IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 15 April 2023, accepted 6 May 2023, date of publication 9 May 2023, date of current version 24 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3274601

RESEARCH ARTICLE

Identifying Neuropeptides via Evolutionary and Sequential Based Multi-Perspective Descriptors by Incorporation With Ensemble Classification Strategy

SHAHID AKBAR[1], HEBA G. MOHAMED[2], HASHIM ALI[1], AAMIR SAEED[3], AFTAB AHMED KHAN[1], SARAH GUL[4], ASHFAQ AHMAD[2], FARMAN ALI[5], YAZEED YASIN GHADI[6], AND MUHAMMAD ASSAM[7]

[1]Department of Computer Science, Abdul Wali Khan University Mardan (AWKUM), Mardan, KP 23200, Pakistan
[2]Department of Computer Science, Muslim Youth University, Islamabad 44000, Pakistan
[3]Department of Computer Science and IT, University of Engineering and Technology at Peshawar, Peshawar 25000, Pakistan
[4]Department of Biological Sciences, FBAS, International Islamic University at Islamabad, Islamabad 44000, Pakistan
[5]Department of Software Engineering, University of Science and Technology, Bannu, KP 28100, Pakistan
[6]Department of Computer Science, Al Ain University, Abu Dhabi, United Arab Emirates
[7]Department of Electrical Engineering, College of Engineering, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia

Corresponding author: Heba G. Mohamed (hegmohamed@pnu.edu.sa)

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023TR140), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

ABSTRACT Neuropeptides (NPs) are a kind of neuromodulator/ neurotransmitter that works as signaling molecules in the central nervous system, and perform major roles in physiological and hormone regulation activities. Recently, machine learning-based therapeutic agents have gained the attention of researchers due to their high and reliable prediction results. However, the unsatisfactory performance of the existing predictors is due to their high execution cost and minimum predictive results. Therefore, the development of a reliable prediction is highly indispensable for scientists to effectively predict NPs. In this study, we presented an automatic and computationally effective model for identifying of NPs. The evolutionary information is formulated using a bigram position-specific scoring matrix (Bi-PSSM) and K-spaced bigram (KSB). Moreover, for noise reduction, a discrete wavelet transform (DWT) is utilized to form Bi-PSSM_DWT and KSB_DWT based high discriminative vectors. In addition, one-hot encoding is also employed to collect sequential features from peptide samples. Finally, a multi-perspective feature set of sequential and embedded evolutionary information is formed. The optimum features are chosen from the extracted features via Shapley Additive exPlanations (SHAP) by evaluating the contribution of the extracted features. The optimal features are trained via six classification models i.e., XGB, ETC, SVM, ADA, FKNN, and LGBM. The predicted labels of these learners are then provided to a genetic algorithm to form an ensemble classification approach. Hence, our model achieved a higher predictive accuracy of 94.47% and 92.55% using training sequences and independent sequences, respectively. Which is ~3% highest predictive accuracy than present methods. It is suggested that our presented tool will be beneficial and may execute a substantial role in drug development and research academia. The source code and all datasets are publicly available at https://github.com/shahidawkum/Target-ensC_NP.

INDEX TERMS Neuropeptides, ensemble classification, multi-perspective vector, discrete wavelet transform, SHAP analysis, bigram-position specific scoring matrix.

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

I. INTRODUCTION

Neuropeptides (NPs) are short peptides, usually less than 100 amino acids [1]. The 3D structures of neuropeptides are smaller and less complex than the proteins. Compared to traditional neurotransmitters, NPs have more receptor recognition sites [2]. As a result, the NPS is highly selective for a specific target and has minimal side effects. In the immune system NPs, act as neurotransmitters and behave like hormones in the endocrine system. NPs perform a key part in developmental processes and other biological activities [3]. Whenever the nervous system adapts to new challenges like stress, injury, and drug abuse NPs are especially important indicated by many studies [4], [5]. NPs induced neural activity as well as numerous other features of non-neuronal cells, i.e. social behavior, food uptake, and energy usage. Matured NPs are retained in closely packed vesicles and released under controlled conditions in response to a stimulus [6], [7]. Binding to a G protein-coupled receptor initiates a signaling pathway [8]. Prepropeptides are the precursor of NPs that undergoes alternative splicing and produce one or many bioactive peptides. To produce functionally active neuropeptides, neuropeptide precursors (NPPs) go through several regulated cleavages. Prominently, these cleavage sites are identified by a cluster of basic amino acids [9]. To treat a variety of neurological disorders these NPs have paved the way for the development of novel therapeutic strategies. In nematodes, approximately 250 neuropeptides have been identified [10]. Echinoderms and cicadas are two well-known sources of neuropeptides. The neurosecretory glands of cicadas are rich source of neuropeptides [11]. In insects, over 30 different families of NPs have been identified with their diverse roles and structure. The majority of these molecules have an impact on the insect's physiology. Though, the activities of these molecules are dependent on the age and the type of species [12].

A unique set of challenges were faced while designing drugs for Alzheimer's disease (AD) [13]. The key biological feature of AD is the altered amyloid-beta biochemistry which is considered one of the potential drug targets for AD. However, due to their low intrinsic toxicity, peptide-based drugs may be a viable option for treating the symptoms of a variety of AD diseases. Neurological conditions such as stroke, pain, brain tumors, psychiatric disorders, and neurodegeneration are treated with effective and precise peptide-based drugs [14]. Many methods have been developed to identify neuropeptides. Liquid chromatography-tandem mass spectrometry, genetic analysis, and receptor binding assay were the traditional approaches for identifying neuropeptides [15]. The experimental process is considered very costly and laborious. Due to the recent success of machine learning applications in drug development and discovery for the effective analysis of bioactive peptides. Therefore, considering its significance, several machine learning based computational models have been developed for the prediction of NPs [9]. Jiang et al. presented a stacking-based ensemble model called NeuroPpred-Fuse for predicting NPs [16]. The peptide samples were trained via six different sequential representation techniques. Further, to decrease the vector size of the hybrid features, a feature selection is also employed. The proposed model reported an accuracy of 90.60%. Similarly, Hasan et al. proposed a NeuroPred-FRL predictor for the prediction of NPs [17]. Whereas, the numerical descriptors were formulated from the peptide samples via evolutionary, physiochemical properties, and sequential descriptors. The formulated features were then passed through a 2-step feature selection to gather the best features. Finally, the formulated vectors were measured via the random forest. Subsequently, Kang et al. presented NeuroPP for discrimination of NPs [6]. NeuroPP used frequency-based extraction schemes namely, AAC, TPC, and DPC to represent training samples. Furthermore, the optimal features were selected via ANOVA-based feature selection. Moreover, Bin et al. used a Binary profile, Composition, and Physicochemical properties-based hybrid feature vector for the prediction of NPs [18]. The predictive results were examined via several hypothesis learners and then an ensemble learning algorithm is utilized to further improve the predictive results.

