Hindawi
Computational Intelligence and Neuroscience
Volume 2022, Article ID 2987407, 8 pages
https://doi.org/10.1155/2022/2987407

Research Article

DBP-iDWT: Improving DNA-Binding Proteins Prediction Using Multi-Perspective Evolutionary Profile and Discrete Wavelet Transform

Farman Ali,[1] **Omar Barukab**,[2] **Ajay B Gadicha**,[3] **Shruti Patil**,[4] **Omar Alghushairy**,[5] **and Akram Y. Sarhan**[6]

[1]*Department of Elementary and Secondary Education, Peshawar, Khyber Pakhtunkhwa, Pakistan*
[2]*Faculty of Computing and Information Technology, King Abdulaziz University, Rabigh, Jeddah 21911, Saudi Arabia*
[3]*Department of Computer Science and Engineering, P.R. Pote, Collage of Engineering and Management, Amravati, India*
[4]*Symbiosis Institute of Technology, Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International University, Pune, India*
[5]*Department of Information Systems and Technology, College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia*
[6]*Department of Information Technology, College of Computing and Information Technology at Khulais, University of Jeddah, Jeddah, Saudi Arabia*

Correspondence should be addressed to Farman Ali; farman335@yahoo.com

Received 29 June 2022; Revised 19 August 2022; Accepted 9 September 2022; Published 28 September 2022

Academic Editor: Xiaolong Zhou

DNA-binding proteins (DBPs) have crucial biotic activities including DNA replication, recombination, and transcription. DBPs are highly concerned with chronic diseases and are used in the manufacturing of antibiotics and steroids. A series of predictors were established to identify DBPs. However, researchers are still working to further enhance the identification of DBPs. This research designed a novel predictor to identify DBPs more accurately. The features from the sequences are transformed by F-PSSM (Filtered position-specific scoring matrix), PSSM-DPC (Position specific scoring matrix-dipeptide composition), and R-PSSM (Reduced position-specific scoring matrix). To eliminate the noisy attributes, we extended DWT (discrete wavelet transform) to F-PSSM, PSSM-DPC, and R-PSSM and introduced three novel descriptors, namely, F-PSSM-DWT, PSSM-DPC-DWT, and R-PSSM-DWT. Onward, the training of the four models were performed using LiXGB (Light eXtreme gradient boosting), XGB (eXtreme gradient boosting, ERT (extremely randomized trees), and Adaboost. LiXGB with R-PSSM-DWT has attained 6.55% higher accuracy on training and 5.93% on testing dataset than the best existing predictors. The results reveal the excellent performance of our novel predictor over the past studies. DBP-iDWT would be fruitful for establishing more operative therapeutic strategies for fatal disease treatment.

1. Introduction

DNA-binding proteins perform many crucial activities like DNA translation, repair, translation, and damage [1]. DBPs are directly encoded into the genome of about 2–5% of the prokaryotic and 6–7% of eukaryotic [2]. Several DBPs are responsible for gene transcription and replication, and some DBPs shape the DNA into a specific structure, called chromatin [3]. The research on DBPs is significant in diverse fatal disease treatment and production of drugs. For instance, nuclear receptors are the key components of tamoxifen and bicalutamide medicines which are used in cancer treatment. Similarly, glucocorticoid receptors participate in the production of dexamethasone, which is utilized in autoimmune diseases and anti-inflammatory, allergies, and asthma treatment [4–6]. Onward, Inhibitor of DNA-binding (ID) proteins are closely related to tumor-associated processes including chemoresistance, tumorigenesis, and angiogenesis. In addition, ID proteins are also directly concerned with lung, cervical, and prostate cancers [7].

Protein sequences are rapidly growing in the online database. A series of predictors were developed for diverse biological problems including iRNA-PseTNC [8], iACP-GAEnsC [9], cACP-2LFS [10], DP-BINDER [11], Deep-AntiFP [12], cACP [13], iAtbP-Hyb-EnC [14], iAFPs-EnC-GA [15], and cACP-DeepGram [16]. It is highly demanding to predict DBPs by computational approaches. Several predictors were introduced using the primary sequential information and structural features. Structured-based predictors produce good prediction results, but structural features for all proteins are unavailable. Some of the structure-based protocols are iDBPs [17], DBD-Hunter [18], and Seq(DNA) [19]. Sequence-based systems have been developed using sequential information, more convenient and easy to employ for large datasets. Therefore, many sequence-based systems were adopted for DNA-binding proteins identification. Among these methods: DBP-DeepCNN [20], DNA-Prot [21], iDNA-Prot [22], iDNA-Prot|dis [23], Kmer1 + ACC [24], Local-DPP [25], DBPPred-PDSD [26], DPP-PseAAC [27], and StackDPPred [28]. Consequently, Li et al. extracted features by a convolutional neural network (CNN) and Bi-LSTM [29]. Onward, Zhao et al. the features of the proteins are analyzed by six methods and classification is performed with XGBoost [30]. Each computational method contributed well to enhancing the prediction of DBPs. However, more efforts are needed to improve prediction of DBPs. Considering this, a new method (DBP-iDWT) is established to identify DBPs accurately. The contribution of our research is as follows:

(i) Designed three new feature descriptors i.e., F-PSSM-DWT, PSSM-DPC-DWT, and R-PSSM-DWT

(ii) LiXGB is applied for model training and prediction

(iii) Constructed a new computational model (DBP-iDWT) for improving DBPs identification

In addition to LiXGB, the features set is fed into three classification algorithms, namely ERT, XGB, and Adaboost. The efficacy of each classifier was assessed with ten-fold test, while the generalization capability was assessed by a testing set. LiXGB using R-PSSM-DWT secured the highest prediction outcomes than past methods. The flowchart of the DBP-iDWT is depicted in Figure 1.

