OXFORD

Briefings in Bioinformatics, 22(6), 2021, 1–11

https://doi.org/10.1093/bib/bbab310
Problem Solving Protocol

NeuroPpred-Fuse: an interpretable stacking model for prediction of neuropeptides by fusing sequence information and feature selection methods

Mingming Jiang, Bowen Zhao, Shenggan Luo, Qiankun Wang, Yanyi Chu, Tianhang Chen, Xueying Mao, Yatong Liu, Yanjing Wang, Xue Jiang, Dong-Qing Wei and Yi Xiong

Corresponding author: Yi Xiong, State Key Laboratory of Microbial Metabolism, and School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China. Email: xiongyi@sjtu.edu.cn

Abstract

Neuropeptides acting as signaling molecules in the nervous system of various animals play crucial roles in a wide range of physiological functions and hormone regulation behaviors. Neuropeptides offer many opportunities for the discovery of new drugs and targets for the treatment of neurological diseases. In recent years, there have been several data-driven computational predictors of various types of bioactive peptides, but the relevant work about neuropeptides is little at present. In this work, we developed an interpretable stacking model, named NeuroPpred-Fuse, for the prediction of neuropeptides through fusing a variety of sequence-derived features and feature selection methods. Specifically, we used six types of sequence-derived features to encode the peptide sequences and then combined them. In the first layer, we ensembled three base classifiers and four feature selection algorithms, which select non-redundant important features complementarily. In the second layer, the output of the first layer was merged and fed into logistic regression (LR) classifier to train the model. Moreover, we analyzed the selected features and explained the feasibility of the selected features.

Mingming Jiang is a master student at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. He works on machine learning and natural language processing in Bioinformatics.
Bowen Zhao is a master student at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. His research interests include molecular representation learning and machine learning.
Shenggan Luo is a Ph. D. candidate at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. He works on activation energy barrier prediction through machine learning methods.
Qiankun Wang is a Ph. D. candidate at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. He works on molecular dynamics simulation through machine learning methods.
Yanyi Chu is a Ph.D. candidate at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. She works on drug discovery through machine learning methods.
Tianhang Chen is an undergraduate student at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. He has expertise in peptide prediction through machine learning.
Xueying Mao is a Ph.D. candidate at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. She has expertise in computer-aided drug design and machine learning.
Yatong Liu is a master student at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. He works on machine learning and deep learning.
Yanjing Wang is conducting the postdoctoral training with the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. He works on big biomedical data analysis through deep learning.
Xue Jiang is conducting the postdoctoral training with the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. Her research interest is artificial intelligence applications in big biomedical data analysis of complex disease.
Dong-Qing Wei is a full professor at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. His main research areas include structural bioinformatics and biomedicine.
Yi Xiong is an associate professor at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. His main research interests focus on machine learning algorithms, and their applications in protein sequence–structure–function relationships and biomedicine.
Submitted: 16 April 2021; **Received (in revised form):** 1 July 2021

Experimental results show that our model achieved 90.6% accuracy and 95.8% AUC on the independent test set, outperforming the state-of-the-art models. In addition, we exhibited the distribution of selected features by these tree models and compared the results on the training set to that on the test set. These results fully showed that our model has a certain generalization ability. Therefore, we expect that our model would provide important advances in the discovery of neuropeptides as new drugs for the treatment of neurological diseases.

Key words: neuropeptide prediction; machine learning; feature selection; stacking; feature analysis

Introduction

Neuropeptides are small peptides composed of approximately 3–100 amino acids in length, acting as signaling molecules in the nervous system of various animals such as invertebrates and mammals [1]. They act as neurotransmitters or peptide hormones to possess a wide range of physiological functions and hormone regulation behaviors [2]. Because of the crucial roles of neuropeptides, many technologies focused on identifying neuropeptides. The traditional methods for the identification of neuropeptides were liquid chromatography–tandem mass spectrometry (LC–MS/MS) based on bioassay, receptor binding assay, and genetic analysis [3–6]. However, the experimental process is time-consuming and expensive. Instead, computational models can provide an alternative to the tediously experimental approaches for predicting neuropeptides.

Recently, several computational tools and databases have been developed to accelerate the discovery and identification of neuropeptides. NeuroPep [7] is the most comprehensive database that contains the neuropeptide entries extracted from Neuropeptide [8] and NeuroPedia [9] databases. In addition, the BLAST programs are widely used tools for searching protein and DNA databases for sequence similarities [10]. However, neuropeptides are still hard to be identified by using sequence homology inference-based methods, since neuropeptides are highly short and diverse in sequences, and challenging to discover using standard sequence-similarity methods. To address these challenges, some studies proposed machine learning-based methods to identify neuropeptides. The development of machine learning-based methods has facilitated the identification of active peptides in drug discovery. Neuropod [11] and NeuroPP [12] are tools, which predict the NP precursors based on machine learning algorithms. Agrawal et al. have presented the support machine learning (SVM) model for identifying insect neuropeptides in 2019 [13]. In 2020, Bin, et al have proposed the predictor named PredNeuroP based on a two-layer stacking method [14].

