iEnhancer-DNN: An Accurate identification of enhancer sites by heterogeneous feature using Deep Neural Network

Ali Raza[1], Maria Kanwal[2], Zafar Iqbal[1], Sanya Chaudhary[1], Sarah Gul[3]
Ashfaq Ahmad[1], Hamza Javed[1], Qadeer Yasin[1], Adil Shah[1]

[1] Department of Computer Science, MY University, Islamabad, Pakistan

[2] Department of Computer Science, National University of Technology (NUTECH), Islamabad, Pakistan

[3] Department of Biological Science, FBAS, International Islamic University, Islamabad, Pakistan

Abstract

Enhancers are the critical regulatory elements in DNA sequences that play an essential role in gene transcription and translation with in a genome. However, identifying enhancers is more complex than coding genes due to their high free scattering and positional variability. To address this challenge, numerous computational studies have been conducted in this field. Despite this, some deficiencies still exist in these prediction models. In this study, we propose a reliable computational approach for efficiently identifying enhancers based on a deep neural network model by incorporating heterogeneous features. The proposed model's effectiveness was evaluated using training and independent datasets through a 5-fold cross-validation approach. The validation results demonstrated that the iEnhancer-DNN model achieved an accuracy of 81.83%, respectively, when utilizing the training dataset. Similarly, when using the independent dataset, the model achieved an accuracy of 80.99%, respectively. Notably, our model outperforms previous methods in performance index, and providing valuable inspiration for the future of enhancer prediction using computer technology.

Keywords: Enhancer, DNA sequence, Identification, heterogeneous features, Deep neural network, Pseudo Dinucleotide Composition, Receiver Operating Characteristic, Matthews correlation coefficient

* *Corresponding Author: Ashfaq Ahmad, Assistant Professor, Head of Computer Science Department, MY University Islamabad, Pakistan.*

1. INTRODUCTION

Gene regulation is a fundamental process that governs the intricate orchestration of gene expression, enabling cells to respond to various stimuli and maintain proper functionality. Among the key regulatory elements in eukaryotic genomes, enhancers hold a crucial position as they play a vital role in modulating gene expression. Enhancers are cis-regulatory DNA elements that interact with transcription factors and other regulatory proteins to boost the transcriptional activity of specific target genes [1-3]. Identifying and characterizing enhancers are of paramount importance in unraveling the complex regulatory networks underlying diverse biological processes, including development, differentiation, and disease. In earlier times, the exploration of enhancers primarily relied on experimental approaches, as exemplified in the pioneering investigations conducted by [4]. The former sought to identify enhancers based on their connection with transcription factors (TFs), like P300 [5, 6]. Nevertheless, this method could potentially overlook or inadequately detect the relevant targets because not all enhancers are bound by TFs, resulting in a considerable number of incorrect rejections [7]. The latter method involved identifying enhancers through DNase I hypersensitivity, which might cause the erroneous or excessive identification of some other DNA segments or non-enhancer regions as enhancers [8, 9], leading to a significant number of false positives [7]. Despite efforts to address the aforementioned limitations in identifying enhancers and improving the detection rate through subsequent methods, such as genome-wide mapping of histone modifications, these approaches still incur substantial expenses and time requirements [10-15]. Consequently, numerous computational approaches have been proposed to predict enhancers, given that biological experimental techniques are costly and time-consuming. The CSI-ANN computation methodology was initially published by [13]. It comprises two main stages: data transformation and feature extraction, followed by classification using a time-delay neural network. In the past few years, several bioinformatics methods have been developed for the prediction of enhancers [16]. Subsequently, the Support Vector Machine (SVM) learning technique led to the creation of two successful systems: iEnhancer-2L [8] by Liu et al. and EnhancerPred [17] by Jia and He. While EnhancerPred utilized bi-profile Bayes and pseudo-nucleotide composition, iEnhancer-2L employed pseudo k-tuple nucleotide composition (PseKNC) for sequence encoding. Despite both techniques yielding relatively low Matthews correlation coefficients (MCCs), they still performed satisfactorily. When comparing EnhancerPred to iEnhancer-2L, EnhancerPred exhibited a slightly better MCC performance, but its efficacy still fell short. An improved version of iEnhancer-2L, called iEnhancer-EL [18], was introduced by Liu et al. in 2018. Notably, iEnhancer-EL showcased a complex structure, comprising two ensemble models constructed from 16 different main classifiers. These crucial classifiers were developed using 171 SVM-based elementary classifiers, which combined PseKNC, subsequence profile, and k-mers characteristics. Despite iEnhancer-EL's current status as one of the most effective methods for identifying enhancers and evaluating their strength, the potential for even better models exists by employing cutting-edge learning algorithms and advanced encoding techniques. Recently a model called iEnhancer-RF [19], which utilizes increased feature representation with random forest, was proposed for enhancer prediction. However, there is still room for enhancing the model's resilience. Despite the current methodologies demonstrating impressive performance in identifying enhancers and assessing their strength, their accuracy still requires improvement. To achieve this goal and enhance the predictive performance of enhancers, this study will employ novel encoding approaches and classification models. In this study, we utilize two different feature extraction techniques and subsequently combine them. By feeding the final fused feature into a Deep Neural Network (DNN) for model training, we can predict enhancers. DNNs have demonstrated exceptional capabilities in processing complex and diverse data, making them well-suited for extracting information from the heterogeneous genetic properties associated with enhancers. Our objective is to leverage DNNs to enhance the accuracy and reliability of enhancer identification, leading to a more comprehensive understanding of gene regulation mechanisms. Our iEnhancer-DNN demonstrates superior generalized efficiency in forecasting both enhancers and non-enhancers. To ensure a fair comparison with prior studies by Liu et al. [8, 18] and Jia and He [17], the same dataset is used for model construction and evaluation. The detailed framework of the iEnhancer-DNN model is illustrated in Figure 1.

