BMC Bioinformatics

RESEARCH Open Access

StackedEnC-AOP: prediction of antioxidant proteins using transform evolutionary and sequential features based multi-scale vector with stacked ensemble learning

Gul Rukh[1], Shahid Akbar[2,3], Gauhar Rehman[1], Fawaz Khaled Alarfaj[4] and Quan Zou[2,5*]

*Correspondence:
zouquan@nclab.net

[1] Department of Zoology, Abdul Wali Khan University Mardan, Mardan 23200, KP, Pakistan
[2] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, People's Republic of China
[3] Department of Computer Science, Abdul Wali Khan University Mardan, Mardan 23200, KP, Pakistan
[4] Department of Management Information Systems (MIS), School of Business, King Faisal University (KFU), 31982 Al-Ahsa, Saudi Arabia
[5] Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou 324000, People's Republic of China

Abstract

Background: Antioxidant proteins are involved in several biological processes and can protect DNA and cells from the damage of free radicals. These proteins regulate the body's oxidative stress and perform a significant role in many antioxidant-based drugs. The current invitro-based medications are costly, time-consuming, and unable to efficiently screen and identify the targeted motif of antioxidant proteins.

Methods: In this model, we proposed an accurate prediction method to discriminate antioxidant proteins namely StackedEnC-AOP. The training sequences are formulation encoded via incorporating a discrete wavelet transform (DWT) into the evolutionary matrix to decompose the PSSM-based images via two levels of DWT to form a Pseudo position-specific scoring matrix (PsePSSM-DWT) based embedded vector. Additionally, the Evolutionary difference formula and composite physiochemical properties methods are also employed to collect the structural and sequential descriptors. Then the combined vector of sequential features, evolutionary descriptors, and physiochemical properties is produced to cover the flaws of individual encoding schemes. To reduce the computational cost of the combined features vector, the optimal features are chosen using Minimum redundancy and maximum relevance (mRMR). The optimal feature vector is trained using a stacking-based ensemble meta-model.

Results: Our developed StackedEnC-AOP method reported a prediction accuracy of 98.40% and an AUC of 0.99 via training sequences. To evaluate model validation, the StackedEnC-AOP training model using an independent set achieved an accuracy of 96.92% and an AUC of 0.98.

Conclusion: Our proposed StackedEnC-AOP strategy performed significantly better than current computational models with a ~ 5% and ~ 3% improved accuracy via training and independent sets, respectively. The efficacy and consistency of our proposed StackedEnC-AOP make it a valuable tool for data scientists and can execute a key role in research academia and drug design.

Keywords: Antioxidant proteins, Transformation, Evolutionary features, Stacked ensemble model, Feature selection, Prediction

Introduction

Oxidation is a chemical reaction available in various biological and non-biological processes. It uses several oxidizing agents to form electrons or hydrogen atoms from substances [1]. Free radicals, hazardous byproducts, and compounds with unpaired, unstable, and extremely reactive electrons are also produced during this process [2]. Whereas, these molecules act as oxidants by pairing or accepting the free-electrons from other molecules [3]. In contrast, a low level of free radicals is significant for several biological activities, including immunity, differentiation, cell death, protein phosphorylation, and transcription factor activation [4]. However, excessive concentrations have the potential to negatively impact cell functions. It caused several reactions with essential biological cellular components i.e., proteins, lipids, RNA/DNA, and carbohydrates. Reactive oxygen species (ROS) are produced of free radicals containing oxygen, that are essential for maintaining cell signaling, and homeostasis [5]. In healthy cells, the ROS ratio is typically low and involves many intricate metabolic procedures. ROS levels in cells are extremely raised and cause internal cell damage, when the organism is subjected to environmental stresses. The damage happened because of an excessive concentration known as oxidative stress. Oxidative stress is an oxidative imbalance stemming from the inability to purify their reactive products, which occurs by the production of ROS during cellular metabolism [6]. Oxidative stress causes many harmful disorders in humans [7], including cardiac failures [8], Parkinson's [9], Alzheimer's [10], hypertension [11], and cancer [12]. To continuously monitor the ROC formation, cells have developed antioxidant system-based procedures to effectively resist the damages that occur because of ROS [13]. By neutralizing free radicals, antioxidants can decrease the responses of oxygen-free radicals [14], which is essential for maintaining the body's redox balance by preventing food deterioration and safeguarding against the aging process [15]. To demonstrate the potent antioxidant activities, several artificial antioxidants have been employed. However, due to numerous health risks, these are considered ineffective in some domains [16]. Antioxidant proteins are also essential for screening and developing antioxidant medications that are used to treat a variety of diseases and issues associated with aging [17]. Additionally, it helps in repairing DNA damage caused by free radicals [18]. Antioxidant proteins have been successfully investigated using a variety of empirical techniques, including spectroscopic analysis [19, 20], electrochemical [21], electrophoresis [22], and chromatography [23]. However, due to their high processing time, high chance of experimental failures, and high costs when handling large amounts of data, these conventional biochemical processes are not considered suitable [24–26]. Recently, prediction of the bioinformatics data using machine learning and deep learning-based computational models have performed significantly due to their reliability, efficiency, and high training and validation performance [27]. The comprehensive literature review of the existing machine learning-based computational models are described as following.

In recent decades, with the huge expansion in genomics sequences, researchers have diverted their directions toward computational model-based alternatives for predicting different protein types. Numerous novel computational methods have been presented for predicting antioxidant proteins (AOPs) [28]. Feng et al. [29] developed an AOD database for antioxidant proteins which is useful for researchers and performs a pivotal role

in identifying antioxidant proteins. Initially, Blanco et al. [30] presented a random forest-based prediction model for AOPs. The protein samples were formulated using topological index-based star graph networks. Later on, Feng et al. [31] proposed an amino acid residual frequency-based sequential model for AOPs. The best feature set was selected using a correlation filter-based method, and then the Naive Bayes model was trained by achieving an accuracy of 66.68% using the jackknife test. Again, Feng et al. [32] developed an AodPred webserver for identifying AOPs. The pseudo-g-gapped dipeptide-based features were trained via the SVM model and achieved an accuracy of 74.79% using the jackknife test. Similarly, Zhang et al. [33] applied a gapped dipeptide and PSSM for encoding primary sequences. Information gain integrated with incremental selection was utilized to select the best features and trained using the RF model. On the other hand, Zhang et al. [34] used an ensemble approach for predicting AOPs. Various best features were chosen from the integrated vector of evolutionary, physiochemical, and structured properties-based methods. The proposed model attained an accuracy of 94% and a sensitivity of 95%. Additionally, Lei et al. [35] proposed the SeqSVM-based computational model for AOPs. SeqSVM used eight different physiochemical properties to formulate sequences, and then MRMD was used for selecting optimal features. Furthermore, Li et al. [13] developed an SVM-based vote9 method for AOPs. Vote9 numerically formulated the amino acid samples using gapped dipeptide and cluster profile-based methods. The optimal set was chosen by applying the ANOVA-IFS. Likewise, Chao et al. presented the AOPs-SVM model using IFS & MRMD-based ensemble feature selection [36]. The selected descriptors were trained via the SVM model and achieved an accuracy of 94.2%. Butt et al. [37] employed the Chou's pseudo amino acid composition and statistical movement features for AOPs. A tenfold CV based multilayer neural network was applied to evaluate the model. Ahmad et al. [38] employed a K-space amino acid pair (KSAAP) with SFS-SVM-based model for predicting AOPs. Likewise, Thanh Lam et al. [39] applied different sequential frequency residue-based feature encoding. The extracted vectors were trained using different machine-learning models and attained an accuracy of 84.6%. Tran et al. [40]; proposed the iAnt method for identifying AOPs using the ensemble training strategy of the CNN and RF training model. In the AOPM model, the 188D vector of physiochemical properties and KSAAP-based pairing sequential features were trained using the RF model [41]. Recently, Meng presented the SVM-based training model called DP-AOP for predicting AOPs [42]. The protein samples were represented samples using secondary structure and evolutionary feature formulations. The optimal features were selected using dynamic programming, choosing from the ranked features using MRMD. Apart from these, several other predictors, i.e., AoP-LSE [43], PredAoDP [44], AnOxPP [45], ANPrAod [46], and AnOxPePred [47] were also recently developed for predicting AOPs.

After thoroughly investigating all the aforementioned computational models, we observed that every model actively and significantly contributes to the prediction of AOPs. However, these approaches still have generalization and reliability issues.

- The current methods employed sequential formulation methods that mainly focus on calculating frequencies of amino acids based on the residue composition by ignoring the sequence order of amino acids.

- Several methods presented conventional evolutionary information, which requires a significant processing time to search similarity matrix for every protein sequence in huge databases.
- Another problem is training with the imbalanced dataset, resulting in biased predictive results towards the majority class by ignoring the minority class, increasing the risk of underfitting.
- The majority of existing methods were trained using conventional classifiers. However, stacked-ensemble training models have recently outperformed classical machine learning models.

Hence, in addressing such concerns, more improvement is needed in presenting alternate computational models that can accurately discriminate AOPs and non-AOPs with high throughput.