Additionally, the existing computational models were developed via conventional machine learning models to examine the predictive performance of the extracted descriptors. Though, these methods were not considered cost-effective and had low prediction performance. Therefore, to handle such situations, it is required to develop an automatic and computationally efficient predictor to correctly discriminate NPs and non-NPs. The evolutionary structure information was explored from the amino acid sequences using novel bigram-PSSM extended with discrete wavelet transform (Bi-PSSM_DWT), and k-spaced bigrams extended with discrete wavelet transform (KSB_DWT). Apart from the evolutionary descriptors, one-hot features were also formulated via (one-hot encoding). Furthermore, to develop a cost-effective model, we applied the SHapley Additive exPlanations (SHAP) approach to choosing optimal features from the multi-perspective hybrid vector of evolutionary features such as Bi-PSSM_DWT + KSB_DWT, and one-hot sequential features [19]. SHAP-bruta interprets the significance of each feature in a multi-perspective vector. To train and evaluate the model, various hypothesis learners such as ETC [20], SVM [21], [22], ADA [23], XGB [24], FKNN [25], and LGBM [26]. The predicted labels of these individual classifiers were provided to the genetic algorithm (GA) to develop an ensemble classifier [27] to improve the predictive outcomes of the model. The graphical abstract of our proposed model is illustrated in Figure 1.

II. MATERIALS AND METHODS

A. DATASET

To develop an automatic predictor, the selection of a valid training dataset is an essential step. To effectively examine the predictive analysis of our predictor, we used the same training samples that were previously presented in the PredNeuroP predictor [17], [18]. Initially, the dataset comprised 5948 laboratory-evaluated positive samples (NPs) from diverse taxa-derived NeuroPep databanks [1]. Among the whole sequences, the samples whose sequence lengths are higher than 100 residues and less than 5 residues are eradicated. Furthermore, on the remaining sequences, a CD-HIT tool was also employed to eradicate homologous peptide samples. Where the threshold value of 0.9 was used to remove peptide samples, whose similarities are greater than 90%. Hence, 2435 NPs are selected. On the other hand, a similar redundancy removal approach was used on negative samples (non-NPs), and finally, 2435 non-NPs are selected. Moreover, 80% of the whole dataset is used for the training dataset (NPs =1940 and non-NPs =1940) and the remaining 20% of peptide sequences are employed for testing. The overfitting and generalization ability of the proposed training model was measured using independent sequences. The independent dataset comprises 495 NPs and 495 non-NPs. Additionally, it was ensured that the sequences of the training data were not used in testing data.

feature set is generated by assigning "1" against matched residue and for unavailable amino acid residues '0' will be placed. The working procedure of one hot encoding using peptide sample "APLMGFQHVR" is graphically illustrated in Figure 2. In addition, to effectively train the machine learning models, a feature vector of equal dimension is required. Hence, the length of the peptide samples is organized by adding some dummy alphabets (Padding) [28]. Keeping the same procedure, the peptide sequences of the whole training dataset are represented in equal length. However, it was also investigated that adding these dummy alphabets has no biological or functional effect on a peptide sequence. In other words, padding allows us to generate a fixed-length input vector of a protein sample regardless of its original length. Additionally, padding can increase the efficiency of the training model by allowing the algorithm to process batches of sequences in parallel, rather than processing each sequence individually.

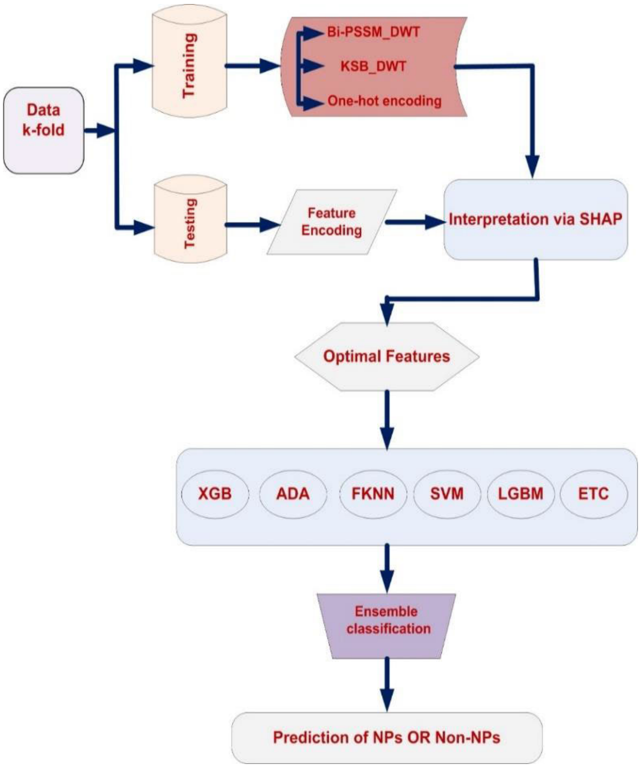
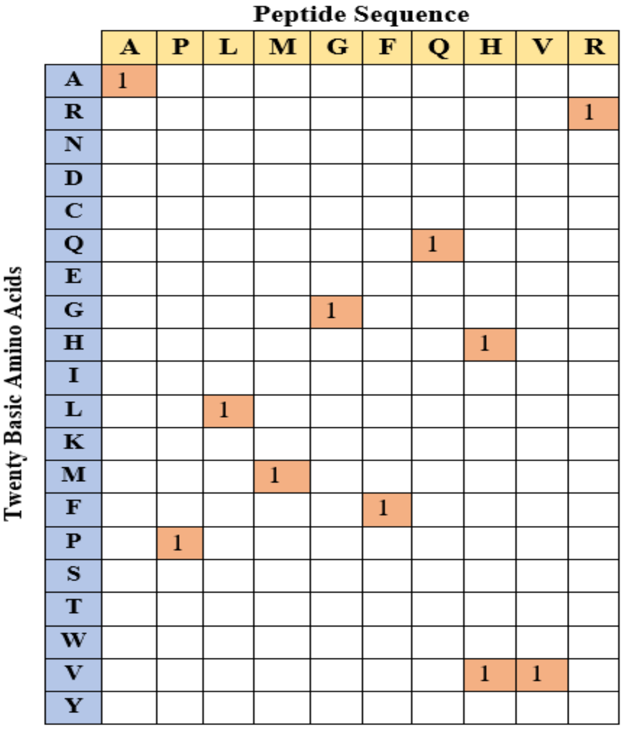


FIGURE 1. Proposed NPs predictive model.


Peptide Sequence
Twenty Basic Amino Acids

	A	P	L	M	G	F	Q	H	V	R
A	1									
R										1
N										
D										
C										
Q							1			
E										
G					1					
H								1		
I										
L			1							
K										
M				1						
F						1				
P		1								
S										
T										
W										
V								1	1	
Y										

FIGURE 2. One-hot encoding for peptide sequence 'APLMGFQHVR'.

B. FEATURE FORMULATION METHODS

1) ONE-HOT ENCODING

One hot encoding is a sparse formulation technique that has been extensively utilized to numerically represent the peptide sequences. Apart from the other techniques, one hot encoding represents binary features without affecting the sequence ordering of the amino acids in a peptide sample. Whereas, each residue of a peptide sequence is numerically converted to a feature vector having dimensions of 20 features. The final

2) K-SEPARATED BIGRAMS (KSB)

KSB was initially presented by Saini et al.which computes the association among those amino acid residues that are non-adjacent in a protein sample. The bigram probabilities are obtained from the sequential evolution probabilities in a PSSM Matrix [29]. Where k-spaced bigrams in a non-adjacent manner are separated by K amino acid residues in the sample, while k represents the positional distance among the bigrams [30]. The complete mechanism can be summarized via the following equation (1); Where R represents the PSSM matrix having H number of rows, H denotes the size of the amino acid sample in matrix R, and 20 columns mean 20 valid amino acids. The transition of the *pth* amino acid to the *qth* amino acid can be represented as follows:

$$T_{p,q}(K) = \sum_{i=1}^{H-K} R_{i,p}R_{i+K,q} \tag{1}$$

where, $1 \leq p \leq 20$, $1 \leq q \leq 20$, and $1 \leq k \leq K$

$$\begin{aligned} T(K) = [&T_{1,1}(K), \ldots\ldots T_{1,20}(K), \\ &\ldots, T_{20,1}(K), \ldots, T_{20,20}(K)] \end{aligned} \tag{2}$$

The above equation represents a matrix T(k) which consists of 400 features for amino acid transition for a single value of K. In the KSB formulation scheme, the protein samples are examined using different values of K, such as K = 1 2, and 3, as shown in Table 1. However, keeping the issue of computational cost and its highest predictive rates of the classifiers, we consider the feature vector using K=2.