However, up to now, these machine learning-based models lacked interpretability, and the accuracy of these models still have remaining room for improvement. In order to overcome these difficulties, we propose a more robust model for identifying neuropeptides. The coding scheme should capture enough sequence information and ensure the diversity and reliability of sequence information. In recent years, a variety of encoding methods on peptide sequences have been proposed, such as adaptive skip dipeptide composition (ASDC) [15], g-gap dipeptide composition (GGAP) [16] and position-specific amino acid composition (PSAAC) [17]. One simple approach to utilize these various features is concatenating them, resulting in bringing the curse of dimensionality causing high computation complexity [18]. These high-dimensional feature vectors also increase the possibility of correlation or redundancy among its feature elements. However, feature selection can help decrease the computational time and complexity of the final prediction model, and also provide more insights into the data abundance. Just one feature selection method is less reliable than the combination of multiple feature selection methods, which filter features based on different evaluation criteria. Furthermore, in the past few years, ensemble learning has widely been used in various bioinformatics applications and data competitions [19, 20]. Through strategically generating and combining multiple weak classifiers, a strong classifier [21] can improve the robustness of the model. Considering the existing multiple machine learning methods and feature selection methods mentioned above, we combine ensemble learning with feature selection algorithms together to reduce the wrong decision of a single learner and the low fault tolerance of a single feature selection method. Motivated by the above factors, we build our workflow as follows: Firstly, peptide sequences were encoded by a variety of features and then were merged. Secondly, a variety of feature selection methods were applied to these features for getting the condensed set of effective features, which were fed into the first layer of the classifier, which is composed of three tree learners. Finally, the predicted class probability of the first layer is used as the input of the second layer to train the second layer (Logistic Regression) to get the final result.

Materials and methods

Datasets

The performance of the model depends on the quantity and quality of the data. The same benchmark dataset as [14] was used in this work in order to fairly compare our method with previous models. The distribution of the training set and test set are shown in Figure S1. Moreover, we used the same ratio to divide the total data into the training and test data sets. 80% positive and negative data sets (3880 sequences in total) were selected to construct the training set, and the remaining 20% data sets (970 sequences in total) were selected to construct the test set. In addition, we split our training set into 10-folds and performing 10-fold cross-validation (CV).

Feature representation

In this study, we used six different feature encoding schemes to encode the peptide sequences into a feature vector, which included AAC, DPC, GGAP, CTD, ASDC and PSAAC. Briefly, each encoding definition is explained in the following subsections:

AAC

AAC consists of the frequency of all 20 natural amino acid types in the peptide sequences. The calculation of AAC is as follows:

$$f(a) = \frac{N(a)}{L}, a \in \{A, C, D, \dots, Y\} \tag{1}$$

where N(a) is the total number of amino acid type a, while L is the length of the peptide.

DPC

DPC is a 400-dimensional feature vector, which is computed as follows:

$$f\left(a,b\right)=\frac{\mathrm{N}\left(a,b\right)}{\mathrm{L}-1},a,b\in\{A,C,D,\ldots\mathrm{Y}\}\tag{2}$$

where *N(a, b)* is the total number of dipeptides denoted by amino acid types *a* and *b*.

GGAP

Dipeptides have been widely used in the field of protein prediction [22, 23]. However, it can only reflect the correlation between two adjacent amino acids. In fact, amino acids separated by two distant residues in the sequence may be adjacent in three-dimensional structure [24]. Therefore, we used G-GAP dipeptide composition (DC) to transform the protein sequence into the characteristic carrier. GGAP can be computed as follows:

$$\mathrm{F}^{g}=\left[p_{1}^{g},p_{2}^{g},\ldots,p_{400}^{g}\right]^{T}\tag{3}$$

where T denotes the transposition of the feature vector. p_{i}^{g} is the frequency of the *i-th* g-gap dipeptide and is defined as:

$$p_{\mathrm{i}}^{g}=\frac{n_{\mathrm{i}}^{g}}{\sum_{\mathrm{i}=1}^{400}n_{\mathrm{i}}^{g}}=\frac{n_{\mathrm{i}}^{g}}{L-g+1}\tag{4}$$

where $n_{\mathrm{i}}^{g}$is the number of the *i-th* g-gap dipeptide along the whole sequence. 0-gap dipeptide describes the correlation between two adjacent residues, and g-Gap dipeptide indicates the correlation between two residues with the interval of g residues. In this paper, due to the short length of neuropeptides, the parameter g is determined as 3.

CTD

The distribution pattern of amino acid physicochemical properties (PCPs) [25] encoded by CTD. As shown in Table S1, 20 standard amino acids were divided into three categories according to seven PCPs. The composition (C) is expressed as a 21-dimensional vector, and it is calculated as follows:

$$\mathrm{C}(a)=\frac{\mathrm{N}(a)}{L},a=1,2,3\tag{5}$$

where N(a) is the total number of amino acid class a. The transition (T) is encoded as a 21-dimensional vector, indicating that type A residues are followed by type B residues or type B residues are followed by type A residues. It can be calculated as follows:

$$\begin{cases}T\left(a,b\right)=\frac{N\left(a,b\right)+N\left(b,a\right)}{L-1}\\T\left(a,\mathrm{c}\right)=\frac{N(a,\mathrm{c})+N\left(b,\mathrm{c}\right)}{L-1}\\T\left(b,\mathrm{c}\right)=\frac{N\left(b,\mathrm{c}\right)+N\left(\mathrm{c},b\right)}{L-1}\end{cases}\tag{6}$$

where *M(a,b)* represents the number of the dipeptide *ab* in the sequence. The distribution (D) descriptor consists of five values of three classes, based on the score of the entire sequence, where the first residual of a given group is located, containing 25, 50, 75 and 100% of events. D is encoded as a 105-dimensional vector. In summary, the CTD encoding is a 147-dimensional vector.

ASDC

The adaptive k-skip-n-gram model is an upgraded version of the skip-*n*-gram model. The method adapts to the different lengths of sequences in the dataset and includes more distance information in the features. ASDC can be computed as follows:

$$FV=\left\{\frac{\mathrm{N}\left(\alpha_{m_1}\alpha_{m_2}\ldots\alpha_{m_n}\right)}{N\left(T_{skipgram}\right)}|1\le m_1\le 20,\le m_1\le 20,\ldots,\le m_n\le 20\right\}\tag{7}$$

$$T_{skipgram}=\cup_{a=1}^{k}Skip\left(DT=a\right)$$

where $Skip(DT=a)=\{\mathrm{A}_i\mathrm{A}_{i+a+1}\ldots\mathrm{A}_{i+a+n\text{-}1}|1\le i\le L-a,1\le a\le k\}$ [26]. This method considers the n residues with distances 1 to k and k is the length of each sequence.

