Methods 230 (2024) 129–139

Contents lists available at ScienceDirect

Methods

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

StackDPPred: Multiclass prediction of defensin peptides using stacked ensemble learning with optimized features

Muhammad Arif [a], Saleh Musleh [a], Ali Ghulam [b], Huma Fida [c], Yasser Alqahtani [d], Tanvir Alam [a,*]

[a] College of Science and Engineering, Hamad Bin Khalifa University, Doha 34110, Qatar
[b] Information Technology Centre, Sindh Agriculture University, Sindh, Pakistan
[c] Department of Microbiology, Abdul Wali Khan University Mardan, 23200, KPK, Pakistan
[d] Independent Researcher, Madinah, Saudi Arabia

ABSTRACT

Host defense or antimicrobial peptides (AMPs) are promising candidates for protecting host against microbial pathogens for example bacteria, virus, fungi, yeast. Defensins are the type of AMPs that act as potential therapeutic drug agent and perform vital role in various biological process. Conventional Experiments to identify defensin peptides (DPs) are time consuming and expensive. Thus, the shortcomings of wet lab experiments are leveraged by computational methods to accurately predict the functional types of DPs. In this paper, we aim to propose a novel multi-class ensemble-based prediction model called StackDPPred for identifying the properties of DPs. The peptide sequences are encoded using split amino acid composition (SAAC), segmented position specific scoring matrix (SegPSSM), histogram of oriented gradients-based PSSM (HOGPSSM) and feature extraction based graphical and statistical (FEGS) descriptors. Next, principal component analysis (PCA) is used to select the best subset of attributes. After that, the optimized features are fed into single machine learning and stacking-based ensemble classifiers. Furthermore, the ablation study demonstrates the robustness and efficacy of the stacking approach using reduced features for predicting DPs and their families. The proposed StackDPPred method improves the overall accuracy by 13.41% and 7.62% compared to existing DPs predictors iDPF-PseRAAC and iDEF-PseRAAC, respectively on validation test. Additionally, we applied the local interpretable model-agnostic explanations (LIME) algorithm to understand the contribution of selected features to the overall prediction. We believe, StackDPPred could serve as a valuable tool accelerating the screening of large-scale DPs and peptide-based drug discovery process.

1. Introduction

Antimicrobial peptides (AMPs) are naturally occurring small peptides found throughout the body, contributing significantly to the innate immune system. Among these, defensins stand out as an evolutionarily ancient class, characterized by their cationic cysteine residues and frequent expression in epithelial or neutrophil cells [1]. They serve multifaceted roles in the host's innate immune response against a variety of infections. Defensins exhibit a broad spectrum of antimicrobial activities, including combating bacteria, viruses, fungi, and even certain cancers [2]. Additionally, they have shown promise in overcoming bacterial drug resistance, making them valuable candidates for therapeutic development. The antibacterial characteristics and distinctive mechanisms of action displayed by defensins have attracted significant attention in the development of a new class of natural antibiotic peptides. These peptides are aimed at combating bacterial infections, especially those that have developed resistance to traditional antibiotics [3]. Defensins derived from various sources share structural and functional resemblances, reflecting phylogenetic relationships among different types of defensins. Despite this, the amino acid sequences of mature defensins exhibit significant variability within each defensin family and subfamily [4]. Precisely identifying the types of defensins proves invaluable for analyzing their specificities towards various microbial targets. Such identification also offers novel insights into their functional roles and aids in the discovery of targets for antimicrobial drug development [5]. However, in postgenomic age, the abundance of sequence information in online databases, traditional laboratory-based methods such as Mass spectrometry [6], AMP Arrays [7] and Nuclear-Magnetic-Resonance Spectroscopy [8] for screening and characterizing DPs are challenging due high cost, long time and resource intensive.

In recent years, there has been a proliferation of designing computational methods predicting antimicrobial peptides that have defensins activity [9]. Some researchers initiated the computational exploration of the defensin family in 2009, employing diversity measures. Another research group introduced DEFENSINPRED, a classifier capable of categorizing human defensin proteins and their types based on pseudo amino acid compositions [10]. In 2015, they proposed the iDPF-PseRAAAC servers, focusing on distinguishing defensin peptides and its subfamilies

* Corresponding author.
E-mail address: talam@hbku.edu.qa (T. Alam).

https://doi.org/10.1016/j.ymeth.2024.08.001
Received 1 July 2024; Received in revised form 30 July 2024; Accepted 13 August 2024
Available online 22 August 2024

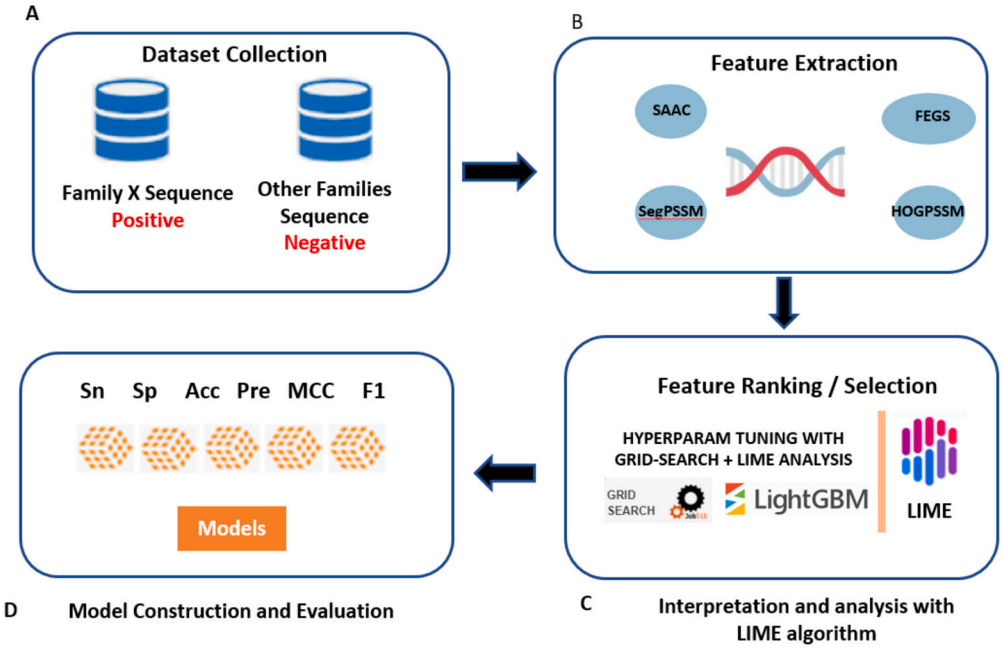

Fig. 1. Schematic workflow of the proposed StackDPPred model.

using Protein Blocks [11]. In another study, they proposed the improved version called iDEF-PseRAAC public method for the identification of the defensin peptide based on reduced amino Acid composition descriptor [12]. The Subsequent refinements were made using support vector machines (SVMs) classifier, leading to the development of a free online predictor for predicting multiple function of DPs from sequence information.

Despite the remarkable advancements, existing DPs-based predictors still exhibit unsatisfactory performances, warranting further research. The challenges can be classified into two folds: feature extraction-based and model construction-based. Firstly, the existing tools used compositional features and failed to explore the graphical, physicochemical and evolutionary properties of peptide sequences. These properties contribute to identify the correct function of particular protein/peptide. Secondly, the proposed available methods used single classifiers i.e., support vector machine to build final prediction model that restrict the accurate prediction of true antimicrobial defensin peptides (ADPs). Third, the previous developed tools used leave-one-out or jackknife cross validation test to train the model and unable to validate their models on test data.

Based on the aforementioned challenges, we aimed to develop a powerful stacking-based ensemble model for the novel ADPs and their types with high accuracy. Fig. 1 shows the designed framework of the proposed StackDPPred method. In essence, several discrete steps are performed in the development StackDDPred model as like: dataset collection, feature encoding, feature ranking and selection, model construction and evaluation. We considered four different types of feature descriptors that extracts multiple properties such as evolutionary-based information, structural and graphical-based, compositional-based, local and global information from peptide sequence. We further fused the extracted features and applied PCA feature selection method to construct the optimal subset from all hybrid features. Then, the selected feature subset were input to five base-classifiers for generating 20D (dimension) probability features (PFs). Furthermore, these 20D PFs were fed into meta classifier using 5-fold CV method to build the final prediction model. In this research, we summarize our contributions as follows:

(a) We encoded the novel graphical information by FEGS descriptor, global and local evolutionary information by SegPSSM and transformer-based HOGPSSM descriptors and compositional information by split amino acid composition (SAAC) descriptor from ADP sequences.
(b) We optimized the hybrid features by applying PCA algorithm to eliminate the redundancy and improved the prediction performance of the developed model.
(c) We proposed StackDPPred, a powerful ML-based model that target multiple classes of ADPs with high accuracy compare to existing methods both on training and testing data.