The rest portion of the manuscript comprises three parts. Section 2 comprises details regarding datasets and methodologies; in Section 3, the performance of classifiers has illustrated; and Section 4 summarizes the conclusion.

2. Materials and Methods

2.1. Selection of Datasets. We selected two datasets from the previous work [31]. One dataset (PDB14189) is employed model training and the other dataset is deployed as a testing dataset. PDB14189 was collected from the UniProt database [32]. To design a standard dataset, they removed more than 25% of similar sequences by CD-HIT toolkit. The final training dataset comprises 7129 DBPs and 7060 non-DBPs. The independent set was retrieved by a procedure explained in reference [33]. The similar sequences with a cutoff value 25% are removed. The final testing dataset contains 1153 DBPs and 1119 non-DBPs.

2.2. Feature Descriptors. In this work, the patterns are discovered with PSSM-DPC-DWT, F-PSSM-DWT, and R-PSSM-DWT. These approaches are elaborated in the following parts.

2.2.1. Position-specific Scoring Matrix (PSSM). Recently, evolutionary features are successfully implemented and improve the prediction results of many predictors [1, 20]. We also implemented PSSM for the formulation of evolutionary patterns. Each sequence is searched against the NCBI database applying the PSI-BLAST program for the alignment of homologous features [34].

The PSSM can be denoted as follows:

$$\begin{aligned} PSSM &= \left(P_1, P_2, \ldots, P_j, \ldots, P_{20}\right)^T, \\ P_{i,j} &= \left(P_{1,j}, P_{2,j}, \ldots, P_{L,j}\right), \; (i = 1, 2, \ldots, L), \end{aligned} \tag{1}$$

where $T$and $P_{i,j}$ indicate the transpose operator and score of j type of amino acid in the i^{th} position of query sequence.

2.2.2. Filtered Position-specific Scoring Matrix (F-PSSM). PSSM transforms the evolutionary patterns into numerical forms. It may comprise some negative scores which can lead to generating similar feature vectors despite different sequences. To cope with this hurdle, F-PSSM filters all the negative scores in the preprocessing step. The detail of dimension formulation is provided in [35].

2.2.3. Position-specific Scoring Matrix-Dipeptide Composition (PSSM-DPC). The local sequence-order patterns contains informative feature which are explored by incorporating DPC into PSSM. DPC calculates the frequency of continuous amino acids and produces a dimension of 400 [36]. DPC is calculated as follows:

$$PSSM - DPC = \left(G_{1,1}, \ldots, G_{1,20}, G_{2,1}, \ldots, G_{2,20}, \ldots, G_{20,1}, \ldots, G_{20,20}\right)^T, \tag{2}$$

where

$$P_{i,j} = \frac{1}{L} \sum_{k=1}^{L-1} G_{k,i} \times G_{k+1,j} \, (1 \le i, j \le 20). \tag{3}$$

2.2.4. Reduced Position Specific Scoring Matrix (R-PSSM). It is believed that there exist several similarities among 20 unique amino acids. Based on these similarities, researchers categorized these residues into groups. Li et al. [37] suggested that according to some specific residue the following groups can be formed:

$$G(i) = \begin{cases} Y, & ifi = F, Y, W; \\ L, & ifi = M, L; \\ V, & ifi = I, V; \\ S, & ifi = A, T, S; \\ N, & ifi = N, H; \\ E, & ifi = Q, E, D; \\ K, & ifi = R, K; \\ i, & \text{otherwise}. \end{cases} \quad (4)$$

Using the Li et al. rule, the $L \times 20$ PSSM is converted to $L \times 10$ matrix by the following equations:

$$\begin{aligned} G_1 &= \frac{F+Y+W}{3}, \\ G_2 &= \frac{M+L}{2}, \\ G_3 &= \frac{I+V}{2}, \\ G_4 &= \frac{A+T+S}{3}, \\ G_5 &= \frac{N+H}{2}, \\ G_6 &= \frac{Q+E+D}{3}, \\ G_7 &= \frac{R+K}{2}, \\ G_8 &= C, \\ G_9 &= G, \\ G_{10} &= P. \end{aligned} \quad (5)$$

If $r_1 r_2 r_3 \ldots\ldots r_L$ is a given protein sequence, then its reduced PSSM (R-PSSM) is indicated as follows:

$$RP = \left[\begin{array}{ccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & \\ r_1 & R_{1,1} & R_{1,2} & R_{1,3} & R_{1,4} & R_{1,5} & R_{1,6} & R_{1,7} & R_{1,8} & R_{1,9} & R_{1,10} \\ r_2 & R_{2,1} & R_{2,2} & R_{2,3} & R_{2,4} & R_{2,5} & R_{2,6} & R_{2,7} & R_{2,8} & R_{2,9} & R_{2,10} \\ \cdots & \cdots & \cdots & & & & & & & & \\ r_L & R_{L,1} & R_{L,2} & R_{L,3} & R_{L,4} & R_{L,5} & R_{L,6} & R_{L,7} & R_{L,8} & R_{L,9} & R_{L,10} \end{array} \right]. \quad (6)$$

We obtain 110 feature vector from RP.