PSAAC

PSAAC is a variant of AAC. Considering the importance of terminal residues for the structure and function of bioactive peptides [27, 28], we extracted the feature codes of the first and last five residues of N-terminal and C-terminal (NT5 and CT5) for the prediction model. It can be computed as follows:

$$f\left(a,\mathrm{i}\right)=\frac{\mathrm{N}\left(a,\mathrm{i}\right)}{\mathrm{M(i)}}\tag{8}$$

where $N\left(a,i\right)$ is the number frequency of amino acid *a* at position i and *M(i)* denotes the number of *i-th* position in all sequences.

Workflow

In this work, the workflow of NeuroPpred-Fuse consists of the following steps (Figure 1): Firstly, we used six types of features to encode the peptide sequences, and then combined these features and performed feature selection. For the combined features, we used five different feature selection algorithms (actually four, one is discarded due to bad performance relatively) to select non-redundant important features and filtered the one of them which had relatively inferior performance. Then, we fused three base classifiers selected from various single classifiers with the chosen feature selection algorithms in the first layer. On the second-layer learning, the outputs of the first-layer are merged and imported into a logistic regression classifier to train the final model, which outputs the final prediction results. The detail of our model was introduced in the following subsections.

Feature selection

Feature selection is an important step in the application of machine learning and there are some reasons for this. The merged data sets are described with far too many variables for building this practical model. Usually, most of these variables are irrelevant to the model performance and their relevance is not known in advance, especially in biology [29]. Feature selection methods are mainly categorized into three types: filtering method, wrapping method, and embedding method. In this study, rather than to the embedding method, we chose the other two types of classic representative algorithms, namely Relief feature selection (Relief) and Boruta algorithm belonging

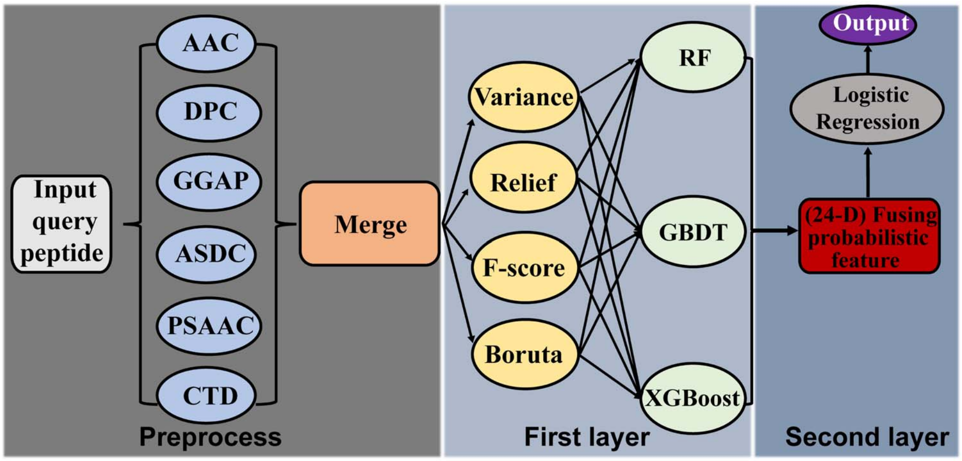

Figure 1. Workflow of NeuroPpred-Fuse. It involves the following steps: (**i**) Neuropeptides was first represented in six different features; (**ii**) in the first layer, six feature encodings were analyzed and merged , and then perform feature selection by four methods. Subsequently, the selected features were used as an input for three classifiers resulting in their corresponding prediction models; (**iii**) the predicted probability of the first layer was fused and inputted to LR for development of the final prediction model in the second layer.

to wrapping method, variance-based and F-score-based algorithms, which are parts of filtering method. Briefly, the Relief algorithm [30] is a feature-weighting algorithm in which the correlation of various features and categories gives different weights to features, and features with weights less than a certain threshold will be removed. The correlation between features and categories in the Relief algorithm is based on the ability of features to distinguish close samples. Boruta algorithm [31] is a wrapper built around the random forest classification; However, both F-score [32] and variance-based [33] feature selection methods use statistics to screen irrelevant features and filter redundant features. At the same time, we also used a feature dimensionality reduction method named Kernel Principal Component Analysis (KPCA) [34] is a nonlinear data processing method, whose core idea is to project the original spatial data into the high-dimensional feature space through a nonlinear map, and then carry out PCA based analysis in the high-dimensional feature space. These feature selections methods or dimension reduction have different metrics and capture different features benefitting to identify neuropeptides. Therefore, we combine their respective strength by fusing them.

Machine learning classifiers

From the view point of the machine learning, the prediction performance not only depends on feature encodings but also the choice of the classifiers. For example, SVM [35], random forest (RF) [36], Gradient Boost [37], XGBoost [38] and K-Nearest Neighbors (KNN) [39] have been widely used for constructing classification models for identifying various types of peptides. In this study, seven different machine learning classifiers were employed (*i.e.,* SVM, RF, Naïve Bayes, GBDT, Artificial Neural Network (ANN) [40], KNN, XGBoost). Specifically, XGBoost, GBDT, and RF are commonly typical tree models, which refer to the model such as the decision-making tree with tree branch structure based on feature space partitioning. As a representative ensemble learning method, the stacking framework integrates different base-learners (first-layer-learners) with different feature selection algorithms based on meta-learner (second-layer-learner) to pull together advantages of various feature selection methods and construct a well-performed predictor. On the first layer, six features are combined as the input feature vector, and the four feature selection methods (the kernel PCA has been discarded) and three base classifiers are combined as the 12 base-learners, which are constructed on the training set. In this work, 10-fold CV was adopted to train the base classifiers and obtain the prediction on the test set in the first layer (Figure 2). On the second layer, a second-layer-learner with LR classifier is trained based on the outputs of the first layer, which is 24-dimensional probabilistic vector, and the output of this layer is the final result.

