Hindawi
Computational and Mathematical Methods in Medicine
Volume 2020, Article ID 8852258, 9 pages
https://doi.org/10.1155/2020/8852258

Research Article

Identification and Classification of Enhancers Using Dimension Reduction Technique and Recurrent Neural Network

Qingwen Li,[1,2] **Lei Xu**,[3] **Qingyuan Li**,[4] **and Lichao Zhang**[5]

[1]*College of Animal Science and Technology, Northeast Agricultural University, Harbin, China*
[2]*Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China*
[3]*School of Electronic and Communication Engineering, Shenzhen Polytechnic, Shenzhen, China*
[4]*Forestry and Fruit Tree Research Institute, Wuhan Academy of Agricultural Sciences, Wuhan, China*
[5]*School of Intelligent Manufacturing and Equipment, Shenzhen Institute of Information Technology, Shenzhen, China*

Correspondence should be addressed to Qingyuan Li; liqingyuan@webmail.hzau.edu.cn
and Lichao Zhang; lczhang5354@szu.edu.cn

Received 25 August 2020; Revised 16 September 2020; Accepted 30 September 2020; Published 19 October 2020

Academic Editor: Hui Ding

Enhancers are noncoding fragments in DNA sequences, which play an important role in gene transcription and translation. However, due to their high free scattering and positional variability, the identification and classification of enhancers have a higher level of complexity than those of coding genes. In order to solve this problem, many computer studies have been carried out in this field, but there are still some deficiencies in these prediction models. In this paper, we use various feature extraction strategies, dimension reduction technology, and a comprehensive application of machine model and recurrent neural network model to achieve an accurate prediction of enhancer identification and classification with the accuracy of was 76.7% and 84.9%, respectively. The model proposed in this paper is superior to the previous methods in performance index or feature dimension, which provides inspiration for the prediction of enhancers by computer technology in the future.

1. Introduction

Enhancers are a small area of DNA that can link with protein, located upstream or downstream of the gene, and gene transcription will be enhanced after they bind with protein [1]. Because of the winding structure of chromatin, enhancers being far apart in the sequence still have the opportunity to contact each other. Therefore, they are not necessarily close to the gene to be affected, or even located on the same chromosome as the gene. Studies have shown that enhancer mutations may lead to a variety of diseases.

Owing to the significance of enhancers, the identification and classification of enhancers have always been the focus of computational biologists and experimental biologists [2, 3]. The fact is that to identify enhancers by biochemical experiments is expensive and time-consuming.

In the past few years, some bioinformatics methods have been developed to predict enhancers [4]. Liu et al. [5] proposed iEnhancer-2L, which extracts features by pseudo *k*-tuple nucleotide composition and achieves the enhancer identification and classification with the accuracy of 73% and 60.5%, respectively. Jia and He [6] suggested EnhancerPred, which extracts features by biprofile Bayes and pseudo *k*-tuple nucleotide composition to support the vector machine and achieves the accuracy of 75% and 55% for the prediction of enhancer identification and classification, respectively, Liu et al. [7] proposed iEnhancer-EL, which applies K-mer, pseudo *k*-tuple nucleotide composition and subsequence profile feature extraction methods and uses the ensemble classifier based on support vector machine to achieve the accuracy of 74.8% for enhancer identification and 61% for enhancer classification [8]. Nguyen et al. [9] proposed iEnhancer-ECNN, which uses a convolutional neural network to achieve the accuracy of 76.9% for enhancer identification and 67.8% for enhancer classification prediction [10]. All of the above methods emphasize the better prediction results but fail to mention the dimensional advantages of the model [11, 12]. Due to the fact that high-dimensional

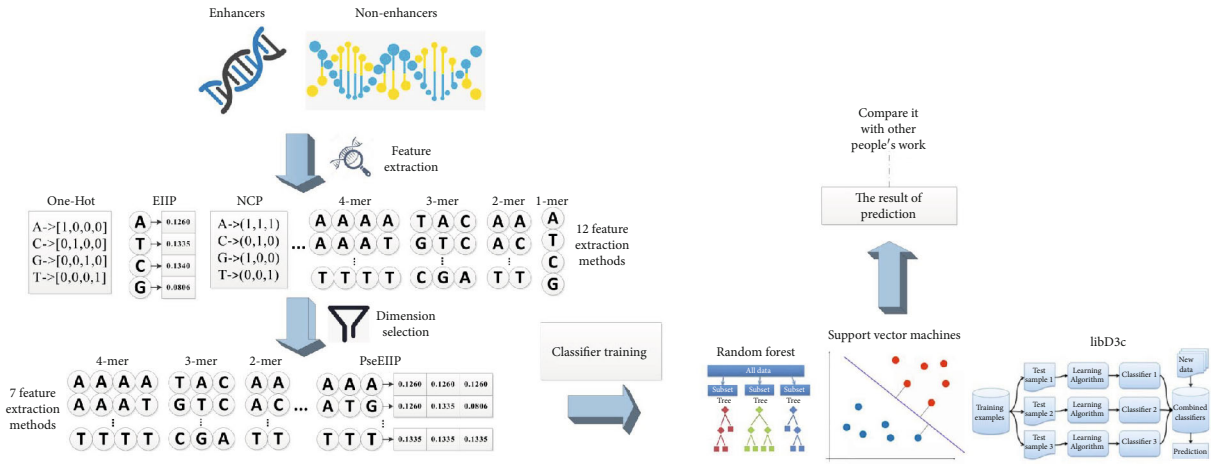

Figure 1: Research process of enhancer identification.

features may lead to an over-fitting and high-dimension disaster or an increase of redundant information, the machine learning model trained by this initial high-dimension feature is often found to be underperforming in practice [13–17].

In this paper, a low dimensional feature model is obtained by using a variety of feature extraction strategies and dimension reduction technology [18–23]. The identification and classification of enhancers have been achieved via the combination of machine learning models and artificial neural network with the accuracy rate of 76.7% and 84.9%, respectively. It also should be noted that the dimension of the feature model used to identify enhancers is only 37, which is much lower than the past methods. And this paper also got an 18-dimension feature model for enhancer identification, and its accuracy reached 76.5% after testing.

