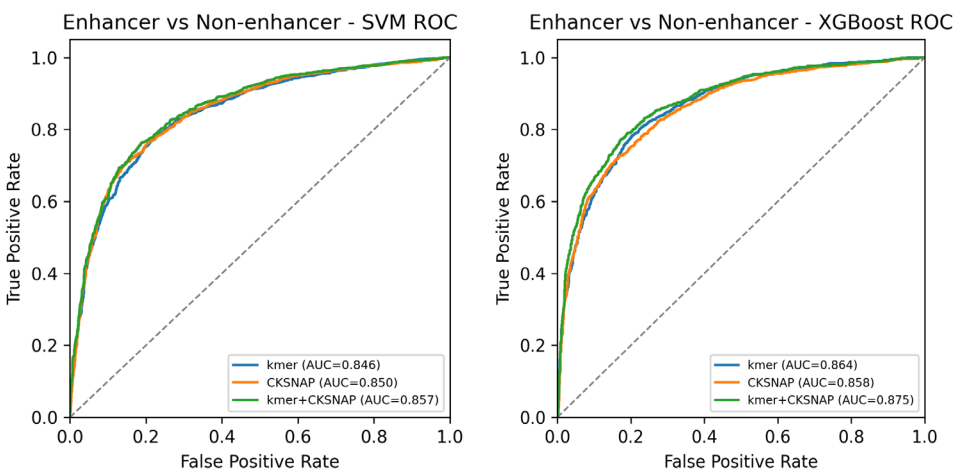
Table 1. Performance of two classification models (SVM and XGBoost) under different feature types

Feature Type	Model	Enhancer-Non-enhancer classification	Strong-Weak Classification
k-mer	SVM	0.775	0.645
k-mer	XGBoost	0.790	0.652
CKSNAP	SVM	0.777	0.631
CKSNAP	XGBoost	0.776	0.666
k-mer + CKSNAP	SVM	0.785	0.657
k-mer + CKSNAP	XGBoost	0.796	0.681

In five-fold cross-validation, the model outputs the probability of the positive class on each validation fold. ROC curves are obtained by varying the threshold, while AUC reflects the overall discriminative ability independent of the threshold.In Figure 1a, for the enhancer identification task, under the SVM model, the AUCs for k-mer, CKSNAP, and the fused features are 0.846, 0.850, and 0.857, respectively. While the three curves are close, the fused features perform better overall, indicating that feature fusion leads to a more stable improvement under the SVM model.

In Figure 1b, the AUCs for the enhancer identification task under XGBoost are 0.864, 0.858, and 0.875, respectively. The fused features achieve the highest AUC, which is consistent with the highest ACC of 0.796 corresponding to the fused features in Table 1.It can also be observed that k-mer performs slightly better than CKSNAP under XGBoost; however, after fusion, the AUC further increases to 0.875. This indicates that continuous local compositional information and short-range dependency information are complementary, and their joint representation enhances the model's overall discriminative power.

In Figures 2a and 2b, the AUCs of XGBoost for the strong-weak enhancer classification tasks are 0.713, 0.724, and 0.756, respectively, which are generally lower than those for the enhancer identification task. The ROC curves also lie closer to the diagonal, reflecting that strong-weak classification constitutes a finer-grained and more challenging discrimination problem.In this task, CKSNAP achieved a higher AUC than k-mer, indicating that short-range dependency information between adjacent dinucleotide pairs is more sensitive to strength differences; feature fusion further boosted the AUC to 0.756, consistent with the highest ACC of 0.681 achieved by feature fusion combined with XGBoost in Table 1, demonstrating that the fusion strategy can yield more stable gains in the strength classification task.



a b

Figure 1. ROC curves for enhancer/non-enhancer classification based on three feature

extraction methods

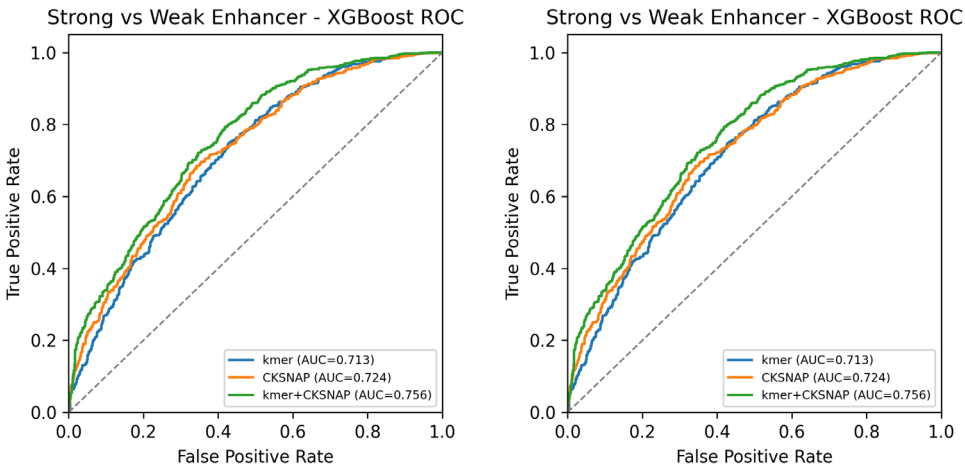


Figure 2. ROC curves for strong enhancer/weak enhancer classification based on three feature extraction methods

4. Conclusions

This paper addresses the tasks of DNA enhancer identification and enhancer strength prediction by constructing a machine learning classification framework that relies solely on sequence information.Under a unified data preprocessing and evaluation workflow, the study extracted k-mer frequency features and CKSNAP-based dinucleotide interval features, and combined them through feature concatenation to obtain a fused representation. Within a hierarchical five-fold cross-validation framework, the performance of two models—SVM and XGBoost—was compared, and their discriminative capabilities were comprehensively evaluated based on accuracy and ROC/AUC. The main conclusions are as follows.