Performance measures

In this study, sensitivity (Sen), specificity (SPE), accuracy (ACC), and Matthew's correlation coefficient (MCC), which are commonly used in binary classification [41–46], were used to evaluate the performance of the model. They are calculated as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{9}$$

$$Specificity = \frac{TN}{TN + FP} \tag{10}$$

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} \tag{11}$$

$$MCC = \frac{TP * TN + FP * FN}{\sqrt{(TP + FN)\,(TP + FP)\,(TN + FP)\,(TN + FN)}} \tag{12}$$

where TP, TN, FP and FN represent the number of true positive, true negative, false positive and false negative samples, respectively. In addition, the area under the receiver operating characteristic curve (AUC-ROC) was established to evaluate the performance of the model. The larger the AUC-ROC, the better the predictive performance of the predictor. On this basis, the 10-fold CV technique is used to evaluate the performance of various models on the training data sets.

Results and discussion

Feature representation comparison

In order to feed the original sequence into different machine learning classifiers, we utilized six different sequence-based

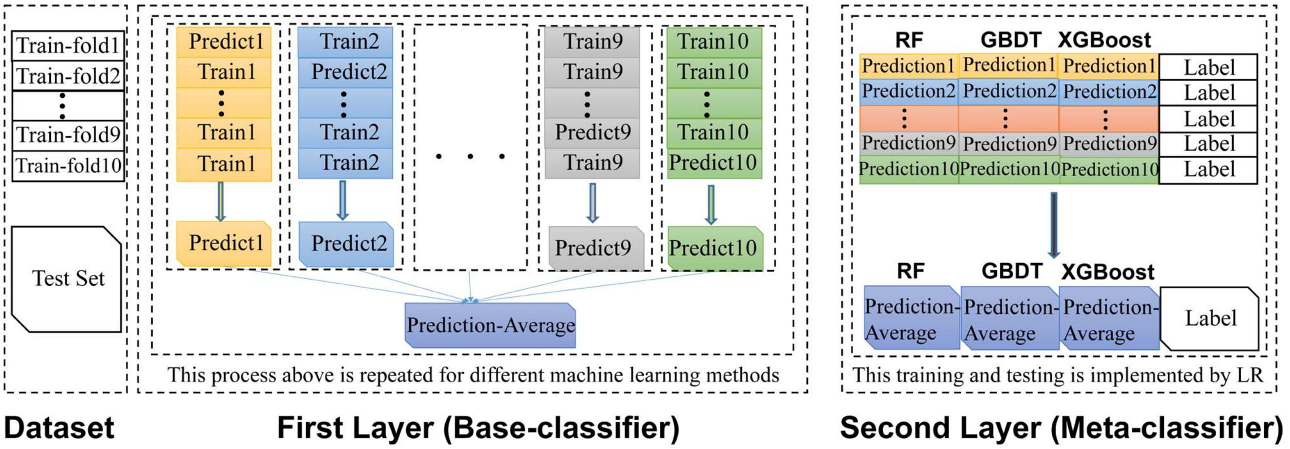

Figure 2. The framework of the stacking strategy used in NeuroPpred-Fuse. (i) The training dataset has been split into 10 parts. The nine of them was used to train the model and one was used to predict corresponding output in the first layer; (ii) the test set was performed the same process as the training set but average the prediction; (iii) through combining different machine learning classifiers, multiple combinations of output from the first layer were fed into the second layer (meta-classifier) and generating the final output.

encodings, including AAC, DPC, GAPC, CTD, PSAAC, ASDC and their merged features and evaluated its ability to predict neuropeptides or non-neuropeptides by integrating them with seven kinds of classification algorithms (SVM, RF, GBDT, XGBoost, KNN, NB and ANN). In total, we generated 36 classifiers (ANN has not been evaluated on single feature encoding as some of them are too sparse and low dimensional) based on individual features and 7 classifiers of merged features. Table 1 showed that SVM, RF, GBDT, XGBoost, KNN, NB and ANN yielded accuracy in the range of 64.0–84.0%, 79.4–88.5%, 76.7–87.5%, 80.8–89.8%, 62.6–78.8%, 66–76%, 78.9%, respectively, with respect to the seven different encodings. As shown in Figure 3 and Table 1, it could be noted that the features of ASDC, PSAAC, GGAP and AAC have relatively high accuracy in the six classification algorithms, especially the ASDC and CTD encoding. This observation indicates that CTD encoding and ASDC encoding have higher discriminative power for the prediction of neuropeptides as compared to other types of features. From Figure 3, we also observed that the three tree-based models (RF, GBDT, XGBoost) achieved similarly high performance relative to other models. In addition, these three tree-based models have excellent performance in merged features, especially XGBoost. They all have achieved accuracy more than 85.0%. On the other hand, we examined the remaining models and noted that these models achieved relatively low accuracy with no more than 85.0%. Therefore, instead of focusing on making use of all models like the traditional stacking method, we selected the three tree-based models (RF, GBDT, XGBoost) as our base classifiers in the first layer and the merged features as our final input. Furthermore, we also applied the ANN (Fully connected network of 1407 × 128 × 2) to fit the merged feature, the result showed the inferior performance of ANN due to fewer data.