In this paper, we developed a stacked-ensemble model, StackEnC-AOP, to predict antioxidant peptides (AOPs). The protein samples are numerically encoded via Evolutionary difference formula features (EEDP) and composite physicochemical properties (CPP). Apart from these, we embedded the level-based discrete wavelet transformation approach into the pseudo-position-specific scoring matrix to generate the enhanced evolutionary features called PsePSSM-DWT. Moreover, the CPP, EEDP, and PsePSSM-DWT vectors are fused to form the multi-informative vector. The minimum redundancy and maximum relevance (mRMR), a filter-based feature selection is then employed to gather optimal features by removing irrelevant and duplicated features. Finally, our proposed training model is passed through two stages. Initially, four baseline classifiers, such as XGBoost (XGB), Decision Tree (DT), RF, and SVM are individually applied for model training. Subsequently, the predicted outcomes of the baseline classifiers are provided to the logistic regression (LR) to develop the stacked-ensemble model [48]. The developed StackEnC-AOP demonstrated remarkable and achieved improved predictions using training and independent samples. The detailed architecture of our StackEnC-AOP model is provided in Fig. 1.

Material and methods

Dataset

In deep learning and bioinformatics, choosing an appropriate training dataset is essential for developing an intelligent prediction model [32, 36, 44]. The choice of benchmark dataset has a major effect on the performance of a computational model. In this study, we used a training dataset, previously created by Feng et al. [31]. The dataset is prepared by following several steps such as: (a) the obtained protein samples have validated antioxidant activities, (b) useless letters i.e., "B", "U", and "X" were removed from the protein samples, (c) to remove overfitting the homologous samples were eradicated via a CD-HIT tool by keeping the threshold of 0.60 [49]. Hence, a training dataset comprised 1805 sequences containing 253 AOPs and 1552 non-AOPs. The same training sequences have been utilized for developing several AOP models i.e., as AodPred [32], ANPrAod [46], and PredAoDP [44]. Furthermore, to assess the reliability of the trained model, an independent set is also employed. The independent set includes unseen samples with 73

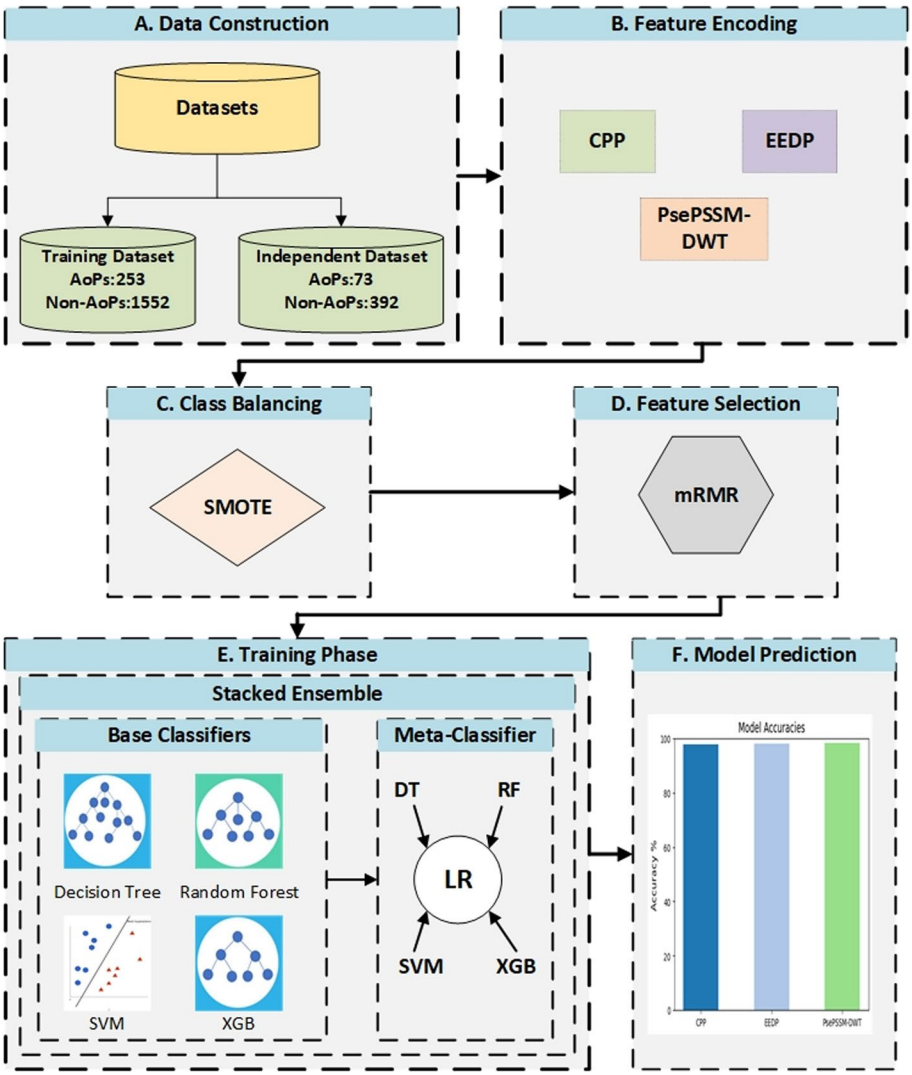

Fig. 1 The proposed architecture of the StackEnC-AOP model

AOPs and 392 non-AOPs [34]. To ensure the generalization of the training model, none of the sequences from the training data were repeated in the independent dataset.

Feature encoding schemes

Pseudo position specific scoring matrix (Pse-PSSM)

Position Specific Scoring Matrix (PSSM) information produces the evolutionary descriptors of every peptide sequence [50]. The key issue of our proposed work is imbalanced training classes and high variation in the protein sequences created in model training [51]. Furthermore, the sequence ordering and correlation characteristics of the protein sequence cannot preserved by the traditional PSSM features [52]. As a result, PsePSSM using a variety of protein sequences creates a consistent vector length. PsePSSM uses the correlation of the amino acid residues separated by 'd' amino acids to compute the mean score of each residue amino acid in the PSSM matrix by determining the correlation between residues separated by 'd' amino acids.

The PsePSSM feature space for a protein sequence can be represented as:

$$A_{PsePSSM} = (\overline{A_1}, \overline{A_2}, ...\overline{A_{20}}, \phi_1^1, \phi_2^1, ...\phi_{20}^1, ...\phi_1^{lag}, \phi_2^{lag}, ...\phi_{20}^{lag})^T \quad (1)$$

where $\overline{A_s} = \sum_{r=1}^{L} {}^{A_{r,s}}/_{L}$ $(r = 1, 2, 3, ...20)$, $\overline{A_q}$ denotes the mean score of all amino acid residues. Which are mutated to s amino acid in peptide sample 'A.'

$$\phi_s^{lag} = \frac{1}{L - lag} \sum_{r=1}^{L-lag} (A_{r,s} - A_{(r+lag),s})^2, \quad s = 1 \rightarrow 20, \quad lag < L \text{ and } lag \neq L \quad (4)$$

where ϕ_s^{lag} is the sequence ordering details of the peptide sample, q represents the amino acid, and lag is the contiguous distance.

Discrete wavelet transform (DWT)

Previously, Nanni et al. introduced DWT approach to represent the residue frequency and internally capture the intrinsic information from the protein-based images [53]. At first, the computed PSSM matrix of each protein sequence is then represented in the form of the image. Subsequently, the DWT-based image denoising and compression technique is employed to divide the PSSM images into different levels to discover their hidden patterns [54]. At each level, the PSSM images are divided into two sub-wavelets such as approximation coefficients, and detailed coefficients [55]. Whereas, the low-frequency components can be represented via approximation coefficients, and high-frequency components are represented using the detailed coefficients. Whereas, it was observed from the previous methods that high-frequency components are less informative than low-frequency components. Therefore, in order to thoroughly assess each decompose each image into further levels to collect high discriminative intrinsic features.

Mathematically DWT can be formulated as:

$$A(r, s) = \frac{1}{\sqrt{r}} \int_0^j y(j) \psi \left(\frac{j - s}{r} \right) d_j \quad (5)$$

where $y(j)$ is the input signal, $A(r, s)$ represents the transform coefficients, $\psi(\frac{j-s}{r})$ is the wavelet function, and r, s denotes the scaling and translation variables, respectively.

In this paper, we employed two-level DWT decomposition to collect noiseless and high discriminative features to propose a novel extraction method for anti-oxidant proteins namely PsePSSM-DWT.

Evolutionary difference formula features (EEDP)

EEDP is an evolutionary feature engineering technique that was originally introduced for discriminating protein structure classes [56]. It measures the residue scores of the adjoining triads of amino acids to retrieve structure features of the protein sequences, particularly the protein sequences with low similarity [57]. As a result, for each protein sample, a feature set of 400D is extracted [58]. Firstly, the average evolutionary score is determined by the adjacent triads, as follows:

$$mean_a = \frac{A_{j-1,m} + A_{j,t}}{2}, mean_b = \frac{A_{j,t} + A_{j+1,n}}{2} \quad 1 \leq j \leq L \quad 1 \leq \text{m,t,n} \leq 20 \tag{6}$$

where $A_{j+1,n}$, $A_{j,t}$, and $A_{j-1,m}$ are the components in M_{PSSM}, $mean_b$ denotes the residue score among $j+1$ and j, and $mean_a$ denotes the residue score among j and $j-1$. The residue scoring matrix "AED" can be computed as below:

$$\text{AED}_{j-1,j+1} = (mean_a - mean_b)^2 = \left(\frac{e_{j-1,m} + e_{j+1,n}}{2}\right)^2 \tag{7}$$

Finally, the EEDP space can be represented using the expression:

$$r_{m,n} = \frac{1}{L-2}\sum_{j=2}^{L} AED_{j-1,j+1}, 1 \leq m,n \leq 20 \quad 1 \leq m,n \leq 20 \tag{8}$$

$$Vec_{EEDP} = \left\{r_{1,1}, .., r_{1,20}, r_{2,1}, ..., r_{2,20}, .., r_{m,n}..., r_{20,1}, .., r_{20,20}\right\} \tag{9}$$

Composite physiochemical properties (CPP)

A peptide sequence is composed of twenty distinct amino acids (AAs), and each AA possesses specific biological and physicochemical attributes [59]. These physiochemical attributes play a pivotal role in understanding the structure and behavior of amino acids. The inclusion of physiochemical properties-based information performs directly or indirectly in predicting different protein function types and their activities [60]. Consequently, we formed a CPP-based feature space that consists of eight different physiochemical properties such as hydrophobic, acidic, aromatic, hydrophilic, aliphatic, tiny, small, and charge, as provided in Table S1 of supplementary materials. Finally, a feature vector of 57*N dimensions is produced against each sample, and N denotes the total number of peptide sequences.