2. Materials and Methods

In this paper, the identification and classification of enhancers are described by Figures 1 and 2, respectively.

2.1. Benchmark Dataset. This paper used a dataset proposed by Liu et al., which was also used in the development of iEnhizer-2L, EnhancerPred, iEnhancer-EL, and iEnhancer-ECNN. In this dataset, enhancer information was collected from 9 different cell lines, and DNA sequences of 200 bp in length were extracted. In order to avoid the deviation of the classifier, enhancers with the similarity of over 90% were deleted from the dataset through CD-HIT [24, 25]. The dataset contains 1484 enhancers and 1484 nonenhancers. Among them, 1484 enhancers include 742 strong enhancers and 742 weak enhancers.

2.2. Feature Extraction. Machine learning algorithms cannot directly perform annotations on continuous nucleotide sequences, so it is necessary to convert nucleotide sequences represented by strings into feature vectors represented by numbers [26–28]. This paper implemented feature extraction through iLearn [29].

2.2.1. K-mer. The K-mer feature extraction strategy refers to calculating the frequency of the unit in the entire sequence with k adjacent nucleotides as a unit [30, 31]. This paper uses 1-mer, 2-mer, 3-mer, and 4-mer feature extraction methods, which are stated by the following formulas:

$$\begin{aligned}
1-\text{mer}: f(a) &= \frac{N_a}{N_t}, a \in (\text{A, T, C, G}),\\
2-\text{mer}: f(a,b) &= \frac{N_a}{(N_t-1)}, a, b \in (\text{A, T, C, G}),\\
3-\text{mer}: f(a,b,c) &= \frac{N_{abc}}{(N_t-2)}, a, b, c \in (\text{A, T, C, G}),\\
4-\text{mer}: f(a,b,c,d) &= \frac{N_{abcd}}{(N_t-3)}, a, b, c, d \in (\text{A, T, C, G}).
\end{aligned} \tag{1}$$

N_t is the length of a DNA sequence and N_a, N_{ab}, N_{abc}, N_{abcd} are the units composed of adjacent K nucleotides.

2.2.2. Reverse Compliment K-mer (RCK-mer). Reverse Compliment K-mer is a variant of K-mer, which ignores the complementary sequences of adjacent nucleotide sequences. For example, there are 16 types of 2-mer: "AA," "AC," "AG," "AT," "CA," "CC," "CG," "CT," "GA," "GC," "GG," "GT," "TA," "TC," "TG," and "TT." Because 'TT' is the reverse completion K-mer of "AA," it can be left out. Therefore, there are only 10 kinds of 2-mer in this method: "AA," "AC," "AG," "AT," "CA," "CC," "CG," "GA," "GC," and "TA." The frequency of each K-mer was calculated in turn.

2.2.3. Enhanced Nucleic Acid Composition (ENAC). Enhanced nucleic acid composition is the frequency of each nucleotide occurring within a fixed sequence window length, which slides continuously from the 5′ end to the 3′ end of each nucleotide sequence and usually used to encode nucleotide sequences of the same length.

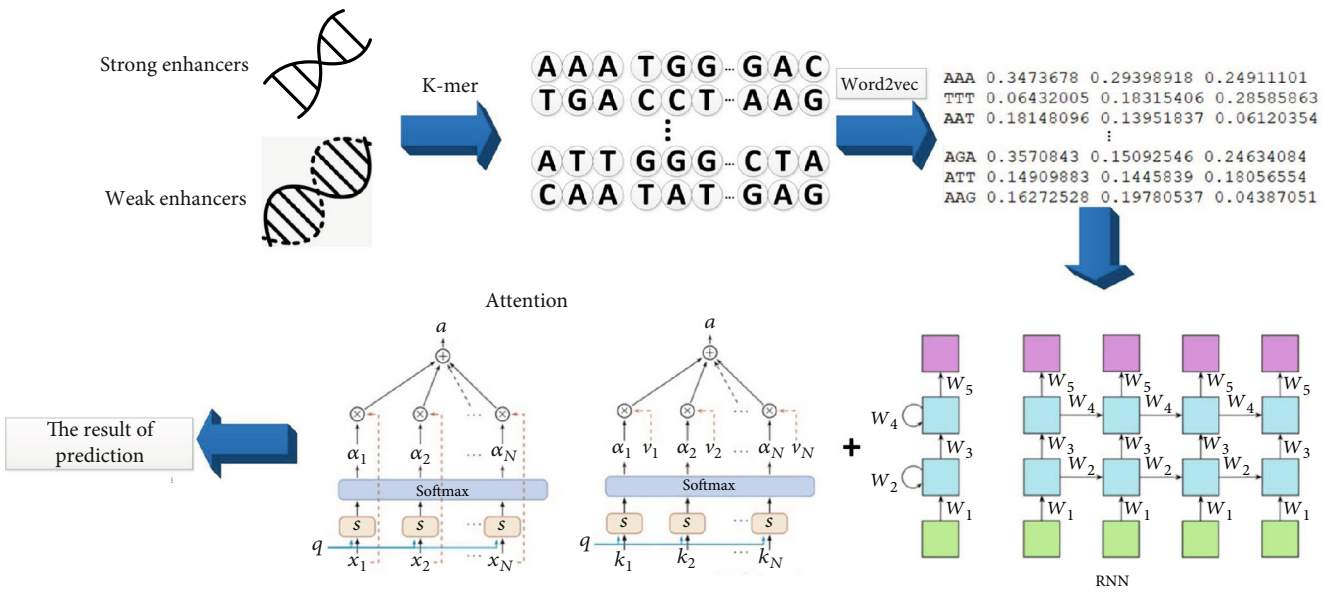

Figure 2: Research process of enhancer classification.

2.2.4. Composition of K-Spaced Nucleic Acid Pairs (CKSNAP). This method calculated the frequency of pairs of nucleotides separated by K nucleotides in the whole sequence. When $k = 0$, it is consistent with the features represented by 2-mer. It should be noted that the frequency of nucleotide pairs is calculated though, when $k = 0, 1, 2, 3, 4$, and 5, the length of sequences should be L-1, L-2, L-3, L-4, L-5, and L-6.