Feature selection comparison

The goal of this section is to compare the performance of various feature selection methods and obtain the filtered features by using these methods. We evaluated the results of the five feature selection algorithms based on the best classifier mentioned above (XGBoost). From Figure 4, we noted that most feature selection methods achieved about 90.0% accuracy, except KPCA. The Relief algorithm achieved the highest accuracy at about 780-dimensions and the feature selection method based on F-score and based on variance also achieved relatively high accuracy at 1200-dimensions and 1140-dimensions respectively in Figure 4. On the other hand, as the results are shown in Table 2, the Boruta algorithm also achieved about 90.33% and the optimal size can be determined without our operations. Considering these feature selections methods having different metrics and capture different features benefit to identify neuropeptides, we selected Boruta, Relief, F-score-based, and variance-based as our final feature selection methods for improving the performance of our kinds of models. Consequently, we fused their advantages into different machine learning models mentioned above. Practically, there are various ways to integrate multiple prediction models, such as the ensemble method [47, 48] and meta-predictor [47, 49–53]. Herein, we employed a meta-predictor approach to develop the final model. Specifically, the predicted probabilities from the above-mentioned base 12 predictors (4 feature selections × 3 classifiers) were concatenated and considered as a new 24-dimensional feature vector representing multi-view information [54]. Hence, the 24D probabilistic vector was considered as the final input fed into the second layer (LR).

As above mentioned, we obtained the features selected by four methods (Relief, Boruta, F-score-based, and Variance-based). In this section, we analyzed statistically, which features are crucial for identifying neuropeptides and whether the feature set selected by one method is a subset of another feature set filtered by another method. As shown in Figure 4 and Table 2, it could be noted that the Boruta algorithm selected fewer dimensions and achieved relatively high accuracy (90.33%). This observation indicated that it selected the features that are most important to the classifier. Therefore, in order to avoid overfitting, this Boruta algorithm could be considered. The feature selection algorithms based on statistics (F-score-based and Variance-based) seemed to remain more features in Figure 4. However, Relief not only achieved the best accuracy but also selected no too many features. In addition, Figure 5 showed that these features selection algorithms all remained fewer features about DPC coding in their total features, respectively. This observation shed a light on us that DPC coding could not provide great information for classifying neuropeptides, due to its sparsity in short and few peptides. By contrast, we noted that

Table 1. Performance comparison of different combinations of features and classification algorithms by 10-fold CV on the training set

Classifier	PSAAC	GGAP	ASDC	DPC	CTD	AAC	Merged
XGBoost	0.808	0.807	0.862	0.840	0.876	0.843	0.898
SVM	0.809	0.807	0.840	0.839	0.810	0.812	0.640
RF	0.827	0.818	0.861	0.846	0.853	0.835	0.885
KNN	0.782	0.713	0.773	0.744	0.784	0.788	0.626
Naïve Bayes	0.753	0.728	0.760	0.750	0.658	0.735	0.741
GBDT	0.794	0.784	0.841	0.816	0.841	0.812	0.875
ANN	\	\	\	\	\	\	0.789

Note: ANN just be trained on combined features as some single features are less dimension.

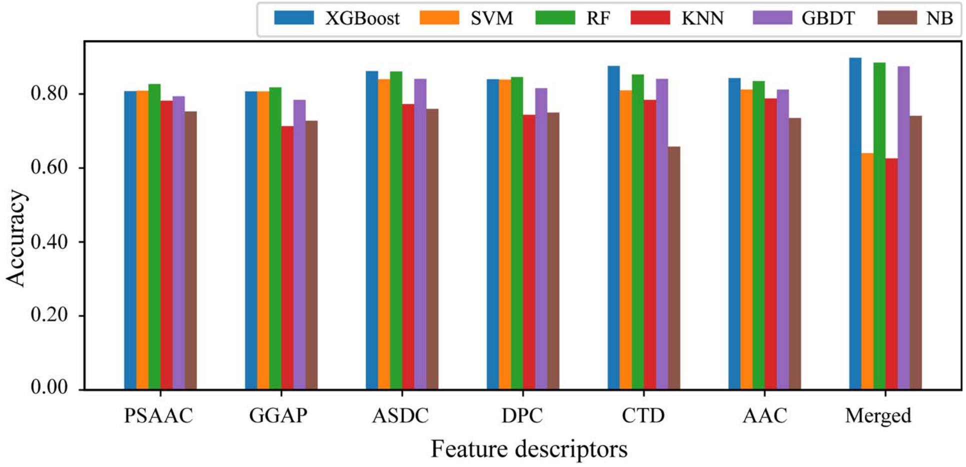

Figure 3. Prediction performance of seven classifiers utilizing seven different coding schemes based on the training dataset.

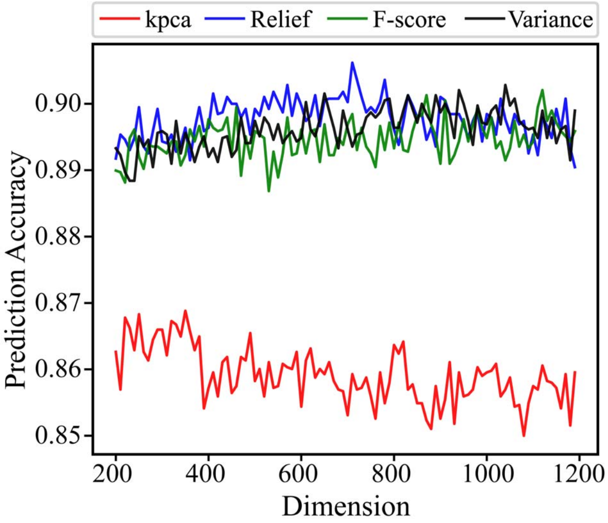

Figure 4. The effect of four different dimensionality reduction algorithms for identifying neuropeptides.

