Methods 226 (2024) 49–53

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

DEEP-EP: Identification of epigenetic protein by ensemble residual convolutional neural network for drug discovery

Farman Ali [a,*], Abdullah Almuhaimeed [b], Majdi Khalid [c], Hanan Alshanbari [c], Atef Masmoudi [d], Raed Alsini [e]

[a] Department of Computer Science, Bahria University Islamabad Campus, Pakistan
[b] Digital Health Institute, King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia
[c] Department of Computer Science and Artificial Intelligence, College of Computing, Umm Al-Qura University, Makkah 21955, Saudi Arabia
[d] College of Computer Science, King Khalid University, Abha, Saudi Arabia
[e] Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

ARTICLE INFO

Keywords:
Deep learning
Dipeptide deviation from expected mean
Ensemble residual convolutional neural network

ABSTRACT

Epigenetic proteins (EP) play a role in the progression of a wide range of diseases, including autoimmune disorders, neurological disorders, and cancer. Recognizing their different functions has prompted researchers to investigate them as potential therapeutic targets and pharmacological targets. This paper proposes a novel deep learning-based model that accurately predicts EP. This study introduces a novel deep learning-based model that accurately predicts EP. Our approach entails generating two distinct datasets for training and evaluating the model. We then use three distinct strategies to transform protein sequences to numerical representations: Dipeptide Deviation from Expected Mean (DDE), Dipeptide Composition (DPC), and Group Amino Acid (GAAC). Following that, we train and compare the performance of four advanced deep learning models algorithms: Ensemble Residual Convolutional Neural Network (ERCNN), Generative Adversarial Network (GAN), Convolutional Neural Network (CNN), and Gated Recurrent Unit (GRU). The DDE encoding combined with the ERCNN model demonstrates the best performance on both datasets. This study demonstrates deep learning's potential for precisely predicting EP, which can considerably accelerate research and streamline drug discovery efforts. This analytical method has the potential to find new therapeutic targets and advance our understanding of EP activities in disease.

1. Introduction

Epigenetic proteins perform critical functions in regulating gene expression without affecting the underlying DNA sequence. They accomplish this by changing the packing and accessibility of DNA within the cell nucleus, eventually regulating which genes are switched on or off. These changes can be inherited through cell division, resulting in long-term changes in gene expression patterns [1].

Aberrant epigenetic modifications have been linked to diverse diseases, such as neurodegenerative diseases, autoimmune disorders, and cancer. Understanding these changes can lead to the development of novel therapies. For instance, Aberrant epigenetic modifications contribute to cancer development by silencing tumor suppressor genes and activating oncogenes [2]. DNA methylation, frequently observed in cancer cells, silences genes associated to apoptosis, DNA repair, and cell cycle control. Additionally, histone modifications can alter chromatin structure and gene accessibility, further promoting cancer progression. Several epigenetic drugs, like DNA methyltransferase (DNMT) inhibitors and histone deacetylase (HDAC), are currently being developed and evaluated for cancer treatment [3].

Alterations in histone modifications and DNA methylation have been linked to neurodegenerative diseases like Parkinson's and Alzheimer's [4]. Abnormal expression of epigenetic proteins may contribute to the accumulation of toxic proteins and neuronal death. Epigenetic drugs hold promise for treating these diseases by reversing these abnormal modifications and protecting neurons from damage. Dysregulation of epigenetic mechanisms can lead to autoimmune disorders, where the immune system attacks healthy tissues. Aberrant DNA methylation patterns in immune cells have been observed in diseases like rheumatoid arthritis and lupus [5]. Epigenetic drugs are being investigated as

[*] Corresponding authors.
E-mail addresses: farman.buic@bahria.edu.pk, farman335@yahoo.com (F. Ali).

https://doi.org/10.1016/j.ymeth.2024.04.004
Received 27 February 2024; Received in revised form 6 April 2024; Accepted 8 April 2024
Available online 14 April 2024

potential therapeutic agents for autoimmune disorders by modulating the immune response and reducing inflammation.

Epigenetic proteins are emerging as powerful players in the realm of drug discovery [6]. These fascinating molecules, by regulating gene expression without altering DNA itself, offer unique therapeutic avenues for treating various diseases.