2.2.5. Nucleotide Chemical Property (NCP). The method took into account different chemical structures and chemical properties of four nucleotides [32, 33]. "A" is presented as (1, 1, 1), "C" as (0, 1, 0), "G" as (1, 0, 0), and "T" as (0, 0, 1).

2.2.6. Accumulated Nucleotide Frequency (ANF). This method combined the approach of nucleotide chemical properties and considers the chemical properties, the location, and the frequency of each nucleotide. For example, for a sequence "TCGTTCATGG," "T" appears in bits 1, 4, 5, and 8, with frequencies corresponding to 1 (1/1), 0.5 (2/4), 0.6 (3/5), and 0.5 (4/8), respectively; "C" appears in bits 2 and 6, with frequencies corresponding to 0.5 (1/2) and 0.33 (2/6), respectively; "G" appears in bits 3, 9, and 10, with frequencies corresponding to 0.33 (1/3), 0.22 (2/9), and 0.3 (3/10), respectively; "A" appears in the 7th position, so its frequency was 0.14 (1/7). Therefore, the sequence can be expressed as {(0, 0, 1, 1), (0, 1, 0, 0.5), (1, 0, 0, 0.33), (0, 0, 1, 0.5), (0, 0, 1, 0.6), (0, 1, 0, 0.33), (1, 1, 1, 0.14), (0, 0, 1, 0.5), (1, 0, 0, 0.22), (1, 0, 0, 0.3)} [34, 35].

2.2.7. Electron-Ion Interaction Pseudopotentials of Trinucleotide (EIIP). Nair and Pillai [36] proposed the Electron-Ion Interaction Pseudopotentials of Trinucleotide (EIIP) of nucleotides A, G, C, and T. The EIIP of the four nucleotides is A: 0.1260, C: 0.1340, G: 0.0806, and T: 0.1335. This method directly used the EIIP to represent the nucleotides in the DNA sequence. Therefore, the dimension of EIIP is the length of the DNA sequence.

2.2.8. Electron-Ion Interaction Pseudopotentials of Trinucleotide (PseEIIP). In these codes, EIIPA, EIIPT, EIIPG, and EIIPC were used to represent the EIIP of nucleotides A, T, G, and C, respectively. Then, the average value of EIIP of the three nucleotides in each sample was used to construct the feature vector, which can be expressed as follows:

$$V = [\mathrm{EIIP}_{\mathrm{AAA}} \times f_{\mathrm{AAA}}, \mathrm{EIIP}_{\mathrm{AAC}} \times f_{\mathrm{AAC}}, \mathrm{EIIP}_{\mathrm{AAG}} \times f_{\mathrm{AAG}}, \cdots, \mathrm{EIIP}_{\mathrm{TTG}} \times f_{\mathrm{TTG}}, \mathrm{EIIP}_{\mathrm{TTT}} \times f_{\mathrm{TTT}}]_{64}. \quad (2)$$

$f_{abc}, a, b, c \in (\mathrm{A}, \mathrm{T}, \mathrm{C}, \mathrm{G})$ is the normalized frequency of a trinucleotide, and $\mathrm{EIIP}_{abc}, a, b, c \in (\mathrm{A}, \mathrm{T}, \mathrm{C}, \mathrm{G})$ is the sum of EIIP values of three nucleotides.

2.2.9. One-Hot. Each enhancer in the dataset is a 200 bp nucleotide sequence, which consists of four nucleotides, namely, adenine (A), guanine (G), cytosine (C) and thymine (T). Each nucleotide is represented by a set of vectors (Table 1) [37, 38].

2.3. Feature Selection. Feature selection is the method of selecting a subset of related features used in model construction [39, 40]. Because the dimension of features will be reduced after selection, this process is called dimension reduction.

2.3.1. MRMD2.0. This paper used MRMD2.0 [41] to achieve dimension reduction. Firstly, MRMD2.0 uses seven main feature ranking methods (ANOVA, MRMD, MIC, Lasso, mRMR, chi-square test, and RFE) to calculate the feature sets, respectively, and then uses the idea of the PageRank algorithm to comprehensively process the results of the seven feature ranking algorithms and get the final feature ranking, Then, using the positive addition strategy, the features arranged in descending order are added to the feature subset for verification, and the best feature subset is finally obtained.

Table 1: One-Hot encoding.

Nucleotides	Code
A	[1,0,0,0]
T	[0,0,0,1]
C	[0,1,0,0]
G	[0,0,1,0]

2.3.2. Evolutionary Search. Evolutionary Search uses evolutionary algorithms for feature selection. An evolutionary algorithm is not a specific algorithm; it includes a variety of algorithms (genetic algorithm, memetic algorithm, and multiobjective evolutionary algorithm). The inspiration of the evolutionary algorithm draws on the evolutionary operations of living things in nature. Compared with traditional optimization algorithms such as calculus-based methods and exhaustive methods, it is a mature global with high robustness and wide applicability. The optimization method has the characteristics of self-organization, self-adaptation, and self-learning. It is not limited by the nature of the problem and can effectively handle complex problems that are difficult to solve by traditional optimization algorithms.

2.4. Classifier

2.4.1. Recurrent Neural Network. This paper also used recurrent neural networks to make predictions on the basis of the memory model. It is expected that the network can remember the previous features and infer the subsequent results according to the features; hence, the overall network structure continues in the cycle. The biggest problem with memory is that it has forgetfulness. We can always remember the recent events more clearly and forget the events that happened long ago. Recurrent neural networks also have this problem. In order to solve this problem, two variants of the network structure have emerged: one is called LSTM, and the other is called GRU. Both of these variants can well solve the problem of long-term dependence.

2.4.2. Random Forest. In this study, a random forest was applied to play a role as a classifier for prediction. Random forest is widely employed in the bioinformatics research [42–52]. This classifier concludes multiple decision trees while the output category is arranged by the mode of the category output by trees individually. This paper implemented a random forest classifier through the weka platform.

2.4.3. Support Vector Machine. As a very powerful machine learning method widely used in biological sequence prediction [53–71], the support vector machine was used for prediction in this research. It is a class of generalized linear classifiers that classify data binary in a supervised learning method, and its decision boundary is the maximum margin hyperplane that is solved for the learning sample. This paper used libSVM to implement support vector machine and adjust parameters c and g using grid to optimize the prediction results.