Syntactic minority over-sampling technique

In the field of computational science, training a model using imbalanced classes poses a challenging task [61]. High variation in data samples of a binary class problem can affect the predictive outcomes of a model by ignoring or neglecting the instances of a minority class [62]. To handle such problems, various techniques, including rescaling data samples, learning-based approaches, and hybrid techniques have been applied [63]. These methods are further categorized based on under-sampling and over-sampling techniques [64].

In this paper, we applied a SMOTE oversampling approach to address the instances of the minority class to develop a reliable predictive model [65]. SMOTE generates the synthetic samples by computing differences between the minority samples and their closest neighbors [66]. In our case, the use of SMOTE not only addresses the samples of the minority training class but also enhances the prediction outcomes through an effective balancing of both classes. Smote can be numerically represented as follows.

$$X_{new} = X_i + (\hat{X}_i - X_i) \times \beta \tag{10}$$

where X_i signifies the instances of the minority class. To create synthetic data $\hat{X}_i$ is a combination of X_i a random number 'β' ranging [0–1].

Feature selection

In computational model development, feature selection is a key step to remove the redundant and less informative features from the extracted vector that can significantly deteriorate the effectiveness of a training model, more specifically model stability. Therefore, we utilized mRMR, a filter-based selection approach. mRMR computes the correlation between two features using its mutual information. The best feature set is obtained by keeping the low redundancy between features and high relevancy with a strong relationship to the predicted classes [67]. The redundancy and relevancy (*R)* between two vectors can be calculated using the below formula:

$$R(i,j) = \iint p(i,j) \log \frac{p(i,j)}{p(i)p(j)} d_i d_j \tag{11}$$

where i and j are the two different features, $p(i,j)$ represent the joint probabilistic density function, and $p(i)\& \ p(j)$ denotes the marginal probability densities. Let us suppose, that R is the extracted features, R_a represents the selected vector included of X features, and R_b denotes the chosen feature set vector comprised of n-features. The relevancy (Rel) among the features Q in R with target L can calculated as:

$$Rel = I(Q,L) \tag{12}$$

The redundancy Red among feature Q in S_b and total features R can be calculated as:

$$Red = \frac{1}{X} \sum_{Q_i \in R} I(Q,Q_i) \tag{13}$$

The Q_k feature in R_b with minimum redundancy and maximum relevancy can be computed as:

$$\max_{x \in R_b} [I(Q_K,L) - \frac{1}{X} \sum_{Q_i \in R} I(Q,Q_i)] \quad k,1,2,,3,\ldots n \tag{14}$$

Stacked-ensemble learning

In this study, we developed a stacking-based ensemble training model to effectively predict AOPs. In the literature, a variety of models using stacked-ensemble models have been applied to produce better prediction outcomes with low generalization errors than conventional models [68–73]. Recently, Stacked ensemble predictors have shown better results using different biological data, such as non-coding RNA [74], DNA-binding proteins [75], and therapeutic peptides [76]. The stacked-based learning integrates the predicted probability scores of several baseline classifiers to develop a consistent predictor [48]. Our proposed stacked-model primarily consists of two steps. Firstly, the baseline classifiers such as XGB [77], DT [78], SVM [79], and RF [80] are trained using the extracted training features. In this paper, we applied CPP, EEDP, and

PsePSSM-DWT-based methods to formulate AOP samples. The optimal feature space is trained using the baseline classifiers using the parameters mentioned in Table 1. The grid searching technique is applied for choosing these optimal hyperparameters. Moreover, the optimal features are selected from the fused vector of EECP + CPP + PsePSSM. On the other hand, to handle model overfitting, a five-fold cross-validation (CV) test with the stratified looping mechanism is employed. Secondly, the probability scores of the baseline models are provided for the logistic regression (LR) to build a meta-classifier. Usually, the probability outcomes are between (0–1), and the threshold = 0.5 is used for predicting the targeted class of an input sample, such as probability_score higher than 0.5 will predict class A, and lower than 0.5 will predict class B. Finally, the development stacking-based ensemble model remarkably enhanced the predictive results of the proposed model than single classification models.

Performance measurement parameters

In computational models, several evaluation metrics are utilized to measure the efficacy of prediction methods [81, 82]. While evaluating the training model, we generate the confusion matrix representing the prediction in the form of true-negative (TN), true-positive (TP), false-positive (FP), and false-negative (FN). To measure the predictive power of the StackedEnC-AOP model, accuracy (ACC) is often considered the most stable parameter for evaluating the training models [83]. However, it may not be sufficient

Table 1 Hyper parameters of baseline classifiers

Classifiers	Parameter-tuning	Selected value
RF	Random_state	42
	No. of estimators	300
	Max_depth	32
	Max features	Auto
	min_samples_split	10
	min_samples_leaf	4
	bootstrap	true
DT	C	10
	Gamma	0.01
	Random_state	42
	Kernel	RBF
XGB	Learning rate	0.001
	reg_lambda	2
	max depth	15
	No. of estimators	300
	Gamma	1
	objective function	binary-logistic
	reg_alpha	1
	booster	gbtree
SVM	Gamma	0.01
	C	10
	Kernel	RBF
	Random_state	42

in certain situations to validate a generalized prediction model [84, 85]. In this study, we focused on additional performance evaluation metrics, including specificity (Sp), sensitivity (Sn), Matthews's correlation coefficient (MCC), and area under the curve (AUC), to thoroughly assess our proposed model.

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{15}$$

$$Sp = \frac{TN}{TN + FP} \tag{16}$$

$$Sn = \frac{TP}{TP + FN} \tag{17}$$

$$MCC = \frac{(TN \times TP) - (FN \times FP)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{18}$$

Predicted result

This paper evaluates the prediction performance of the AOPs samples using the CV test in the individual baseline classifiers and stacked ensemble model. We computed the mean value of the CV test by repeating the stratified loop process 100 times [54, 86]. Whereas, to obtain reliable predictive outcomes, the training data is distributed in each fold randomly. Firstly, the protein samples are formulated via CPP and EEDP-based physiochemical characteristics and evolutionary-based features. Additionally, to achieve a high discriminative vector, the irrelevant and noisy features are irradiated from the PSSM profile matrix using the DWT transformation, which generates the embedded evolutionary vector. Furthermore, the multi-perspective feature vector is formed by fusing the extracted features of CPP, PsePSSM-DWT, and EEDP. The computational cost of the multi-perspective feature vector is reduced by selecting the highly relevant feature set via mRMR feature selection. All the extracted vectors (individual, hybrid, and selected) are evaluated via the individual classifiers and the stacked ensemble model. In the below subsections, the detailed predictive results of training features and independent features are discussed, comprehensively.

Analysis of baseline classifiers via different training features

Table 2 presents the performance of individual feature sets using the proposed baseline classifiers. As mentioned earlier, we numerically encoded the protein sequences using three distinct extraction methods: CPP, PsePSSM-DWT, and EEDP. To address the class imbalance issue in our training set and to reduce the majority bias in predictions, we employed SMOTE oversampling on the minority class. Next, we utilized SVM, RF, DT, and XGB classifiers to analyze the oversampled feature sets using the hyperparameter values listed in Table 1. Additionally, we provided the probability scores from these individual classifiers into a LR model to create a stacked ensemble model. Prior to oversampling, the extracted features were examined (details in supplementary information Table S1). However, due to class imbalance, significant variation was

Table 2 Prediction results of the baseline models via oversampled training features