Table 1
Sample Tables for reference – Summary of Datasets.

Label	Family type	No. of peptides
P1	Insect Defensins	60
P2	Invertebrate Denensins	31
P3	Plant Denensins	42
P4	Unclassified Denensins	38
P5	Vertebrate Denensins	157

2. Material and methods

2.1. Benchmark dataset

A curated and stringent benchmark dataset is essential for developing an intelligent predictive model [13–16]. Herein, we collected a high quality dataset from previously published method for predicting definsin peptides iDPF-PseRAAC [11]. Initially, the authors extracted the experimentally verified DPs from definsin knowledge-base [17]. Further, to overcome the redundancy, CD-HIT program was used with cut-off similarity at less or equal to 80%. The final dataset composed of five sub-families of DPs primary sequences in which 60 peptides are insect defensins denoted by P1, 31 samples are invertebrate defensins denoted as P2, 42 peptides are plant defensins denoted by P3, 38 peptides are non-classified defensins denoted by P4 and 157 peptides are vertebrate defensins denoted by P5 in Table 1. Finally, we partitioned these samples with 80:20 ratio for training and testing the proposed model.

2.2. Feature encoding methods

Encoding the biological peptides into numerical feature vectors is a challenging but important step in developing machine learning models [18–21]. Because the accurate prediction of a model highly depends on the hidden concealed information in a peptide sequence. In this research, we developed novel PSSM-based, graphical-based and compositional-based feature extraction algorithms for DPs prediction.

2.3. Graphical-based statistical features

In scientific research, FEGS is a powerful feature encoding method proposed by Zhengchao et al. [22]. The RAAC method, extract the compositional features from protein sequence; however, lose of sequence information is a major problem that remain unsolved. FEGS descriptor, tackle this issue by extracting the graphical features of peptides or protein sequences based on physicochemical (PC) characteristics. We describe the working process of this method in the subsequent steps.

First, we need to map the twenty AAs to twenty points in 3D space based on the selected PC properties from amino acid index (AAindex). The AAindex is an online database repository denoting the biochemical and PC characteristics of twenty AA residues [23,24]. The current updated version contain 566 indexes. After removing redundant indexes for all duplicate values, we selected 158 indexes for DPs. For detailed information on these selected indexes, the readers are referred to supplementary Table T1. Next we arrange the twenty AAs in ascending order to for the effective utilization of their PC indices by the give mathematical expression Eqn. (1)

$$\phi(\delta i) = \left(\cos\left(\frac{2\pi i}{20}\right), \sin\left(\frac{2\pi i}{20}\right), 1\right), \quad i = 1, 2, \ldots, 20 \tag{1}$$

where δi denotes the twenty AAs on the circumference at the bottom of a straight cone with a height of 1. To encode the high order information we used pair combination of AAs by adopting dipeptide composition (DPC) by the following equation Eqn. (2).

$$\phi(\delta_i, \delta_j) = \phi(\delta_i) + \frac{1}{4}(\phi(\delta_j) - \phi(\delta_i)), \quad i, j = 1, 2, \ldots, 20 \tag{2}$$

where δ_i, δ_j denotes the 20 by 20 features of AAs. Next to build the 3D curve for the given peptide sequence P having length L $P = p_1 p_2 \ldots p_L$, the corresponding coordinates x_k, y_k, z_k for point $S_k(x_k, y_k, z_k)$ can be computed by the given mathematical formula Eqn. (3)

$$\psi(P_i) = \psi(P_{i-1}) + \phi(P_i) + \sum_{\delta_1, \delta_2 \in \{A,C,D,\ldots,Y\}} f_{\delta_1\delta_2} \cdot \varphi(\delta_1\delta_2) \tag{3}$$

Finally, the 578D feature vector is generated of peptide sequence P.

2.4. Split amino acid composition

The Split amino acid composition (SAAC) developed by [25], has been successfully utilized to predict numerous protein activities, including membrane protein types [26], outer membrane protein [27], enzyme family class [28] and prediction of heat shock proteins. Saac-based encoding scheme split a short length peptide/protein sequence into parts and calculating the composition of each part individually. For our investigation, we partitioned the sequence of the DPs into three segments: (i) the first 5 amino acids at the N terminus, (ii) the last 5 amino acids at the C terminus, and (iii) the region between these two termini. In contrast to conventional amino-acid composition, the resulting feature vector has a dimension of 60D instead of 20D. The mathematical expression for this can be represented as like ourś previous paper [29]. This paper the critical factors influencing symbols N and C denote the N-terminus and C-terminus, respectively, whereas the word “integral segment” refers to a specific part of the molecule. This work utilizes (SAAC), which involves calculating the composition of the N-termini, C-termini, and the remaining portion of the protein separately in Eqn. (4).

$$S = \left[S_1, S_2, S_3, S_4, \ldots, S_{20}, S_{21}, \ldots, S_{20+\lambda}\right]^T \tag{4}$$

2.5. Segmentation-based position specific scoring matrix method

Position-specific scoring matrix (PSSM) is widely used and effective feature representation technique in numerous bioinformatics problems such as anticancer peptides [30], DNA-binding proteins prediction [31] etc. PSSM basically extract the evolutionary information that are helpful to discover and analyze the function of proteins or peptides [32]. By this inspiration, we encode the multiple types of DP sequences to PSSM. PSSM construct M by 20 dimension matrix for each peptide sample using Swiss-Prot and BSI-BLAST program [33]. Let’s assume that L denotes the length of peptide sample and M denotes the twenty residues of amino acid in the given peptide sequence then normalized PSSM matrix with fixed length feature space can be defined as Eqn (5):

$$\text{PSSM} = \begin{bmatrix} M_{1,1} & M_{1,2} & \cdots & M_{1,20} \\ M_{2,1} & M_{2,2} & \cdots & M_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ M_{L,1} & M_{L,2} & \cdots & M_{L,20} \end{bmatrix}_{L\times 20} \tag{5}$$

However, PSSM failed to retain the sequence-order information. To cope with this challenge, we introduce segmentation-based feature encoding (SegPSSM) method to capture the significant patterns hidden in the PSSM. We considered SagPSSM descriptor to encode the local and global complementary features buried in amino acid constituents. The SegPSSM based framework method divided the generated PSSM into several equal segments by row wise. Each fixed size segment (S-PSSM) of the PSSM matrix can be formulated as follows Eqn (6):

$$\text{SegPSSM}(\xi) = \begin{bmatrix} S_{p+1,1} & S_{p+1,2} & \cdots & S_{p+1,20} \\ S_{p+2,1} & S_{p+2,2} & \cdots & S_{p+2,20} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p+L(\xi),1} & S_{p+L(\xi),2} & \cdots & S_{p+L(\xi),20} \end{bmatrix}_{L(\xi)\times 20} \tag{6}$$

In the above equation s = (ξ - 1) ·L(ξ), $\xi = 1, 2, 3, \ldots$ where K represents the SPSSM and L(ξ) the number of tuples in each segment S_PSSM Eqn. (7).

$$N(\xi) = \begin{cases} \left\lfloor \frac{L}{\xi} \right\rfloor, & \xi = 1, 2, \ldots, K-1 \\ L - \left\lfloor \frac{L}{\xi} \right\rfloor, & \xi = K \end{cases} \tag{7}$$

Here $\left\lfloor \frac{L}{\xi} \right\rfloor$ denotes the (K-1)$^{\text{th}}$ and L - $\left\lfloor \frac{L}{\xi} \right\rfloor$ denotes the K$^{\text{th}}$ S-PSSM.

Since DPs are short length peptides ranging from (5-50) residues. Therefore, based on the experimental outcomes we kept the value of K = 2,3. For further detail, the readers are refereed to our published work [34].