TABLE 1. Prediction analysis of KSB formulation method using values of K.

Method	Classifier		ACC %	Sen %	Spe %	MCC	AUC
KSB	K=1	ETC	72.11	70.23	75.45	0.55	0.81
		FKNN	72.34	80.52	69.02	0.54	0.81
		Ada	68.02	67.23	74.76	0.51	0.74
		XGB	70.31	68.85	73.77	0.53	0.78
		LGBM	71.21	69.42	76.98	0.54	0.79
		SVM	71.98	70.38	75.93	0.55	0.8
	K=2	ETC	80.55	72.11	82.64	0.63	0.87
		FKNN	76.93	81.83	73.23	0.6	0.84
		Ada	70.11	69.09	76.41	0.55	0.81
		XGB	79.31	74.33	81.31	0.62	0.86
		LGBM	78.36	82.96	72.54	0.61	0.83
		SVM	75.53	80.33	73.83	0.59	0.82
	K=3	ETC	76.51	79.44	73.65	0.58	0.82
		FKNN	71.33	78.53	70.23	0.56	0.78
		Ada	70.03	67.39	74.81	0.55	0.78
		XGB	78.11	74.34	86.23	0.61	0.82
		LGBM	77.33	75.93	81.79	0.6	0.81
		SVM	74.23	72.44	81.84	0.59	0.81

3) BIGRAM POSITION-SPECIFIC SCORING MATRIX (BI-PSSM)

Bi-PSSM examines the evolutionary feature extraction strategy that computes the intrinsic pattern of the protein samples via different alignments of various protein families [31]. Along a protein sequence, Bi-PSSM replaces the occurring frequencies of the amino acid residues at a particular position [22], [32]. The resultant vector of a Bi-PSSM matrix contains the negative and positive scores of the amino acid residues. The negative scoring value shows the low occurring frequency of the amino acids and the positive value represents the high frequent occurrence of the amino acid substitution in an alignment [33]. The resultant PSSM feature space "K" can be shown as follows:

$$K = \begin{bmatrix} k_{1,1} & k_{1,2} & \cdots & k_{1,20} \\ k_{2,1} & k_{2,2} & \cdots & k_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ k_{L,1} & k_{L,2} & \cdots & k_{L,20} \end{bmatrix}_{20\times L} \tag{3}$$

where $k_{i,j}$ denotes the i^{th} residue of the j^{th} amino acid along a sequence. L is the length of a biological sample and twenty are the number of amino acids in a protein.

4) DISCRETE WAVELET TRANSFORM (DWT)

DWT is a transformation filter-based noise compression and elimination approach. DWT divides an input signal of the protein into two sub-parts namely wavelets [34]. While, one wavelet contains high-frequency coefficients namely detailed coefficients, and 2nd wavelet keeps low-frequency coefficients called the approximation coefficients [35]. Moreover, it is also observed from recent studies, that low-frequency wavelet is more informative than the high-frequency wavelet. Hence, to extract highly effective information, the low-frequency wavelet is further divided into several levels as given in Figure 3. Where H_F represents high-frequency coefficients, and L_F are the low-frequency coefficients. In DWT, the input is divided into several scales (levels). Whereas, the detailed coefficients and approximation coefficients of a signal can be represented by 2^k, where k denotes the number of decomposed levels. DWT can be formulated as follows:

$$B(s,t) = \sqrt{\frac{1}{s}\int_{0}^{a} z(a)\psi\left(\frac{a-t}{s}\right)d_a} \tag{4}$$

where $z(a)$ denotes the input signal, $B(s, t)$ show the transform values/ coefficients for the specific position on the wavelet periods and signal. $\psi(\frac{a-t}{s})$ represents the wavelet function, while s is the scaling variable and t denotes the translation variable.

The detailed coefficients and approximation coefficients for a signal $z(a)$ can be formulated as:

$$W_{i,L}[x] = \sum_{K=1}^{N} c[m]S[2x-m] \tag{5}$$

$$W_{i,H}[x] = \sum_{K=1}^{N} c[m]R[2x-m] \tag{6}$$

where $c[m]$, S, and R represent the input signal of the peptide sequence, low pass filter, and high pass filter, respectively. $W_{i,L}[x]$, and $W_{i,H}[x]$ denotes the detailed coefficient, and approximation coefficient of input samples, respectively.

In this computational model, the Bi-PSSM features are transformed using DWT for signal de-noising. We evaluated

DWT up to 5 levels and generated 260, 520, 780, 1040, and 1300 feature vectors for level-1, level-2, level-3, level-4, and level-5, respectively. The predictive results of the model were examined via five different level of features. Among which level-3 features shows produce significant results. The features of level-1 and level-2 achieved lower performance due to less informative patterns as compared to level-3. Similarly, level-4 and level-5 also depicted lower results due to redundant motifs that impair the model performance. Therefore, we select DWT upto level-3 to form a novel feature extraction approach called Bi-PSSM-DWT. It was also observed that further decomposition of DWT levels leads to similar and redundant features which may affect the model performance.

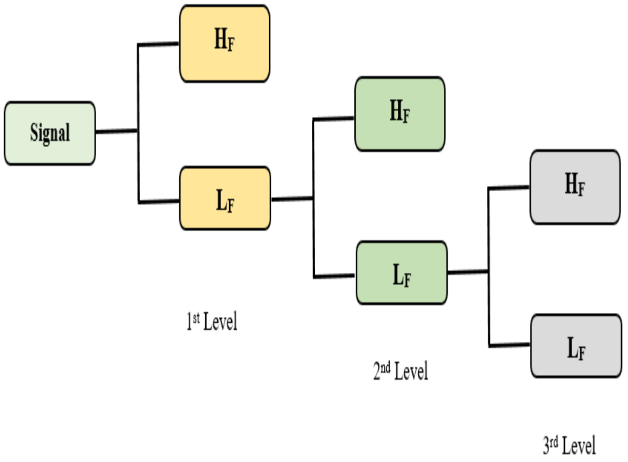


FIGURE 3. Three levels decomposition of DWT.