Method	Model	ACC	Sp	Sn	MCC	AUC
CPP	DT	81.83	79.57	83.90	0.62	0.81
	XGB	82.71	81.18	89.12	0.71	0.85
	RF	87.83	89.93	86.14	0.83	0.89
	SVM	84.54	99.87	69.22	0.72	0.88
	Stacked-ensemble	89.33	90.23	88.44	0.86	0.91
EEDP	DT	77.99	72.15	83.92	0.56	0.82
	XGB	81.30	75.43	87.17	0.63	0.89
	RF	87.34	85.82	88.89	0.82	0.90
	SVM	88.18	82.49	91.87	0.82	0.91
	Stacked-ensemble	90.33	91.20	89.46	0.88	0.93
PsePSSM-DWT	DT	73.04	56.96	89.38	0.48	0.81
	XGB	88.96	85.13	92.80	0.84	0.91
	RF	89.25	92.78	86.67	0.85	0.92
	SVM	91.45	91.06	92.05	0.90	0.92
	Stacked-ensemble	93.34	95.53	94.13	0.92	0.95

observed in the evaluation parameters. The Sn values were more prone to suffer than Sp values because the model prioritized the majority class (non-AOP) during training. This resulted in a bias towards correctly predicting non-AOPs at the expense of accurately predicting AOPs. Furthermore, the predicted MCC values were unsatisfactory. Therefore, to address the bias caused by the majority class, SMOTE oversampling was employed for the minority class. The EEDP feature set using the RF and SVM classifiers achieved improved results than XGB and DT by reporting an ACC of 87.34%, and 88.18%, respectively. On the other hand, the stacked-Meta classifier using EEDP features achieved a Sp of 91.20%, and ACC and AUC of 90.33%, and 0.93, respectively. Likewise, the stacked meta-model using the CPP feature set, obtained an ACC of 89.33%, and an AUC of 0.91. The embedded evolutionary features of PsePSSM-DWT achieved the Sn of 92.05%, ACC of 91.45%, with SP, AUC, and MCC of 91.06%, 0.92, and 0.90, respectively. The stacking model via PsePSSM-DWT vector achieved better results with SP of 95.33%, Sn of 94.13%, ACC of 93.34%, and MCC and AUC of 0. 92, and 0.95, respectively. Instead of individual vectors, our applied baseline classifiers and meta-model are examined via different feature integration methods such as CPP + PsePSSM-DWT, PsePSSM-DWT + EEDP, CPP + EEDP, and CPP + EEDP + PsePSSM-DWT. The predictive outcomes of the hybrid encoding schemes via baseline classifiers and meta-model are listed in Table 3. After examining all the feature sets, our stacking-based meta-model using fused features of CPP + EEDP + PsePSSM-DWT yielded better performance rates, with a Sn of 98.64%, SP of 96.11%, ACC of 96.45%, and AUC of 0.98. After selecting features from the fused vector (CPP + PsePSSM-DWT + EEDP) using mRMR, the computational cost of the training model is reduced by choosing highly relevant features with low duplicated features. Hence, the training vector is reduced to 195D from 682D. The selected features have efficiently proved their role by predicting the input protein towards predicted labels. The prediction rates of the optimal feature set using the baseline model and the proposed meta-model are listed in Table 4. The stacked meta-model has shown more reliability by demonstrating the Sp of 98.91%, ACC of 98.40, Sn, AUC, and MCC

Table 3 Prediction results of the baseline models using different hybrid vector schemes

Encoding vector	Model	ACC	Sp	Sn	MCC	AUC
CPP + PsePSSM-DWT	DT	82.69	85.11	80.25	0.80	0.91
	XGB	91.04	88.54	92.65	0.90	0.94
	RF	90.56	93.30	87.61	0.80	0.93
	SVM	93.68	94.34	92.97	0.89	0.95
	Stacked-ensemble	94.45	94.58	93.32	0.93	0.96
CPP + EEDP	DT	85.28	82.72	88.83	0.75	0.89
	XGB	89.71	89.16	90.31	0.85	0.94
	RF	90.92	91.50	88.29	0.83	0.93
	SVM	91.29	92.82	89.64	0.90	0.95
	Stacked-ensemble	93.97	94.89	91.98	0.91	0.94
PsePSSM-DWT + EEDP	DT	83.07	84.60	80.53	0.74	0.87
	XGB	91.62	87.57	95.98	0.83	0.94
	RF	90.79	92.88	89.61	0.86	0.93
	SVM	92.29	93.20	91.31	0.91	0.94
	Stacked-ensemble	94.11	95.20	93.98	0.92	0.96
EEDP + CPP + PsePSSM-DWT	DT	86.75	89.16	85.35	0.82	0.92
	XGB	90.68	89.78	92.65	0.85	0.96
	RF	90.89	93.44	86.29	0.82	0.91
	SVM	93.65	94.27	91.98	0.92	0.95
	Stacked-ensemble	**96.45**	**96.11**	**98.64**	**0.94**	**0.98**

Bold values indicate best evaluation results as compared to other classification models

Table 4 Prediction of mRMR-based selected training features

Method	Classifiers	ACC	Sp	Sn	MCC	AUC
mRMR + hybrid features	DT	88.93	91.63	86.29	0.86	0.94
	XGB	93.49	95.93	91.16	0.91	0.96
	RF	92.11	94.63	90.54	0.88	0.98
	SVM	94.85	96.82	91.47	0.93	0.98
	Stacked-ensemble	**98.40**	**98.91**	**97.29**	**0.96**	**0.99**

Bold values indicate best evaluation results as compared to other classification models

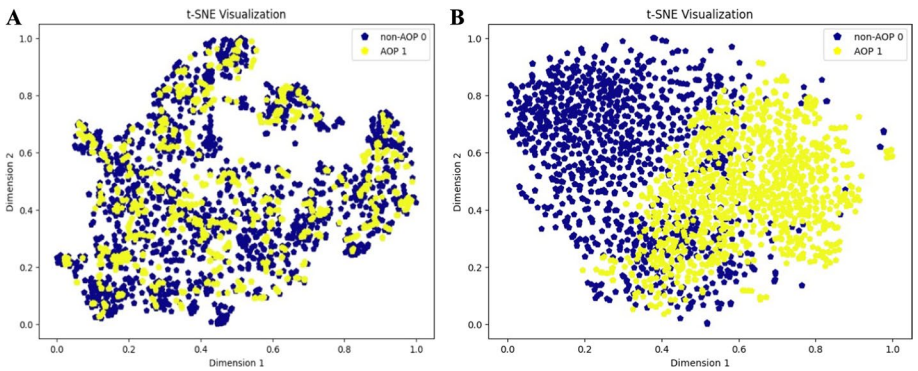

Fig. 2 t-SNE visualization of training dataset **A** hybrid features, **B** mRMR based hybrid features

of 97.29%, 0.99, and 0,96, respectively. In comparison with baseline models, the SVM and XGB yielded an ACC of 94.85%, and 93.49%, respectively. The extracted features are further visualized using the t-SNE approach to convert the high-dimension vector into

2D space, as shown in Fig. 2. In Fig. 2A, the hybrid features show some degree of overlap between positive and negative samples, which is somewhat effective but does not accurately classify the targeted classes (Fig. 3). However, in Fig. 2B, the data samples of both classes are clearly separable, demonstrating the effectiveness of mRMR-based optimal features in predicting between AOPs and non-AOPs compared to the hybrid features in Fig. 2A. Furthermore, the consistency of the training model is further assessed using Precision-recall (PR), and ROC analysis using all extracted spaces using fused vector and optimal vector as illustrated in Fig. 4.

Validation of StackEnC-AOP Model via independent sequences

The consistency and generalization of our StackEnC-AOP model are further validated using an independent set. As per the description provided in the dataset section, the independent dataset consists of 392 non-AOPs and 73 AOPs sequences. The detailed predictive outcomes of the independent dataset using the baseline classifiers and stacking-based ensemble model are listed in Table 5. Our proposed encoding scheme (hybrid features + mRMR) with stacking model yielded superior predictive rates of Sp of 97.44%, Sn of 95.79%, ACC of 96.92%, AUC and MCC of 0.98, and 0.94, respectively as provided in Fig. 5. Furthermore, to evaluate the instance-based investigation of the proposed model using the independent data, an AUC-ROC and PR-analysis are plotted as given in Fig. 4.

Visualization of StackedEnC-AOP method via SHAP and LIME interpretation

The Shapley Additive Explanation Algorithm (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) interpolation are used in our StackedEnC-AOP model to evaluate and interpret the contribution of the extracted features [48, 77, 87]. These methods use machine learning models to evaluate the contribution of each extracted feature and display the high contributory features. SHAP is a global visualization method to analyze the contribution features via aggregating its shapely values [88]. In this study, we highlighted the top 10 highly contributory features from the mRMR based selected features based on their shapely values as displayed in Fig. 6. The SHAP value distribution of each feature is represented using a row. Data point colors show the importance of each feature, the blue indicates the low values and the red is higher values. Hence, the model output can be represented by displaying the impact of each

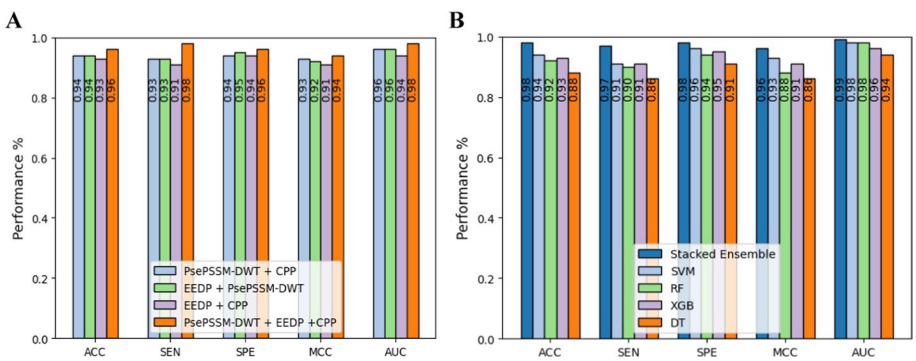

Fig. 3 Performance of training dataset using **A** hybrid vectors, **B** mRMR selected features