2.4.4. libD3C. This paper also applied the libD3C classifier [72] to test the performance of models. The classifier adopts a selective ensemble strategy, based on the hybrid ensemble pruning model combining k-means clustering and function selection cycle framework and sequential search, by training multiple candidate classifiers and then selecting a set of accurate and different classifiers to settle the problem.

2.5. Evaluation of Prediction. This paper used sensitivity (Sn), specificity (Sp), total accuracy (Acc), and Mathew (Mcc) correlation coefficients to evaluate the performance of the model [73–83].

$$\begin{aligned} \mathrm{Sn} &= \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})}, \\ \mathrm{Sp} &= \frac{\mathrm{TN}}{(\mathrm{TN}+\mathrm{FP})}, \\ \mathrm{Acc} &= \frac{(\mathrm{TP}+\mathrm{TN})}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})}, \\ \mathrm{Mcc} &= \frac{(\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})}}. \end{aligned} \tag{3}$$

TP is true positive; FN is false negative; FP is false positive; TN is true negative.

3. Results and Discussion

3.1. Identification of Enhancers. Feature vectors of enhancers and nonenhancers were obtained by K-mer, RCK-mer, ENAC, CKSNAP, NCP, ANF, EIIP, PseEIIP, and One-Hot feature extraction methods. In order to determine which feature extraction methods were suitable for the identification of enhancers, the random forest was adopted through ten-fold cross-validation for each method. After testing (Figure 3), this paper believed that 2-mer, 3-mer, 4-mer, CKSNAP, ENAC, PseEIIP, and RCK-mer, the seven feature extraction methods, were more effective. Since the dimension of the feature model obtained through the seven extraction methods was rather high, which could cause the classifier overfitting the training set and lead to a less effective performance in practical applications. This paper expected to get a low-dimension and excellent performance feature model; hence, the seven feature models were merged after individual dimension reduction through MRMD2.0; then, we found that the dimension was 1049, which was still relatively high. Therefore, the merged model went through 5 consecutive dimension reductions by MRMD2.0, and a 37-dimension feature model was achieved eventually. At this time, the dimension can no longer be reduced further (Figure 4). Using the random forest classifier, the 37-dimension feature model was tested through ten-fold cross-validation (Table 2), and the accuracy reached 76.7%; the running time of the method is 2.14 seconds.

At the same time, this paper used Evolutionary Search to reduce the dimension of the merged 1049-dimensional

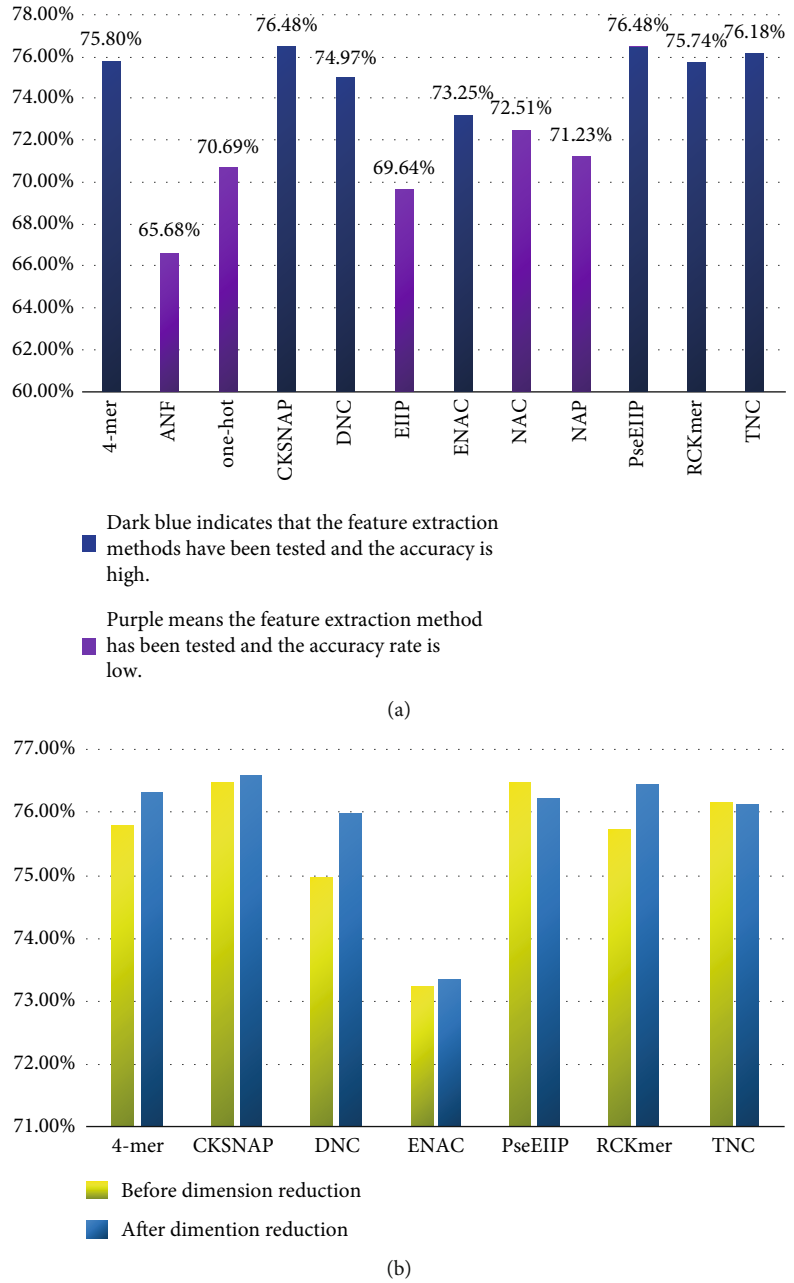

Figure 3: (a) The accuracy of different feature extraction methods after verification. Through analysis, this article believed that the method represented by dark blue had higher accuracy, while the method represented by purple had lower accuracy. (b) Changes in accuracy of different extraction methods before and after dimensionality reduction. Through analysis, this paper believed that accuracy has improved after dimensionality reduction.

model to compare the differences between different dimension reduction tools. After 8-dimension reductions, an 18-dimension model was obtained in this paper, and the accuracy rate reached 76.5% after 10-fold cross-validation. Although this feature model is inferior to the model obtained by MRMD2.0 in performance, it has obvious advantages in dimension. The 18-dimensional feature model may imply that it is an important marker for distinguishing enhancers. These 18-dimension features come from 4-mer, 2-mer, CKSNAP, RCK-mer, and PseEIIP, respectively, indicating that specific dinucleotides, trinucleotides, and their electronic-ion interactions play an important role in enhancer sequences. By using two tools, we can find that Evolutionary Search has an advantage in dimension after dimension reduction, and MRMD2.0 has more advantages in terms of performance parameters after dimension reduction.