Accurate identification of EP is one of the challenges in this field. Identification of EP via experimental methods is expensive and time-consuming. This research establishes Deep-EP, a novel computational predictor for EP prediction based on deep learning. In this investigation, we introduced a novel sequence-based computational predictor utilizing deep learning algorithms to precisely forecast the Epigenetic protein. First, we constructed two sequence-based datasets i.e., the training dataset and testing dataset. Second, numerical features were extracted from primary sequences using diverse feature representative techniques, encompassing group amino acid composition, dipeptide composition, and dipeptide deviation from expected mean. Third, model training employed deep learning methodologies, including convolutional neural network, gated recurrent unit, generative adversarial network, and ensemble residual convolutional neural network. The efficacy of each model is assessed via 5-fold cross-validation (CV) with evaluation metrics including accuracy, sensitivity, specificity, and Mathew coefficient correlation DDE-based ERCNN attained the most robust predictive outcomes, surpassing competing predictors on both the training and testing datasets. Fig. 1 depicts the methodizes of the proposed research.

2. Material and methods

2.1. Dataset collection

We need to develop a computational model that can distinguish between two types of proteins: epigenetic protein and non-epigenetic protein. To achieve this goal, we collected protein sequences from the UniProt database (https://www.uniprot.org/) [7]. We used the search terms “epigenetic protein” and “non-epigenetic protein” to find sequences of both types. We then used a tool called CD-hit [8,9] to remove redundant sequences and sequences that were too short. This resulted in a final dataset containing 1113 epigenetic protein and 1143 non-epigenetic protein. We divided this dataset into two parts: a training set and a testing set. The training set contains 979 epigenetic protein and 1029 non-epigenetic protein, while the testing set contains 134 epigenetic protein and 114 non-epigenetic protein. We train our model on the training set and then test its performance on the testing set to see how well it can distinguish between epigenetic protein and non-epigenetic protein.

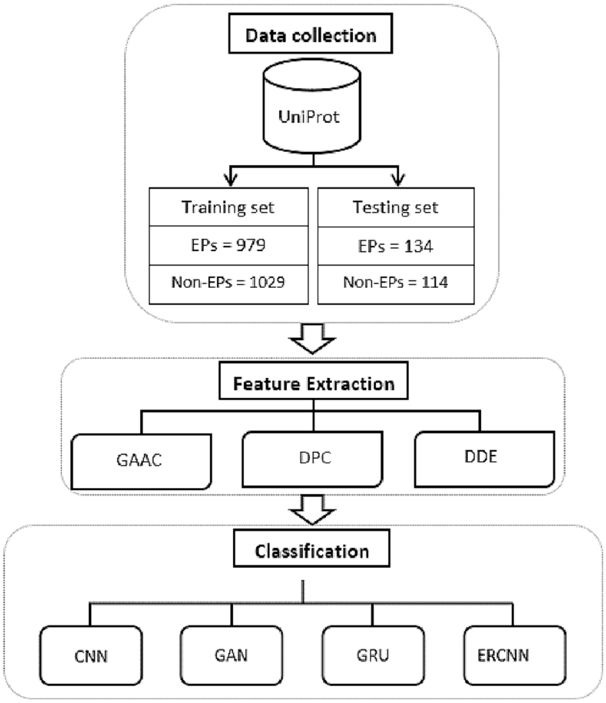

Fig. 1. Schematic view of the study.

2.2. Feature representation methods

Raw protein sequences cannot be directly processed by deep/machine learning algorithms [10,11]. To facilitate their analysis, feature extraction methods are utilized to convert these sequences into numerical representations. In this investigation, DDE has been applied for the conversion of sequences into numerical format. The specifics of this method are outlined below.

2.2.1. Dipeptide Deviation from Expected mean

DDE refers to a computational approach used in the analysis of protein sequences [12,13]. Proteins, comprised of amino acids, are the building blocks of biological structures and act active roles in different cellular activities [14]. Dipeptides are pairs of adjacent amino acids within a protein sequence. The concept of “Deviation from Expected Mean” involves examining the frequency of dipeptides in a protein sequence and comparing it to an expected average [15].