CTD coding, PSAAC coding, and AAC coding with low dimensions have a significant proportion of the selected features. Instead of making efforts to determine the best feature selection algorithm among these methods, we fused their advantages into our model. This further suggests that these features (CTD, PSAAC, AAC) are better in the case of small data and short peptides.

Table 2. Best performance of four feature selections by 10-fold CV on the training set

Feature selection	Dimension	Accuracy
Relief	710	90.62%
F-score	1200	89.23%
Variance	1140	90.24%
Boruta	76	90.33%

Note: The accuracy KPCA was not shown in this table but shown in Figure 6.

Furthermore, we analyzed whether there are a lot of common features after these four feature selection algorithms. If there are a great many of repeated features, discarding some feature selection algorithms should be considered. From Figure 6, we noticed that these four feature selection algorithms exactly do not choose the same feature among each other. Herein, although the most of features selected by Boruta are contained by the other three algorithms, we do not throw it away as it is a very neat feature selection algorithm avoiding overfitting.

Feature analysis

To identify the most important features beneficial for classifying neuropeptides, we ranked the feature importance and analyzed the distribution of them. The features screened by these tree-based models showed a similar distribution (Figure S2). As shown in Figure 7 the feature score ranking of the first ten features of all the models selected where the 704th dimension

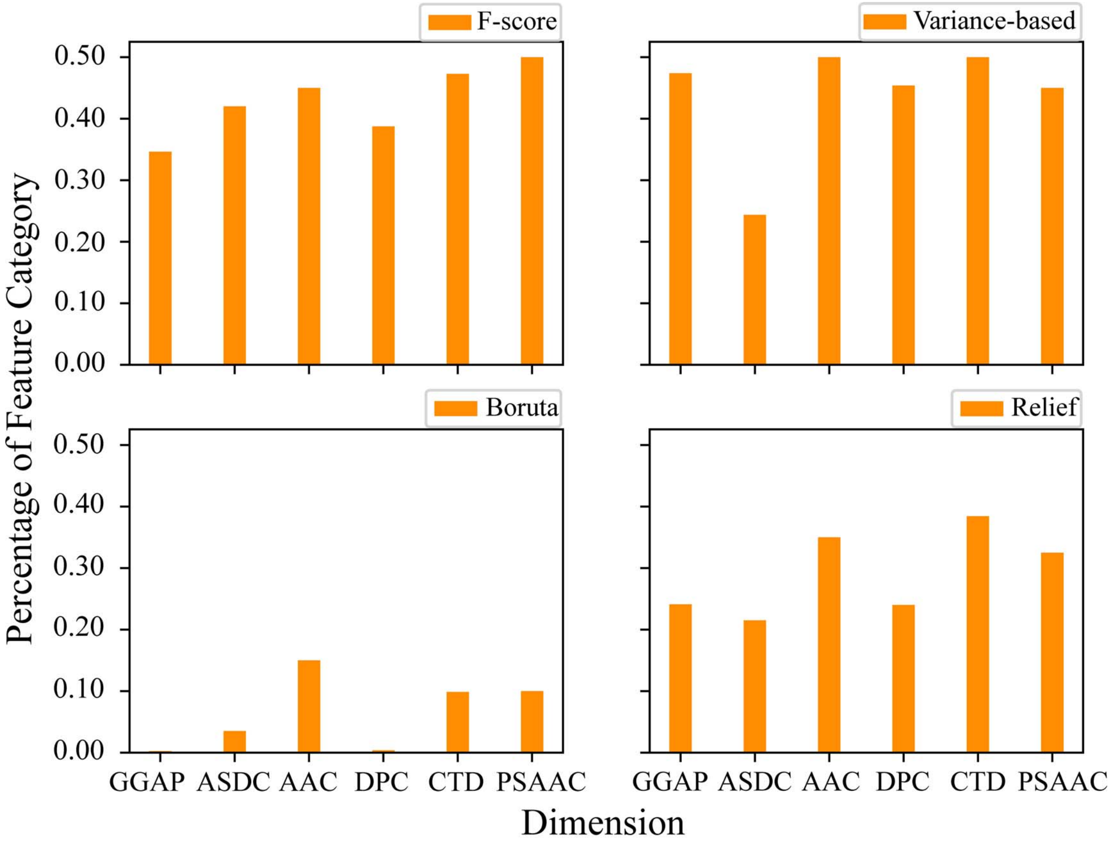

Figure 5. The feature analysis of four different dimensionality reduction algorithms on the prediction accuracy of the training dataset.

belonging to ASDC coding achieved the highest score. Furthermore, Figure 8 showed that the distribution of the top ten features that were finally screened. First of all, we noticed that GGAP (0–399) coding and DPC (820–1219) coding ratio is relatively low, corresponding to selected features distribution by four feature selection methods before. On the other hand, we can observe from Figure 6 that in all the selection of the top ten features, GGAP coding and DPC coding is so little, which suggested that the two types of features contribute less to our model for identifying neuropeptides. Secondly, as shown in Figures 7 and 8, the proportion of ASDC coding and AAC (800–819) coding is relatively moderate and still have a certain proportion while CTD coding and PSAAC (1367–1406) coding are the two most important types of features, which have small dimension. This again indicates that they are beneficial for classifying neuropeptides.