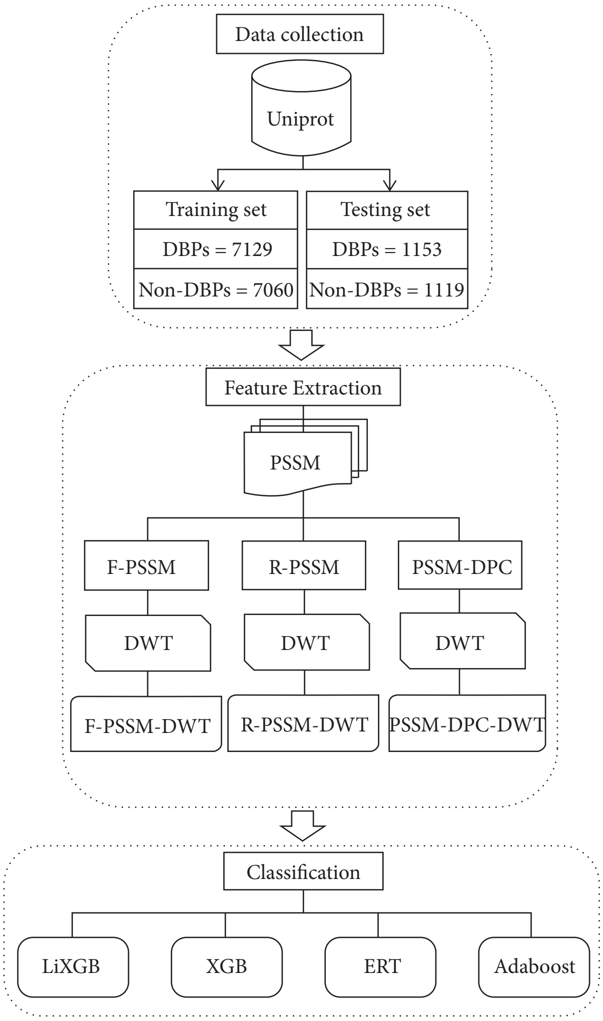

FIGURE 1: Architecture of the proposed model.

2.2.5. Discrete Wavelet Transform. To achieve only salient information, some compression approaches like DWT is applied in research areas. DWT is used for compression of signals and denoising [38, 39]. DWT divides a signal into low-frequency and high-frequency components [40]. Low frequencies are more important than high-frequencies [41]. The Low frequencies are onward split into low and high levels to achieve discriminative patterns. DWT is computed as follows:

$$X(m,n) = \sqrt{\frac{1}{m}} \int_0^y f(y) \Psi\left(\frac{y-n}{m}\right) d_y, \quad (7)$$

where $m$represents the scale variable and n shows the translation variable. $X(m,n)$ is the transform coefficient. The low and high frequencies of a signal *f(t)* is computed as follows:

$$C_{i,low}[a] = \sum_{k=1}^{N} s[k]L[2a-k],$$
$$C_{i,high}[a] = \sum_{k=1}^{N} s[k]H[2a-k], \quad (8)$$

where $C_{i,high}[a]$ and $C_{i,low}[a]$ are the high and low frequencies of the signal. H, $s[k]$, and L, represent the high pass filter, discrete signal, and low pass filter, respectively.

To obtain only important features and eliminate the less informative and noisy patterns, DWT is extended into F-PSSM, PSSM-DPC, and R-PSSM to split into low and high frequencies up to two levels. Finally, PSSM-DPC-DWT, R-PSSM-DWT, and F-PSSM-DWT novel feature descriptors are constructed. The dimension of each feature set is 512 after applying DWT. Figure 2 depicts the schematic view of Two-level DWT.

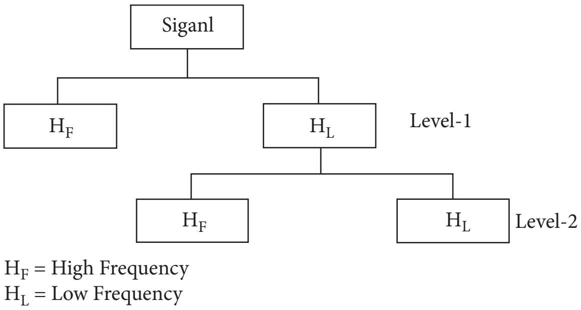

Figure 2: 2-level structure of DWT.

2.3. Light eXtreme Gradient Boosting. During the establishment of the predictor, the model training is performed by a classifier. Gradient Boosting Machine (GBM) classifier uses decision trees for the construction of a model. The model performance is improved with loss function [42]. Unlike GBM, eXtreme Gradient Boosting (XGB) employs an objective function. XGB concatenates loss function and regularization for regulating the model complexity. It performs parallel computations to optimize the computational speed. Due to these benefits of XGB, Light eXtreme Gradient Boosting (LiXGB) was proposed [43]. LiXGB possesses many additional features like lower memory, higher efficiency, and fast model training speed that improve the model performance. LiXGB minimizes the model training time of the large datasets. We utilized the hyperparameters like max depth, estimator, eta, lambda, and alpha. The "eta" maintains the learning rate, "estimator" constructs trees, "max depth" is used for controlling the tree depth, "alpha" shrinks the high dimension of the dataset, and "lambda" avoids the overfitting. Other parameters have been kept as default. These hyperparameters are also summarized in Table 1.

Table 1: Applied parameters with values.