2.6. Transformed histogram of oriented gradient-based PSSM method

Histogram of oriented gradient (HOG) [34] is widely used descriptor in computer vision and image analysis. Recently, we utilized HOG feature encoding method in detecting protein or peptide function characterization. Transforming evolutionary-based PSSM matrix into image-based feature set called HOGPSSM can be generated in the following steps.

After calculating the PSSM for each DP sequence, we need to calculate the vertical $H_y(a, b)$ and horizontal gradients $H_x(a, b)$ of the extracted PSSM matrix by the following formula (Eqn (8), (9)):

$$H_x(a, b) = \begin{cases} PSSM(a+1, b) - 0, a = 1, \\ PSSM(a+1, b) - PSSM(a-1, b), 1 < a < 20, \\ 0 - PSSM(a-1, b), a = 20 \end{cases} \tag{8}$$

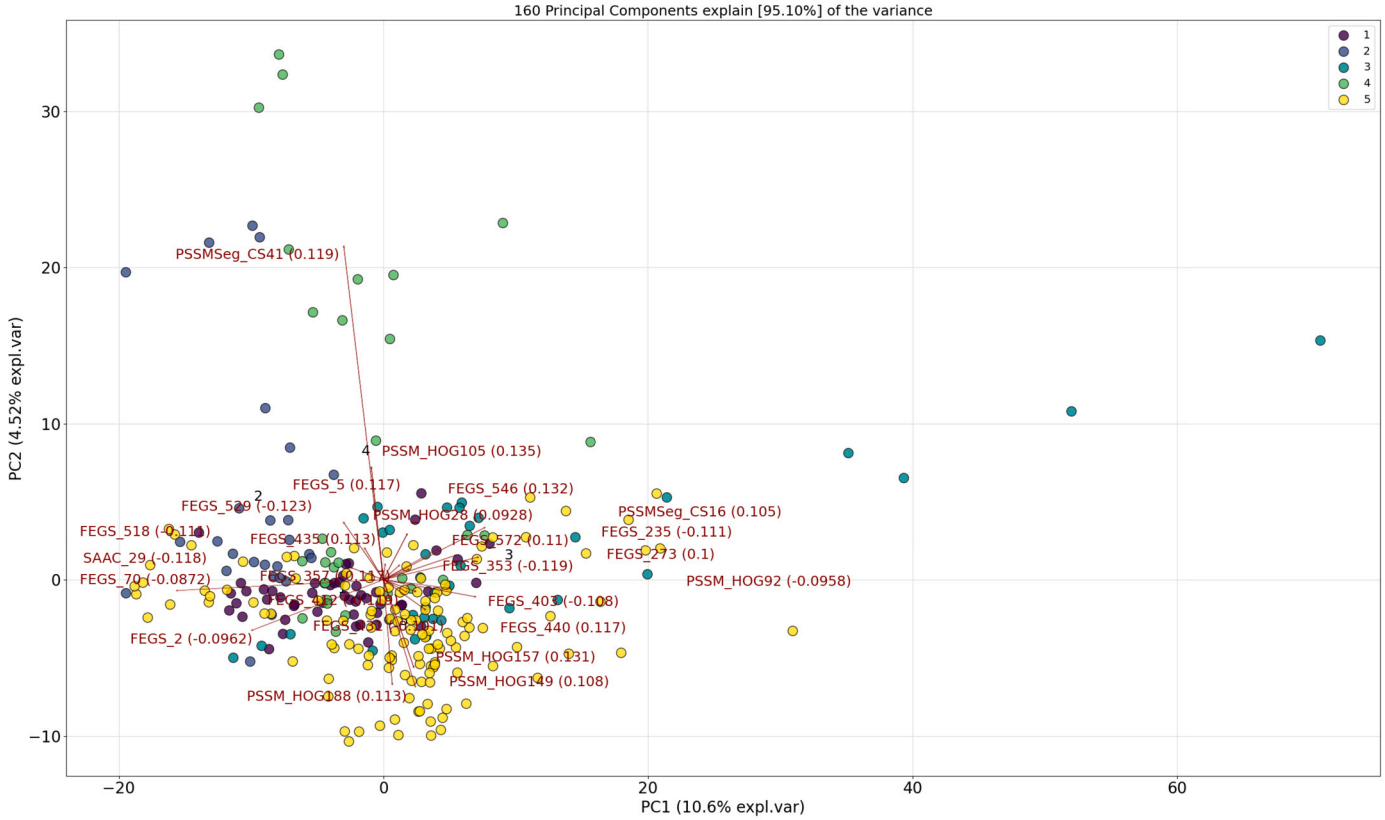

Fig. 2. PCA-based optimal feature selection.

$$G_y(a,b)=\begin{cases} PSSM(a,b+1)-0, b=1, \\ PSSM(a,b+1)-PSSM(a,b-1), 1<b<L, \\ 0-PSSM(a,b-1), b=L \end{cases} \quad (9)$$

Subsequently, the gradient's direction and magnitude can be calculated by the below mathematical expression (Eqn (10), (11)):

$$H(a,b)=\sqrt{H_x(a,b)^2+H_y(a,b)^2}, \quad (10)$$

$$\Theta(a,b)=\tan^{-1}\left[\frac{H_x(a,b)}{H_y(a,b)}\right], \quad (11)$$

where denotes the gradient magnitude $H(a,b)$ and $\Theta(a,b)$ gradient direction of the PSSM matrix. For the 3rd step, the image is segmented into 16 by 16 size connected areas known as cells. Each cell encompasses the feature set compressing gradient magnitude and direction within the sub-matrix (Eqn (12), (13)).

$$H_{i,j}(s,t)=H\left(5\times i+1+s, j\times\frac{L}{4}+1+t\right) \quad (12)$$

$$\Theta_{i,j}(s,t)=\Theta\left(5\times i+1+, n\times\frac{L}{4}+1+t\right) \quad (13)$$

here i,j represents the sub-matrix subscripts $(0\le i\le 2, 0\le j\le 2)$ and the subscripts inside the sub-matrix locations $(0\le v\le 9, 0\le v\le L/2-1)$ are denoted by s,t. Each sub-matrix produces sixteen different histogram channels on the basis of gradient direction. As a result, for each peptide sample HOG-PSSM generates 16*16 = 256-D (dimensions) feature vector.

2.7. Feature selection-based on principal component analysis

We extracted the compositional-based, graphical-based, image-based and evolutionary-based features from the raw peptide sequence. However, sometimes these techniques unable to dig out wide rang of valuable features that are helpful for defensin peptide family prediction. Therefore, feature selection methods perform an integral role in developing an accurate predictor. Herein, we implemented principal component analysis (PCA), a well-known feature selection algorithm proposed by Kirbby et al. for the compression of images in face recognition [35]. PCA not only reduce the dimension of noisy and irrelevant features but also help to decrease the time complexity of the proposed model. The core idea of PCA is the transformation of high dimension feature space to low dimension features so that to preserve the maximum variance of the extracted attributes. When we applied the PCA algorithm on the hybrid features (FEGS+SAAC+SegPSSM and HOGPSSM), it performed the following operation to reduce the feature dimension. Firstly, transformed the given two dimension feature vector into one dimension, next, the eigenvalues and corresponding eigenvectors are obtained by decomposing the eigenvalues of the covariance matrix. These eigenvectors form the principal components of the data (training and testing samples), and the eigenvalues represent the magnitude of the variance in the direction of the corresponding principal components [36]. After the entire process, 160 best features were selected as shown in Fig. 2. Among the top ranked features 98 are selected from FEGS descriptor, 27 from SAAC descriptor, 28 from HOGPSSM and 7 from SegPSSM descriptors as illustrated in Fig. 3. The ranked features and its three dimensional structure generated by PCA are provided in the Supplementary Table T2 and Fig. S1.