In order to further determine the stability of the feature model, this paper used support vector machine and libD3C to test the 37-dimension model at the same time (Table 2). Through the support vector machine combined with the grid search method (c 8192.0, g 0.001953125), the accuracy reached 76.5%. Using the libD3C classifier, the accuracy

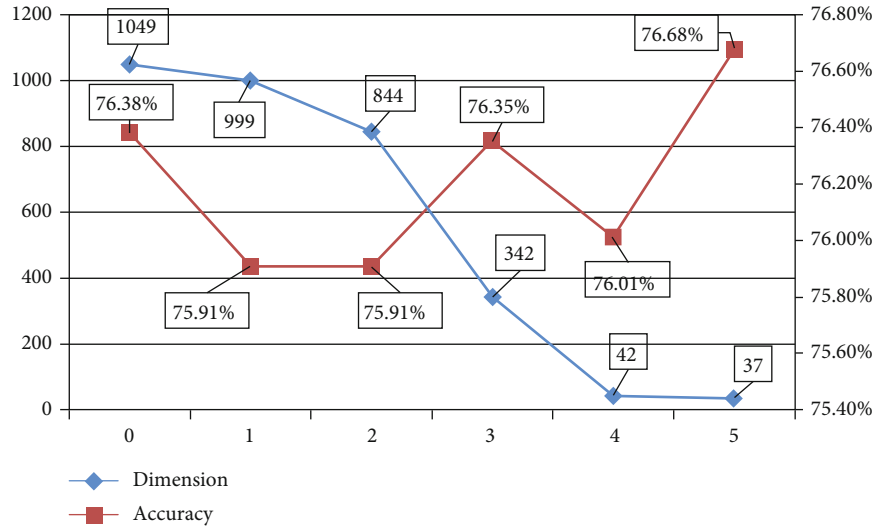

Figure 4: The relationship between accuracy change and dimension change. According to trends, this paper believed that dimension and accuracy are negatively correlated. Using MRMD2.0, when the dimension was 37, the accuracy reached 76.68%, and the dimension reduction continued; the accuracy cannot be improved.

Table 2: The comparison between this paper and the previous work on enhancer identification.

	Acc	AUC	SN	SP	MCC	Dimension
iEnhancer-2L	0.730	0.806	0.710	0.750	0.460	
EnhancerPred	0.740	0.801	0.735	0.745	0.480	
iEnhancer-EL	0.748	0.817	0.710	0.785	0.496	
iEnhancer-ECNN	0.769	0.832	0.785	0.752	0.537	2400
Our method	0.767	0.837	0.733	0.801	0.535	37

Table 3: The comparison between this paper and the previous work on enhancer classification.

	Acc	SN	SP	MCC
iEnhancer-2L	0.605	0.470	0.740	0.218
EnhancerPred	0.550	0.45	0.65	0.102
iEnhancer-EL	0.61	0.540	0.68	0.222
iEnhancer-ECNN	0.678	0.791	0.564	0.368
Our method	0.849	0.858	0.84	0.699

reached 75.5%. The prediction accuracy of the three classifiers for the feature model all exceeded 75%, indicating a very stable feature model. Meanwhile, in addition to the excellent performance of the feature model examined in this paper, it also has a very low dimension compared with a previous work (Table 2), which can effectively avoid dimensional disasters.

3.2. Classification of Enhancers. For the feature extraction of strong enhancers and weak enhancers, the same methods as enhancer identification were adopted, and then, the random forest was used through ten-fold cross-validation to examine the performance. After testing, this paper believed that also 2-mer, 3-mer, 4-mer, CKSNAP, ENAC, PseEIIP, and RCK-mer, the seven feature extraction methods, perform slightly better than other methods, but were not satisfactory. Therefore, this paper attempted to improve accuracy through dimension reduction techniques. After reducing the dimensions of the seven feature models that performed slightly better, they were merged to continue the dimension reduction. After four dimension reductions, an 82-dimension feature model was obtained. At this time, it was impossible to continue the further dimension reduction. The 82-dimension model was cross-validated with a random forest classifier, and the accuracy of 62.3% was still not ideal.

Next, this paper used the voting mechanism to output the prediction results of the 82 feature model of the three classifiers libSVM, random forest, and libD3C and retained the prediction results with the highest confidence based on the given confidence of each classifier result. After statistics, the final accuracy was 63.1%, the result was still not ideal.

As the recurrent neural network has contributed a lot in the fields of sequence problems and natural language processing with a limited capacity of memory, the variant of recurrent neural network—Long Short-Term Memory—was applied in this research to predict biological sequences. This paper used the 3-mer method to segment the sequence and then trained the word embedding through word to vector. Next, this study used the LSTM model based on the attention mechanism to predict the word segmentation file. When the model was a two-layer neuron, hidden_dim was 100, the learning rate was 0.005, and the adam optimizer was used; the accuracy of ten-fold cross-validation reached 84.9%. After comparison (Table 3), this paper has achieved ideal results in the classification of enhancers.

4. Conclusions

In this paper, a 37-dimension feature model for identifying enhancers was obtained through multiple dimension

reductions. After testing, the performance of the model was sound and stable. At the same time, this paper has achieved ideal results in the classification of enhancers through 3-mer methods, word to vector techniques, and RNN models. It is expected that the method proposed in this paper can provide a certain reference for the future research on enhancers in the academic world.

Data Availability

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Authors' Contributions

Qingwen Li and Lei Xu contributed equally to this work.