DPC represents the frequency of occurrence of pairs of amino acids along a protein sequence. Nonetheless, earlier research [16] has found differences in DPC. In response to these findings, Saravanan et al. [17] proposed Dipeptide Deviation from Expected Mean. DDE is a more effective feature descriptor than other approaches since it considers correlation factors and sequence-order information [18]. Notably, DDE considers global motifs, which are important features of proteins. Recognizing its effectiveness, many researchers have recently implemented DDE, significantly improving model performance in various domains, such as identification of angiogenic protein [19], antifreeze protein prediction [20], predicting ubiquitin proteins [21], and identifying anticancer peptides [22].

DDE has three parameters: theoretical variance (Tv), dipeptide composition measure ($D_{c(i)}$), theoretical mean ($T_{m(i)}$). $D_{c(i)}$ is formulated as

$$D_{c(i)} = \frac{f_i}{N} \tag{1}$$

Here, f_i denotes the frequency of dipeptide i and N represent the number of dipeptide in a protein sequence. The theoretical mean $T_{m(i)}$ can be computed as

$$T_{m(i)} = \frac{C_{i1}}{C_N} \times \frac{C_{i2}}{C_N} \tag{2}$$

C_{i1}, C_{i2}, and C_N are the number of codons for first, second, and total amino acids in the given dipeptide i is the total codons. The theoretical variance $T_{v(i)}$ of the dipeptide i can be expressed as

$$T_{v(i)} = \frac{T_{m(i)}(1 - T_{m(i)})}{N} \tag{3}$$

N represents residues in a sequence P, while $T_{m(i)}$ is calculated by Eq. (3). The DDE_i is computed as

$$DDE_i = \frac{D_{c(i)} - T_{m(i)}}{\sqrt{T_{v(i)}}} \tag{4}$$

The dimension of DDE_d is calculated by equation below.

$$DDE_d = (DDE_{(1)}, DDE_{(2)}, DDE_{(3)}, \cdots\cdots, DDE_{(400)}) \tag{5}$$

The total attributes of DDE are 400.

2.3. Model training algorithms

Following the feature extraction stage, the training of a model using a robust algorithm is pivotal for achieving promising performance [14,23]. In this context, we incorporated various deep learning and machine learning techniques, including GAN, GRU, CNN, ERT, RF, XGB, and ERCNN. Among these frameworks, the ERCNN demonstrates notable contributions to the performance of the proposed study. Further details regarding the ERCNN are elucidated below.

2.3.1. Ensemble residual convolutional neural network

Ensemble Residual Convolutional Neural Network (ERCNN) is an advanced deep learning architecture that integrates principles from ensemble learning, residual neural networks (ResNet), and convolutional neural networks (CNN) [24,25]. This innovative approach is tailored to enhance model accuracy, robustness, and generalization. ResNet, introduced by Kaiming He et al. [26] in 2015, serves as a crucial component of ERCNN, addressing the issue of vanishing gradients in CNN through shortcut connections. These shortcuts facilitate a more direct flow of information, mitigating the risk of performance deterioration with increasing network depth [27].

Ensemble learning [28,29], a machine learning technique that leverages multiple models to improve prediction accuracy, is employed in ERCNN to enhance model performance by ensembling multiple ResNet-CNN models. The combination of diverse sub-models helps counter overfitting and captures a broader range of features, particularly beneficial for small or noisy datasets [30]. Concatenating predictions from multiple sub-models ensures a more reliable final decision based on a consensus among multiple networks [31].

In this study, the ERCNN model is constructed using two types of residual building blocks (RBB): RBB-1 and RBB-2. Both RBBs consist of three convolutional and batch normalization layers, with a padding layer added before the convolutional layers to preserve input feature spatial dimensions. Different numbers of filters (32, 64, 128, 256) are tested for each convolutional layer, with 64 filters showing optimal prediction performance [32,33]. A stride size of 3 is applied uniformly across all convolutional layers. RBB-1 incorporates a shortcut denoted by 'x,' while RBB-2 includes one convolutional and batch normalization layer in its shortcut path.