2.8. Stack-ensemble framework

Over the past years, several prior research in bioinformatics and pattern recognition reveal that ensemble models can achieve superior predictive performance compared to individual models or classifiers [37]. Generally, ensemble learning (EL) can be categorized into three types from broad spectrum: majority voting-based EL, average-based EL and stacking-based EL. Here in, we adopted stack EL method, that leverage the discriminating power of the baseline models to enhance the prediction performance of the system [38–40]. The framework of Stack

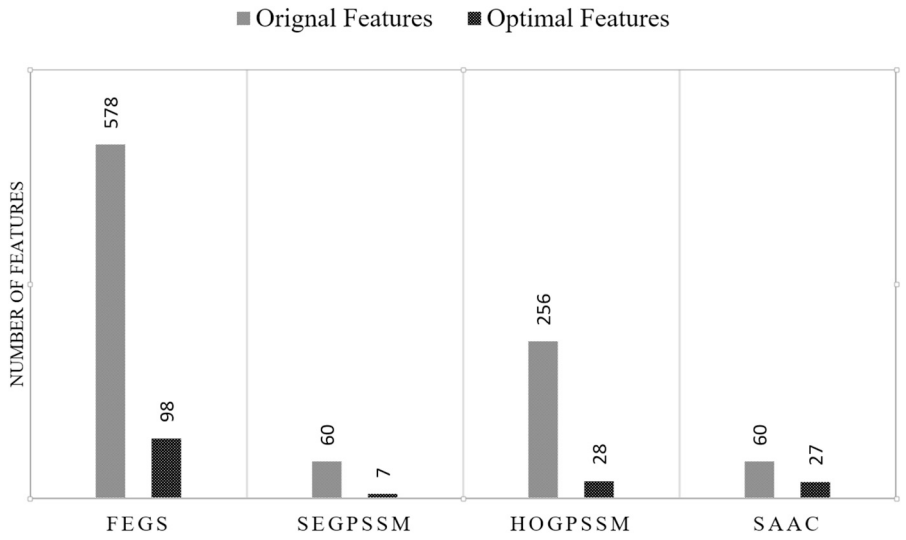

Fig. 3. The statistics of PCA-based optimal and original features.

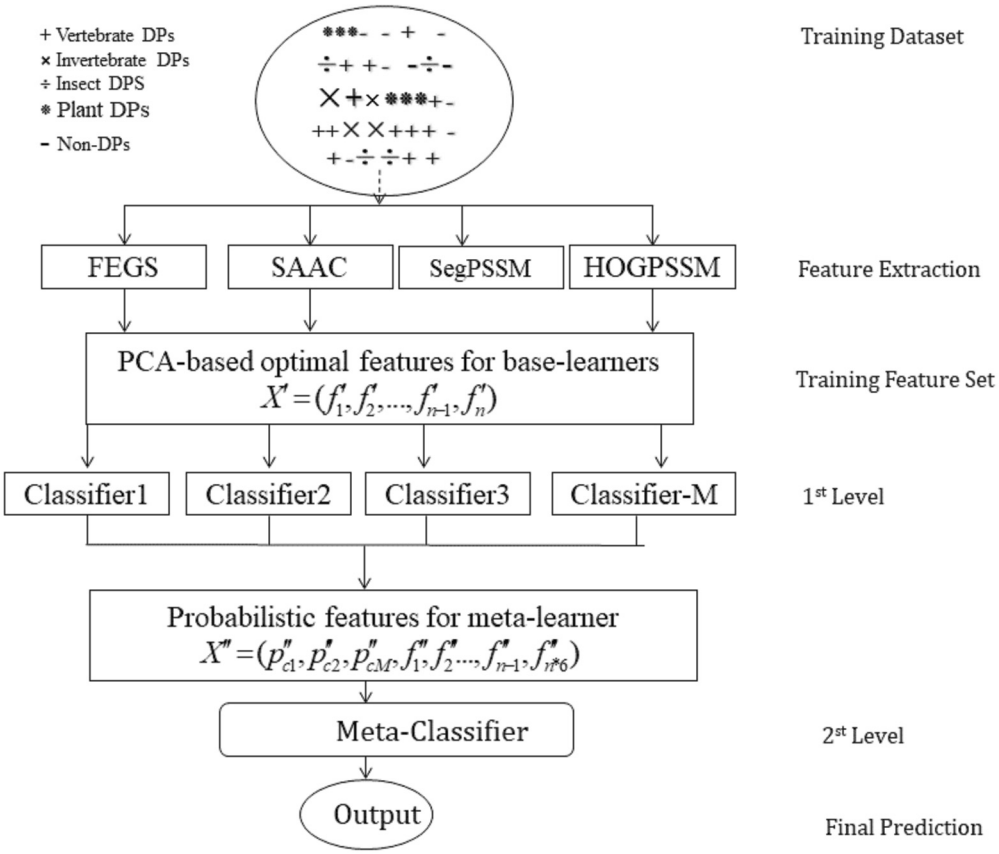

Fig. 4. Diagrammatic representation of stacked ensemble classifier.

EL process is depicted Fig. 4. The stacking-based EL models work in two phases (layers). The first-layer classifiers are called base-learners and second-layer classifier is called meta-learner. The base-classifiers are trained by using the optimal feature vectors and generate the probability features (PFs). Then meta-classifier use these M number of PFs as input and predict the final output. In this research, we used five base classifiers i.e., CatBoost, decision tree (DT), Gaussian naive base (GNB), multilayer perceptran (MLP) and extreme gradient boost (xGB) with four types of feature encoders i.e. FEGS, SegPSSM, SAAC and HOGPSSM in the first layer. In the second layer, we used Catboost algorithm as meta classifier that takes the generated PFs as input and build the final prediction model. The description of each step in stacking EL strategy is elaborated in the given pseudo code Algorithm 1 and Eqn (14).

$$nFeat(S) = \left[f(BC1(S)), \quad f(BC2(S)), \quad \cdots, \quad f(BC20(S)) \right]^T \tag{14}$$

where the $f(BC1(S))$ represents the probability feature (PF) generated by the baseclassifier (BC_i) of DP sequence S.

2.9. Performance measure and model evaluation

In case of multi-class classification scenario, some classes are imbalanced. To tackle this situation, we used five evaluation measures i.e.,

Algorithm 1 Pseudo-code of the proposed Stacking-based ensemble learning.

Input: Data set $M = \{(a_1, b_1), (a_2, b_2), \ldots, (a_m, b_m)\}$;
Base-learners $B_1, B_2, \ldots, B_T$;
Meta-learner $\mathcal{A}$.
Process:
for $t = 1, 2, \ldots, T$ **do** %Train the base-model by employing the
$h_t = B_t(M)$; *%1st-layer model* B_t;
end for
$M' = \emptyset$; %Create a new data set
for $i = 1, \ldots, m$ **do**
for $t = 1, \ldots, T$ **do**
$z_{it} = h_t(a_i)$;
end for
$M' = M' \cup ((z_{i1}, \ldots, z_{iT}), b_i)$;
end for
$h' = A(M')$; *%Train the* 2^{nd} *layer model* h' *by employing*;
%The meta-model A to the new data set M';
Output: $H(a) = h'(h_1(a), \ldots, h_T(a))$

Table 2
Ablation Study of base-classifiers with FEGS descriptor using 5-Fold CV.

Model/Class	Pr	Sn	Acc	Sp	F1
CatBoost					
S1	1.00	0.98	1.00	1.00	0.99
P2	1.00	1.00	1.00	1.00	1.00
P3	1.00	0.97	1.00	1.00	0.99
P4	1.00	0.71	0.97	1.00	0.83
P5	0.92	1.00	0.96	0.92	0.96
avg	0.98	0.93	0.98	0.98	0.95
DT					
P1	0.81	0.86	0.94	0.96	0.83
P2	0.76	0.85	0.96	0.97	0.80
P3	0.71	0.71	0.93	0.96	0.71
P4	0.48	0.61	0.88	0.91	0.54
P5	0.89	0.79	0.85	0.91	0.83
avg	0.73	0.76	0.91	0.94	0.74
GNB					
P1	1.00	0.94	0.99	1.00	0.97
P2	1.00	1.00	1.00	1.00	1.00
P3	1.00	0.57	0.95	1.00	0.73
P4	0.93	0.90	0.98	0.99	0.92
P5	0.87	1.00	0.93	0.87	0.93
avg	0.96	0.88	0.97	0.97	0.91
MLP					
P1	1.00	0.98	1.00	1.00	0.99
P2	0.96	1.00	1.00	1.00	0.98
P3	0.97	0.94	0.99	1.00	0.96
P4	1.00	0.90	0.99	1.00	0.95
P5	0.96	0.99	0.98	0.96	0.98
avg	**0.98**	**0.96**	**0.99**	**0.99**	**0.97**
xGB					
P1	0.94	0.96	0.98	0.99	0.95
P2	1.00	0.92	0.99	1.00	0.96
P3	0.94	0.86	0.97	0.99	0.90
P4	0.92	0.74	0.96	0.99	0.82
P5	0.93	1.00	0.96	0.93	0.96
avg	0.95	0.90	0.98	0.98	0.92

precision (Pr), F1-score(F1), overall accuracy (Acc), sensitivity (Sn), and specificity (Sp) to evaluate the predictive efficacy our proposed Stack-DPPred method [41–44]. The equation of the evaluation metrics are shown below (Eqn (15), (16), (17), (18), (19)):

Table 3
Ablation Study of base-classifiers with SAAC descriptor using 5-Fold CV.