Comparison of the proposed predictor and existing predictors

At last, we compare our models with other state-of-art models. The prediction result showed that the proposed NeuroPpred-Fuse achieved MCC, ACC, Sn, Sp and ROC of 81.3, 90.6. 88.2, 93.0 and 95.8%, respectively. In order to determine whether our approach could improve the performance and outperform the existing models. First, the test data set was independently evaluated by the NeuroPpred-Fuse predictive model, and the performance indicators were shown in Table 3. In addition, as can be seen from Table S3, the similar performance of the NeuroPpred-Fuse model on the training data set indicates that its generalization ability is relatively great, and it is an effective tool to distinguish between neuropeptides and non-neuropeptides. More than that, we also compared our model with other single ML algorithms without feature selection. As shown in Figure S3 and Table S3, our model has the best performance and generality. (All the hyperparameter adjustments are shown in Table S2), while the other models are either less accurate on the test set or overfitting. At last, we compared the performance of our method with the NeuroPIpred [13] and PredNeuroP [14], two existing tools for insect neuropeptides identification. NeuroPIpred did not perform well in the test dataset because it included neuropeptides from all phyla, not only insect neuropeptides, the performance of NeuroPIpred is no more than 90.0% in the test dataset. It is observed that NeuroPpred-Fuse achieved higher accuracy in discriminating neuropeptides from non-neuropeptides.

Conclusion

Accurate identification of neuropeptides can help accelerate peptide-based drug discovery in search of newly effective therapeutic peptides. Therefore, in this study, we proposed a two-layer predictive framework, namely NeuroPpred-Fuse, to improve the identification of neuropeptides based on sequence information. In order to build an effective prediction model, a novel stacking scheme based on three different tree ML models and four feature selection algorithms in conjunction with six feature encoding covering comprehensive sequence information was built to

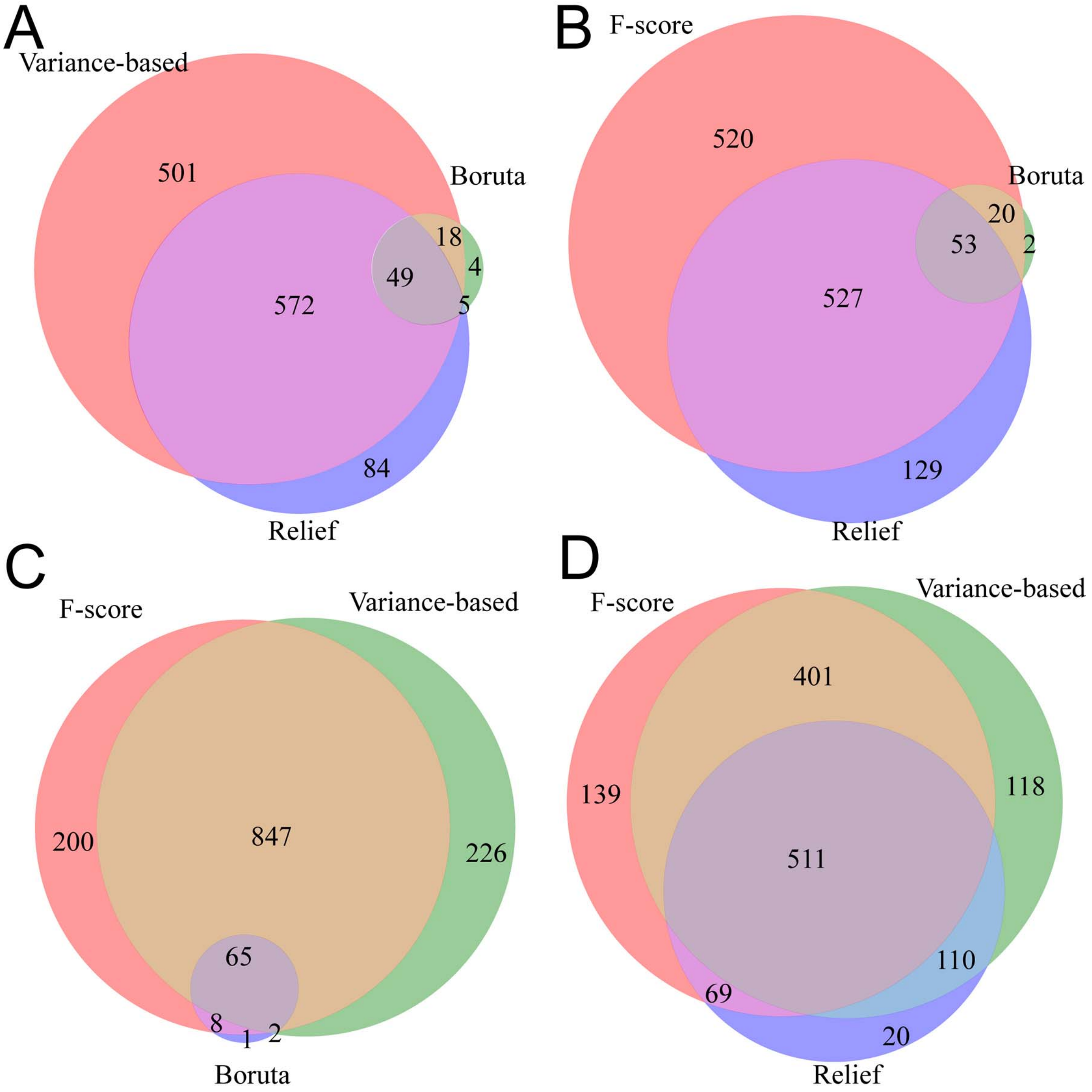

Figure 6. The proportion of selected features in the four feature selection methods. (**A**) The proportion of selected features in the three feature selection methods (variance, Boruta, relief); (**B**) the proportion of selected features in the three feature selection methods (F-score, Boruta, relief); (**C**) the proportion of selected features in the three feature selection methods (variance, Boruta, F-score); (**D**) the proportion of selected features in the three feature selection methods (variance, F-score, relief).