First, in the enhancer identification task, both k-mer-based and CKSNAP-based features provided effective discriminative information, indicating that sequence statistical features can reliably distinguish enhancers from non-enhancers without relying on additional omics experimental data.Compared to SVM, XGBoost performed slightly better or was comparable under most feature configurations, demonstrating the strong nonlinear fitting and generalization capabilities of gradient-boosted tree models when handling high-dimensional statistical features.

Second, in the strong-weak enhancer classification task, the overall performance was significantly lower than in the enhancer identification task, with ROC curves closer to the diagonal and lower AUC values, indicating that strong-weak prediction is a finer-grained and more challenging classification problem.In this task, the AUC of CKSNAP features under XGBoost outperformed that of k-mers, suggesting that short-range inter-nucleotide dependency information is more sensitive to strength

differences; simultaneously, k-mers and CKSNAP are complementary in terms of information content, and their combined representation can more comprehensively characterize the sequence differences between strong and weak enhancers.

Third, feature fusion yields consistent improvements in both tasks. Whether measured by accuracy or AUC, the fused features outperform any single feature.Particularly under the XGBoost model, the fused features achieved the best results in the experiments presented in Chapter 3 of this paper: the average accuracy for the enhancer identification task reached 0.796, and the average accuracy for the strong-weak enhancer classification task reached 0.681, with AUCs of 0.875 and 0.756, respectively. This indicates that the fusion of continuous local composition information with short-range interval-dependent information can enhance the model's discriminative power and stability.

In summary, this study demonstrates that, under a unified evaluation framework of hierarchical five-fold cross-validation, a modeling strategy based on k-mer and CKSNAP fusion features combined with XGBoost can effectively predict the presence and strength levels of enhancers relying solely on DNA sequences. This provides a reproducible baseline method and experimental basis for subsequent studies incorporating more complex feature representations, ensemble learning strategies, or validation on larger-scale datasets.

References

[1] Pennacchio L A, Bickmore W, Dean A, et al. Enhancers: five essential questions[J]. Nature Reviews. Genetics, 2013, 14(4): 288-295.

[2] Maston G A, Evans S K, Green M R. Transcriptional regulatory elements in the human genome[J]. Annual Review of Genomics and Human Genetics, 2006, 7: 29-59.

[3] Whyte W A, Orlando D A, Hnisz D, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes[J]. Cell, 2013, 153(2): 307-319.

[4] Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions[J]. Nature Reviews. Genetics, 2014, 15(4): 272-286.

[5] Lee J Y. The principles and applications of high-throughput sequencing technologies[J]. Development & Reproduction, 2023, 27(1): 9-24.

[6] Kheradpour P, Ernst J, Melnikov A, et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay[J]. Genome Research, 2013, 23(5): 800-811.

[7] Liu B, Fang L, Long R, et al. iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition[J]. Bioinformatics (Oxford, England), 2016, 32(3): 362-369.

[8] Ghandi M, Mohammad-Noori M, Beer M A. Robust k-mer frequency estimation using gapped k-mers.[J]. Journal of mathematical biology, 2014, 69(2): 469-500.

[9] Chen Z, Zhao P, Li C, et al. iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization[J]. Nucleic Acids Research, 2021, 49(10): e60-e60.

[10] Chen Y Z, Tang Y R, Sheng Z Y, et al. Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs[J]. BMC Bioinformatics, 2008, 9: 101.

[11] Liu B, Liu F, Fang L, et al. repDNA: a python package to generate various modes of

feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects[J]. Bioinformatics (Oxford, England), 2015, 31(8): 1307-1309.

[12] H. T. Lin and C. J. Lin, "A Study on Sigmoid Kernels for SVM and the Training of Non-PSD Kernels by SMO- Type Methods," Department of Computer Science and Information Engineering, National Taiwan University, Taiwan, 2003. - References - Scientific Research Publishing[EB/OL]. [2025-08-11]. https://www.scirp.org/reference/referencespapers?referenceid=133464.

[13] Chen T, Guestrin C. XGBoost: a scalable tree boosting system[J]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and DATA Mining, 2016.

[14] Chen W, Feng P M, Lin H, et al. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition[J]. Nucleic Acids Research, 2013, 41(6): e68.

[15] Xu Y, Shao X J, Wu L Y, et al. iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins[J]. PeerJ, 2013, 1: e171.

[16] Arlot S, Celisse A. A survey of cross-validation procedures for model selection[J]. Statistics Surveys, 2009, 4: 40-79.