Model/Class	Pr	Sn	Acc	Sp	F1
CatBoost					
P1	0.80	0.94	0.95	0.95	0.86
P2	0.59	0.38	0.92	0.97	0.47
P3	0.88	0.86	0.97	0.98	0.87
P4	0.43	0.10	0.88	0.98	0.16
P5	0.81	0.97	0.88	0.80	0.89
avg	0.70	0.65	0.92	0.94	0.65
DT					
P1	0.63	0.60	0.86	0.92	0.61
P2	0.26	0.23	0.86	0.93	0.24
P3	0.72	0.74	0.93	0.96	0.73
P4	0.21	0.23	0.81	0.89	0.22
P5	0.80	0.81	0.81	0.82	0.81
avg	0.52	0.52	0.86	0.90	0.52
GNB					
P1	0.93	0.54	0.91	0.99	0.68
P2	0.23	0.81	0.72	0.71	0.35
P3	0.61	0.57	0.90	0.95	0.59
P4	0.18	0.13	0.84	0.93	0.15
P5	0.83	0.61	0.75	0.89	0.70
avg	0.56	0.53	0.82	0.89	0.50
MLP					
P1	0.86	0.84	0.95	0.97	0.85
P2	0.67	0.54	0.93	0.97	0.60
P3	0.71	0.77	0.93	0.95	0.74
P4	0.44	0.35	0.88	0.94	0.39
P5	0.86	0.92	0.89	0.86	0.89
avg	0.71	0.68	0.91	0.94	0.69
xGB					
P1	0.80	0.86	0.93	0.95	0.83
P2	0.50	0.35	0.90	0.96	0.41
P3	0.83	0.86	0.96	0.97	0.85
P4	0.37	0.23	0.87	0.95	0.28
P5	0.82	0.92	0.86	0.82	0.87
avg	0.66	0.64	0.91	0.93	0.65

$$Acc = \frac{(tp + tn)}{(tp + tn + fp + fn)} \tag{15}$$

$$Sen = \frac{tp}{tp + fn} \tag{16}$$

$$Spe = \frac{tn}{tn + fp} \tag{17}$$

$$Pr = \frac{tp}{tp + fp} \tag{18}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{19}$$

In the above-mentioned notation tp denotes the ADPs, and tn denotes the peptides with non-ADPs. Similarly, fp denotes the number of incorrect samples that have no ADP properties and fn means the number of incorrect samples having ADPs activity. The aforementioned assessment metrics are threshold dependent. Furthermore, we used the receiver operating characteristic (ROC) curve, along with the area under the ROC curve (AUC) as threshold-independent indexes to evaluate the overall effectiveness of the proposed method [45]. The closer the prediction value is to 1, the better the predictive performance of the classification algorithm and vice versa. We adopted 5-fold CV method to build the robust DPs-based predictor and then validate the generalization power by test data.

Table 4
Ablation Study of base-classifiers with HOGPSSM descriptor using 5-Fold CV.

Model/Class	Pr	Sn	Acc	Sp	F1
CatBoost					
P1	0.92	0.90	0.97	0.98	0.91
P2	0.88	0.27	0.93	1.00	0.41
P3	0.96	0.66	0.95	1.00	0.78
P4	0.95	0.68	0.96	1.00	0.79
P5	0.75	0.98	0.84	0.70	0.85
avg	0.89	0.70	0.93	0.93	0.75
DT					
P1	0.55	0.52	0.84	0.91	0.54
P2	0.27	0.27	0.86	0.92	0.27
P3	0.56	0.51	0.89	0.94	0.54
P4	0.51	0.71	0.89	0.91	0.59
P5	0.70	0.67	0.71	0.74	0.69
avg	0.52	0.54	0.84	0.88	0.52
GNB					
P1	0.85	0.82	0.94	0.97	0.84
P2	0.48	0.42	0.90	0.95	0.45
P3	0.90	0.77	0.96	0.99	0.83
P4	0.48	0.35	0.88	0.95	0.41
P5	0.80	0.91	0.85	0.79	0.85
avg	0.70	0.66	0.91	0.93	0.67
MLP					
P1	0.87	0.92	0.96	0.97	0.89
P2	0.68	0.58	0.93	0.97	0.62
P3	0.96	0.77	0.97	1.00	0.86
P4	0.84	0.87	0.97	0.98	0.86
P5	0.89	0.94	0.92	0.89	0.91
avg	**0.85**	**0.82**	**0.95**	**0.96**	**0.83**
xGB					
P1	0.88	0.86	0.95	0.97	0.87
P2	0.85	0.42	0.94	0.99	0.56
P3	0.81	0.63	0.93	0.98	0.71
P4	0.85	0.74	0.96	0.98	0.79
P5	0.79	0.95	0.85	0.77	0.86
avg	0.84	0.72	0.93	0.94	0.76

Table 5
Ablation Study of base-classifiers with SegPSSM3 descriptor using 5-Fold CV.

Model/Class	Pr	Sn	Acc	Sp	F1
CatBoost					
P1	0.94	0.92	0.97	0.99	0.93
P2	1.00	0.54	0.96	1.00	0.70
P3	0.97	0.80	0.97	1.00	0.87
P4	0.85	0.74	0.96	0.98	0.79
P5	0.84	0.99	0.91	0.83	0.91
avg	0.92	0.80	0.95	0.96	0.84
DT					
P1	0.78	0.80	0.92	0.95	0.79
P2	0.47	0.54	0.90	0.94	0.50
P3	0.73	0.77	0.93	0.96	0.75
P4	0.68	0.61	0.92	0.96	0.64
P5	0.81	0.79	0.81	0.83	0.80
avg	0.69	0.70	0.90	0.93	0.70
GNB					
P1	0.91	0.82	0.95	0.98	0.86
P2	0.68	0.65	0.94	0.97	0.67
P3	1.00	0.77	0.97	1.00	0.87
P4	0.60	0.58	0.91	0.95	0.59
P5	0.85	0.95	0.89	0.85	0.90
avg	0.81	0.75	0.93	0.95	0.78
MLP					
P1	0.92	0.90	0.97	0.98	0.91
P2	0.94	0.65	0.96	1.00	0.77
P3	0.94	0.91	0.98	0.99	0.93
P4	0.81	0.81	0.96	0.98	0.81
P5	0.91	0.98	0.94	0.91	0.94
avg	**0.90**	**0.85**	**0.96**	**0.97**	**0.87**
xGB					
P1	0.94	0.90	0.97	0.99	0.92
P2	0.94	0.58	0.96	1.00	0.71
P3	0.91	0.83	0.97	0.99	0.87
P4	0.79	0.71	0.95	0.98	0.75
P5	0.86	0.98	0.91	0.85	0.91
avg	0.89	0.80	0.95	0.96	0.83

3. Results and discussion

3.1. Ablation experiment of individual features and baseline models

In this subsection, we conducted an ablation study of five baseline models (CatBoost, DT, xGB, MLP and GNB) using four different types of feature descriptors (FEGS, SAAC, HOGPSSM and SegPSSM) using 5-fold CV method. In order to compare the effectiveness of individual features using different ML models, the performances are reported in Tables 2–5. We used average Acc and F1 as performance indicators to evaluate the best classification model. We can observe from Tables 2–5, that amongst the 20 base-models, MLP is the best performer obtained the highest Acc of 0.990 and F1 score 0.970 using FEGS descriptor, Acc of 0.91 and F1 of 0.690 using SAAC descriptor, Acc of 0.950 and F1 of 0.830 using HOGPSSM encoding method, Acc of 0.960 and F1 of 0.870 respectively. CatBoost learning model, was reported the second best performer in predicting all families of definsin peptides. In contrast, DT base classifier was reported the worst prediction model. The obtained Acc of DT model are 0.910, 0.860, 0.840 and 0.900 using FEGS, SAAC, HOGPSSM and SegPSSM feature representation methods. So, from the foregoing discussion about the individual features we can observe that each encoding method behaves differently in characterizing the families of DPs. In other context, we can say that one descriptor can perform better than the other using the same learning model due to some circumstances.