The computation of RBB-1 is expressed as

$$y = F(x) + x \tag{6}$$

where F is the nonlinear function for the convolutional path. Similarly, RBB-2 is computed as

$$y = F(x) + H(x) \tag{7}$$

where H represents the shortcut path. The structure of both RBBs is depicted in Fig. 2. Dropout layers are strategically placed after fully-connected layers to improve classification performance and prevent overfitting [34]. The output layer comprises a flatten layer for transforming features into a vector shape and a sigmoid layer for outputting the probability of each possible outcome. The study implements various hyperparameters, detailed in Table 1.

2.4. Model validation methods

To ensure the dependability of the proposed model, it undergoes validation through established methods [35,36]. A widely employed technique in bioinformatics [37–40] is the 5-fold cross-validation (CV). In this approach, our data is divided into five sets, and the model undergoes testing on the tenth set after being trained on the other four sets. This process is repeated five times, each time testing with a distinct set. The ultimate prediction is determined as the mean of the outcomes from the 5-fold CV [41]. Furthermore, the model's performance is evaluated using assessment metrics, including Matthews Correlation Coefficient (MCC), sensitivity (Sn), specificity (Sp), and accuracy (Acc) [42]. The formulation of these metrics relies on the utilization of a confusion matrix, and the computation is performed using the following equations.

$$Acc = 1 - \frac{E_-^+ + E_+^-}{E^+ + E^-} \tag{3}$$

$$Sn = 1 - \frac{E_-^+}{E^+} \tag{4}$$

$$Sp = 1 - \frac{E_+^-}{E^-} \tag{5}$$

$$MCC = \frac{1 - \left(\frac{E_-^+ + E_+^-}{E^+ + E^-}\right)}{\sqrt{\left(1 + \frac{E_+^- + E_-^+}{E^+}\right)\left(1 + \frac{E_-^+ + E_+^-}{E^-}\right)}}$$

The notation E^+ signifies instances of correctly predicted EP, whereas E^- represents instances correctly identified as non-EP. Similarly, E_-^+ denotes instances where EP has been incorrectly identified as non-EP, while E_+^-

Table 1
List of hyperparameters of the ERCNN model.

Hyperparameter	Value
Number of epochs	80
Batch size	100
Learning rate	0.001
Convolution layers	3
Dropout	0.4
Activation function	Sigmoid
Optimizer	Adam

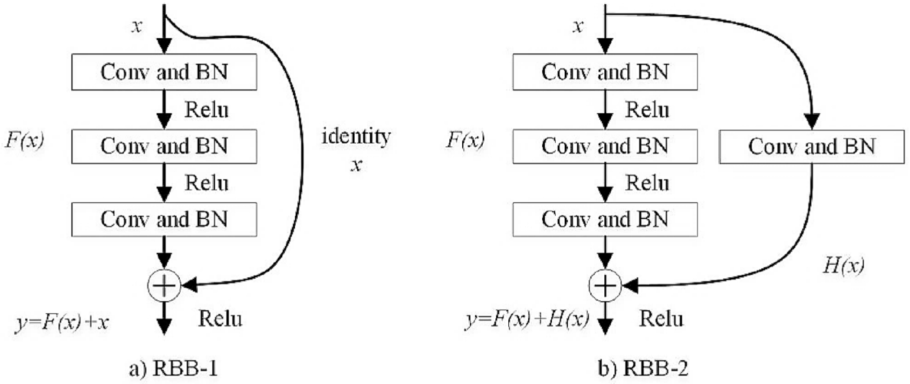

Fig. 2. Schematic view of RBB-1 and RBB-2.

represents the incorrect prediction of non-EP as EP. To assess the model's generalization capabilities, we employed a separate testing or independent dataset.

3. Results and discussion

This section presents the performance of different classifiers utilizing various feature descriptors. A detailed comparison is provided in the following subsections.