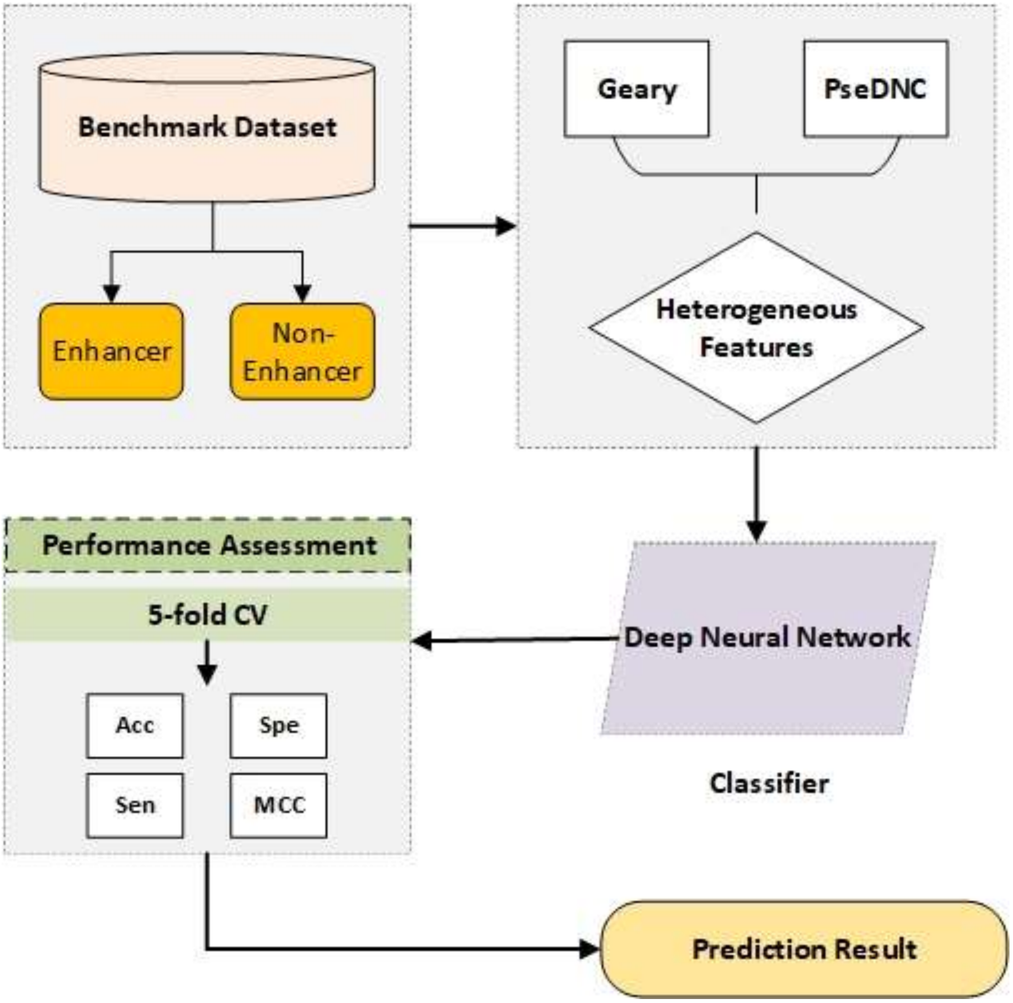


Fig 1. Detailed framework of the iEnhancer-DNN

2. MATERIALS AND METHODS

2.1. Dataset

This paper used a dataset proposed by Liu et al. [8] was utilized in this investigation and in the development of iEnhancer-2L [8], EnhancerPred [17], iEnhancer-EL [18], and iEnhancer-RF [19]. This dataset comprises 200bp-long DNA sequences extracted from 9 different cell lines, along with corresponding enhancer data. To ensure the classifier's accuracy, enhancers with a similarity of more than 90% were removed from the dataset using CD-HIT [20]. The final dataset consists of 1484 enhancers and 1484 non-enhancers. The training and independent datasets are accessible in Refs. [8].

2.2. Feature Extraction

Machine learning or deep learning methods are not possible to directly annotate continuous sequences of nucleotides. To complete this task, it is vital to transform the sequence representation of nucleotide sequences into feature vectors that are generated with numerical values [21, 22]. In this investigation, feature extraction was performed using iLearn [23].

2.2.1. Geary

The Geary autocorrelation descriptors for a protein or peptide sequence are defined as

$$C(d)=\frac{\frac{1}{2(N-d)}\sum_{i=1}^{N-d}(P_i-P_{i+d})^2}{\frac{1}{N-1}\sum_{i=1}^{N}(P_i-\bar{P})^2},d=1,2,..,nlag$$

In this equation, d represents the lag, P denotes the property value for the ith residue, P_i represents the mean of the property values over the entire sequence, P_{i+d} is the mean of the property values for residues at a lag distance d from the ith residue, and $nlag$ indicates the total number of lags considered [24].

2.2.2. PseDNC

The Pseudo Dinucleotide Composition (PseDNC) encoding is used to integrates both contiguous local sequence-order information and global sequence-order information into the feature vector of a nucleotide sequence [25]. The PseDNC encoding is defined as follows:

$$D=[d_1,d_2,....,d_{16},d_{16+1},...,d_{16+1},..,d_{16+\lambda}]^T,$$

$$d_k=\left\{\begin{array}{c}\dfrac{f_k}{\sum_{i=1}^{16}f_i+w\sum_{j=1}^{\lambda}\theta_j},(1\le k\le 16)\\ \hline \dfrac{w\theta_{k-16}}{\sum_{i=1}^{16}f_i+w\sum_{j=1}^{\lambda}\theta_j},(17\le k\le 16+\lambda)\end{array}\right\}$$

In this equation, $f_k(k=1,2,...,16)$ represents the normalized occurrence frequency of dinucleotides in the nucleotide sequence. λ denotes the highest counted rank (or tie) of the correlation along the nucleotide sequence. w is a weight factor ranging from 0 to 1, and $\theta(j=1,2,...,\lambda)$ represents the j-tier correlation factor defined as:

$$\left\{\begin{array}{c} \theta_1 = \frac{1}{L-2}\sum_{i=1}^{L-2}\theta(R_iR_{i+1}, R_{i+1}R_{i+2}) \\ \theta_2 = \frac{1}{L-3}\sum_{i=1}^{L-3}\theta(R_iR_{i+1}, R_{i+2}R_{i+3}) \\ \theta_3 = \frac{1}{L-4}\sum_{i=1}^{L-2}\theta(R_iR_{i+1}, R_{i+3}R_{i+4}), (\lambda < L) \\ \cdots \\ \theta_\lambda = \frac{1}{L-4}\sum_{i=1}^{L-1-\lambda}\theta(R_iR_{i+1}, R_{i+\lambda}R_{i+\lambda+4}) \end{array}\right\}$$

where the correlation function is defined:

$$\theta(R_iR_{i+1}, R_{j+1}R_{j+1}) = \frac{1}{\mu}\sum_{u=1}^{\mu}[P_u(R_iR_{i+1}) - P_u(R_jR_{j+1})]^2$$

In this definition, μ is the number of physicochemical indices. $P_u(R_iR_{i+1})$ is the numerical value of the u-th $(u=1,2,...,\mu)$ physicochemical index of the dinucleotide R_iR_{i+1} at position i, and $P_u(R_jR_{j+1})$ represents the corresponding value of the dinucleotide R_jR_{j+1} at position j. The PseDNC descriptor has shown successful applications in recombination spot identification, making it a valuable tool for analyzing nucleotide sequences and capturing important structural and positional information of dinucleotides [25].