3.2. Ablation experiment of baseline and stacked models on hybrid features

In this section, we performed another ablation study on hybrid features, to demonstrate the effectiveness of using various models. It is known that, Prominent features are crucial in designing an intelligent predictor using ML models. We enhanced the overall performance of the proposed StackDPPred model for predicting five families of defensin peptides, by serially combining four types of compositional, evolutionary, and graphical features extracted from raw sequences. We experimentally validated the performance of base-classifiers and stacked ensemble classifier on fused features. The empirical outcomes of the base-classifiers and stacking model on test dataset are reported in Table 6. We can observe from Table 6 that StackDDPred accurately predicted DPs families on test dataset with Acc of 0.980 and F1 of 0.920. However, MLP is nominated as the best model amongst the base-classifiers, predicted more false positives than ourś model and achieved Acc of 0.974 and F1 of 0.912. Similarly, xGB mode also produced slightly similar results which are Acc of 0.974 and F1 of 0.910. In contrast, DT produced the worst outcomes on the hybrid features. This example demonstrates that, stacked ensemble learning has the superior ability to predict DPs families as compare to baseline classifiers. The underline reason is the stacking strategy leverages multiple individual predictors to deliver more stable and accurate predictions [46–48].

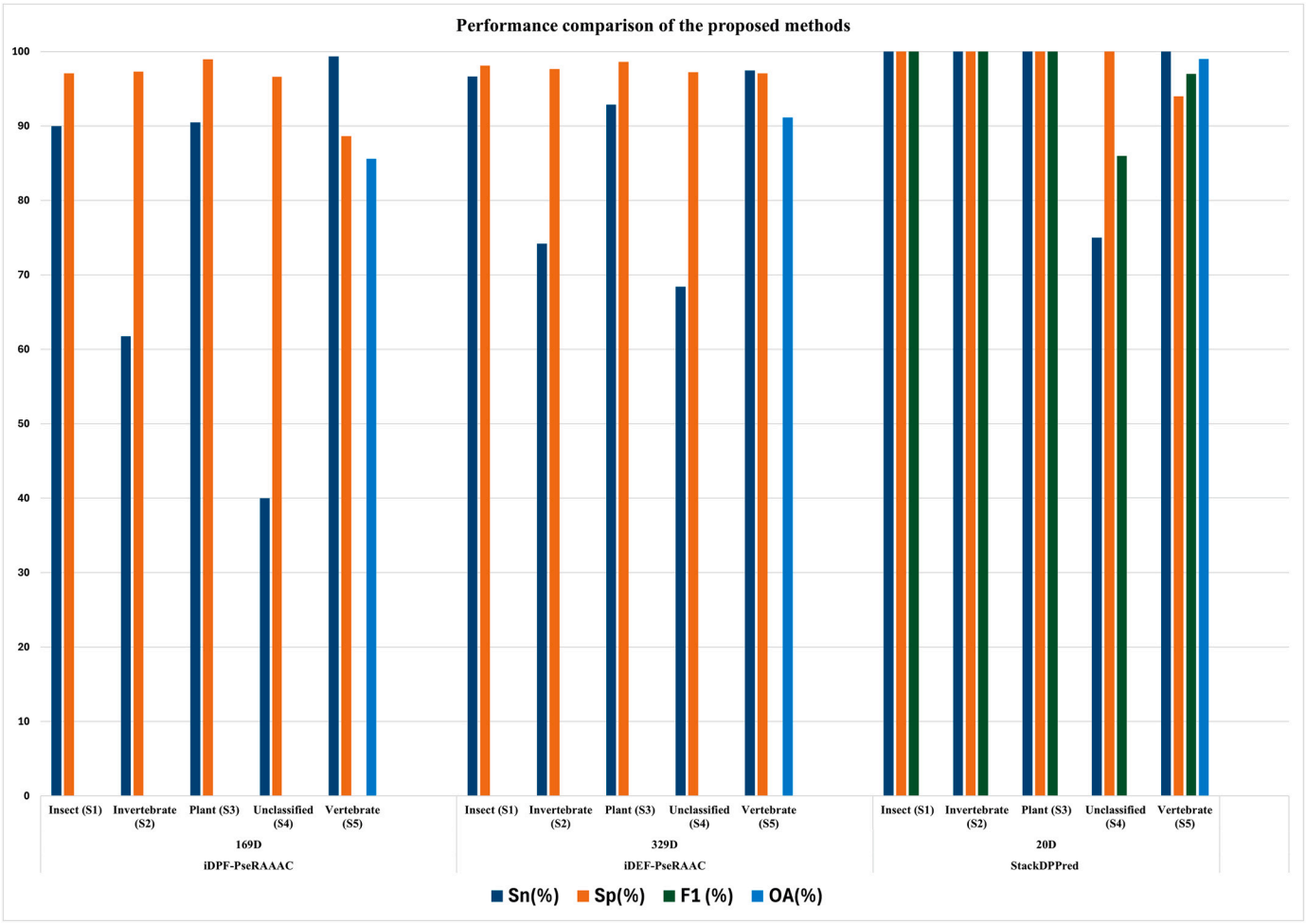

Fig. 5. Performance comparison of the proposed methods.

3.3. Feature selection improves the prediction performance

Feature selection is an essential step in machine learning for selecting the prominent features [49,50]. The extracted features might contain redundant information that affects the performance of the trained model predicting multiple types of DPs. Since, we trained the proposed model by single and hybrid features. The hybrid feature (954D) is a serial combination of four types of single features i.e., SAAC (60D), FEGS (578D), SegPSSM (60D) and HOGPSSM (256D) that might cause the overfitting problem due curse of dimension disaster and noise. To cope with this dilemma, we applied PCA on the hybrid features to select enrich subset of features (160D) for training our proposed ensemble model. The ratio of optimal subset of features, selected by PCA algorithm is FEGS (98/578), SAAC (30/60), SegPSSM (27/60), HOGPSSM (28/60) as shown in Fig. 3.

This statistical calculation shows that the graphical-based features (FEGS) and evolutionary-based (SegPSSM) significantly contribute in predicting DPs and its family types. In order to investigate whether the PCA-based features are effective or not in enhancing the overall performance of the proposed method for DPs prediction. We report the validation success rates of single and ensemble classifiers in Table 7 to analyze the prediction efficacy before and after feature selection. We compare the prediction efficacy of the baseline and stack ensemble model using the optimal features consisting of properties of FEGS, SAAC, SegPSSM, and HOGPSSM. The contribution of each feature descriptor to the identification of DPs and their families are listed in Supplementary Table T2. From Table 7 we can observe that, the proposed StackDPPred model exhibited the superior prediction performance with respected to all performance indicators (Pr, Sn, Acc, Sp and F1) as compared to baseline models CatBoost, DT, MLP, xGB and GNB on the test set. The achieved validation Acc and F1 of StackDPPred model is 0.990 and 0.970 respectively. The prediction power of the best baseline models CatBoost, MLP and xGB did not reach that of the stack ensemble learning model. The Acc of CatBosst, MLP and xGB are 0.984, 0.972 and 0.966 and F1 scores are 0.966,0.926 and 0,906 respectively. Thus, based on the aforementioned discussion, we conclude that feature selection particularly FEGS descriptor mostly Incorporated in discriminating multi-functions of DPs from sequence information. The enriched features selected by PCA as elaborated in the above section, the top 160 high ranked correlated features are shown in Fig. 2.

3.4. StackDPPred performance comparison with existing methods

We compare the prediction performance of our proposed multi-class StackDPPred method with advanced existing methods for identifying different families of defensing peptides. In order to avoid bias and demonstrate the effectiveness of the proposed method [51,52], we used the same dataset as adopted by iDPF-PseRAAAC [11] and iDEF-PseRAAAC [12]. As shown in Table 8 and Fig. 5, it is apparent that StackDPPred outperformed all the available methods by all performance measures i.e. overall accuracy (OA), F1, Sp and Sn. Our proposed model beat the second best predictor by 7.84% and 13.41% OA on testing data. This impressive success rates reflect the robustness and high generalization power of our developed definsin peptide-based model.