3.1. Performance of deep learning frameworks

The outcomes achieved by deep learning models employing various feature-representative approaches are detailed in Table 2. The GAN model, coupled with the GAAC feature descriptor, attains an accuracy of 68.52 %, sensitivity of 70.45 %, and specificity of 66.72 %. An AUC value of 77.43 % and MCC of 0.374 signify a moderate overall performance, effectively balancing true positives and true negatives. In contrast, the GRU model with the GAAC descriptor exhibits less impressive performance, featuring a slightly low accuracy of 67.12 % and sensitivity of 61.36 %. The specificity of 72.83 % and AUC value of 68.91 % reflects the model's limited ability to distinguish between classes. The CNN model with GAAC demonstrates promising performance, achieving 70.78 % accuracy, 71.71 % sensitivity, and 69.53 % specificity. The AUC (79.96 %) and MCC (0.468) suggest a robust ability to distinguish between the two classes. ERCNN on GAAC outperforms CNN, GRU, and GAN with an accuracy of 76.33 %, sensitivity of 78.54 %, specificity of 74.84 %, AUC of 85.12 %, and MCC of 0.524, indicating higher overall performance.

In the evaluation of the DPC approach, GAN on DPC secures 69.52 % accuracy. GRU demonstrates improved performance across all parameters, while CNN exhibits optimal outcomes with DPC in terms of accuracy, sensitivity, specificity, AUC, and MCC. The highest results are achieved by ERCNN on the DDE descriptor, with accuracy, sensitivity, specificity, AUC, and MCC values of 82.14 %, 85.15 %, 85.21 %, 88.90 %, and 0.665, respectively. ERCNN consistently demonstrates promising performance.

With the DDE [43] feature representative approach, GAN performs impressively, achieving an accuracy of 73.41 %, surpassing GAAC and DPC by 4.89 % and 3.85 %, respectively. GRU also outperforms the GAAC descriptor in terms of accuracy, specificity, AUC, and MCC. Similarly, CNN with DDE boosts performance, acquiring 85.70 % accuracy, 85.90 % sensitivity, 85.51 % specificity, 90.45 % AUC, and 0.682 MCC. Among all classification algorithms, ERCNN reflects superior performance, leveraging its advantages for tasks involving sequence data, such as capturing hierarchical features, parameter efficiency, translation invariance, and parallelization capacity. DDE feature representation approach demonstrates promising efficacy across all deep learning frameworks compared to other feature encoders, extracting local and global features effectively while considering sequence residues' correlations, making it a unique and advantageous technique.

Table 2
Results comparison of the deep learning approaches.

Classifier	Feature descriptor	Acc (%)	Sn (%)	Sp (%)	AUC (%)	MCC
GAN	GAAC	68.52	70.45	66.72	77.43	0.374
GRU		67.12	61.36	72.83	68.91	0.395
CNN		70.78	71.71	69.53	79.96	0.468
ERCNN		76.33	78.54	74.84	85.12	0.524
GAN	DPC	69.56	70.43	68.22	78.44	0.465
GRU		74.55	75.41	73.75	83.71	0.558
CNN		79.54	77.24	80.86	80.66	0.613
ERCNN		82.14	85.15	85.21	88.90	0.655
GAN	DDE	73.41	72.70	76.11	81.85	0.523
GRU		76.03	74.20	78.61	83.15	0.543
CNN		85.70	85.90	85.51	90.45	0.682
ERCNN		87.59	87.28	87.94	92.74	0.701

3.2. Comparison with machine learning models

We conduct additional experiments on machine learning algorithms, employing the same feature methods, validation, and evaluation approaches, with summarized outcomes presented in Table 3. To build an efficient model, hyperparameters for all machine learning classifiers are determined through a grid search strategy. As observed in Table 3, RF exhibits sensitivity of 76.84 %, specificity of 77.14 %, yielding accuracy, AUC, and MCC values of 78.63 %, 85.68 %, and 0.589, respectively. Similarly, ERT improves the performance with all evaluation metrics.

XGB outperforms RF and ERT classifiers, securing an accuracy of 82.14 %, sensitivity of 83.57 %, specificity of 81.10 %, AUC of 88.56 %, and MCC of 0.647. Our model achieves an accuracy of 87.59 %, indicating its proficiency in correctly classifying the data. The sensitivity of 87.28 % is notably high, signifying its effectiveness in identifying positive instances (EP). The specificity of 87.94 % is also high, resulting in a well-balanced performance. The outstanding AUC value of 92.74 % underscores the model's strong ability to discriminate EP from non-EP. Additionally, the MCC value of 0.701 suggests an excellent overall performance in effectively balancing true positives and true negatives.