2.3. Classification Model

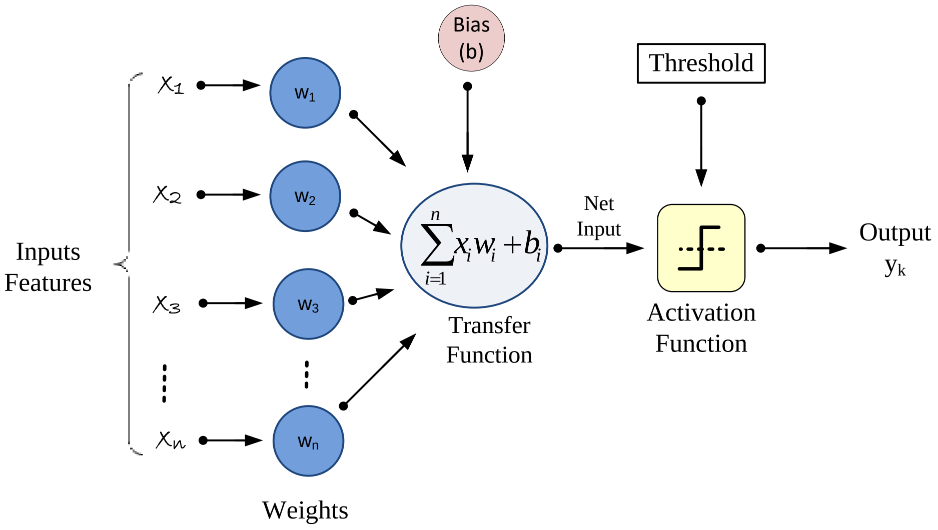
2.3.1. Deep Neural Network

The Deep Neural Network (DNN) is a fundamental type of artificial neural network with widespread applications in machine learning. Our DNN architecture consists of 5 layers, containing 64, 32, 32, 16, and 16 neurons, respectively, utilizing the ReLU activation function for the hidden layers. The last layer comprises two neurons, enabling the prediction of two classes: enhancer and non-enhancer, using the sigmoid activation function. To prevent overfitting, we implemented dropout regularization with a dropout rate and additionally experimented with a dropout value of 0.2. In our optimization process, we explored an algorithm, namely "Adam." The hyper parameters for our deep neural network and other pertinent details are provided in Table 1. This experimentation and fine-tuning of hyper parameters aim to optimize the model's performance for the specific classification task.

Table 1: The hyper parameter of the DNN Model

Parameter	**Value**
Number of Layers	5
Neurons per layer	64,32,32,16,16
Learning rate	0.001
Dropout rate	0.2
Loss function	Binary cross entropy
Batch size	64
Epochs	40
Optimizer	Adam

Figure 2. Mechanism of transmission and activation function in DNN



2.3.2. Performance measure method

We assessed the efficacy of our DNN models through 5-fold cross-validation on the training dataset. In addition, we used separate data sets to evaluate the results of the best model for each category. The predictive model's effectiveness was gauged using five widely-used metrics in bioinformatics classification tasks [26-29]: accuracy, sensitivity, specificity, Matthews correlation coefficient (MCC), and the area under the Receiver Operating Characteristic (ROC) curve are the evaluation metrics used in this study.

2.4. Evaluation parameters

The subsequent four criteria are frequently employed to assess the effectiveness of a predictor: accuracy (Acc), specificity (Sp), sensitivity (Sn), and Matthew's correlation coefficient (MCC). Extensive discussions on these metrics can be found in the literature [30-34], and their mathematical formulations are as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sen = \frac{TP}{TP + FN}$$

$$Spe = \frac{TN}{TN + FP}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Table 2. Result of Enhancer Identification on Benchmark dataset

Feature Encoding	Classifier	Acc (%)	Spe (%)	Sen (%)	MCC
Geary	DNN	79.05	80.94	77.31	0.58
PseDNC		77.76	80.92	73.96	0.54
Hybrid		**81.83**	**82.89**	**81.08**	**0.63**