Parameter	Value
Era	0.1
No. of estimator	500
Alpha	1
Lambda	1
Max depth	8

2.4. Proposed Model Validation Methodologies. The model performance is examined by different validation approaches The commonly used validation methods are k-fold and jackknife [44–47]. However, the jackknife is time-consuming and costly [48–50]. During 10-fold cross validation, training set is split into 10-folds. The 9 folds are used for model training and 1 fold is used for model validation. This process is repeated 10 times so that each fold is used for the test exactly once. The final prediction is the average of all tested folds [51–54]. The current work performance is evaluated with 10-fold and five indexes, i.e., specificity (Sp), F-measure, sensitivity (Sn), accuracy (Acc), and Mathew's correlation coefficient (MCC) for evaluating the model performance [55–58]. These parameters are computed as follows:

$$Acc = 1 - \frac{H_-^+ + H_+^-}{H^+ + H^-},$$
$$Sn = 1 - \frac{H_-^+}{H^+},$$
$$Sp = 1 - \frac{H_+^-}{H^-},$$
$$MCC = \frac{1 - \left(H_-^+ + H_+^-/H^+ + H^-\right)}{\sqrt{\left(1 + \left(H_-^+ + H_+^-/H^+\right)\right)\left(1 + \left(H_-^+ + H_+^-/H^-\right)\right)}},$$
$$F1\,Score = \frac{(2 * \text{precision} * \text{recall})}{(\text{precision} + \text{recall})},$$
$$\text{Precision} = \frac{H^+}{H_-^+ + H^+},$$
$$\text{Recall} = \frac{H^+}{H_+^- + H^+}, \quad (9)$$

where H^+ is used to denote the DBPs, H^- is the non-DBPs, H_+^- shows the prediction of non-DBPs which the model predicted mistakenly as DBPs, and H_-^+ represents the DBPs which are classified by the model as non-DBPs.

3. Results and Discussion

After performing experiments on the models, In this part, we will elaborate the obtained results of the learning algorithms via the extracted feature sets of the training and testing sequences.

Table 2: Results of encoders before DWT.

Model	Encoder	Acc (%)	Sn (%)	Sp (%)	MCC (%)
Adaboost	F-PSSM	71.52	80.42	62.54	43.67
	PSSM-DPC	80.05	78.44	81.69	60.15
	R-PSSM	80.07	76.15	84.02	60.35
ERT	F-PSSM	75.18	84.74	58.97	44.56
	PSSM-DPC	79.22	73.18	85.31	58.42
	R-PSSM	79.56	74.99	84.18	59.40
XGB	F-PSSM	74.57	82.17	66.90	49.67
	PSSM-DPC	81.53	76.15	86.97	63.47
	R-PSSM	81.63	76.48	86.84	63.64
LiXGB	F-PSSM	76.60	82.47	66.01	48.75
	PSSM-DPC	83.54	84.61	82.46	67.10
	R-PSSM	83.62	82.30	84.96	67.27

3.1. Results of Feature Encoders before DWT. In this section, we have reported the outcomes of F-PSSM, PSSM-DPC, and R-PSSM in Table 2. The performance of the individual descriptor is analyzed by 10-fold test and assessment indices. On F-PSSM, the accuracies secured by LiXGB, XGB, ERT, and Adaboost are 76.60%, 74.57%, 75.18%, and 71.52%, respectively. Among all classifiers, LiXGB achieved the best accuracy. On PSSM-DPC, all classifiers enhanced the prediction results and generated 83.62%, 81.63%, 79.56%, and 80.07% accuracies by LiXGB, XGB, ERT, and Adaboost, respectively. Similarly, the classifiers also improved the performance on the R-PSSM descriptor using all evaluation parameters. LiXGB attained the highest (83.62%) accuracy. The predictions indicate that LiXGB possesses higher learning power comparatively XGB, ERT, and Adaboost.

3.2. Results of Feature Encoders after DWT. The features extracted by representative methods may contain some noisy, redundant, or less informative features. To avoid such features, DWT is applied to F-PSSM, PSSM-DPC, and R-PSSM. DWT considers the informative patterns and improves the performance of the model. After applying DWT, we achieve F-PSSM-DWT, PSSM-DPC-DWT, and R-PSSM-DWT. Each feature is fed into Adaboost, ERT, XGB, and LiXGB in order to examine the performance over these feature descriptors and results are summarized in Table 3. With 10-fold test, Adaboost, ERT, XGB, and LiXGB produced 73.20%, 77.26%, 75.37%, and 79.40% accuracies which are 1.68%, 2.08%, 0.80%, and 2.80% than F-PSSM, PSSM-DPC, and R-PSSM, respectively. Similarly, the classifiers also boosted the performance on PSSM-DPC-DWT on all evaluation parameters. Furthermore, with R-PSSM-DWT, Adaboost, ERT, XGB, and LiXGB have enhanced the accuracies by 2.16%, 3.49%, 1.98%, and 3.22% than R-PSSM. These results demonstrate that all classifiers show improvement in performance after applying DWT. Among all feature descriptors, the best results are secured by R-PSSM-DWT.

LiXGB has constantly depicted better achievement than other classifiers. LiXGB enhanced the performance and generated 3.23%, 3.79%, and 4.61% higher accuracies than XGB, ERT, and Adaboost with R-PSSM-DWT. It is concluded that the performance of LiXGB is superior to other classifiers.

Table 3: Results of feature encoders after DWT.