3.3. Comparison of testing dataset

Examining the validation of the proposed predictor on the training dataset, we proceeded to assess its generalization capability using an independent dataset. GAN, GRU, CNN, and ERCNN were implemented on the testing dataset, and the outcomes are detailed in Table 4. GAN exhibits commendable performance across all evaluation parameters. Similarly, the GRU model boosted the performance in terms of all parameters. CNN model outperformed both GRU and GAN models in terms of accuracy, sensitivity, specificity, and MCC when compared to both GRU and GAN. The proposed predictor, denoted as DEEP-EP, achieves a notable 79.03 % accuracy, 77.61 % sensitivity, 80.70 % specificity, and MCC of 0.581. These results signify that DEEP-EP excels in predicting Epigenetic protein with high precision. Our study secured the highest success rate and surpassed other deep learning approaches across all evaluation parameters on the testing dataset, reflecting its remarkable performance and promising generalization efficacy.

4. Conclusion

Epigenetic proteins perform a pivotal role in gene expression regulation and controlling cellular functions, contributing significantly to various biological processes. By understanding the roles of epigenetic proteins in diseases, we gain valuable insights into their underlying mechanisms, which facilitates the identification of promising therapeutic targets. Additionally, epigenetic proteins are integral to drug discovery, as targeting them can modify gene expression patterns and potentially lead to the development of novel treatments for various diseases. In this study, we presented the first computational predictor for the identification of Epigenetic protein using deep learning.

Our study evaluated the efficacy of four deep learning architectures (CNN, GAN, ERCNN, and GRU) for predicting EP using two separate datasets. Protein sequences were encoded using three distinct techniques (DDE, DPC, and GAAC). The DDE-encoded ERCNN model

Table 3
Results comparison of the machine learning approaches.

Classifier	Acc (%)	Sn (%)	Sp (%)	AUC (%)	MCC
RF	78.63	76.84	77.14	85.68	0.589
ERT	80.90	73.56	87.86	88.85	0.611
XGB	82.14	83.57	81.10	89.56	0.647
DEEP-EP	87.59	87.28	87.94	92.74	0.701

Table 4
Results comparison of classifiers on the testing set.

Classifier	Acc (%)	Sn (%)	Sp (%)	MCC
GAN	68.62	70.88	71.92	0.458
GRU	71.45	69.94	70.90	0.496
CNN	72.28	71.39	75.67	0.515
DEEP-EP	79.03	77.61	80.70	0.581

consistently demonstrated superior performance, highlighting the potential of deep learning for accurate EP prediction and its potential to accelerate research and streamline drug discovery efforts.

The best prediction of DEEP-EP is due to the application of an appropriate feature descriptor and an effective deep-learning model. Investigating and manipulating epigenetic proteins hold immense potential for understanding biology, disease causes, and developing new treatments.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through large group Research Project under grant number RGP. 2/171/44.

References

[1] E. Gibney and C.J.H. Nolan, “Epigenetics and gene expression,” vol. 105, no. 1, pp. 4-13, 2010.

[2] M.J.B. j. o. c. Esteller, “Epigenetics provides a new generation of oncogenes and tumour-suppressor genes,” vol. 94, no. 2, pp. 179-183, 2006.

[3] F. Lyko and R.J. J. o. t. N. C. I. Brown, “DNA methyltransferase inhibitors and the development of epigenetic cancer therapies,” vol. 97, no. 20, pp. 1498-1506, 2005.

[4] K.-X. Wen *et al.*, “The role of DNA methylation and histone modifications in neurodegenerative diseases: a systematic review,” vol. 11, no. 12, p. e0167201, 2016.

[5] M. Vecellio, H. Wu, Q. Lu, C.J.C.R. Selmi, “The multifaceted functional role of DNA methylation in immune-mediated rheumatic diseases,” vol. 40, pp. 459-476, 2021.