C. SHAP FEATURE SELECTION

Undoubtedly, determining the biological importance of the formulated numerical descriptors is not an easy task. The classification methods are also called "black boxes" owing to their intricate internal mechanisms. To comprehend and recognize the significance of individual features of the extracted space is a challenging task for a machine learning model [36]. Shapley Additive exPlanations (SHAP) interpretation is a global technique to evaluate the significance of each numerical feature based on aggregations of SHAP values [37]. The interpretable evaluation using a classification model also deals the issues that occur due to the lack of feature directivity [38]. In this paper, the predictive results of the XGB model are higher as compared to other models [39]. The procedure for selecting the optimal features via the SHAP algorithm can be described as follows:

Initially, we select an objective function 'K', then the Shapley value 'δ_D' of each extracted descriptor D ∈ F was computed. Finally, only high-ranked features 'R' were selected, where R<d = |F|. The resultant Boruta-SHAP plot showing the high-ranked features was summarized in Figure 4. Where each row denotes the ranked feature and each point corresponds to the SHAP value of each instance. The red point indicates the high-ranked features, while the bluer points indicate the smaller value of the feature; the abscissae represent the SHAP values. Keeping the same procedure, the entire visualization of the entire model was measured via SHAP interpolation. The positive SHAP-value of a feature predicts that features diving towards NPs class and the negative SHAP-value represents the prediction to the non-NPs class.

D. ENSEMBLE LEARNING

Ensemble learner is an optimized classification algorithm that has been extensively employed for computing and predicting biological sequences via machine learning and deep learning because of its high generalization abilities and prediction results. The key objective of ensemble learning is to concatenate the predicted labels of the individual classification algorithms to develop an ensemble learning model that can enhance the predicted outcomes of a classifier with a minimal error rate. Instead of traditional classifiers, ensemble classification is highly reliable due to its minimum variance happens due to erroneous results of classical machine learners. Therefore, different ensemble learning models have been utilized for the prediction of different biological types i.e., anticancer peptide [40], subcellular localization [41], antifreeze proteins [42], antiviral peptides [34], Recombination spots [27], antifungal peptides [35], nucleosome positioning [43], and enhancer functions [44]. Consequently, we developed a genetic algorithm (GA) based ensemble learning model to further examine the predicted labels of the individual classifiers obtained via different extracted feature vectors. It is a heuristic learning approach that has been effectively applied in the bioinformatics area to solve different prediction problems with significant predictive results [45]. In a GA-based ensemble model that randomly chooses a specific population from the whole chromosomes and then different operators of genetic algorithm are employed to obtain the best performance results [46], [47].

At first, we calculated the prediction labels of five different traditional machine learning models, such as ETC, LGBM, SVM, XGB, and ADA. Then all the predicted labels are then provided to GA to develop an ensemble model as follows:

$$EnC_i = ADA \oplus LGBM \oplus SVM \oplus ETC \oplus XGB \tag{7}$$

In eq. (6), EnC_i denotes the proposed ensemble model, and ⊕ signifies the fusing operator utilized to combine the predicted labels of the single learner. The EnC_i model using different classification learners can be formulated as follows:

Let us consider a machine learning model 'ML' for a protein sample 'R' is:

$$\{ML_1, ML_2, ML_3, ML_4, ML_5\} \in \{C_1, C_2\} \tag{8}$$

where $ML_1, ML_2, ML_3, ML_4, ML_5$ represents the individual classifier and C_1, C_2 denotes the predicted classes

(NPs, non-NPs).

$$E_i = \sum_{i=1}^{5} \delta(ML_i C_r) \quad where \;\; r = 1, 2 \tag{9}$$

$$\delta(ML_i, C_r) = \left\{ \begin{matrix} 1, & if \; ML_i \in C_r \\ 0, & otherwise \end{matrix} \right\} \tag{10}$$

Finally, the predictive results $\mathrm{EnC_i}$ using GA are measured as:

$$GA\_EnC_i = Max(W_1E_1, W_2E_2, W_3E_3, W_4E_4, W_5E_5) \tag{11}$$

where $\mathrm{GA\_EnC_i}$ shows the GA-based ensemble classification model, 'Max' denotes the higher predictive result, and $W_1, \ldots, W_5$ represents the optimal weight adjustment for an individual classifier.

E. FRAMEWORK OF THE PROPOSED MODEL

In this study, we presented an ensemble learning-based prediction model for the prediction of neuropeptides. Initially, the training sequences are formulated using one-hot encoding-based sequential features. In addition to the sequential features, Bi-PSSM_DWT, and KSB_DWT are also applied to extract embedded evolutionary features from the peptide sequences. Additionally, we used a Multi-perspective Descriptors strategy, by combining the feature vectors of the aforementioned methods. The Multi-perspective vector consists of 1140 features, representing 20 features of one hot encoding, 720 features of Bi-PSSM_DWT, and 400 features of KSB_DWT. The training cost of the hybrid vector will be high due to the feature dimension of the training vector. Therefore, to reduce the computational cost of the proposed approach, we applied XGB-based SHAP feature selection to obtain 158 optimal features from the whole vector. In the next phase, six different machine-learning models are trained using the extracted features. While training the models, a train and split ratio of 80:20 is used to divide the training dataset. Moreover, to form an ensemble learning model, the predicted labels of the individual classifiers are provided to the genetic algorithm to boost the prediction results. Our predictor reported the highest prediction results than existing models using the training as well as the independent dataset. The framework of our proposed model is illustrated in Figure 1.

TABLE 2. Optimal parameters of genetic algorithm for ensemble classifier.

Parameter	Optimal Values
Population size	100
Population Type	Bit string
Number of Iteration	50
Selection	Tournament
Mutation function	Uniform Distribution
Crossover function	Scattered
Crossover Fraction	0.7
Elite count	2
Plot	Best Fitness

F. PREDICTION MEASUREMENT PARAMETERS

In learning algorithms, several parameters are utilized to evaluate the prediction abilities of a learning model [48]. Where the evaluation parameters determine a prediction model and whether the required objectives of a research problem are effectively addressed or not [22], [25]. Therefore, various prediction metrics have been applied in the literature to calculate the predictive results of a machine-learning model [49], [50], [51]. However, collecting the optimal parameters highly depends on the distribution of samples in a classifier. In order to compute the performance rates, initially, a confusion matrix is generated. Where a table is maintained by keeping the actual labels of a problem and its predicted labels. By properly maintaining the confusion matrix, the predictive results of our study are examined via the following evaluation parameters.

$$Accuracy = 1 - \frac{NP_{-}^{+} + NP_{+}^{-}}{NP^{+} + NP^{-}} \tag{12}$$

$$Sensitivity = 1 - \frac{NP_{-}^{+}}{NP^{+}} \tag{13}$$

$$Specificity = 1 - \frac{NP_{+}^{-}}{NP^{-}} \tag{14}$$

$$Mcc = \frac{1 - \left(\frac{NP_{-}^{+}+NP_{+}^{-}}{NP^{+}+NP^{-}}\right)}{\sqrt{\left(1 + \frac{NP_{+}^{-}+NP_{-}^{+}}{NP^{+}}\right)\left(1 + \frac{NP_{-}^{+}+NP_{+}^{-}}{NP^{-}}\right)}} \tag{15}$$

where NP^{+} denotes the positive sequences and NP^{-} denotes non-negative sequences. Similarly, NP_{+}^{-} shows the false-negative predictions, and NP_{-}^{+}, represents an error of false positive, the model falsely determines the true instances as false.

III. RESULTS AND DISCUSSIONS

In this study, the predictive rates of the extracted vectors are examined via a k-fold CV test. Where the 10-fold CV is applied to randomly divide the features into 10 folds of equal size [49]. Among the 10 folds, 9 folds are employed for model training and the samples of one fold are kept for model testing. Additionally, a Stratified looping method is also used by randomly splitting training 100 times, and then the mean results are calculated to achieve reliable outcomes. In the below subsections, the predicted outcomes of the numerically formulated vectors using training samples and test samples using various classification models.

TABLE 3. Prediction rates of training dataset via hybrid descriptors.