Table 3. Result of Enhancer Identification an independent dataset

Feature Encoding	Classifier	Acc (%)	Spe (%)	Sen (%)	MCC
Geary	DNN	71.47	70.89	73.70	0.44
PseDNC		77.20	81.78	72.57	0.54
Hybrid		**80.99**	**83.16**	**78.81**	**0.62**

3. RESULT AND DISCUSSION

3.1 Comparative performance among different feature encoding

In this study, we conducted an experiment involving two distinct feature encodings: Geary and PseDNC. These individual features were integrated. Subsequently, a Deep Neural Network (DNN) model was applied to the integrated features. The performance results for both benchmark and independent datasets are presented in Tables 2 and 3, respectively. Upon analyzing the results, it becomes evident that the integration of both Geary and PseDNC features led to an enhancement in performance, as indicated in Tables 2 and 3.

3.2. Independent test

An independent evaluation is crucial to assess the model's performance, determine its capacity to avoid overfitting and achieve consistent results with new, unseen data. Our model, iEnhancer-DNN, achieved impressive performance compared to the other existing methods listed in Table 4. Meanwhile, our proposed model achieved a sensitivity of 78.81%, specificity of 83.16%, accuracy of 80.99%, MCC of 0.62, and an AUC of 0.87, which were relatively promising and balanced. Our model iEnhancer-DNN is an optimal model with high performance.

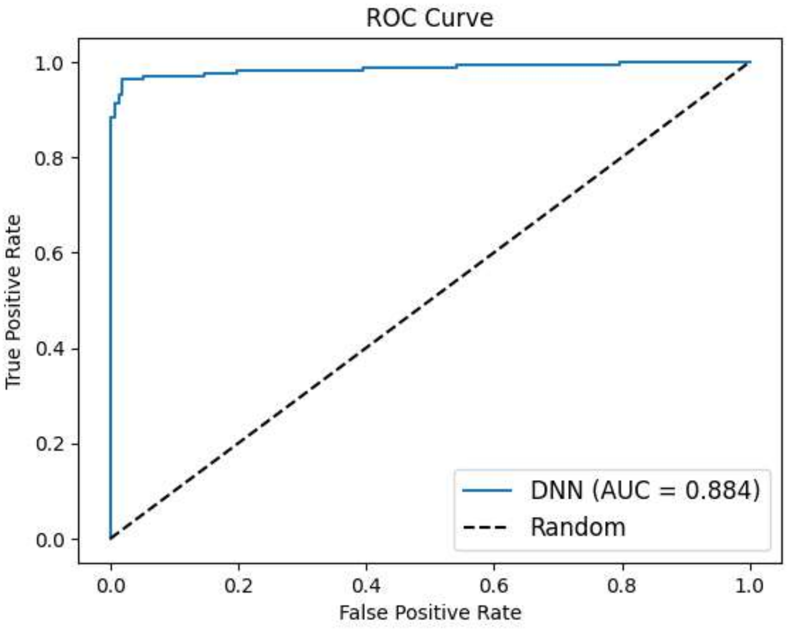


Fig 2. ROC curves for iEnhancer-DNN on Benchmark dataset

3.3. Comparison with Existing State-of-the-Art Methods

Numerous remarkable techniques are available for enhancer prediction, and some well-known ones include iEnhancer-2L [8], EnhancerPred [17], iEnhancer-EL [18], and iEnhancer-RF [19]. Table 4 illustrates the comparison results with existing state-of-the-art methods for enhancer identification. Upon observing Table 4, which involves distinguishing enhancers from non-enhancers, the proposed predictor outperforms the existing state-of-the-art predictors regarding all the metrics. It is crucial to emphasize that, among the four metrics, Acc and MCC are of particular significance. The former measures the overall accuracy of a predictor, while the latter evaluates its stability. In this context, iEnhancer-DNN demonstrates superior performance compared to other methods based on the Acc and MCC metrics. To visually represent the performance, ROC curves and graphical representation of proposed model with existing methods are presented in Figure 2 and 3, respectively.