Model	Encoder	Acc (%)	Sn (%)	Sp (%)	MCC (%)
Adaboost	F-PSSM-DWT	73.20	82.35	56.67	40.21
	PSSM-DPC-DWT	81.81	80.45	83.19	63.66
	R-PSSM-DWT	82.23	77.68	86.81	64.77
ERT	F-PSSM-DWT	77.26	79.91	74.59	54.58
	PSSM-DPC-DWT	81.53	76.15	86.97	63.47
	R-PSSM-DWT	83.05	81.30	84.82	66.15
XGB	F-PSSM-DWT	75.37	83.43	60.81	45.31
	PSSM-DPC-DWT	82.45	83.65	81.25	64.91
	R-PSSM-DWT	83.61	82.66	84.56	67.23
LiXGB	F-PSSM-DWT	79.40	83.11	75.65	58.94
	PSSM-DPC-DWT	84.74	84.30	85.19	69.49
	R-PSSM-DWT	86.84	86.60	87.08	73.69

3.3. Comparison with Existing Predictors Using Training Set. Several methods have been implemented for the identification of DBPs. The proposed work is compared with past studies including iDNA-Prot [22], iDNA-Prot|dis [23], TargetDBP [59], MsDBP [60], PDBP-CNN [29], and XGBoost [30] and summarized the results in Table 4. Our proposed study improved the accuracy by 4.82%, sensitivity by 10.58%, and MCC by 0.09 than the best predictor (PDBP-CNN). Similarly, The DBP-iDWT enhanced 5.42% Acc, 2.49% Sn, 8.65% Sp, and 0.11 MCC than the second best study (XGBoost). In the same fashion, our predictor performance is superior to past studies using all four assessment parameters. The outcomes verified that DBP-iDWT can discriminate DBPs with high precision.

3.4. Comparison with Past Predictors Using Independent Set. A method is considered effective if it has high generality for the new sequences. We also evaluated the proposed work using a testing dataset. The results compared with past studies like PseDNA-Pro, iDNAPro-PseAAC, iDNAProt-ES, DPP-PseAAC, TargetDBP, MsDBP, and PDBP-Fusion as noted in Table 5. It is noted that our predictor (DBP-iDWT) raised 5.06% Acc, 17.06% Sn, 8.22% Sp, and 0.10 MCC than PDBP-Fusion. Similarly, DBP-iDWT improved 6.14% Acc, 14.02% Sn, and 0.13 MCC than TargetDBP. Onward, the proposed study also secured higher prediction results than other past methods in Table 5.

These results analysis confirm that the incorporation of DWT into R-PSSM in conjunction with LiXGB can identify DBPs more accurately. Past studies have reported that the selection of the best features can improve the model performance [61–63]. In this study, we also implemented feature selection approach including mRmR and SVM-RFE, however, no improvement in the model performance is observed.

Table 4: Comparative analysis with past work on the training set.

Predictor	Acc (%)	Sn (%)	Sp (%)	MCC
iDNA-prot	75.40	83.81	64.73	0.50
iDNA-prot|dis	77.30	79.40	75.27	0.54
TargetDBP	79.71	79.56	79.85	0.59
MsDBP	80.29	80.87	79.72	0.60
PDBP-CNN	82.02	87.49	76.50	0.64
XGBoost	81.42	84.11	78.43	0.62
DBP-iDWT	86.84	86.60	87.08	0.73

Table 5: Comparative analysis with past work using testing dataset.

Predictor	Acc (%)	Sn (%)	Sp (%)	MCC
PseDNA-pro	67.23	78.38	56.08	0.35
iDNAPro-PseAAC	66.22	78.37	54.05	0.33
iDNAProt-ES	68.58	95.95	41.22	0.44
DPP-PseAAC	61.15	55.41	66.89	0.22
TargetDBP	76.69	76.35	77.03	0.53
MsDBP	66.99	70.69	63.18	0.33
PDBP-fusion	77.77	73.31	66.85	0.56
DBP-iDWT	82.83	90.37	75.07	0.66

4. Conclusion and Future Vision

DBPs play an active role in many biological functions and drug designing. We have designed a predictor for improving DBPs prediction with high precision. The global information, local features, sequence-order patterns, and correlated factors are explored by PSSM-DPC-DWT, R-PSSM-DWT, and PSSM-DPC-DWT.

The models are trained with LiXGB, XGB, ERT, and Adaboost. It is concluded that R-PSSM-DWT with LiXGB has effectively attained superlative performance than other predictors. The successful outcomes of the proposed study is due to factors like utilization of effective descriptors, application of a compression scheme, and appropriate classifier.

DBP-iDWT will be effective for the identification of DBPs due to its promising prediction power than other predictors and perform an active role in drug development. DBP-iDWT would be fruitful for establishing more operative therapeutic strategies for fatal disease treatment. In addition, we will apply advanced deep learning frameworks [64–67] in our future work to further improve the DBPs prediction.

Data Availability

The data and code are freely available at https://github.com/Farman335/DBP-DWTPred.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Authors' Contributions

Farman Ali: Conceptualization, Methodology. Harish Kumar and Shruti Patil: data collection, writing-original draft preparation. Omar Barukab and Ajay B Gadicha: Visualization, performed experiments suggested by reviewers. Omar Alghushairy and Akram Y Sarhan: Code writing, editing, and reviewed the paper.

Acknowledgments

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work under Grant no. RGP.2/198/43.

References

[1] S. Ahmed, M. Kabir, Z. Ali, M. Arif, F. Ali, and D.-J. Yu, "An integrated feature selection algorithm for cancer classification using gene expression data," *Combinatorial Chemistry and High Throughput Screening*, vol. 21, pp. 631–645, 2018.

[2] N. M. Luscombe, S. E. Austin, H. M. Berman, and J. M. Thornton, "An overview of the structures of protein-DNA complexes," *Genome Biology*, vol. 1, 2000.

[3] K. Sandman, S. L. Pereira, and J. N. Reeve, "Diversity of prokaryotic chromosomal proteins and the origin of the nucleosome," *Cellular and Molecular Life Sciences*, vol. 54, no. 12, pp. 1350–1364, 1998.

[4] J. P. Overington, B. Al-Lazikani, and A. L. Hopkins, "How many drug targets are there," *Nature Reviews Drug Discovery*, vol. 5, no. 12, pp. 993–996, 2006.