[6] C.H. Arrowsmith, C. Bountra, P.V. Fish, K. Lee, M.J.N.R. D.D. Schapira, “Epigenetic protein families: a new frontier for drug discovery,” vol. 11, no. 5, pp. 384-400, 2012.

[7] R. Alsini *et al.*, “Deep-VEGF: deep stacked ensemble model for prediction of vascular endothelial growth factor by concatenating gated recurrent unit with two-dimensional convolutional neural network,” pp. 1-11, 2024.

[8] A. Adnan, W. Hongya, F. Ali, M. Khalid, O. Alghushairy, and R. Alsini, “A bi-layer model for identification of piwiRNA using deep neural learning,” *Journal of Biomolecular Structure and Dynamics,* pp. 1-9.

[9] F. Ali, A. Ghulam, Z.A. Maher, M.A. Khan, S.A. Khan, W. Hongya, Deep-PCL: A deep learning model for prediction of cancerlectins and non cancerlectins using optimized integrated features, Chemom. Intell. Lab. Syst. 221 (2022) 104484.

[10] 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,* 2020.

[11] F. Ali, M. Arif, Z.U. Khan, M. Kabir, S. Ahmed, 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, Anal. Biochem. 589 (2020) 113494.

[12] O. Barukab, F. Ali, W. Alghamdi, Y. Bassam, 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 Syst. Appl.* (2022).

[13] S. Akbar, S. Khan, F. Ali, M. Hayat, M. Qasim, S. Gul, iHBP-DeepPSSM: Identifying hormone binding proteins using PsePSSM based evolutionary features and deep learning approach, Chemom. Intell. Lab. Syst. 204 (2020) 104103.

[14] F. Ali, S. Ahmed, Z.N.K. Swati, S. Akbar, DP-BINDER: machine learning model for prediction of DNA-binding proteins by fusing evolutionary and physicochemical information, J. Comput. Aided Mol. Des. 33 (7) (2019) 645–658.

[15] S. Akbar *et al.*, “Prediction of Amyloid Proteins using Embedded Evolutionary & Ensemble Feature Selection based Descriptors with eXtreme Gradient Boosting Model,” 2023.

[16] J. Chen, H. Liu, J. Yang, K.-C. Chou, Prediction of linear B-cell epitopes using amino acid pair antigenicity scale, Amino Acids 33 (3) (2007) 423–428.

[17] V. Saravanan, N. Gautham, Harnessing computational biology for exact linear B-cell epitope prediction: a novel amino acid composition-based feature descriptor, OMICS 19 (10) (2015) 648–658.

[18] R. Sikander, A. Ghulam, F. Ali, XGB-DrugPred: computational prediction of druggable proteins using eXtreme gradient boosting and optimized features set, Sci. Rep. 12 (1) (2022) 1–9.

[19] F. Ali, W. Alghamdi, A. O. Almagrabi, O. Alghushairy, A. Banjar, and M. J. I. J. o. B. M. Khalid, “Deep-AGP: Prediction of angiogenic protein by integrating two-dimensional convolutional neural network with discrete cosine transform,” p. 125296, 2023.

[20] A. Khan *et al.*, “AFP-SPTS: An Accurate Prediction of Antifreeze Proteins Using Sequential and Pseudo-Tri-Slicing Evolutionary Features with an Extremely Randomized Tree,” 2023.

[21] S. Rahu *et al.*, “UBI-XGB: Identification of ubiquitin proteins using machine learning model,” vol. 8, pp. 14-26, 2022.

[22] A. Ghulam, F. Ali, R. Sikander, A. Ahmad, A. Ahmed, S. Patil, ACP-2DCNN: Deep learning-based model for improving prediction of anticancer peptides using two-dimensional convolutional neural network, Chemom. Intell. Lab. Syst. 226 (2022) 104589.

[23] O. Barukab, F. Ali, S.A. Khan, DBP-GAPred: An intelligent method for prediction of DNA-binding proteins types by enhanced evolutionary profile features with ensemble learning, J. Bioinform. Comput. Biol. (2021) 2150018.