Table 4. Comparison of the proposed predictor with the state-of-the-art predictors in enhancer’s identification on Benchmark and Independent Dataset

Dataset	Model	ACC%	Sen%	Spe%	MCC
Benchmark	**iEnhancer-DNN**	**81.83**	**81.08**	**82.89**	**0.63**
	iEnhancer-RF	76.18	73.64	78.71	0.52
	iEnhancerPred	73.18	72.57	73.79	0.46
	iEnhancer-EL	78.03	76.67	80.39	0.56
	iEnhancer-2L	76.89	78.09	75.88	0.54
Independent	**iEnhancer-DNN**	**80.99**	**78.81**	**83.16**	**0.62**
	iEnhancer-RF	79.75	78.50	81.00	0.59
	iEnhancerPred	74.00	73.50	74.50	0.48
	iEnhancer-EL	74.75	71.00	78.50	0.49
	iEnhancer-2L	73.00	71.00	75.00	0.46

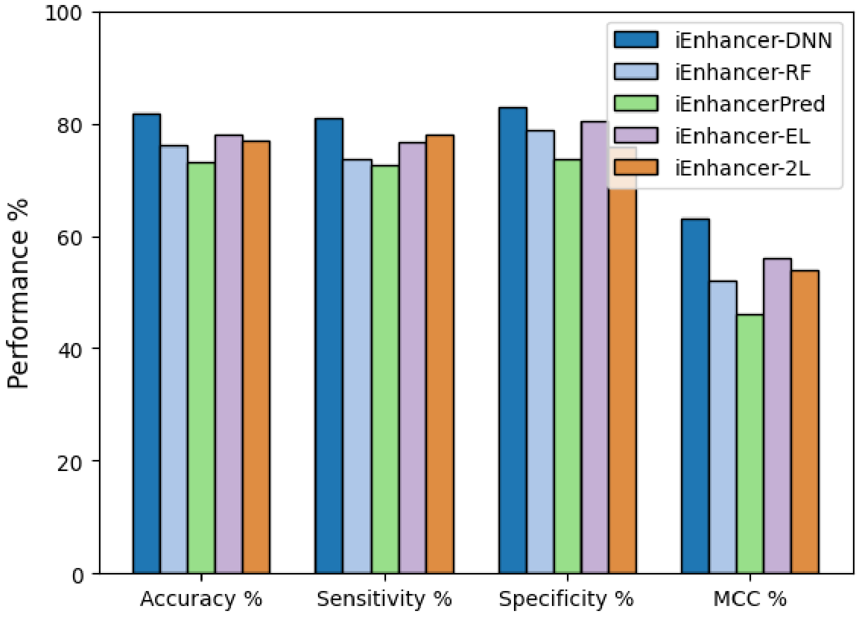


Fig 3. Graphical representation of the iEnhancer-DNN with existing methods on Benchmark dataset

4.CONCLUSION

This paper presents an efficient and effective method known as iEnhancer-DNN for identifying enhancers. Our proposed DNN-based approach demonstrated significant advancements in enhancer prediction compared to traditional methods. The fusion of feature extraction strategies and the utilization of deep neural network techniques were crucial factors in enhancing the predictive performance of our model. Integrating heterogeneous data allowed us to create a comprehensive and informative representation of enhancer regions, leading to a more refined and accurate identification process. Our DNN-based model consistently outperformed previous methods in accuracy, sensitivity, specificity and MCC, demonstrating its superiority in enhancer identification tasks.

References

[1] M. Bulger and M. Groudine, "Functional and mechanistic diversity of distal transcription enhancers," *Cell,* vol. 144, no. 3, pp. 327-339, 2011.

[2] E. Calo and J. Wysocka, "Modification of enhancer chromatin: what, how, and why?," *Molecular cell,* vol. 49, no. 5, pp. 825-837, 2013.

[3] J. L. Plank and A. Dean, "Enhancer function: mechanistic and genome-wide insights come together," *Molecular cell,* vol. 55, no. 1, pp. 5-14, 2014.

[4] N. D. Heintzman and B. Ren, "Finding distal regulatory elements in the human genome," *Current opinion in genetics & development,* vol. 19, no. 6, pp. 541-549, 2009.

[5] N. D. Heintzman *et al.*, "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome," *Nature genetics,* vol. 39, no. 3, pp. 311-318, 2007.

[6] A. Visel *et al.*, "ChIP-seq accurately predicts tissue-specific activity of enhancers," *Nature,* vol. 457, no. 7231, pp. 854-858, 2009.

[7] J. Chen, H. Liu, J. Yang, and K.-C. Chou, "Prediction of linear B-cell epitopes using amino acid pair antigenicity scale," *Amino acids,* vol. 33, pp. 423-428, 2007.