[5] H. Gronemeyer, J. A Gustafsson, and V. Laudet, "Principles for modulation of the nuclear receptor superfamily," *Nature Reviews Drug Discovery*, vol. 3, no. 11, pp. 950–964, 2004.

[6] W. H. Hudson, I. M. S. d. Vera, J. C. Nwachukwu et al., "Cryptic glucocorticoid receptor-binding sites pervade genomic NF-κB response elements," *Nature Communications*, vol. 9, no. 1, p. 1337, 2018.

[7] H. A. Sikder, M. K. Devlin, S. Dunlap, B. Ryu, and R. M. Alani, "Id proteins in cell growth and tumorigenesis," *Cancer Cell*, vol. 3, no. 6, pp. 525–530, 2003.

[8] S. Akbar, M. Hayat, M. Iqbal, and M. Tahir, "iRNA-PseTNC: identification of RNA 5-methylcytosine sites using hybrid vector space of pseudo nucleotide composition," *Frontiers of Computer Science*, vol. 14, no. 2, pp. 451–460, 2020.

[9] S. Akbar, M. Hayat, M. Iqbal, and M. A. Jan, "iACP-GAEnsC: evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space," *Artificial Intelligence in Medicine*, vol. 79, pp. 62–70, 2017.

[10] S. Akbar, M. Hayat, M. Tahir, and K. T. Chong, "cACP-2LFS: classification of anticancer peptides using sequential discriminative model of KSAAP and two-level feature selection approach," *IEEE Access*, vol. 8, pp. 131939–131948, 2020.

[11] F. Ali, S. Ahmed, Z. N. K. Swati, and S. Akbar, "DP-BINDER: machine learning model for prediction of DNA-binding proteins by fusing evolutionary and physicochemical information," *Journal of Computer-Aided Molecular Design*, vol. 33, no. 7, pp. 645–658, 2019.

[12] A. Ahmad, S. Akbar, S. Khan et al., "Deep-AntiFP: prediction of antifungal peptides using distanct multi-informative features incorporating with deep neural networks," *Chemometrics and Intelligent Laboratory Systems*, vol. 208, Article ID 104214, 2021.

[13] S. Akbar, A. U. Rahman, M. Hayat, and M. Sohail, "CACP: classifying anticancer peptides using discriminative intelligent model via Chou's 5-step rules and general pseudo components," *Chemometrics and Intelligent Laboratory Systems*, vol. 196, Article ID 103912, 2020.

[14] S. Akbar, A. Ahmad, M. Hayat, A. U. Rehman, S. Khan, and F. Ali, "iAtbP-hyb-EnC: prediction of antitubercular peptides via heterogeneous feature representation and genetic algorithm based ensemble learning model," *Computers in Biology and Medicine*, vol. 137, Article ID 104778, 2021.

[15] A. Ahmad, S. Akbar, M. Tahir, M. Hayat, and F. Ali, "iAFPs-EnC-GA: Identifying Antifungal Peptides Using Sequential and Evolutionary Descriptors Based Multi-Information Fusion and Ensemble Learning Approach," *Chemometrics and Intelligent Laboratory Systems*, vol. 222, Article ID 104516, 2022.

[16] S. Akbar, M. Hayat, M. Tahir, S. Khan, and F. K. Alarfaj, "cACP-DeepGram: classification of anticancer peptides via deep neural network and skip-gram-based word embedding model," *Artificial Intelligence in Medicine*, vol. 131, Article ID 102349, 2022.

[17] G. Nimrod, M. Schushan, A. Szilágyi, C. Leslie, and N. Ben-Tal, "iDBPs: a web server for the identification of DNA binding proteins," *Bioinformatics*, vol. 26, no. 5, pp. 692-693, 2010.

[18] M. Gao and J. Skolnick, "DBD-Hunter: a knowledge-based method for the prediction of DNA–protein interactions," *Nucleic Acids Research*, vol. 36, no. 12, pp. 3978–3992, 2008.

[19] H. Zhao, J. Wang, Y. Zhou, and Y. Yang, "Predicting DNA-binding proteins and binding residues by complex structure prediction and application to human proteome," *PLoS One*, vol. 9, no. 5, Article ID e96694, 2014.

[20] F. Ali, H. Kumar, S. Patil, A. Ahmed, A. Banjar, and A. Daud, "DBP-DeepCNN: Prediction of DNA-Binding Proteins Using Wavelet-Based Denoising and Deep Learning," *Chemometrics and Intelligent Laboratory Systems*, vol. 229, Article ID 104639, 2022.

[21] K. K. Kumar, G. Pugalenthi, and P. N. Suganthan, "DNA-Prot: identification of DNA binding proteins from protein sequence information using random forest," *Journal of Biomolecular Structure and Dynamics*, vol. 26, no. 6, pp. 679–686, 2009.

[22] W.-Z. Lin, J.-A. Fang, X. Xiao, and K.-C. Chou, "IDNA-Prot: identification of DNA binding proteins using random forest with grey model," *PLoS One*, vol. 6, no. 9, Article ID e24756, 2011.

[23] B. Liu, J. Xu, X. Lan et al., "IDNA-Prot| dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition," *PLoS One*, vol. 9, Article ID e106691, 2014.

[24] Q. Dong, S. Wang, K. Wang, X. Liu, and B. Liu, "Identification of DNA-binding proteins by auto-cross covariance transformation," in *Proceedings of the Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 470–475, Washington DC, USA, November 2015.

[25] L. Wei, J. Tang, and Q. Zou, "Local-DPP: an improved DNA-binding protein prediction method by exploring local evolutionary information," *Information Sciences*, vol. 384, pp. 135–144, 2017.