Method	Classifier	ACC%	Sen%	Spe%	MCC	AUC
KSB_DWT	ETC	83.09	83.21	82.99	0.66	0.92
	FKNN	84.02	91.66	75.54	0.68	0.90
	Ada	80.02	79.03	80.96	0.60	0.87
	XGB	82.62	80.69	84.58	0.65	0.91
	LGBM	83.19	82.32	84.06	0.66	0.91
	SVM	82.80	81.24	84.37	0.66	0.90
	Ensemble-GA	91.08	92.36	90.81	0.83	0.92
Bi-PSSM_DWT	ETC	85.35	86.34	84.32	0.71	0.93
	FKNN	80.02	85.29	74.18	0.60	0.87
	Ada	79.25	78.41	80.10	0.59	0.87
	XGB	82.88	82.48	83.29	0.66	0.91
	LGBM	84.02	84.63	83.40	0.68	0.92
	SVM	82.37	82.26	82.47	0.65	0.90
	Ensemble-GA	91.93	92.46	91.40	0.84	0.93
One-hot	ETC	85.23	86.21	83.82	0.72	0.92
	FKNN	76.80	81.61	71.47	0.54	0.92
	Ada	81.88	82.06	81.70	0.64	0.89
	XGB	83.11	85.10	81.13	0.66	0.91
	LGBM	84.64	85.67	83.61	0.69	0.92
	SVM	84.94	84.58	85.31	0.70	0.92
	Ensemble-GA	89.56	91.12	87.99	0.79	0.92
Hybrid Vector	ETC	86.56	89.07	84.02	0.74	0.94
	FKNN	81.02	88.50	72.74	0.67	0.90
	Ada	82.92	82.63	83.01	0.64	0.89
	XGB	84.09	84.41	83.76	0.68	0.91
	LGBM	85.15	85.99	84.05	0.71	0.92
	SVM	84.47	83.67	85.97	0.68	0.92
	Ensemble-GA	92.36	93.22	91.89	0.84	0.94

A. PARAMETER SETTING OF GENETIC ALGORITHM

In GA, choosing the best parameters is a crucial step that leads to achieving the maximum predictive outcomes of a machine learning problem. Initially, the chromosomes of GA were represented in bit-string form. A population of "80" was randomly chosen from the whole size to obtain the best results. Whereas the high population may boost predictive rates, it can directly increase the execution time of the model. Additionally, a tournament-selection approach was used to pick potential parents from the existing population. For the production of off-springs, we used a rank scaling parameter. Additionally, an intermediate crossover method is used having a value of 0.7 and a uniform distribution is selected to mutate the genetic diversity for the next generations. Hence, an improved GA model was ended with the optimal parameters. The complete details of the optimal parameters selection are provided in Table 2. After applying these parameters higher predictive rates of the training model are achieved as given in Figure 5.

B. RESULTS ANALYSIS OF CLASSIFICATION MODELS BEFORE FEATURE SELECTION

The predictive outcomes of the learning models using formulated feature spaces are given in Table 3. The performance of all feature vectors is examined by each learning by computing its predictive accuracy (ACC), sensitivity (Sen), specificity (Spe), AUC, and MCC. ETC with KSB_DWT descriptor obtained an accuracy of 83.09%, a sensitivity of 83.21%, a specificity of 82.99%, an MCC of 0.66%, and an AUC of 0.92. FKNN achieved 84.02% accuracy, 91.66% sensitivity, 75.54% specificity, 0.68 MCC, and 0.90 AUC. Compare with ETC, FKNN boosted the accuracy, sensitivity, AUC, and MCC, while the specificity of the model reduced. Ada classifier achieved lower performance than ETC and FKNN. XGB showed better performance than Ada and achieved 82.62% accuracy, 80.69% specificity, 84.58% specificity, 0.65 MCC, and 0.91 AUC. The LXGB achieved 83.19% accuracy while SVM decreased the performance. The best results were reflected by ensemble learning and attained 91.08% ACC, 92.36% Sen, 90.81%, Spe, 0.83 MCC, and 0.92 AUC. On the other hand, Bi-PSSM_DWT descriptors, showed good performance while some classifiers decreased the prediction results. The accuracies secured by ETC, FKNN, Ada, XGB, LXGB, SVM, and Ensemble-GA are 85.35%, 80.02%, 79.25%, 82.88%, 84.02%, 82.37%, and 91.93%, respectively. Compared with the KSB_DWT descriptor, ETC, LXGB, and Ensemble-GA with Bi-PSSM_DWT improved the results, while the remaining classifiers reduced the performance. Further analyzing the performance of classifiers over the one-hot sequential features, ETC consistently shows better performance than other classifiers. The accuracy of ETC is 85.23% while FKNN is 76.80%. Similarly, Ada obtained an accuracy of 81.88%, while 83.11% is secured by XGB. Among all classifiers, Ensemble-GA attained the best accuracy with 89.56%. Normally, single descriptor features are less informative. To boost the performance of a model, features are fused to gain a multi-perspective feature vector. On the hybrid feature set, all classifiers enhanced the performance and attained 86.56%, 81.02%, 82.92%, 84.09%, 85.15%, 84.47%, and 92.36% accuracies by ETC, FKNN, Ada, XGB, LGBM, SVM, and Ensemble-GA, respectively. These results verified that fused features can enhance the model's performance.

C. RESULTS ANALYSIS OF CLASSIFIERS AFTER THE SHAP FEATURE SELECTION APPROACH

Previous models have proved that learning models increase performance outcomes via an optimized feature vector [52], [53], [54]. In this regard, the SHAP feature selection approach is implemented to choose the optimal feature space and summarized the results in Table 4. The optimal feature vector is then evaluated by ETC, FKNN, Ada, XGB, LGBM, SVM, and Ensemble-GA classifiers using the 10-fold CV. ETC yielded an accuracy of 86.34%, a sensitivity of 87.93%, a specificity of 84.74%, an MCC of 0.73, and an AUC of 0.94. Compared with the ETC classifier, the performance outcomes of FKNN and Ada are not satisfactory, while XGB reflected better performance. Onward, LGBM achieved 86.08% accuracy and SVM attained 83.17% accuracy. On the other hand, our proposed ensemble learning model achieved the highest prediction results having an accuracy of 94.47%, a sensitivity of 97.32%, a specificity of 93.81%, an MCC of 0.91, and an AUC of 0.97. These results confirm that the Ensemble strategy can discriminate NPs more accurately.

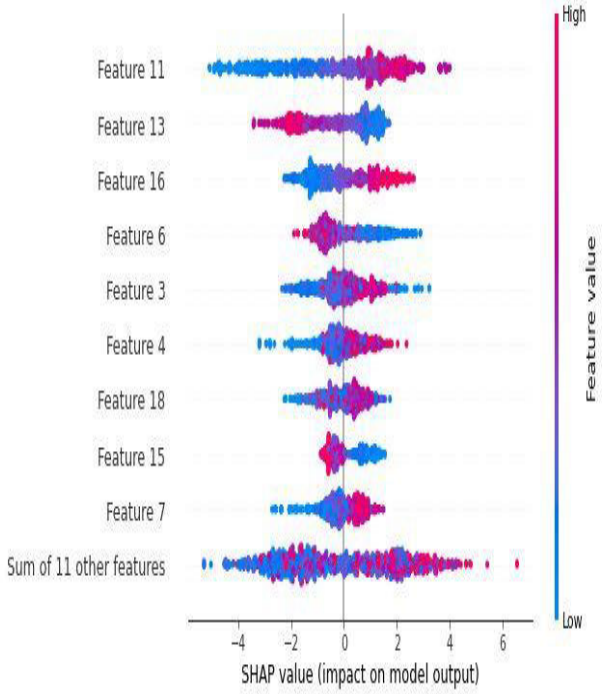


FIGURE 4. SHAP interpolation using hybrid descriptors.