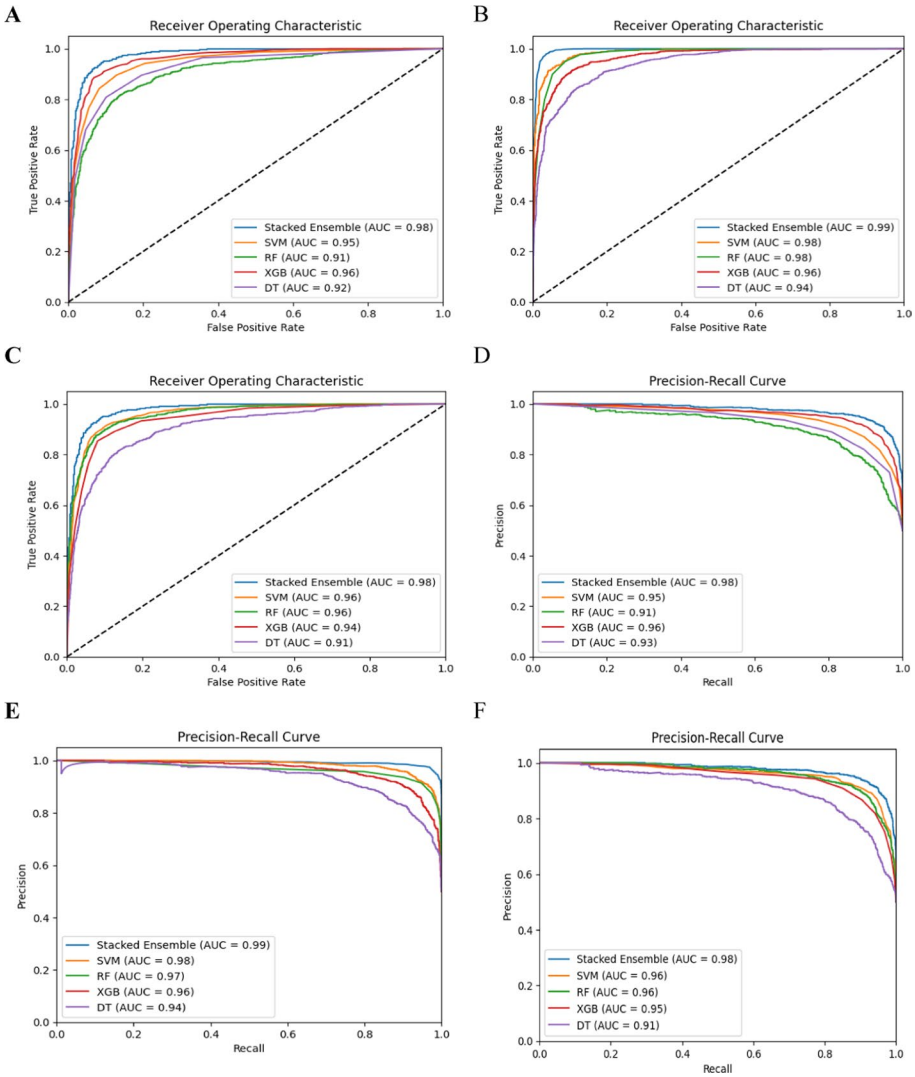

Fig. 4 ROC Analysis of **A** hybrid training features, **B** selected training features, **C** independent samples precision-recall analysis of **D** hybrid training features, **E** selected training features, **F** independent samples

Table 5 Prediction results of the StackEnC-AOP method via independent set

Models	ACC	Sp	Sn	MCC	AUC
DT	84.51	81.92	87.82	0.69	0.91
XGB	92.17	94.59	91.59	0.90	0.94
RF	91.71	92.77	90.54	0.83	0.96
SVM	93.54	95.18	92.94	0.91	0.96
Stacked-ensemble	**96.92**	**97.44**	**95.79**	**0.94**	**0.98**

Bold values indicate best evaluation results as compared to other classification models

feature via these colored dots. The SHAP value < 0 predicts the negative class (non-AOPs) and the value > 0 the positive class AOPs. Our predictive results highlight the importance of the selected features from the Hybrid vector in order to predict the targeted labels. On the other hand, the significance of instance-based prediction is

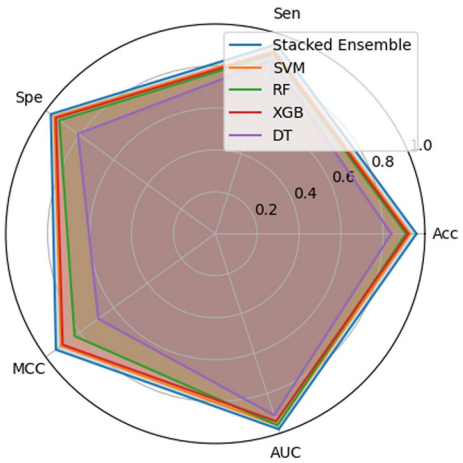

Fig. 5 Comparison of the baseline classifiers via Independent set

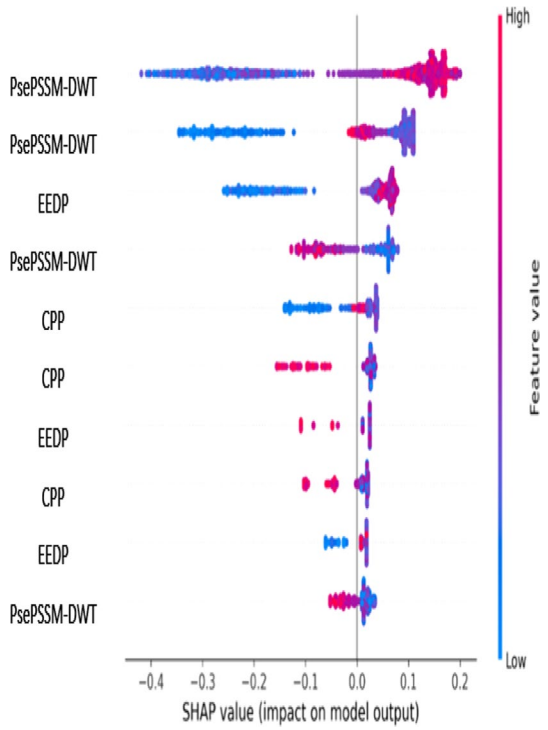

Fig. 6 SHAP interpolation of contributory features

illustrated using the LIME analysis [89]. LIME uses the feature vector permutations to simplify models. A key concern of LIME analysis is to develop the similarity matrix by determining distances between query samples and perturbed samples. It is the interpretable insights into model predictions by illustrating the contributions of the feature enhancing model effectiveness. Lime analysis is also useful for model validation, debugging, and improving decision-making by providing how specific features influence outcomes. In this study, we performed the LIME interpolation of the

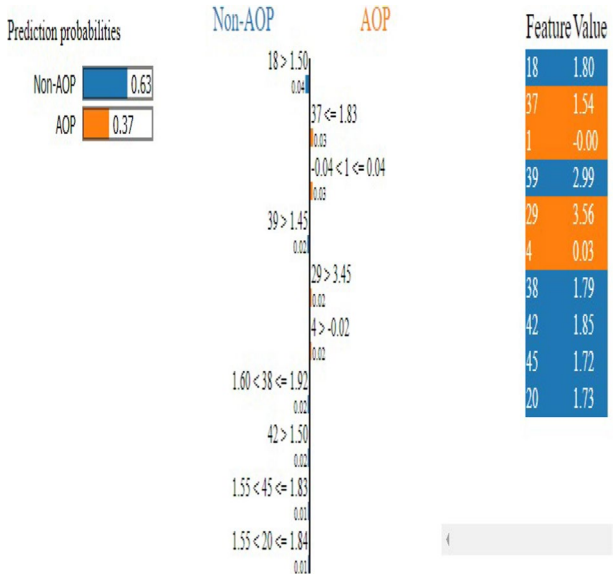

Fig. 7 LIME analysis of StackEnC-AOP model

Table 6 Comparison of StackEnC-AOP method with existing state of the art models

Dataset	Method	ACC	Sp	Sn	MCC	AUC
Training data	Feng et al. [31]	66.88	66.05	72.04	–	0.85
	AodPred [32]	74.79	74.48	75.09	–	–
	UniDL4BioPep [90]	80.40	79.90	81	0.61	0.87
	AoP-LSE [43]	82.40	84.90	67.40	0.43	–
	Ao et al. [80]	83.91	96.30	66.50	–	–
	Thanh-Lam et al. [39]	84.50	85.10	81.50	–	–
	ANPrAod [46]	87.53	98.33	92.92	–	–
	DP-AOP [42]	91.07	85.80	96.40	0.82	–
	AOPM [41]	92	94.20	87.30	0.81	0.97
	PredAoDP [44]	93.18	96.77	71.65	0.71	0.84
	StackEnC-AOP	**98.40**	**98.91**	**97.29**	**0.96**	**0.99**
Independent data	iAnt [40]	95.20	94.90	97.30	0.85	0.98
	Zhang et al.[34]	86.3	86.0	87.8	0.61	0.94
	Thanh-Lam et al. [39]	94.2	94.1	94.6	0.81	0.98
	Ahmad et al. [38]	93.71	88.15	94.14	0.92	–
	StackEnC-AOP	**96.92**	**97.44**	**95.79**	**0.94**	**0.98**

Bold values indicate best evaluation results as compared to other classification models

independent set as given in Fig. 7. LIME analysis predicts the input instance using its correlation with AOPs (red) and non-AOPs (blue).