Acknowledgments

This work is supported by the Funding of Shenzhen Polytechnic (No. 6020320002K). This manuscript used iLearn online tool to extract features, used classifiers through Weka platform, and used MRMD2.0 and Evolutionary Search to reduce dimensions. Dongyuan Yu contributed to the language editing of this article. Dongyuan Yu is from Northeast Agricultural University.

References

[1] Y. H. Li, C. Y. Yu, X. X. Li et al., "Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics," *Nucleic Acids Research*, vol. 46, no. D1, pp. D1121–D1127, 2018.

[2] B. Li, J. Tang, Q. Yang et al., "NOREVA: normalization and evaluation of MS-based metabolomics data," *Nucleic Acids Research*, vol. 45, no. W1, pp. W162–W170, 2017.

[3] J. Fu, J. Tang, Y. Wang et al., "Discovery of the consistently well-performed analysis chain for SWATH-MS based pharmacoproteomic quantification," *Frontiers in Pharmacology*, vol. 9, p. 681, 2018.

[4] H. Bu, J. Hao, J. Guan, and S. Zhou, "Predicting Enhancers from Multiple Cell Lines and Tissues across Different Developmental Stages Based On SVM Method," *Current Bioinformatics*, vol. 13, no. 6, pp. 655–660, 2018.

[5] B. Liu, L. Fang, R. Long, X. Lan, and K. C. Chou, "iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition," *Bioinformatics*, vol. 32, no. 3, pp. 362–369, 2016.

[6] C. Jia and W. He, "EnhancerPred: a predictor for discovering enhancers based on the combination and selection of multiple features," *Scientific Reports*, vol. 6, no. 1, p. 38741, 2016.

[7] B. Liu, K. Li, D. S. Huang, and K. C. Chou, "iEnhancer-EL:identifying enhancers and their strength with ensemble learning approach," *Bioinformatics*, vol. 34, no. 22, pp. 3835–3842, 2018.

[8] J. Tang, J. Fu, Y. Wang et al., "Simultaneous improvement in the precision, accuracy, and robustness of label-free proteome quantification by optimizing data manipulation chains," *Molecular & Cellular Proteomics*, vol. 18, no. 8, pp. 1683–1699, 2019.

[9] Q. H. Nguyen, T. H. Nguyen-Vo, N. Q. K. le, T. T. T. Do, S. Rahardja, and B. P. Nguyen, "iEnhancer-ECNN: identifying enhancers and their strength using ensembles of convolutional neural networks," *BMC Genomics*, vol. 20, Suppl 9, p. 951, 2019.

[10] W. Xue, F. Yang, P. Wang et al., "What contributes to serotonin-norepinephrine reuptake inhibitors' dual-targeting mechanism? The key role of transmembrane domain 6 in human serotonin and norepinephrine transporters revealed by molecular dynamics simulation," *ACS Chemical Neuroscience*, vol. 9, no. 5, pp. 1128–1140, 2018.

[11] Q. Yang, Y. Wang, Y. Zhang et al., "NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data," *Nucleic Acids Research*, vol. 48, no. W1, pp. W436–W448, 2020.

[12] J. Yin, W. Sun, F. Li et al., "VARIDT 1.0: variability of drug transporter database," *Nucleic Acids Research*, vol. 48, no. D1, pp. D1042–D1050, 2020.

[13] Q. Li, W. Zhou, D. Wang, S. Wang, and Q. Li, "Prediction of anticancer peptides using a low-dimensional feature model," *Frontiers in Bioengineering and Biotechnology*, vol. 8, 2020.

[14] L. Cheng, "Omics data and artificial intelligence: new challenges for gene therapy," *Current Gene Therapy*, vol. 20, no. 1, p. 1, 2020.

[15] B. Liu, X. Gao, and H. Zhang, "BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA, and protein sequences at sequence level and residue level based on machine learning approaches," *Nucleic Acids Research*, vol. 47, no. 20, article e127, 2019.

[16] L. Xu, G. Liang, S. Shi, and C. Liao, "SeqSVM: A Sequence-Based Support Vector Machine Method for Identifying Antioxidant Proteins," *International Journal of Molecular Sciences*, vol. 19, no. 6, p. 1773, 2018.

[17] Y. Xu, Y. Wang, J. Luo, W. Zhao, and X. Zhou, "Deep learning of the splicing (epi) genetic code reveals a novel candidate mechanism linking histone modifications to ESC fate decision," *Nucleic Acids Research*, vol. 45, no. 21, pp. 12100–12112, 2017.

[18] Y. Wang, S. Zhang, F. Li et al., "Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics," *Nucleic Acids Research*, vol. 48, no. D1, pp. D1031–D1041, 2020.

[19] L. Yu, F. Xu, and L. Gao, "Predict new therapeutic drugs for hepatocellular carcinoma based on gene mutation and expression," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 8, 2020.

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

[21] J. Shao, K. Yan, and B. Liu, "FoldRec-C2C: protein fold recognition by combining cluster-to-cluster model and protein similarity network," *Briefings in Bioinformatics*, 2020.

[22] Q. Zou, L. Chen, T. Huang, Z. Zhang, and Y. Xu, "Machine-learning and analytics in biomedicine," *Artificial Intelligence in Medicine*, vol. 83, p. 1, 2017.

[23] Q. Zou, D. Mrozek, Q. Ma, and Y. Xu, "Scalable data mining algorithms in computational biology and biomedicine," *BioMed Research International*, vol. 2017, 3 pages, 2017.

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

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

[26] B. Liu and K. Li, "iPromoter-2L2.0: identifying promoters and their types by combining Smoothing Cutting Window algorithm and sequence-based features," *Molecular Therapy-Nucleic Acids*, vol. 18, pp. 80–87, 2019.

[27] Y. H. Yang, C. Ma, J. S. Wang et al., "Prediction of N7-methylguanosine sites in human RNA based on optimal sequence features," *Genomics*, vol. 112, no. 6, pp. 4342–4347, 2020.

[28] M. L. Liu, W. Su, Z. X. Guan et al., "An overview on predicting protein subchloroplast localization by using machine learning methods," *Current Protein & Peptide Science*, vol. 21, 2020.

[29] Z. Chen, P. Zhao, F. Li et al., "iLearn: an integrated platform and meta-learner for feature engineering, machine learning analysis and modeling of DNA, RNA and protein sequence data," *Briefings in Bioinformatics*, vol. 21, no. 3, pp. 1047–1057, 2020.