TABLE 4. Prediction rates of hybrid vector using SHAP feature selection.

Classifier	ACC%	Sen%	Spe%	MCC	AUC
ETC	86.34	87.93	84.74	0.73	0.94
FKNN	84.41	91.42	76.63	0.69	0.91
ADA	83.14	82.98	83.29	0.66	0.90
XGB	85.16	85.61	84.53	0.70	0.93
LGBM	86.08	86.96	85.20	0.72	0.93
SVM	83.17	81.95	84.38	0.66	0.91
Ensemble-GA	**94.47**	**97.32**	**93.81**	**0.91**	**0.98**

D. PREDICTION COMPARISON OF OUR MODEL WORK WITH PRESENT STUDIES

The predictive outcomes of our predictor and its comparison with the existing models via training and independent set are given in Table 5. Our proposed model via training sequence achieved an accuracy of 94.47%, a sensitivity of 97.32%, a specificity of 93.81%, an AUC of 0.98%, and an MCC of 0.91, respectively via training sequences. Our model reported significant results by improving the accuracy by 2.57%, sensitivity by 7.82%, MCC by 0.07%, and AUC by 0.02 then NeuroPred-FRL [17]. A prediction model is said to be reliable if it has a high generalization power for unseen data (independent data). In this

TABLE 5. Comparison of proposed study with existing methods.

Dataset	Model	ACC %	Sen %	Spe %	MCC
Training Sequence	NeuroPred-FRL[17]	91.90	89.50	94.30	0.84
	Proposed Model	**94.47**	**97.32**	**93.81**	**0.91**
Independent Sequences	NeuroPred-FRL [17]	91.10	92.70	89.50	0.82
	NeuroPpred-Fuse[16]	90.60	88.20	93.00	0.81
	NeuroPIpred [9]	53.60	33.10	73.60	0.74
	PredNeuroP [18]	89.70	88.60	90.70	0.79
	Proposed Model	**92.55**	**93.84**	**91.23**	**0.87**

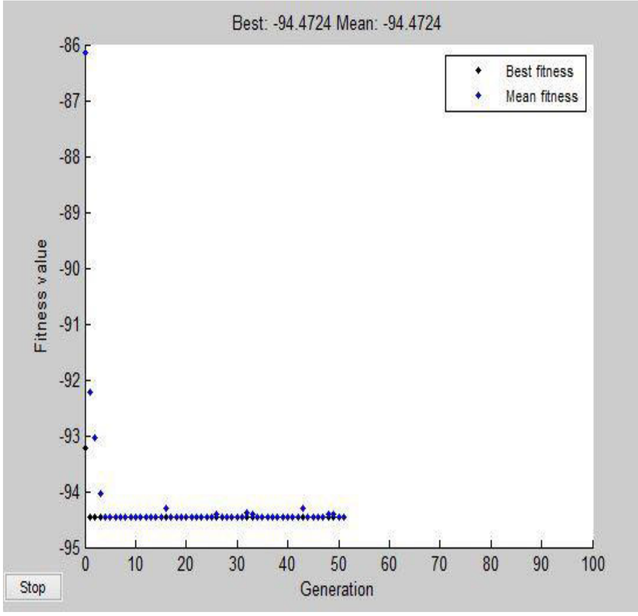


FIGURE 5. Performance of GA via training dataset.

regard, we used an independent dataset to validate the effectiveness of the proposed study, the detailed predictive result of the independent set is illustrated in Table 5. It shows that our proposed predictor obtained the highest outcomes than existing approaches. Our model improved by ~2.15% accuracy, ~5.64% sensitivity, ~ 6% MCC, and ~0.02 % AUC than NeuroPred-FRL [17]. Similarly, our model boosted 2.15% accuracy, 5.64% sensitivity, 6% MCC, and 0.01% AUC than NeuroPpred-Fuse [16]. The current study surpassed all other existing tools on all evaluation parameters. The achieved outcomes demonstrate the efficiency of the proposed study.

IV. CONCLUSION AND FUTURE INSIGHTS

NPs play critical roles in a variety of biological processes and the pharmacological industry. In this study, a successful attempt has been performed for the accurate prediction of NPs using GA-based Ensemble learner and SHAP interpretation for the selection of optimal features from the heterogeneous feature set. The proposed approach reported remarkable predictive rates over the existing machine learning models applied for the prediction of NPs. The improved results of our predictive model are due to various reasons i.e. the suitable sequence formulation technique, selection of optimal descriptors using novel Shap analysis, and an effective model training algorithm. The achieved results confirm that our proposed model will be effectively performed a key role in identifying NPs in drug development due to their superior discriminative and generalization abilities. In our future model, we will try to establish a publically available web server for the proposed work and make further efforts to develop more capable approaches, such as feature selection or advanced deep neural networks to further improve the predictive results of NPs.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGMENT

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023TR140), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

REFERENCES

[1] Y. Wang, M. Wang, S. Yin, R. Jang, J. Wang, Z. Xue, and T. Xu, "NeuroPep: A comprehensive resource of neuropeptides," *Database*, vol. 2015, Apr. 2015, Art. no. bav038.

[2] T. Hökfelt, T. Bartfai, and F. Bloom, "Neuropeptides: Opportunities for drug discovery," *Lancet Neurol.*, vol. 2, no. 8, pp. 463–472, Aug. 2003.

[3] D. R. Nässel and M. Zandawala, "Recent advances in neuropeptide signaling in drosophila, from genes to physiology and behavior," *Prog. Neurobiol.*, vol. 179, Aug. 2019, Art. no. 101607.

[4] L. Carniglia, D. Ramírez, D. Durand, J. Saba, J. Turati, C. Caruso, T. N. Scimonelli, and M. Lasaga, "Neuropeptides and microglial activation in inflammation, pain, and neurodegenerative diseases," *Mediators Inflammation*, vol. 2017, pp. 1–23, Jan. 2017.

[5] J. Gonçalves, J. Martins, S. Baptista, A. F. Ambrósio, and A. P. Silva, "Effects of drugs of abuse on the central neuropeptide y system," *Addiction Biol.*, vol. 21, no. 4, pp. 755–765, Jul. 2016.

[6] J. Kang, Y. Fang, P. Yao, N. Li, Q. Tang, and J. Huang, "NeuroPP: A tool for the prediction of neuropeptide precursors based on optimal sequence composition," *Interdiscipl. Sci., Comput. Life Sci.*, vol. 11, no. 1, pp. 108–114, Mar. 2019.

[7] L. Funkelstein, M. Beinfeld, A. Minokadeh, J. Zadina, and V. Hook, "Unique biological function of cathepsin L in secretory vesicles for biosynthesis of neuropeptides," *Neuropeptides*, vol. 44, no. 6, pp. 457–466, Dec. 2010.

[8] G. Jékely, "Global view of the evolution and diversity of metazoan neuropeptide signaling," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 21, pp. 8702–8707, May 2013.