[8] 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.

[9] B. Liu, F. Yang, D.-S. Huang, and K.-C. Chou, "iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC," *Bioinformatics,* vol. 34, no. 1, pp. 33-40, 2018.

[10] J. Ernst *et al.*, "Mapping and analysis of chromatin state dynamics in nine human cell types," *Nature,* vol. 473, no. 7345, pp. 43-49, 2011.

[11] G. D. Erwin *et al.*, "Integrating diverse datasets improves developmental enhancer prediction," *PLoS computational biology,* vol. 10, no. 6, p. e1003677, 2014.

[12] M. Fernandez and D. Miranda-Saavedra, "Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines," *Nucleic acids research,* vol. 40, no. 10, pp. e77-e77, 2012.

[13] H. A. Firpi, D. Ucar, and K. Tan, "Discover regulatory DNA elements using chromatin signatures and artificial neural network," *Bioinformatics,* vol. 26, no. 13, pp. 1579-1586, 2010.

[14] D. Kleftogiannis, P. Kalnis, and V. B. Bajic, "DEEP: a general computational framework for predicting enhancers," *Nucleic acids research,* vol. 43, no. 1, pp. e6-e6, 2015.

[15] N. Rajagopal *et al.*, "RFECS: a random-forest based algorithm for enhancer identification from chromatin state," *PLoS computational biology,* vol. 9, no. 3, p. e1002968, 2013.

[16] 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.

[17] 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.

[18] 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.

[19] D. Y. Lim, J. Khanal, H. Tayara, and K. T. Chong, "iEnhancer-RF: Identifying enhancers and their strength by enhanced feature representation using random forest," *Chemometrics and Intelligent Laboratory Systems,* vol. 212, p. 104284, 2021.

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

[21] 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.

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

[23] Z. Chen *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.

[24] R. R. Sokal and B. A. Thomson, "Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population," *American Journal of Physical Anthropology: The*

Official Publication of the American Association of Physical Anthropologists, vol. 129, no. 1, pp. 121-131, 2006.

[25] W. Chen, P.-M. Feng, H. Lin, and K.-C. Chou, "iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition," *Nucleic acids research,* vol. 41, no. 6, pp. e68-e68, 2013.

[26] T.-T.-D. Nguyen, N.-Q.-K. Le, Q.-T. Ho, D.-V. Phan, and Y.-Y. Ou, "TNFPred: Identifying tumor necrosis factors using hybrid features based on word embeddings," *BMC Medical Genomics,* vol. 13, pp. 1-11, 2020.

[27] Q.-T. Ho, D.-V. Phan, and Y.-Y. Ou, "Using word embedding technique to efficiently represent protein sequences for identifying substrate specificities of transporters," *Analytical biochemistry,* vol. 577, pp. 73-81, 2019.

[28] R. M. I. Kusuma and Y.-Y. Ou, "Prediction of ATP-binding sites in membrane proteins using a two-dimensional convolutional neural network," *Journal of Molecular Graphics and Modelling,* vol. 92, pp. 86-93, 2019.

[29] S. W. Taju, T.-T.-D. Nguyen, N.-Q.-K. Le, R. M. I. Kusuma, and Y.-Y. Ou, "DeepEfflux: a 2D convolutional neural network model for identifying families of efflux proteins in transporters," *Bioinformatics,* vol. 34, no. 18, pp. 3111-3117, 2018.

[30] S.-H. Guo *et al.*, "iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition," *Bioinformatics,* vol. 30, no. 11, pp. 1522-1529, 2014.

[31] H. Lin, E.-Z. Deng, H. Ding, W. Chen, and K.-C. Chou, "iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition," *Nucleic acids research,* vol. 42, no. 21, pp. 12961-12972, 2014.

[32] W. Chen, P.-M. Feng, E.-Z. Deng, H. Lin, and K.-C. Chou, "iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition," *Analytical biochemistry,* vol. 462, pp. 76-83, 2014.

[33] A. Razaa *et al.*, "iAFP-ET: A robust approach for accurate identification of antifungal peptides using extra tree classifier and multi-view fusion."

[34] A. Ahmad, A. Raza, and M. F. Waqas, "iRhm5BiRNN: Identification of RNA 5-Hydroxymethylcytosine Modifications Using Bidirectional-Recurrent Neural Network."