Comparison with current methods using training and independent set

In Table 6 a detailed comparison of the StackEnC-AOP model is performed with existing studies using a training dataset as well as an independent set. Feng et al. used a correlation filter-based feature space by applying sequential encoding to formulate the training sequences [31]. The naïve Bayes model was trained using the extracted vector and yielded an ACC of 66.88%, Sn of 72.04%, Sp of 66.05%, and AUC of 0.85, respectively.

Likewise, the AodPred model applying the SVM training model and pseudo-gapped dipeptide vector achieved an ACC, Sp, and Sn 74.79%, 74.48%, and 75.09%, respectively [32]. In the UniDL4BioPep predictor, the bioactivities of eighteen different peptide classes are predicted using the pre-trained biological language model [90]. In the case of antioxidant sequences, the embedded features of the UniDL4BioPep model obtained an acc of 80.40%, with a Sn value of 81%, and Sp of 79.90%. The AoP-LSE model obtained a Sp of 84.90% and ACC of 82.40% by applying the KSAAP encoding and neural network-based deep latent features [43]. Ao et al., model formulated the protein samples using a hybrid space of embedding-evolutionary, physiochemical properties, and sequential encoding methods. The optimal features are chosen from the extracted vector by applying three feature selection approaches [80]. The proposed model obtained an ACC of 83.91%, Sn of 66.50, and SP of 96.30%. Furthermore, Lam et al. attained an accuracy of 84.50% by applying different sequential frequency residue-based feature encoding [39]. Recently, using the DP-AOP training model, the dynamic programming-based secondary structure and evolutionary features reported an ACC of 91.07%, Sn of 96.40%, and Sp of 85.80% [42]. In the AOPM model, the multi-perspective vector was trained using the RF classifier and reported a Sp, Sn, ACC, and AUC of 94.20%, 87.30%, 92%, and 0.97, respectively [41]. Likewise, the ANPrAod method obtained a Sp of 98.33%, Sn of 92.92%, and an ACC of 87.53% via reduced amino acid features and with ANOVA-based feature selection [46]. In contrast, the PredAoDP model trained the SVM model using diverse variants of evolutionary descriptors and reported a Sp, ACC, and Sn of 96.77%, 93.18%, and 71.65%, respectively [44]. Finally, our developed StackEnC-AOP model performed better, achieving a Sp of 98.91%, ACC of 98.40%, Sn of 97.29%, AUC of 0.99, and an MCC of 0.96, respectively, as listed in Table 6.

Discussion

Antioxidant proteins are small fragments or molecules that defend against the disease initiated due free radicals. Keeping the importance of AOPs in biological processes, many in-vitro and computational methods have been proposed to provide alternatives. Nevertheless, existing models have several issues. Our proposed StackEnC-AOP model uses the local evolutionary features, by representing each protein in the form of a 2D image and then decomposing each image into several levels. Our applied two-level DWT-based decomposition has provided improved performance rates by capturing the intrinsic and hidden local features that are difficult to access using traditional sequential encoding methods. In addition to the evolutionary features, EEDP and CPP-based improved sequential and physiochemical structure-based encoding methods are applied to form a hybrid vector. The hybrid feature strategy has significantly performed well and achieved the predictive ACC of 96.4%. Our compact hybrid feature vector performed better by compensating for the limitations of the individual feature vector. To reduce the executing time of the training model, the 195 mRMR-based selected features further enhanced the training accuracy to 98.40%, and AUC to 0.99.

To compare the predictive outcomes of our StackEnC-AOP with existing state-of-the-art predictors, as illustrated in Table 6. To the best of our knowledge, ten predictors have been developed using the same training samples. We categorized these methods as sequence-based, evolutionary-based, and deep features-based. In sequence-based

encoding methods, the models presented by Feng et al. [31], AodPred [32], Thanh-Lam et al. [39], ANPrAod [46], and AOPM [41] employed traditional machine learning classifiers and achieved predictive accuracies of 66.88%, 74.79%, 84.50%, and 92%, respectively. However, these sequence encoding methods only focus on the frequencies of the amino acid sequences without keeping the sequence order information. Similarly, in evolutionary-based methods, the models proposed by Ao et al. [80], DP-AOP [42], and PredAoDP [44] achieved accuracies of 83.91%, 91.07%, and 93.18%, respectively. In the DP-AOP training model, the executing cost was effectively reduced by selecting 17 optimal features. In the AoP-LSE model, the deep latent features were employed [43]. However, the deep training model features achieved an ordinary accuracy of 82.40%. Recently, in the UniDL4BioPep predictor, a protein language model-based embedding features obtained an ACC of 80.40% [90]. In contrast, our StackEnC-AOP model achieved a higher predictive accuracy of 98.40%, and an AUC of 0.99, which is approximately 5% higher accuracy than existing predictors. On the other hand, the generalization power of our StackEnC-AOP model is validated using independent samples by achieving an ACC of 96.92% and an AUC of 0.98. Which is higher than the recent four predictors such as Zhang et al.[34], Ahmad et al. [38], Thanh-Lam et al. [39], and iAnt [40]. Hence, the remarkable predictive results of the StackEnC-AOP are due to the incorporation of novel PsePSSM-DWT encoding and leveraging the powerful training abilities of the stacked-ensemble model. The PsePSSM-DWT based transformed local evolutionary features not only cover the sequence ordering issue of the existing traditional encoding schemes using PsePSSM matrix but also represent the PSSM matrix of each sequence in the form of an image using DWT, and its two-level decomposition leads to capturing hidden informative features that are easily accessible using sequential encoding methods. Finally, in comparison with traditional classifiers, the proposed stacked ensemble model provides more flexibility in selecting baseline models, and its generalization capabilities, leading to achieve robust and reliable predictions.

Conclusion

In this paper, we introduced a StackEnC-AOP training model to effectively identify antioxidant sequences using the stacking ensemble strategy. The accurate prediction of AOPs holds paramount significance in drug development and the pharmaceutical industry due to their key roles in treating various diseases. Addressing the flaws of existing feature encoding schemes, we applied a compact multi-perspective vector of the novel evolutionary, sequential, and structured features to handle the drawbacks of individual feature encoding methods. Whereas, the inclusion of the novel two-level decomposition of the evolutionary features-based images using PsePSSM-DWT performed effectively to highlight the intrinsic hidden information. Moreover, to develop a bias-free model, we oversampled the low instances class of the training dataset using the SMOTE technique. At last, the training cost of the proposed model is minimized by choosing optimal features using mRMR-based selection. Our developed StackEnC-AOP method reported the higher prediction accuracies of 98.40%, and 96.92% for training and independent sequences, respectively. Moreover, the remarkable performance of our training model using the unseen independent sequences validated the generalization power and potential biases. The consistent improvement of our stacking ensemble-based model

compared to existing computational models has a substantial impact on the drug development pipeline. Pharmaceutical industries can enhance their ability to identify, design, and develop new antioxidant-based drugs more effectively.

Abbreviations

AAs	Amino acids
ACC	Accuracy
AOPs	Antioxidant proteins
AUC	Area under the curve
CPP	Composite physicochemical properties
CV	Cross-validation
DT	Decision tree
DWT	Discrete wavelet transform
EEDP	Evolutionary difference formula features
FP	False positive
FN	False negative
LR	Logistic regression
MCC	Matthews's correlation coefficient
mRMR	Minimum redundancy and maximum relevance
PR	Precision-recall
Pse-PSSM	Pseudo Position Specific Scoring Matrix
RF	Random Forest
Sn	Sensitivity
Sp	Specificity
SVM	Support vector machine
SMOTE	Syntactic minority over-sampling technique
TN	True negative
TP	True positive
XGB	XGBoost

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05884-6.

Additional file1

Acknowledgements
The authors would like to thank the Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, China for its computational support and Funding.

Author contributions
GR performed Writing, methodology, implementation, and visualization. SA performed Writing, Model Creation, methodology, Supervision, and Validation. GR performed proof-reading, Formal analysis, Supervision and Visualization. FKA performed writing, software, visualization, and data curation. QZ performed Supervision, Idea, Writing, and Proof-Reading.

Funding
The work was supported by the National Natural Science Foundation of China (No. 62131004), and the National Key R&D Program of China (2022ZD0117700).