Table 6
Baseline and Stack models evaluation on the hybrid features.

Model/Class	Pr	Sn (%)	Acc (%)	Sp (%)	F1 (%)
CatBoost					
P1	1	1	1	1	1
P2	1	1	1	1	1
P3	1	1	1	1	1
P4	0.94	0.63	0.95	1	0.67
P5	0.91	1	0.95	0.91	0.93
Avg	0.97	0.926	0.971	0.982	0.913
DT					
P1	0.85	0.92	0.95	0.96	0.88
P2	0.71	0.83	0.95	0.97	0.77
P3	0.89	1	0.98	0.98	0.94
P4	0.5	0.25	0.88	0.97	0.33
P5	0.88	0.91	0.89	0.88	0.89
Avg	0.766	0.782	0.93	0.952	0.762
GNB					
P1	1	0.92	0.98	1	0.96
P2	0.67	1	0.95	0.95	0.8
P3	1	1	1	1	1
P4	1	0.63	0.95	1	0.77
P5	0.97	1	0.98	0.97	0.98
Avg	0.928	0.91	0.972	0.984	0.902
MLP					
P1	0.92	1	0.98	0.98	0.96
P2	0.75	1	0.97	0.97	0.86
P3	1	1	1	1	1
P4	1	0.63	0.95	1	0.77
P5	0.97	0.97	0.97	0.97	0.97
Avg	0.928	0.92	0.974	0.984	0.912
xGB					
P1	1	1	1	1	1
P2	0.86	1	0.98	0.98	0.92
P3	1	1	1	1	1
P4	1	0.5	0.94	1	0.67
P5	0.91	1	0.95	0.91	0.96
Avg	0.954	0.9	0.974	0.978	0.91
Stack Ensemble					
P1	1	1	1	1	1
P2	0.75	1	0.97	0.97	0.86
P3	1	1	1	1	1
P4	1	0.63	0.95	1	0.77
P5	0.97	1	0.98	0.97	0.98
Avg	0.94	0.93	0.98	0.99	0.92

Table 7
Classifiers evaluation on the optimal features.

Model/Class	Pr	Sn	Acc	Sp	F1
CatBoost					
P1	1	1	1	1	1
P2	1	1	1	1	1
P3	1	1	1	1	1
P4	0.95	0.75	0.95	1	0.86
P5	0.94	1	0.97	0.94	0.97
Avg	0.978	0.95	0.984	0.988	0.966
DT					
P1	0.85	0.92	0.95	0.96	0.88
P2	0.67	1	0.95	0.95	0.8
P3	0.67	0.75	0.92	0.95	0.71
P4	0.67	0.5	0.91	0.97	0.57
P5	0.97	0.88	0.92	0.97	0.92
Avg	0.766	0.81	0.93	0.96	0.776
GNB					
P1	0.91	0.83	0.95	0.98	0.87
P2	0.67	0.67	0.94	0.97	0.67
P3	1	0.88	0.98	1	0.93
P4	0.86	0.75	0.95	0.98	0.8
P5	0.89	0.97	0.92	0.88	0.93
Avg	0.866	0.82	0.948	0.962	0.84
MLP					
P1	1	0.92	0.98	1	0.96
P2	1	0.83	0.98	1	0.91
P3	0.89	1	0.98	0.98	0.94
P4	1	0.75	0.97	1	0.86
P5	0.91	1	0.95	0.91	0.96
Avg	0.96	0.9	0.972	0.978	0.926
xGB					
P1	1	0.92	0.98	1	0.96
P2	0.86	1	0.98	0.98	0.92
P3	0.89	1	0.98	0.98	0.94
P4	1	0.63	0.95	1	0.77
P5	0.91	0.97	0.94	0.91	0.94
Avg	0.932	0.904	0.966	0.974	0.906
Stack Ensemble					
P1	1	1	1	1	1
P2	1	1	1	1	1
P3	1	1	1	1	1
P4	1	0.75	0.97	1	0.86
P5	0.94	1	0.97	0.94	0.97
Avg	0.99	0.95	0.99	0.99	0.97

Table 8
Performance comparison of the proposed methods.

Method	Features	Family type	Sn (%)	Sp (%)	F1 (%)	OA (%)
iDPF-PseRAAAC	169D	Insect (P1)	90	97.07	-	85.59
		Invertebrate (P2)	61.76	97.32	-	
		Plant (P3)	90.48	98.97	-	
		Unclassified (P4)	40	96.63	-	
		Vertebrate (P5)	99.36	88.64	-	
iDEF-PseRAAC	329D	Insect (P1)	96.67	98.13	-	91.16
		Invertebrate (P2)	74.19	97.64	-	
		Plant (P3)	92.86	98.6	-	
		Unclassified (P4)	68.42	97.23	-	
		Vertebrate (P5)	97.45	97.08	-	
StackDPPred	20D	Insect (P1)	100	100	100	**99.00**
		Invertebrate (P2)	100	100	100	
		Plant (P3)	100	100	100	
		Unclassified (P4)	75	100	86	
		Vertebrate (P5)	100	94	97	

4. Conclusion

In this study, we developed an accurate machine learning-based tool, StackDPPred, for predicting the functional types of DPs from sequence only. The proposed method uses compositional-based (SAAC), graphical-based (FEGS) and evolutionary-based (HOGPSSM and SegPSSM) properties as feature descriptors. Then, PCA algorithm was employed to remove the noisy features. Finally, the optimal features are input to stack ensemble model to predict DP's families with high accuracy. Further, we analyzed the interpretability of ML classifiers by LIME algorithm. Thus, StackDPPred protocol can provide valuable insights in accelerating the discovery of novel DPs in particularly and other therapeutic peptides in general. Despite the superior performance, the StackDPPred method also has some research gaps that needs to address in near future: (a) small number of samples (b) reliance on handcrafted feature descriptors (c) imbalance phenomena among different classes (d) deploying an interpretable ensemble deep learning algorithms using pre-trained protein language models.

Funding

The open access publication of this article was funded by Qatar National Library (QNL), Doha, Qatar.

Declaration of competing interest

I, Muhammad Arif, hereby declare that the authors have no conflict of interest.

Data availability

Dataset is publicly available.

Acknowledgements

This work was supported by the College of Science and Engineering, Hamad Bin Khalifa University (HBKU), Doha 34110, Qatar.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ymeth.2024.08.001.