[9] P. Agrawal, S. Kumar, A. Singh, G. P. S. Raghava, and I. K. Singh, "NeuroPIpred: A tool to predict, design and scan insect neuropeptides," *Sci. Rep.*, vol. 9, no. 1, pp. 1–12, Mar. 2019.

[10] J. J. Knickelbine, C. J. Konop, I. R. Viola, C. B. Rogers, L. A. Messinger, M. M. Vestling, and A. O. W. Stretton, "Different bioactive neuropeptides are expressed in two sub-classes of GABAergic RME nerve ring motorneurons in *Ascaris suum*," *ACS Chem. Neurosci.*, vol. 9, no. 8, pp. 2025–2040, Aug. 2018.

[11] C. E. Jones, C. B. Otara, N. D. Younan, J. H. Viles, and M. R. Elphick, "Bioactivity and structural properties of chimeric analogs of the starfish SALMFamide neuropeptides S1 and S2," *Biochimica Biophysica Acta (BBA)-Proteins Proteomics*, vol. 1844, no. 10, pp. 1842–1850, Oct. 2014.

[12] S. Chowa, J. Lubawy, A. Urba, and G. Rosi, "Cardioregulatory functions of neuropeptides and peptide hormones in insects," *Protein Peptide Lett.*, vol. 23, no. 10, pp. 913–931, Sep. 2016.

[13] B. Jobke, T. McBride, L. Nevin, L. Peiperl, A. Ross, C. Stone, and R. Turner, "Setbacks in Alzheimer research demand new strategies, not surrender," *PLOS Med.*, vol. 15, no. 2, Feb. 2018, Art. no. e1002518.

[14] C. S. Mocanu, M. Niculaua, G. Zbancioc, V. Mangalagiu, and G. Drochioiu, "Novel design of neuropeptide-based drugs with β-sheet breaking potential in amyloid-beta cascade: Molecular and structural deciphers," *Int. J. Mol. Sci.*, vol. 23, no. 5, p. 2857, Mar. 2022.

[15] K. Boonen, B. Landuyt, G. Baggerman, S. J. Husson, J. Huybrechts, and L. Schoofs, "Peptidomics: The integrated approach of MS, hyphenated techniques and bioinformatics for neuropeptide analysis," *J. Separat. Sci.*, vol. 31, no. 3, pp. 427–445, Mar. 2008.

[16] M. Jiang, B. Zhao, S. Luo, Q. Wang, Y. Chu, T. Chen, X. Mao, Y. Liu, Y. Wang, X. Jiang, D.-Q. Wei, and Y. Xiong, "NeuroPpred-Fuse: An interpretable stacking model for prediction of neuropeptides by fusing sequence information and feature selection methods," *Briefings Bioinf.*, vol. 22, no. 6, Nov. 2021, Art. no. bbab310.

[17] M. M. Hasan, M. A. Alam, W. Shoombuatong, H.-W. Deng, B. Manavalan, and H. Kurata, "NeuroPred-FRL: An interpretable prediction model for identifying neuropeptide using feature representation learning," *Briefings Bioinf.*, vol. 22, no. 6, Nov. 2021, Art. no. bbab167.

[18] Y. Bin, W. Zhang, W. Tang, R. Dai, M. Li, Q. Zhu, and J. Xia, "Prediction of neuropeptides from sequence information using ensemble classifier and hybrid features," *J. Proteome Res.*, vol. 19, no. 9, pp. 3732–3740, Sep. 2020.

[19] J. Chen, S. Yuan, D. Lv, and Y. Xiang, "A novel self-learning feature selection approach based on feature attributions," *Expert Syst. Appl.*, vol. 183, Nov. 2021, Art. no. 115219.

[20] L. Abhishek, "Optical character recognition using ensemble of SVM, MLP and extra trees classifier," in *Proc. Int. Conf. Emerg. Technol. (INCET)*, Jun. 2020, pp. 1–4.

[21] 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.

[22] 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," *J. Comput.-Aided Mol. Des.*, vol. 33, no. 7, pp. 645–658, Jul. 2019.

[23] L. Dou, X. Li, L. Zhang, H. Xiang, and L. Xu, "IGlu_AdaBoost: Identification of lysine glutarylation using the AdaBoost classifier," *J. Proteome Res.*, vol. 20, no. 1, pp. 191–201, Jan. 2021.

[24] N. Wang, M. Zeng, Y. Li, F.-X. Wu, and M. Li, "Essential protein prediction based on node2vec and XGBoost," *J. Comput. Biol.*, vol. 28, no. 7, pp. 687–700, Jul. 2021.

[25] 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 Comput. Sci.*, vol. 14, no. 2, pp. 451–460, Apr. 2020.

[26] L. Deng, J. Pan, X. Xu, W. Yang, C. Liu, and H. Liu, "PDRLGB: Precise DNA-binding residue prediction using a light gradient boosting machine," *BMC Bioinf.*, vol. 19, no. S19, pp. 135–145, Dec. 2018.

[27] M. Kabir and M. Hayat, "IRSpot-GAEnsC: Identifing recombination spots via ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples," *Mol. Genet. Genomics*, vol. 291, no. 1, pp. 285–296, Feb. 2016.

[28] S. D. Ali, H. Tayara, and K. T. Chong, "Identification of piRNA disease associations using deep learning," *Comput. Struct. Biotechnol. J.*, vol. 20, pp. 1208–1217, Jan. 2022.

[29] J. Wang, B. Yang, J. Revote, A. Leier, T. T. Marquez-Lago, G. Webb, J. Song, K.-C. Chou, and T. Lithgow, "POSSUM: A bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles," *Bioinformatics*, vol. 33, no. 17, pp. 2756–2758, Sep. 2017.

[30] D. Sun, Z. Liu, X. Mao, Z. Yang, C. Ji, and Y. Liu, "ANOX: Predicting the antioxidant proteins based on multi-source heterogeneous features," *Anal. Biochem.*, vol. 2021, Jan. 2021, Art. no. 114257.

[31] M. Waris, K. Ahmad, M. Kabir, and M. Hayat, "Identification of DNA binding proteins using evolutionary profiles position specific scoring matrix," *Neurocomputing*, vol. 199, pp. 154–162, Jul. 2016.

[32] J. Luo, L. Yu, Y. Guo, and M. Li, "Functional classification of secreted proteins by position specific scoring matrix and auto covariance," *Chemometric Intell. Lab. Syst.*, vol. 110, no. 1, pp. 163–167, Jan. 2012.

[33] M. Kabir, M. Arif, F. Ali, S. Ahmad, Z. N. K. Swati, and D.-J. Yu, "Prediction of membrane protein types by exploring local discriminative information from evolutionary profiles," *Anal. Biochem.*, vols. 564–565, pp. 123–132, Jan. 2019.

[34] S. Akbar, F. Ali, M. Hayat, A. Ahmad, S. Khan, and S. Gul, "Prediction of antiviral peptides using transform evolutionary & SHAP analysis based descriptors by incorporation with ensemble learning strategy," *Chemometric Intell. Lab. Syst.*, vol. 230, Nov. 2022, Art. no. 104682.

[35] 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," *Chemometric Intell. Lab. Syst.*, vol. 222, Mar. 2022, Art. no. 104516.

[36] T. Chen, X. Wang, Y. Chu, Y. Wang, M. Jiang, D.-Q. Wei, and Y. Xiong, "T4SE-XGB: Interpretable sequence-based prediction of type IV secreted effectors using eXtreme gradient boosting algorithm," *Frontiers Microbiol.*, vol. 11, p. 2228, Sep. 2020.