[26] F. Ali, M. Kabir, M. Arif et al., "DBPPred-PDSD: machine learning approach for prediction of DNA-binding proteins using Discrete Wavelet Transform and optimized integrated features space," *Chemometrics and Intelligent Laboratory Systems*, vol. 182, pp. 21–30, 2018.

[27] M. S. Rahman, S. Shatabda, S. Saha, M. Kaykobad, and M. S. Rahman, "DPP-PseAAC: a DNA-binding protein prediction model using Chou's general PseAAC," *Journal of Theoretical Biology*, vol. 452, pp. 22–34, 2018.

[28] A. Mishra, P. Pokhrel, and M. T. Hoque, "StackDPPred: a stacking based prediction of DNA-binding protein from sequence," *Bioinformatics*, vol. 35, no. 3, pp. 433–441, 2018.

[29] G. Li, X. Du, X. Li, L. Zou, G. Zhang, and Z. Wu, "Prediction of DNA binding proteins using local features and long-term dependencies with primary sequences based on deep learning," *PeerJ*, vol. 9, Article ID e11262, 2021.

[30] Z. Zhao, W. Yang, Y. Zhai, Y. Liang, and Y. Zhao, "Identify DNA-binding proteins through the extreme gradient boosting algorithm," *Frontiers in Genetics*, vol. 12, Article ID 821996, 2021.

[31] X. Du, Y. Diao, H. Liu, and S. Li, "MsDBP: exploring DNA-binding proteins by integrating multi-scale sequence information via chou's 5-steps rule," *J Proteome res*, vol. 18, 2019.

[32] X. Ma, J. Guo, and X. Sun, "DNABP: identification of DNA-binding proteins based on feature selection using a random forest and predicting binding residues," *PLoS One*, vol. 11, no. 12, Article ID e0167345, 2016.

[33] C. Zou, J. Gong, and H. Li, "An improved sequence based prediction protocol for DNA-binding proteins using SVM and comprehensive feature analysis," *BMC Bioinformatics*, vol. 14, no. 1, p. 90, 2013.

[34] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[35] J. Zahiri, O. Yaghoubi, M. Mohammad-Noori, R. Ebrahimpour, and A. Masoudi-Nejad, "PPIevo: protein–protein interaction prediction from PSSM based evolutionary information," *Genomics*, vol. 102, no. 4, pp. 237–242, 2013.

[36] F. Ali and M. Hayat, "Machine learning approaches for discrimination of Extracellular Matrix proteins using hybrid feature space," *Journal of Theoretical Biology*, vol. 403, pp. 30–37, 2016.

[37] T. Li, K. Fan, J. Wang, and W. Wang, "Reduction of protein sequence complexity by residue grouping," *Protein Engineering Design and Selection*, vol. 16, no. 5, pp. 323–330, 2003.

[38] R. Moshrefi, M. G. Mahjani, and M. Jafarian, "Application of wavelet entropy in analysis of electrochemical noise for corrosion type identification," *Electrochemistry Communications*, vol. 48, pp. 49–51, 2014.

[39] X. Wang, J. Wang, C. Fu, and Y. Gao, "Determination of corrosion type by wavelet-based fractal dimension from electrochemical noise," *International Journal of Electrochemical Science*, vol. 8, pp. 7211–7222, 2013.

[40] B. Yu, S. Li, C. Chen et al., "Prediction subcellular localization of Gram-negative bacterial proteins by support vector machine using wavelet denoising and Chou's pseudo amino acid composition," *Chemometrics and Intelligent Laboratory Systems*, vol. 167, pp. 102–112, 2017.

[41] M. Hayat, A. Khan, and M. Yeasin, "Prediction of membrane proteins using split amino acid and ensemble classification," *Amino Acids*, vol. 42, no. 6, pp. 2447–2460, 2012.

[42] B. Ma, F. Meng, G. Yan, H. Yan, B. Chai, and F. Song, "Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data," *Computers in Biology and Medicine*, vol. 121, Article ID 103761, 2020.

[43] X. Wang, Y. Zhang, B. Yu et al., "Prediction of protein-protein interaction sites through eXtreme gradient boosting with kernel principal component analysis," *Computers in Biology and Medicine*, vol. 134, Article ID 104516, 2021.

[44] S. Akbar, S. Khan, F. Ali, M. Hayat, M. Qasim, and S. Gul, "iHBP-DeepPSSM: identifying hormone binding proteins using PsePSSM based evolutionary features and deep learning approach," *Chemometrics and Intelligent Laboratory Systems*, vol. 204, Article ID 104103, 2020.

[45] F. Ali, S. Akbar, A. Ghulam, Z. A. Maher, A. Unar, and D. B. Talpur, "AFP-CMBPred: computational identification of antifreeze proteins by extending consensus sequences into multi-blocks evolutionary information," *Computers in Biology and Medicine*, vol. 139, Article ID 105006, 2021.

[46] I. A. Khan, D. Pi, N. Khan et al., "A Privacy-Conserving Framework Based Intrusion Detection Method for Detecting and Recognizing Malicious Behaviours in Cyber-Physical Power Networks," *Applied Intelligence*, vol. 51, pp. 1–16, 2021.

[47] P. Chaudhari, H. Agrawal, and K. Kotecha, "Data augmentation using MG-GAN for improved cancer classification on gene expression data," *Soft Computing*, vol. 24, no. 15, pp. 11381–11391, 2020.

[48] F. Ali and M. Hayat, "Classification of membrane protein types using voting feature interval in combination with chou′ s pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 384, pp. 78–83, 2015.