Availability of data and materials
The data and source code are available in the public repository: https://github.com/shahidawkum/StackedEnC-AOP.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 19 April 2024 Accepted: 29 July 2024
Published online: 04 August 2024

References

1. Phaniendra A, Jestadi DB, Periyasamy L. Free radicals: properties, sources, targets, and their implication in various diseases. Indian J Clin Biochem. 2015;30:11–26.
2. Sundaram Sanjay S, Shukla AK. Free radicals versus antioxidants. In: Sanjay SS, Shukla AK, editors. Potential therapeutic applications of nano-antioxidants. Springer: Singapore; 2021. p. 1–17.
3. Nimse SB, Pal D. Free radicals, natural antioxidants, and their reaction mechanisms. RSC Adv. 2015;5(35):27986–8006.
4. Rajendran P, Nandakumar N, Rengarajan T, Palaniswami R, Gnanadhas EN, Lakshminarasaiah U, Gopas J, Nishigaki I. Antioxidants and human diseases. Clin Chim Acta. 2014;436:332–47.
5. Jomova K, Raptova R, Alomar SY, Alwasel SH, Nepovimova E, Kuca K, Valko M. Reactive oxygen species, toxicity, oxidative stress, and antioxidants: chronic diseases and aging. Arch Toxicol. 2023;97(10):2499–574.
6. Kıran TR, Otlu O, Karabulut AB. Oxidative stress and antioxidants in health and disease. J Lab Med. 2023;47(1):1–11.
7. He P, Zhang Y, Zhang Y, Zhang L, Lin Z, Sun C, Wu H, Zhang M. Isolation, identification of antioxidant peptides from earthworm proteins and analysis of the structure–activity relationship of the peptides based on quantum chemical calculations. Food Chem. 2024;431:137137.
8. Pagan LU, Gomes MJ, Gatto M, Mota GA, Okoshi K, Okoshi MP. The role of oxidative stress in the aging heart. Antioxidants. 2022;11(2):336.
9. Chang K-H, Chen C-M. The role of oxidative stress in Parkinson's disease. Antioxidants. 2020;9(7):597.
10. Sun Q, Kong W, Mou X, Wang S. Transcriptional regulation analysis of Alzheimer's disease based on FastNCA algorithm. Curr Bioinform. 2019;14(8):771–82.
11. Liguori I, Russo G, Curcio F, Bulli G, Aran L, Della-Morte D, Gargiulo G, Testa G, Cacciatore F, Bonaduce D. Oxidative stress, aging, and diseases. Clin Interv Aging. 2018;13:757–72.
12. Reddy VP. Oxidative stress in health and disease. Biomedicines. 2023;11(11):2925.
13. Li X, Tang Q, Tang H, Chen W. Identifying antioxidant proteins by combining multiple methods. Front Bioeng Biotechnol. 2020;8:858.
14. Chaudhary P, Janmeda P, Docea AO, Yeskaliyeva B, Abdull Razis AF, Modu B, Calina D, Sharifi-Rad J. Oxidative stress, free radicals and antioxidants: potential crosstalk in the pathophysiology of human diseases. Front Chem. 2023;11:1158198.
15. Dhalaria R, Verma R, Kumar D, Puri S, Tapwal A, Kumar V, Nepovimova E, Kuca K. Bioactive compounds of edible fruits with their anti-aging properties: a comprehensive review to prolong human life. Antioxidants. 2020;9(11):1123.
16. Moulahoum H, Ghorbanizamani F, Timur S, Zihnioglu F. Beyond natural antioxidants in cancer therapy: novel synthetic approaches in harnessing oxidative stress. In: Chakraborti S, editor. Handbook of oxidative stress in cancer: therapeutic aspects. Springer: Singapore; 2022. p. 1–17.
17. Rojas-Fernandez CH, Tyber K. Benefits, potential harms, and optimal use of nutritional supplementation for preventing progression of age-related macular degeneration. Ann Pharmacother. 2017;51(3):264–70.
18. Mishra N, Tripathi S, Nahar L, Sarker SD, Kumar A: **Mitigation of arsenic poisoning induced oxidative stress and genotoxicity by Ocimum gratissimum L**. *Toxicon* 2024:107603.
19. Pisoschi AM, Negulescu GP. Methods for total antioxidant activity determination: a review. Biochem Anal Biochem. 2011;1(1):106.
20. Wachirattanapongmetee K, Katekaew S, Weerapreeyakul N, Thawornchinsombut S. Differentiation of protein types extracted from tilapia byproducts by FTIR spectroscopy combined with chemometric analysis and their antioxidant protein hydrolysates. Food Chem. 2024;437:137862.
21. Madhani Mohammed Sadhakathullah AH, Paulo Mirasol S, Molina García BG, Torras Costa J, Armelín Diggroc EA. PLA-PEG-cholesterol biomimetic membrane for electrochemical sensing of antioxidants. Electrochim Acta. 2024;476:143716.
22. Chen L, Chen S, Rong Y, Zeng W, Hu Z, Ma X, Feng S. Identification and evaluation of antioxidant peptides from highland barley distiller's grains protein hydrolysate assisted by molecular docking. Food Chem. 2024;434:137441.
23. Li W, Zhu L, Zhang F, Han C, Li P, Jiang J. A novel strategy by combining foam fractionation with high-speed countercurrent chromatography for the rapid and efficient isolation of antioxidants and cytostatics from *Camellia oleifera* cake. Food Res Int. 2024;176:113798.
24. Lv Z, Cui F, Zou Q, Zhang L, Xu L. Anticancer peptides prediction with deep representation learning features. Brief Bioinform. 2021;22(5):bbab008.
25. Lv Z, Zhang J, Ding H, Zou Q. RF-PseU: a random forest predictor for RNA pseudouridine sites. Front Bioeng Biotechnol. 2020;8:134.
26. Lv H, Dao F-Y, Zulfiqar H, Lin H. DeepIPs: comprehensive assessment and computational identification of phosphorylation sites of SARS-CoV-2 infection using a deep learning-based approach. Brief Bioinform. 2021;22(6):bbab244.
27. Olawoye B, Fagbohun OF, Popoola-Akinola O, Akinsola JET, Akanbi CT. A supervised machine learning approach for the prediction of antioxidant activities of *Amaranthus viridis* seed. Heliyon. 2024;10:e24506.
28. Meng C, Pei Y, Bu Y, Zou Q, Ju Y. Machine learning-based antioxidant protein identification model: progress and evaluation. J Cell Biochem. 2023;124:1825–34.
29. Feng P, Ding H, Lin H, Chen W. AOD: the antioxidant protein database. Sci Rep. 2017;7(1):7449.
30. Fernández-Blanco E, Aguiar-Pulido V, Munteanu CR, Dorado J. Random Forest classification based on star graph topological indices for antioxidant proteins. J Theor Biol. 2013;317:331–7.
31. Feng P-M, Lin H, Chen W. Identification of antioxidants from sequence information using naive Bayes. Comput Math Methods Med. 2013;2013:567529.
32. Feng P, Chen W, Lin H. Identifying antioxidant proteins by using optimal dipeptide compositions. Interdiscip Sci Comput Life Sci. 2016;8:186–91.
33. Zhang L, Zhang C, Gao R, Yang R. Incorporating g-gap dipeptide composition and position specific scoring matrix for identifying antioxidant proteins. In: 2015 IEEE 28th Canadian conference on electrical and computer engineering (CCECE). IEEE; 2015. p. 31–6.
34. Zhang L, Zhang C, Gao R, Yang R, Song Q. Sequence based prediction of antioxidant proteins using a classifier selection strategy. PLoS ONE. 2016;11(9):e0163274.