[37] C. S. Kumar, M. N. S. Choudary, V. B. Bommineni, G. Tarun, and T. Anjali, "Dimensionality reduction based on SHAP analysis: A simple and trustworthy approach," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, Jul. 2020, pp. 558–560.

[38] D. Fryer, I. Strumke, and H. Nguyen, "Shapley values for feature selection: The good, the bad, and the axioms," *IEEE Access*, vol. 9, pp. 144352–144360, 2021.

[39] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, and S.-I. Lee, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature Biomed. Eng.*, vol. 2, no. 10, pp. 749–760, Oct. 2018.

[40] 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," *Artif. Intell. Med.*, vol. 79, pp. 62–70, Jun. 2017.

[41] W. Wattanapornprom, C. Thammarongtham, A. Hongsthong, and S. Lertampaiporn, "Ensemble of multiple classifiers for multilabel classification of plant protein subcellular localization," *Life*, vol. 11, no. 4, p. 293, Mar. 2021.

[42] X. Xiao, M. Hui, and Z. Liu, "IAFP-ense: An ensemble classifier for identifying antifreeze protein by incorporating grey model and PSSM into PseAAC," *J. Membrane Biol.*, vol. 249, no. 6, pp. 845–854, Dec. 2016.

[43] M. Tahir, M. Hayat, and S. A. Khan, "iNuc-ext-PseTNC: An efficient ensemble model for identification of nucleosome positioning by extending the concept of Chou's PseAAC to pseudo-tri-nucleotide composition," *Mol. Genet. Genomics*, vol. 294, no. 1, pp. 199–210, Feb. 2019.

[44] 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, Nov. 2018.

[45] B. Chowdhury and G. Garai, "A review on multiple sequence alignment from the perspective of genetic algorithm," *Genomics*, vol. 109, nos. 5–6, pp. 419–431, Oct. 2017.

[46] J. Lin, H. Chen, S. Li, Y. Liu, X. Li, and B. Yu, "Accurate prediction of potential druggable proteins based on genetic algorithm and bagging-SVM ensemble classifier," *Artif. Intell. Med.*, vol. 98, pp. 35–47, Jul. 2019.

[47] 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," *Comput. Biol. Med.*, vol. 137, Oct. 2021, Art. no. 104778.

[48] 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," *Artif. Intell. Med.*, vol. 131, Sep. 2022, Art. no. 102349.

[49] S. Akbar and M. Hayat, "iMethyl-STTNC: Identification of N^6-methyladenosine sites by extending the idea of SAAC into Chou's PseAAC to formulate RNA sequences," *J. Theor. Biol.*, vol. 455, pp. 205–211, Oct. 2018.

[50] 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," *Chemometric Intell. Lab. Syst.*, vol. 196, Jan. 2020, Art. no. 103912.

[51] 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," *Comput. Biol. Med.*, vol. 139, Dec. 2021, Art. no. 105006.

[52] M. Arif, S. Ahmad, F. Ali, G. Fang, M. Li, and D.-J. Yu, "TargetCPP: Accurate prediction of cell-penetrating peptides from optimized multi-scale features using gradient boost decision tree," *J. Comput.-Aided Mol. Des.*, vol. 34, no. 8, pp. 841–856, Aug. 2020.

[53] 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," *J. Bioinf. Comput. Biol.*, vol. 19, no. 4, Aug. 2021, Art. no. 2150018.

[54] J. Hu, X. Zhou, Y. Zhu, D. Yu, and G. Zhang, "TargetDBP: Accurate DNA-binding protein prediction via sequence-based multi-view feature learning," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 4, pp. 1419–1429, Jul. 2020.

SHAHID AKBAR received the bachelor's degree in computer science and information technology from the Islamic University of Technology, Bangladesh, in 2011, the M.S. and Ph.D. degrees in computer science from Abdul Wali Khan University Mardan (AWKUM), Pakistan, in 2015. His research interests include bioinformatics, biomedical engineering, and machine learning.

HEBA G. MOHAMED was born in Alexandria, Egypt, in 1984. She received the B.Sc. and M.Sc. degrees in electrical engineering from Arab Academy for Science and Technology, in 2007 and 2012, respectively, the Ph.D. degree in electrical engineering from the University of Alexandria, Egypt, in 2016, and the Associate Professor degree from Egypt, in 2022. In 2016, she was an Assistant Professor with the Alexandria Higher Institute of Engineering and Technology, Ministry of Higher Education, Egypt. Since 2019, she has been an Assistant Professor with the Faculty of Engineering, Communication Department, Princess Nourah bint Abdulrahman University, Saudi Arabia. Her research interests include cryptography, wireless communications, mobile data communications, the Internet of Things, and computer vision.

HASHIM ALI received the Ph.D. degree in computer science from Abdul Wali Khan University Mardan (AWKUM), Pakistan. He is currently an Assistant Professor with the Department of Computer Science, AWKUM. He is proficient in computer systems both theoretically and practically. His research interests include cloud computing, software testing, agile processes, energy-efficient systems, and enterprise systems.

AAMIR SAEED received the Ph.D. degree in wireless communication from Aalborg University, Denmark. He is currently an Assistant Professor with the Department of Computer Science and IT, University of Engineering and Technology at Peshawar. His research interests include big data structures (LSM and Bloom filters), microservices architecture, and the IoT with security in focus.

AFTAB AHMED KHAN received the Ph.D. degree in electronics engineering from the University of York, U.K., in 2019. He is currently a Lecturer with the Department of Computer Science, Abdul Wali Khan University Mardan (AWKUM), Pakistan. His research work is related to improvement in performance in ultra-dense high-capacity networks. His research interests include radio resource management, topology management to improve system performance, and overall energy efficiency in ultra-dense high-performance wireless networks.

SARAH GUL received the Ph.D. degree in biological science from the University of Ulm, Germany. He is currently an Assistant Professor with the Department of Biological Sciences, FBAS, International Islamic University Islamabad, Pakistan. His research interests include cancer genetics, molecular medicine, and machine learning.

ASHFAQ AHMAD, photograph and biography not available at the time of publication.

FARMAN ALI received the master's degree in computer science from the University of Peshawar, and the M.S. degree in computer science from Abdul Wali Khan University Mardan (AWKUM), Pakistan. He is currently a subject Specialist of computer science in elementary and secondary education with KPK, Pakistan. His research interests include machine learning, bioinformatics, and image processing.

YAZEED YASIN GHADI received the Ph.D. degree in electrical and computer engineering from The University of Queensland. He is currently an Assistant Professor of software engineering with Al Ain University. He was a Postdoctoral Researcher with The University of Queensland, before joining Al Ain University. He has published more than 80 peer-reviewed journal and conference papers and he holds three pending patents. His current research interests include developing novel electro-acousto-optic neural interfaces for large scale high resolution electrophysiology and distributed optogenetic stimulation. He is the recipient of several awards. His dissertation on developing novel hybrid plasmonic photonic on chip biochemical sensors received the Sigma Xi Best Ph.D. Thesis Award.

MUHAMMAD ASSAM received the B.Sc. degree in computer software engineering from UET Peshawar, Pakistan, and the M.Sc. degree in software engineering from UET Taxila, Pakistan. He is currently a Lecturer with the Department of Software Engineering, University of Science and Technology at Bannu, KP, Pakistan. His research interests include brain–computer interface, computer vision, artificial intelligence, natural language processing, and medical image processing.