[49] F. Ali, H. Kumar, S. Patil, A. Ahmad, A. Babour, and A. Daud, "Deep-GHBP: improving prediction of Growth Hormone-binding proteins using deep learning model," *Biomedical Signal Processing and Control*, vol. 78, Article ID 103856, 2022.

[50] F. Ali, H. Kumar, S. Patil, K. Kotecha, A. Banjar, and A. Daud, "Target-DBPPred: an intelligent model for prediction of DNA-binding proteins using discrete wavelet transform based compression and light eXtreme gradient boosting," *Computers in Biology and Medicine*, vol. 145, Article ID 105533, 2022.

[51] O. Barukab, F. Ali, W. Alghamdi, Y. Bassam, and S. A. Khan, "DBP-CNN: deep learning-based prediction of DNA-binding proteins by coupling discrete cosine transform with two-dimensional convolutional neural network," *Expert Systems with Applications*, vol. 197, Article ID 116729, 2022.

[52] O. Barukab, F. Ali, and S. A. Khan, "DBP-GAPred: an intelligent method for prediction of DNA-binding proteins types by enhanced evolutionary profile features with ensemble learning," *Journal of Bioinformatics and Computational Biology*, vol. 19, no. 04, Article ID 2150018, 2021.

[53] A. Ghulam, F. Ali, R. Sikander, A. Ahmad, A. Ahmed, and S. Patil, "ACP-2DCNN: deep learning-based model for improving prediction of anticancer peptides using two-dimensional convolutional neural network," *Chemometrics and Intelligent Laboratory Systems*, vol. 226, Article ID 104589, 2022.

[54] A. Ghulam, R. Sikander, F. Ali, Z. N. K. Swati, A. Unar, and D. B. Talpur, "Accurate prediction of immunoglobulin proteins using machine learning model," *Informatics in Medicine Unlocked*, vol. 29, Article ID 100885, 2022.

[55] Z. U. Khan, F. Ali, I. Ahmad, M. Hayat, and D. Pi, "iPredCNC: computational prediction model for cancerlectins and non-cancerlectins using novel cascade features subset selection," *Chemometrics and Intelligent Laboratory Systems*, vol. 195, Article ID 103876, 2019.

[56] Z. U. Khan, F. Ali, I. A. Khan, Y. Hussain, and D. Pi, "iRSpot-SPI: deep learning-based recombination spots prediction by incorporating secondary sequence information coupled with physio-chemical properties via Chou's 5-step rule and pseudo components," *Chemometrics and Intelligent Laboratory Systems*, vol. 189, pp. 169–180, 2019.

[57] Z. U. Khan, D. Pi, S. Yao, A. Nawaz, F. Ali, and S. Ali, "piEnPred: a bi-layered discriminative model for enhancers and their subtypes via novel cascade multi-level subset feature selection algorithm," *Frontiers of Computer Science*, vol. 15, no. 6, pp. 156904–156911, 2021.

[58] M. Ullah, A. Iltaf, Q. Hou, F. Ali, and C. Liu, "A foreground extraction approach using convolutional neural network with graph cut," in *Proceedings of the 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC)*, pp. 40–44, Chongqing, China, June 2018.

[59] J. Hu, X.-G. Zhou, Y.-H. Zhu, D.-J. Yu, and G.-J. Zhang, "TargetDBP: accurate DNA-binding protein prediction via sequence-based multi-view feature learning," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 4, pp. 1419–1429, 2020 Jul-Aug.

[60] X. Du, Y. Diao, H. Liu, and S. Li, "MsDBP: exploring DNA-binding proteins by integrating multiscale sequence information via Chou's five-step rule," *Journal of Proteome Research*, vol. 18, no. 8, pp. 3119–3132, 2019.

[61] G. Sanghani and K. Kotecha, "Incremental personalized E--mail spam filter using novel TFDCR feature selection with dynamic feature update," *Expert Systems with Applications*, vol. 115, pp. 287–299, 2019.

[62] A. Ahmad, S. Akbar, M. Hayat, F. Ali, and M. Sohail, "Identification of Antioxidant Proteins Using a Discriminative Intelligent Model of K-Space Amino Acid Pairs Based Descriptors Incorporating with Ensemble Feature Selection," *Biocybernetics and Biomedical Engineering*, vol. 42, 2020.

[63] S. Akbar, M. Hayat, M. Kabir, and M. Iqbal, "iAFP-gap-SMOTE: an efficient feature extraction scheme gapped dipeptide composition is coupled with an oversampling technique for identification of antifreeze proteins," *Letters in Organic Chemistry*, vol. 16, no. 4, pp. 294–302, 2019.

[64] G. Joshi, R. Walambe, and K. Kotecha, "A review on explainability in multimodal deep neural nets," *IEEE Access*, vol. 9, pp. 59800–59821, 2021.

[65] F. Ali, M. Arif, Z. U. Khan, M. Kabir, S. Ahmed, and D.-J. Yu, "SDBP-Pred: prediction of single-stranded and double-stranded DNA-binding proteins by extending consensus sequence and K-segmentation strategies into PSSM," *Analytical Biochemistry*, vol. 589, Article ID 113494, 2020.

[66] F. Ali, F. Ali, A. Ghulam et al., "Deep-PCL: a deep learning model for prediction of cancerlectins and non cancerlectins using optimized integrated features," *Chemometrics and Intelligent Laboratory Systems*, vol. 221, Article ID 104484, 2022.

[67] R. Sikander, A. Ghulam, and F. Ali, "XGB-DrugPred: computational prediction of druggable proteins using eXtreme gradient boosting and optimized features set," *Scientific Reports*, vol. 12, pp. 5505–5509, 2022.