References

[1] M. Erdem Büyükkiraz, Z. Kesmen, Antimicrobial peptides (amps): a promising class of antimicrobial compounds, J. Appl. Microbiol. 132 (2022) 1573–1596.
[2] X. Gao, et al., Defensins: the natural peptide antibiotic, Adv. Drug Deliv. Rev. 179 (2021) 114008.
[3] B.S. Fazly Bazzaz, S. Seyedi, N. Hoseini Goki, B. Khameneh, Human antimicrobial peptides: spectrum, mode of action and resistance mechanisms, Int. J. Pept. Protein Res. 27 (2021) 801–816.
[4] T.M. Shafee, F.T. Lay, T.K. Phan, M.A. Anderson, M.D. Hulett, Convergent evolution of defensin sequence, structure and function, Cell. Mol. Life Sci. 74 (2017) 663–682.
[5] B. He, Z. Huang, C. Huang, E.C. Nice, Clinical applications of plasma proteomics and peptidomics: towards precision medicine, Proteomics Clin. Appl. 16 (2022) 2100097.
[6] H. Steen, M. Mann, The ABC's (and XYZ's) of peptide sequencing, Nat. Rev. Mol. Cell Biol. 5 (2004) 699–711.
[7] K. Hilpert, et al., Screening and characterization of surface-tethered cationic peptides for antimicrobial activity, Chem. Biol. 16 (2009) 58–69.
[8] K. Wüthrich, Nmr with proteins and nucleic acids, Europhys. News 17 (1986) 11–13.
[9] M. Nedyalkova, A.S. Paluch, D.P. Vecini, M. Lattuada, Progress and future of the computational design of antimicrobial peptides (amps): bio-inspired functional molecules, Digit. Discov. 3 (2024) 9–22.
[10] S. Ramya Kumari, R. Badwaik, V. Sundararajan, V.K. Jayaraman, Defensinpred: defensin and defensin types prediction server, Prot. Peptide Lett. 19 (2012) 1318–1323.
[11] Y. Zuo, et al., iDPF-PseRAAAC: a web-server for identifying the defensin peptide family and subfamily using pseudo reduced amino acid alphabet composition, PLoS ONE 10 (2015) e0145541.
[12] Y. Zuo, et al., iDPF-PseRAAAC: identifying the defensin peptide by using reduced amino acid composition descriptor, Evol. Bioinform. 15 (2019) 1176934319867088.
[13] M. Arif, et al., iMRSAPred: improved prediction of anti-mrsa peptides using physicochemical and pairwise contact-energy properties of amino acids, ACS Omega 9 (2024) 2874–2883.
[14] S. Musleh, M. Arif, N.M. Alajez, T. Alam, Unified mrna subcellular localization predictor based on machine learning techniques, BMC Genomics 25 (2024) 151.
[15] M. Arif, G. Fang, A. Ghulam, S. Musleh, T. Alam, Dpi_cdf: druggable protein identifier using cascade deep forest, BMC Bioinform. 25 (2024) 145.
[16] F. Ge, et al., Vpatho: a deep learning-based two-stage approach for accurate prediction of gain-of-function and loss-of-function variants, Brief. Bioinform. 24 (2023) bbac535.
[17] S. Seebah, et al., Defensins knowledgebase: a manually curated database and information source focused on the defensins family of antimicrobial peptides, Nucleic Acids Res. 35 (2007) D265–D268.
[18] J. Hu, et al., Improving DNA-binding protein prediction using three-part sequence-order feature extraction and a deep neural network algorithm, J. Chem. Inf. Model. 63 (2023) 1044–1057.
[19] F. Ge, et al., Review of computational methods and database sources for predicting the effects of coding frameshift small insertion and deletion variations, ACS Omega 9 (2024) 2032–2047.
[20] F. Ge, J. Hu, Y.-H. Zhu, M. Arif, D.-J. Yu, Targetmm: accurate missense mutation prediction by utilizing local and global sequence information with classifier ensemble, Comb. Chem. High Throughput Screen. 25 (2022) 38–52.
[21] F. Ge, et al., Mmpatho: leveraging multilevel consensus and evolutionary information for enhanced missense mutation pathogenic prediction, J. Chem. Inf. Model. 63 (2023) 7239–7257.
[22] Z. Mu, et al., Fegs: a novel feature extraction model for protein sequences and its applications, BMC Bioinform. 22 (2021) 1–15.
[23] K. Nakai, A. Kidera, M. Kanehisa, Cluster analysis of amino acid indices for prediction of protein structure and function, Protein Eng. Des. Sel. 2 (1988) 93–100.
[24] S. Kawashima, et al., Aaindex: amino acid index database, progress report 2008, Nucleic Acids Res. 36 (2007) D202–D205.
[25] M. Hayat, A. Khan, Memhyb: predicting membrane protein types by hybridizing SAAC and pssm, J. Theor. Biol. 292 (2012) 93–102.
[26] M. Arif, M. Hayat, Z. Jan, iMem-2LSAAC: a two-level model for discrimination of membrane proteins and their types by extending the notion of SAAC into chou's pseudo amino acid composition, J. Theor. Biol. 442 (2018) 11–21.
[27] M. Hayat, A. Khan, Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition, J. Theor. Biol. 271 (2011) 10–17.
[28] Z.U. Khan, M. Hayat, M.A. Khan, Discrimination of acidic and alkaline enzyme using chou's pseudo amino acid composition in conjunction with probabilistic neural network model, J. Theor. Biol. 365 (2015) 197–203.
[29] M. Arif, et al., Targetcpp: accurate prediction of cell-penetrating peptides from optimized multi-scale features using gradient boost decision tree, J. Comput.-Aided Mol. Des. 34 (2020) 841–856.
[30] M. Kabir, et al., Intelligent computational method for discrimination of anticancer peptides by incorporating sequential and evolutionary profiles information, Chemom. Intell. Lab. Syst. 182 (2018) 158–165.
[31] F. Ali, et al., 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.
[32] M. Arif, et al., Pred-bvp-unb: fast prediction of bacteriophage virion proteins using un-biased multi-perspective properties with recursive feature elimination, Genomics 112 (2020) 1565–1574.
[33] S.F. Altschul, et al., Gapped blast and psi-blast: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[34] M. Arif, et al., Deepcppred: a deep learning framework for the discrimination of cell-penetrating peptides and their uptake efficiencies, IEEE/ACM Trans. Comput. Biol. Bioinform. 19 (2021) 2749–2759.
[35] C.E. Thomaz, G.A. Giraldi, A new ranking method for principal components analysis and its application to face image analysis, Image Vis. Comput. 28 (2010) 902–913.
[36] L. Li, S. Liu, Y. Peng, Z. Sun, Overview of principal component analysis algorithm, Optik 127 (2016) 3935–3944.
[37] M. Arif, et al., Stackacpred: prediction of anticancer peptides by integrating optimized multiple feature descriptors with stacked ensemble approach, Chemom. Intell. Lab. Syst. 220 (2022) 104458.
[38] N. Schaduangrat, N. Homdee, W. Shoombuatong, Stacker: a novel smiles-based stacked approach for the accelerated and efficient discovery of erα and erβ antagonists, Sci. Rep. 13 (2023) 22994.
[39] C. Lei, Z. Lu, M. Wang, M. Li, Stackcpa: a stacking model for compound-protein binding affinity prediction based on pocket multi-scale features, Comput. Biol. Med. 164 (2023) 107131.
[40] M. Harun-Or-Roshid, K. Maeda, B. Manavalan, H. Kurata, et al., Stack-dhupred: advancing the accuracy of dihydrouridine modification sites detection via stacking approach, Comput. Biol. Med. 169 (2024) 107848.
[41] X.-W. Liu, et al., iPADD: a computational tool for predicting potential antidiabetic drugs using machine learning algorithms, J. Chem. Inf. Model. 63 (2023) 4960–4969.
[42] H. Lin, Computational methods and resources in biological and medical data, Curr. Med. Chem. 29 (2022) 786–788.

[43] H. Zulfiqar, et al., Deep-stp: a deep learning-based approach to predict snake toxin proteins by using word embeddings, Front. Med. 10 (2023).
[44] C.-Y. Ma, et al., Predicting coronary heart disease in Chinese diabetics using machine learning, Comput. Biol. Med. 169 (2024) 107952.
[45] J. Hu, Z. Li, B. Rao, M.A. Thafar, M. Arif, Improving protein-protein interaction prediction using protein language model and protein network features, Anal. Biochem. 115550 (2024).
[46] P. Charoenkwan, et al., StackIL6: a stacking ensemble model for improving the prediction of il-6 inducing peptides, Brief. Bioinform. 22 (2021) bbab172.
[47] P. Charoenkwan, C. Nantasenamat, M.M. Hasan, W. Shoombuatong, iTTCA-Hybrid improved and robust identification of tumor t cell antigens by utilizing hybrid feature representation, Anal. Biochem. 599 (2020) 113747.
[48] P. Charoenkwan, et al., Stackdppiv: a novel computational approach for accurate prediction of dipeptidyl peptidase iv (dpp-iv) inhibitory peptides, Methods 204 (2022) 189–198.
[49] H. Zulfiqar, et al., Identification of potential inhibitors against sars-cov-2 using computational drug repurposing study, Curr. Bioinform. 16 (2021) 1320–1327.
[50] S. Ahmed, et al., An integrated feature selection algorithm for cancer classification using gene expression data, Comb. Chem. High Throughput Screen. 21 (2018) 631–645.
[51] F. Ge, A. Muhammad, D.-J. Yu, Deepnssnps: accurate prediction of non-synonymous single-nucleotide polymorphisms by combining multi-scale convolutional neural network and residue environment information, Chemom. Intell. Lab. Syst. 215 (2021) 104326.
[52] M.M. Hussein, S. Musleh, H.R. Al-Absi, M. Arif, T. Alam, Dtbapred: improved prediction drug-target binding affinity using machine learning approach, in: 2023 3rd International Conference on Computing and Information Technology (ICCIT), IEEE, 2023, pp. 319–324.