35. Xu L, Liang G, Shi S, Liao C. SeqSVM: a sequence-based support vector machine method for identifying antioxidant proteins. Int J Mol Sci. 2018;19(6):1773.
36. Meng C, Jin S, Wang L, Guo F, Zou Q. AOPs-SVM: a sequence-based classifier of antioxidant proteins using a support vector machine. Front Bioeng Biotechnol. 2019;7:224.
37. Butt AH, Rasool N, Khan YD. Prediction of antioxidant proteins by incorporating statistical moments based features into Chou's PseAAC. J Theor Biol. 2019;473:1–8.
38. Ahmad A, Akbar S, Hayat M, Ali F, Khan S, Sohail M. Identification of antioxidant proteins using a discriminative intelligent model of k-space amino acid pairs based descriptors incorporating with ensemble feature selection. Biocybern Biomed Eng. 2022;42(2):727–35.
39. Ho Thanh Lam L, Le NH, Van Tuan L, Tran Ban H, Nguyen Khanh Hung T, Nguyen NTK, Huu Dang L, Le NQK. Machine learning model for identifying antioxidant proteins using features calculated from primary sequences. Biology. 2020;9(10):325.
40. Tran HV, Nguyen QH. iAnt: combination of convolutional neural network and random forest models using PSSM and BERT features to identify antioxidant proteins. Curr Bioinform. 2022;17(2):184–95.
41. Zhai Y, Zhang J, Zhang T, Gong Y, Zhang Z, Zhang D, Zhao Y. AOPM: application of antioxidant protein classification model in predicting the composition of antioxidant drugs. Front Pharmacol. 2022;12:818115.
42. Meng C, Pei Y, Zou Q, Yuan L. DP-AOP: a novel SVM-based antioxidant proteins identifier. Int J Biol Macromol. 2023;247:125499.
43. Usman M, Khan S, Park S, Lee J-A. AoP-LSE: antioxidant proteins classification using deep latent space encoding of sequence features. Curr Issues Mol Biol. 2021;43(3):1489–501.
44. Ahmed S, Arif M, Kabir M, Khan K, Khan YD. PredAoDP: Accurate identification of antioxidant proteins by fusing different descriptors based on evolutionary information with support vector machine. Chemom Intell Lab Syst. 2022;228:104623.
45. Qin D, Jiao L, Wang R, Zhao Y, Hao Y, Liang G. Prediction of antioxidant peptides using a quantitative structure—activity relationship predictor (AnOxPP) based on bidirectional long short-term memory neural network and interpretable amino acid descriptors. Comput Biol Med. 2023;154:106591.
46. Xi Q, Wang H, Yi L, Zhou J, Liang Y, Zhao X, Zuo Y. ANPrAod: identify antioxidant proteins by fusing amino acid clustering strategy and-peptide combination. Comput Math Methods Med. 2021;2021:1–10.
47. Olsen TH, Yesiltas B, Marin FI, Pertseva M, García-Moreno PJ, Gregersen S, Overgaard MT, Jacobsen C, Lund O, Hansen EB. AnOxPePred: using deep learning for the prediction of antioxidative properties of peptides. Sci Rep. 2020;10(1):21471.
48. Ahmad S, Charoenkwan P, Quinn JM, Moni MA, Hasan MM, Lio' P, Shoombuatong W. SCORPION is a stacking-based ensemble learning framework for accurate prediction of phage virion proteins. Sci Rep. 2022;12(1):4106.
49. Chen Q, Wan Y, Lei Y, Zobel J, Verspoor K. Evaluation of CD-HIT for constructing non-redundant databases. In: 2016 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE; 2016. p. 703–6.
50. Ullah M, Akbar S, Raza A, Zou Q. DeepAVP-TPPred: identification of antiviral peptides using transformed image-based localized descriptors and binary tree growth algorithm. Bioinformatics. 2024;40:btae305.
51. Akbar S, Khan S, Ali F, Hayat M, Qasim M, Gul S. iHBP-DeepPSSM: identifying hormone binding proteins using Pse-PSSM based evolutionary features and deep learning approach. Chemom Intell Lab Syst. 2020;204:104103.
52. Yu B, Li S, Qiu W, Wang M, Du J, Zhang Y, Chen X. Prediction of subcellular location of apoptosis proteins by incorporating PsePSSM and DCCA coefficient based on LFDA dimensionality reduction. BMC Genom. 2018;19:1–17.
53. Nanni L, Brahnam S, Lumini A. Wavelet images and Chou's pseudo amino acid composition for protein classification. Amino Acids. 2012;43:657–65.
54. Ahmad A, Akbar S, Tahir M, Hayat M, Ali F. iAFPs-EnC-GA: identifying antifungal peptides using sequential and evolutionary descriptors based multi-information fusion and ensemble learning approach. Chemom Intell Lab Syst. 2022;222:104516.
55. Lu W, Song Z, Ding Y, Wu H, Cao Y, Zhang Y, Li H. Use Chou's 5-step rule to predict DNA-binding proteins with evolutionary information. BioMed Res Int. 2020;2020:1–9.
56. Zhang L, Zhao X, Kong L. Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou' s pseudo amino acid composition. J Theor Biol. 2014;355:105–10.
57. Sun D, Liu Z, Mao X, Yang Z, Ji C, Liu Y, Wang S. ANOX: a robust computational model for predicting the antioxidant proteins based on multiple features. Anal Biochem. 2021;631:114257.
58. Wang J, Yang B, Revote J, Leier A, Marquez-Lago TT, Webb G, Song J, Chou K-C, Lithgow T. POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. Bioinformatics. 2017;33(17):2756–8.
59. Hayat M, Khan A. Mem-PHybrid: hybrid features-based prediction system for classifying membrane protein types. Anal Biochem. 2012;424(1):35–44.
60. Hayat M, Khan A. WRF-TMH: predicting transmembrane helix by fusing composition index and physicochemical properties of amino acids. Amino Acids. 2013;44:1317–28.
61. Suvarna Vani K, Durga Bhavani S. SMOTE based protein fold prediction classification. In: Advances in computing and information technology: proceedings of the second international conference on advances in computing and information technology (ACITY) July 13–15, 2012, Chennai, India-Volume 2. Springer; 2013. p. 541–50.
62. Akbar S, Hayat M, Kabir M, Iqbal M. iAFP-gap-SMOTE: an efficient feature extraction scheme gapped dipeptide composition is coupled with an oversampling technique for identification of antifreeze proteins. Lett Org Chem. 2019;16(4):294–302.
63. Hu J, He X, Yu D-J, Yang X-B, Yang J-Y, Shen H-B. A new supervised over-sampling algorithm with application to protein-nucleotide binding residue prediction. PLoS ONE. 2014;9(9):e107676.
64. Elreedy D, Atiya AF. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. Inf Sci. 2019;505:32–64.
65. Sun Y, Robinson M, Adams R, Te Boekhorst R, Rust AG, Davey N. Using sampling methods to improve binding site predictions. In: Proceedings of the 14th European symposium on artificial neural networks, ESANN 2006; 2006.

66. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
67. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005;27(8):1226–38.
68. Charoenkwan P, Chiangjong W, Nantasenamat C, Hasan MM, Manavalan B, Shoombuatong W. StackIL6: a stacking ensemble model for improving the prediction of IL-6 inducing peptides. Brief Bioinform. 2021;22(6):bbab172.
69. Mishra A, Pokhrel P, Hoque MT. StackDPPred: a stacking based prediction of DNA-binding protein from sequence. Bioinformatics. 2019;35(3):433–41.
70. Basith S, Lee G, Manavalan B. STALLION: a stacking-based ensemble learning framework for prokaryotic lysine acetylation site prediction. Brief Bioinform. 2022;23(1):bbab376.
71. Liang X, Li F, Chen J, Li J, Wu H, Li S, Song J, Liu Q. Large-scale comparative review and assessment of computational methods for anti-cancer peptide identification. Brief Bioinform. 2021;22(4):bbaa312.
72. Jiang M, Zhao B, Luo S, Wang Q, Chu Y, Chen T, Mao X, Liu Y, Wang Y, Jiang X. NeuroPpred-Fuse: an interpretable stacking model for prediction of neuropeptides by fusing sequence information and feature selection methods. Brief Bioinform. 2021;22(6):bbab310.
73. Guo Y, Yan K, Lv H, Liu B. PreTP-EL: prediction of therapeutic peptides based on ensemble learning. Brief Bioinform. 2021;22(6):bbab358.
74. Cao Z, Pan X, Yang Y, Huang Y, Shen H-B. The lncLocator: a subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier. Bioinformatics. 2018;34(13):2185–94.
75. Zhang Q, Liu P, Wang X, Zhang Y, Han Y, Yu B. StackPDB: predicting DNA-binding proteins based on XGB-RFE feature optimization and stacked ensemble classifier. Appl Soft Comput. 2021;99:106921.
76. Akbar S, Raza A, Zou Q. Deepstacked-AVPs: predicting antiviral peptides using tri-segment evolutionary profile and word embedding based multi-perspective features with deep stacking model. BMC Bioinform. 2024;25(1):102.
77. Akbar S, Ali H, Ahmad A, Sarker MR, Saeed A, Salwana E, Gul S, Khan A, Ali F. Prediction of amyloid proteins using embedded evolutionary & ensemble feature selection based descriptors with extreme gradient boosting model. IEEE Access; 2023.
78. Bukhari SNH, Webber J, Mehbodniya A. Decision tree based ensemble machine learning model for the prediction of Zika virus T-cell epitopes as potential vaccine candidates. Sci Rep. 2022;12(1):7810.
79. Akbar S, Rahman AU, Hayat M, Sohail M. cACP: classifying anticancer peptides using discriminative intelligent model via Chou's 5-step rules and general pseudo components. Chemom Intell Lab Syst. 2020;196:103912.
80. Ao C, Zhou W, Gao L, Dong B, Yu L. Prediction of antioxidant proteins using hybrid feature representation method and random forest. Genomics. 2020;112(6):4666–74.
81. Dwivedi AK. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Comput Appl. 2018;29:685–93.
82. Ali F, Akbar S, Ghulam A, Maher ZA, Unar A, Talpur DB. AFP-CMBPred: computational identification of antifreeze proteins by extending consensus sequences into multi-blocks evolutionary information. Comput Biol Med. 2021;139:105006.
83. Akbar S, Zou Q, Raza A, Alarfaj FK. iAFPs-Mv-BiTCN: predicting antifungal peptides using self-attention transformer embedding and transform evolutionary based multi-view features with bidirectional temporal convolutional networks. Artif Intell Med. 2024;151:102860.
84. Raza A, Uddin J, Almuhaimeed A, Akbar S, Zou Q, Ahmad A. AIPs-SnTCN: predicting anti-inflammatory peptides using fastText and transformer encoder-based hybrid word embedding with self-normalized temporal convolutional networks. J Chem Inf Model. 2023;63:6537–54.
85. Raza A, Uddin J, Akbar S, Alarfaj FK, Zou Q, Ahmad A. Comprehensive analysis of computational methods for predicting anti-inflammatory peptides. Arch Comput Methods Eng. 2024. https://doi.org/10.1007/s11831-024-10078-7.
86. Akbar S, Hayat M. iMethyl-STTNC: identification of N6-methyladenosine sites by extending the idea of SAAC into Chou's PseAAC to formulate RNA sequences. J Theor Biol. 2018;455:205–11.
87. Charoenkwan P, Ahmed S, Nantasenamat C, Quinn JM, Moni MA, Lio' P, Shoombuatong W. AMYPred-FRL is a novel approach for accurate prediction of amyloid proteins by using feature representation learning. Sci Rep. 2022;12(1):7697.
88. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56–67.
89. Garreau D, Luxburg U. Explaining the explainer: a first theoretical analysis of LIME. In: International conference on artificial intelligence and statistics. PMLR; 2020. p. 1287–96.
90. Du Z, Ding X, Xu Y, Li Y. UniDL4BioPep: a universal deep learning architecture for binary classification in peptide bioactivity. Brief Bioinform. 2023;24(3):1–10.

Publisher's Note