[30] H. Yang, W. Yang, F. Y. Dao et al., "A comparison and assessment of computational method for identifying recombination hotspots in Saccharomyces cerevisiae," *Briefings in Bioinformatics*, vol. 21, no. 5, pp. 1568–1580, 2020.

[31] H. Y. Lai, Z. Y. Zhang, Z. D. Su et al., "iProEP: a computational predictor for predicting promoter," *Molecular therapy. Nucleic acids*, vol. 17, pp. 337–346, 2019.

[32] F. Y. Dao, H. Lv, Y. H. Yang, H. Zulfiqar, H. Gao, and H. Lin, "Computational identification of N6-methyladenosine sites in multiple tissues of mammals," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 1084–1091, 2020.

[33] H. Lv, Z. M. Zhang, S. H. Li, J. X. Tan, W. Chen, and H. Lin, "Evaluation of different computational methods on 5-methylcytosine sites identification," *Briefings in Bioinformatics*, vol. 21, no. 3, pp. 982–995, 2020.

[34] Y. Ding, J. Tang, and F. Guo, "Identification of Protein–Proteininteractions via a Matrix-Basedsequence model with acid Information," *International Journal of Molecular Sciences*, vol. 17, no. 10, p. 1623, 2016.

[35] Y. Ding, J. Tang, and F. Guo, "Predicting protein-protein interactions via multivariate mutual information of protein sequences," *BMC Bioinformatics*, vol. 17, no. 1, p. 398, 2016.

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

[37] H. Lv, F. Y. Dao, D. Zhang et al., "iDNA-MS: integrated tool for DNA sites in Genomes," *iScience*, vol. 23, no. 4, article 100991, 2020.

[38] F. Y. Dao, H. Lv, H. Zulfiqar et al., "A computational platform to identify origins of replication sites in eukaryotes," *Briefings in Bioinformatics*, 2020.

[39] Z. Y. Zhang, Y. H. Yang, H. Ding, D. Wang, W. Chen, and H. Lin, "Design powerful predictor for mRNA subcellular location prediction in Homo sapiens," *Briefings in Bioinformatics*, 2020.

[40] W. Yang, X. J. Zhu, J. Huang, H. Ding, and H. Lin, "A brief survey of machine learning methods in protein sub-Golgi localization," *Current Bioinformatics*, vol. 14, no. 3, pp. 234–240, 2019.

[41] S. He, F. Guo, Q. Zou, and H. Ding, "MRMD2.0:a Python tool for machine learning features ranking and reduction," *Current Bioinformatics*, vol. 15, 2020.

[42] Z. Lv, J. Zhang, H. Ding, and Q. Zou, "RF-PseU: A Random Predictor for RNA Sites," *Frontiers in Bioengineering and Biotechnology*, vol. 8, 2020.

[43] X. Q. Ru, L. H. Li, and Q. Zou, "Incorporating distance-based top-n-gram and random forest to identify electron transport proteins," *Journal of Proteome Research*, vol. 18, no. 7, pp. 2931–2939, 2019.

[44] L. Xu, G. Liang, C. Liao, G. D. Chen, and C. C. Chang, "k-Skip-n-Gram-RF: random method for Alzheimer's Identification," *Frontiers in Genetics*, vol. 10, 2019.

[45] Z. Lv, S. Jin, H. Ding, and Q. Zou, "A random forest sub-Golgi protein classifier optimized via dipeptide and amino acid composition features," *Frontiers in Bioengineering and Biotechnology*, vol. 7, p. 215, 2019.

[46] L. Cheng, Y. Jiang, H. Ju et al., "InfAcrOnt: calculating cross-ontology term similarities using information flow by a random walk," *BMC Genomics*, vol. 19, Suppl 1, p. 919, 2018.

[47] L. Yu, S. Yao, L. Gao, and Y. Zha, "Conserved disease modules extracted from multilayer heterogeneous disease and gene networks for understanding disease mechanisms and predicting disease treatments," *Frontiers in Genetics*, vol. 9, 2019.

[48] L. Yu, R. Su, B. Wang et al., "Prediction of novel drugs for hepatocellular carcinoma based on multi-source random walk," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 4, pp. 966–977, 2017.

[49] B. Liu, S. Chen, K. Yan, and F. Weng, "iRO-PsekGCC: identify DNA replication origins based on pseudo k-tuple GC composition," *Frontiers in Genetics*, vol. 10, p. 842, 2019.

[50] M. Wang, L. Yue, X. Cui et al., "Prediction of extracellular matrix proteins by fusing multiple feature information, elastic net, and random forest algorithm," *Mathematics*, vol. 8, no. 2, p. 169, 2020.

[51] X. Wang, B. Yu, A. Ma, C. Chen, B. Liu, and Q. Ma, "Protein–protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique," *Bioinformatics*, vol. 35, no. 14, pp. 2395–2402, 2019.

[52] H. Shi, S. Liu, J. Chen, X. Li, Q. Ma, and B. Yu, "Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure," *Genomics*, vol. 111, no. 6, pp. 1839–1852, 2019.

[53] C. Meng, F. Guo, and Q. Zou, "CWLy-SVM: a support vector machine-based tool for identifying cell wall lytic enzymes," *Computational Biology and Chemistry*, vol. 87, p. 107304, 2020.

[54] Y. Wang, F. Shi, L. Cao et al., "Morphological segmentation analysis and texture-based support vector machines classification on mice liver fibrosis microscopic images," *Current Bioinformatics*, vol. 14, no. 4, pp. 282–294, 2019.

[55] L. Chao, L. Wei, and Q. Zou, "SecProMTB: a SVM-based classifier for secretory proteins of Mycobacterium tuberculosis with imbalanced data set," *Proteomics*, vol. 19, article e1900007, 2019.

[56] N. Zhang, Y. Sa, Y. Guo, W. Lin, P. Wang, and Y. Feng, "Discriminating Ramos and Jurkat cells with image textures from diffraction imaging flow cytometry based on a support vector

machine," *Current Bioinformatics*, vol. 13, no. 1, pp. 50–56, 2018.