[24] X. Fan *et al.*, “Deep learning for intelligent traffic sensing and prediction: recent advances and future challenges,” vol. 2, pp. 240-260, 2020.

[25] O. Alghushairy *et al.*, “Machine learning-based model for accurate identification of druggable proteins using light extreme gradient boosting,” pp. 1-12, 2023.

[26] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[27] F. Ali, H. Kumar, S. Patil, A. Ahmed, A. Banjar, A. Daud, DBP-DeepCNN: Prediction of DNA-binding proteins using wavelet-based denoising and deep learning, Chemom. Intell. Lab. Syst. (2022), p. 104639.

[28] M. Khalid *et al.*, “An ensemble computational model for prediction of clathrin protein by coupling machine learning with discrete cosine transform,” pp. 1-9, 2024.

[29] F. Ali, S. Akbar, A. Ghulam, Z.A. Maher, A. Unar, D.B. Talpur, AFP-CMBPred: Computational identification of antifreeze proteins by extending consensus sequences into multi-blocks evolutionary information, Comput. Biol. Med. 139 (2021) 105006.

[30] F. Ali, H. Kumar, W. Alghamdi, F. A. Kateb, and F. K. J. A. o. C. M. i. E. Alarfaj, “Recent Advances in Machine Learning-Based Models for Prediction of Antiviral Peptides,” pp. 1-12, 2023.

[31] A. Banjar, F. Ali, O. Alghushairy, A. Daud, iDBP-PBMD: A machine learning model for detection of DNA-binding proteins by extending compression techniques into evolutionary profile, Chemom. Intell. Lab. Syst. (2022), p. 104697.

[32] A. Khan, et al., Prediction of antifreeze proteins using machine learning, Sci. Rep. 12 (1) (2022) 1–10.

[33] F. Ali, H. Kumar, S. Patil, K. Kotecha, A. Banjar, A. Daud, Target-DBPPred: An intelligent model for prediction of DNA-binding proteins using discrete wavelet transform based compression and light eXtreme gradient boosting, Comput. Biol. Med. 145 (2022) 105533.

[34] A. Khan, J. Uddin, F. Ali, A. Banjar, A. Daud, Comparative analysis of the existing methods for prediction of antifreeze proteins, Chemom. Intell. Lab. Syst. (2022).

[35] F. Ali, M. Hayat, Classification of membrane protein types using Voting Feature Interval in combination with Chou′ s Pseudo Amino Acid Composition, J. Theor. Biol. 384 (2015) 78–83.

[36] Z.U. Khan, F. Ali, I. Ahmad, M. Hayat, D. Pi, iPredCNC: computational prediction model for cancerlectins and non-cancerlectins using novel cascade features subset selection, Chemom. Intell. Lab. Syst. 195 (2019) 103876.

[37] F. Ali, M. Hayat, Machine learning approaches for discrimination of Extracellular Matrix proteins using hybrid feature space, J. Theor. Biol. 403 (2016) 30–37.

[38] A. Ghulam, R. Sikander, and F. Ali, “AI and Machine Learning-based practices in various domains: A Survey,” 2022.

[39] A. Ahmad, S. Akbar, M. Tahir, M. Hayat, F. Ali, iAFPs-EnC-GA: Identifying antifungal peptides using sequential and evolutionary descriptors based multi-information fusion and ensemble learning approach, Chemom. Intell. Lab. Syst. (2022), p. 104516.

[40] F. Ali, et al., DBPPred-PDSD: Machine learning approach for prediction of DNA-binding proteins using Discrete Wavelet Transform and optimized integrated features space, Chemom. Intell. Lab. Syst. 182 (2018) 21–30.

[41] Z.U. Khan, F. Ali, I.A. Khan, Y. Hussain, 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, Chemom. Intell. Lab. Syst. 189 (2019) 169–180.

[42] Z.U. Khan, D. Pi, S. Yao, A. Nawaz, F. Ali, S. Ali, piEnPred: a bi-layered discriminative model for enhancers and their subtypes via novel cascade multi-level subset feature selection algorithm, Front. Comp. Sci. 15 (6) (2021) 1–11.

[43] 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,* p. 100885, 2022.