[57] C. Meng, S. Jin, L. Wang, F. Guo, and Q. Zou, "AOPs-SVM: A Sequence-Based Classifier of Antioxidant Proteins Using a Support Vector Machine," *Frontiers in Bioengineering and Biotechnology*, vol. 7, 2019.

[58] L. Yu and L. Gao, "Human pathway-based disease network," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 4, pp. 1240–1249, 2019.

[59] B. Liu, C. Li, and K. Yan, "DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks," *Briefings in Bioinformatics*, vol. 21, no. 5, pp. 1733–1741, 2020.

[60] T. D. Capellini, G. Vaccari, E. Ferretti et al., "Scapula development is governed by genetic interactions of Pbx1 with its family members and with Emx2 via their cooperative control of Alx1," *Development*, vol. 137, no. 15, pp. 2559–2569, 2010.

[61] H. Wang, Y. Ding, J. Tang, and F. Guo, "Identification of membrane protein types via multivariate information fusion with Hilbert-Schmidt Independence Criterion," *Neurocomputing*, vol. 383, pp. 257–269, 2020.

[62] Y. Shen, J. Tang, and F. Guo, "Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC," *Journal of Theoretical Biology*, vol. 462, pp. 230–239, 2019.

[63] Y. Shen, Y. Ding, J. Tang, Q. Zou, and F. Guo, "Critical evaluation of web-based prediction tools for human protein subcellular localization," *Briefings in Bioinformatics*, vol. 21, no. 5, pp. 1628–1640, 2020.

[64] Y. Ding, J. Tang, and F. Guo, "Identification of drug-target interactions via multiple information integration," *Information Sciences*, vol. 418, pp. 546–560, 2017.

[65] L. Xu, G. Liang, C. Liao, G. D. Chen, and C. C. Chang, "An efficient classifier for Alzheimer's disease genes identification," *Molecules*, vol. 23, no. 12, p. 3140, 2018.

[66] L. Xu, G. Liang, L. Wang, and C. Liao, "A novel hybrid sequence-based model for identifying anticancer peptides," *Genes*, vol. 9, no. 3, p. 158, 2018.

[67] L. Xu, G. Liang, B. Chen, X. Tan, H. Xiang, and C. Liao, "A computational method for the identification of endolysins and autolysins," *Protein & Peptide Letters*, vol. 26, 2019.

[68] X. Zhu, J. He, S. Zhao, W. Tao, Y. Xiong, and S. Bi, "A comprehensive comparison and analysis of computational predictors for RNA N6-methyladenosine sites of Saccharomyces cerevisiae," *Briefings in Functional Genomics*, vol. 18, no. 6, pp. 367–376, 2019.

[69] Y. Xiong, Q. Wang, J. Yang, X. Zhu, and D. Q. Wei, "PredT4SE-stack: prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method," *Frontiers in Microbiology*, vol. 9, p. 2571, 2018.

[70] J. He, T. Fang, Z. Zhang, B. Huang, X. Zhu, and Y. Xiong, "PseUI: pseudouridine sites identification based on RNA sequence information," *BMC Bioinformatics*, vol. 19, no. 1, p. 306, 2018.

[71] J. Kang, Y. Fang, P. Yao, N. Li, Q. Tang, and J. Huang, "NeuroPP: A Tool for the Prediction of Neuropeptide Precursors Based on Optimal Sequence Composition," *Interdisciplinary Sciences*, vol. 11, no. 1, pp. 108–114, 2019.

[72] C. Lin, W. Chen, C. Qiu, Y. Wu, S. Krishnan, and Q. Zou, "LibD3C: ensemble classifiers with a clustering and dynamic selection strategy," *Neurocomputing*, vol. 123, pp. 424–435, 2014.

[73] L. Cheng, H. Zhao, P. Wang et al., "Computational methods for identifying similar diseases," *Molecular therapy. Nucleic acids*, vol. 18, pp. 590–604, 2019.

[74] L. Cheng, C. Qi, H. Zhuang, T. Fu, and X. Zhang, "gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions," *Nucleic Acids Research*, vol. 48, no. D1, pp. D554–D560, 2020.

[75] Y. J. Tang, Y. H. Pang, and B. Liu, "IDP-Seq2Seq: Identification of Intrinsically Disordered Regions based on Sequence to Sequence Learning," *Bioinformatics*, 2020.

[76] Z. Wang, W. He, J. Tang, and F. Guo, "Identification of highest-affinity binding sites of yeast transcription factor families," *Journal of Chemical Information and Modeling*, vol. 60, no. 3, pp. 1876–1883, 2020.

[77] J. Li, Y. Pu, J. Tang, Q. Zou, and F. Guo, "DeepAVP: a dual-channel deep neural network for identifying variable-length antiviral peptides," *IEEE Journal of Biomedical and Health Informatics*, 2020.

[78] C. Shen, Y. Ding, J. Tang, L. Jiang, and F. Guo, "LPI-KTASLP: prediction of lncRNA-protein interaction by semi-supervised link learning with multivariate information," *IEEE Access*, vol. 7, pp. 13486–13496, 2019.

[79] Y. Ding, J. Tang, and F. Guo, "Identification of drug-side effect association via multiple information integration with centered kernel alignment," *Neurocomputing*, vol. 325, pp. 211–224, 2019.

[80] Y. Ding, J. Tang, and F. Guo, "Identification of drug-target interactions via fuzzy bipartite local model," *Neural Computing & Applications*, vol. 32, no. 14, pp. 10303–10319, 2020.

[81] X. Shan, X. Wang, C. D. Li et al., "Prediction of CYP450 enzyme-substrate selectivity based on the network-based label space division method," *Journal of Chemical Information and Modeling*, vol. 59, no. 11, pp. 4577–4586, 2019.

[82] Y. Chu, A. C. Kaushik, X. Wang et al., "DTI-CDF: a cascade deep forest model towards the prediction of drug-target interactions based on hybrid features," *Briefings in Bioinformatics*, 2019.

[83] A. Al-Ajlan and A. El Allali, "CNN-MGP: Convolutional Neural Networks for Gene Prediction," *Interdisciplinary Sciences*, vol. 11, no. 4, pp. 628–635, 2019.