generate 24 probabilistic features. Subsequently, these multi-view probabilistic features were concatenated to construct the final prediction model. Furthermore, in order to find the features that are meaningful to the recognition of neuropeptides, we also do some feature analysis and model explanatory analysis. Experiments have proved that for short peptide sequences with less data, DPC encoding and GGAP encoding, such as high-dimensional sparse feature coding, are not the optimal coding scheme. On the contrary, low-dimensional coding methods such as CTD coding and PSSAAC encoding, which capture more significant information in less data relative to other features have better performance. Rigorous cross-validation and independent test demonstrated that NeuroPpred-Fuse significantly outperformed existing methods and other conventional ML algorithms. It is anticipated that our proposed predictor, NeuroPpred-Fuse will serve as a useful service, high-throughput, and cost-effective tool for large-scale analysis of therapeutic peptides and also in the timely identification of neuropeptides. In comparison with the state-of-the-art models, our proposed method performs multiple feature selection and screens important features beneficial for identifying neuropeptides. Meanwhile, we develop a more robust model, which fuses the three base classifiers with four feature selection methods. We expect that NeuroPpred-Fuse could offer an important advancement for the research communities on the discovery of neuropeptides as new drugs or targets for the treatment of nervous-system disorders in different phyla [1, 55]. Furthermore, it would be expected that integrating other feature encodings [56, 57] and ML algorithms [58–62] might further improve the performance.

Table 3. Best performance of four feature selections on independent dataset

Method	ACC	Sp	Sn	MCC	AUC-ROC
NeuroPpred-Fuse	0.906	0.930	0.882	0.813	0.958
NeuroPIpred[1]	0.536	0.736	0.331	0.074	0.581
PredNeuroP[2]	0.897	0.907	0.886	0.794	0.954

[1] Agrawal, P., et al., *NeuroPIpred: a tool to predict, design and scan insect neuropeptides*. Scientific Reports, 2019. 9.
[2] Bin, Y., et al., Prediction of Neuropeptides from Sequence Information Using Ensemble Classifier and Hybrid Features. J Proteome Res, 2020. 19(9): 3732–3740.

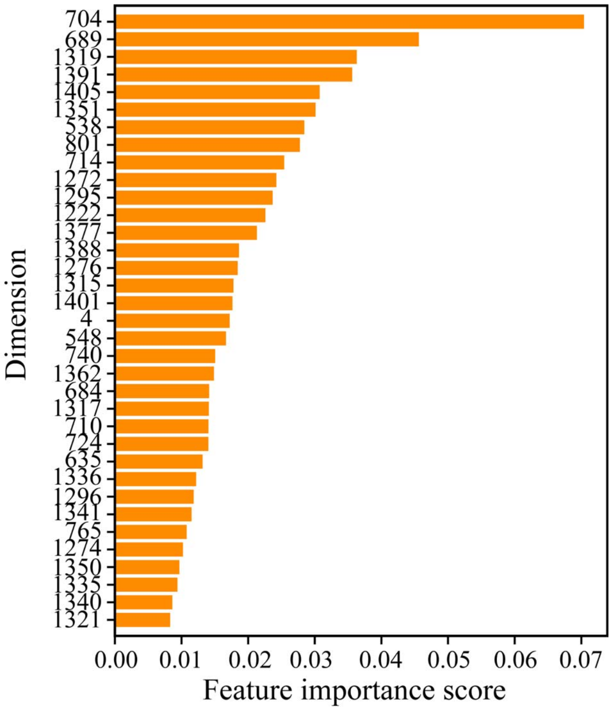

Figure 7. Feature importance score sorting.

> **Key Points**
>
> - This study designs a novel ensemble strategy by fusing feature selection algorithms, which select a multi-view feature from different evaluation metrics and three tree-based models in identifying neuropeptides
> - Our model consists of two layers and the second layer is fed from the first layer, which is multi-view probabilistic features.
> - This study analyzes important features beneficial for predicting neuropeptides and non-neuropeptides in common features. Non-sparse low dimensional features are significant for peptide prediction in small data.
> - The results show that the proposed NeuroPpred-Fuse model has achieved satisfactory results in comparison with the state-of-the-art neuropeptide prediction methods.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*.

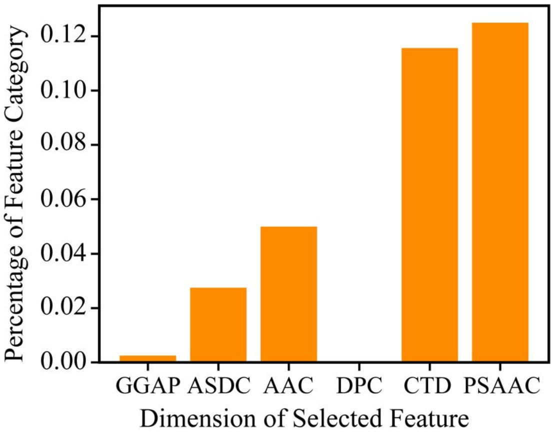

Figure 8. Distribution of top 10 features of tree-based models.

Data availability

The data and source code for this project is freely available at: https://github.com/mingmingjiang1/NeuroPpred-Fuse.

Funding

The National Science Foundation of China [grant numbers 61872094, 61832019, 32070662, 32030063]; the Key Research Area Grant [2016YFA0501703] of the Ministry of Science and Technology of China; the Science and Technology Commission of Shanghai Municipality [grant number 19430750600]; as well as SJTU JiRLMDS Joint Research Fund and Joint Research Funds for Medical and Engineering and Scientific Research at Shanghai Jiao Tong University [ZH2018QNA41, YG2019GD01, YG2019ZDA12, YG2021ZD02].

References

1. Nässel D, Zandawala M. Recent advances in neuropeptide signaling in drosophila, from genes to physiology and behavior. *Prog Neurobiol* 2019;**179**:101607.
2. Mendel HC, Kaas Q, Muttenthaler M. Neuropeptide signalling systems - an underexplored target for venom drug discovery. *Biochem Pharmacol* 2020;**181**:114129.
3. Boonen K, Landuyt B, Baggerman G, *et al.* Peptidomics: the integrated approach of MS, hyphenated techniques and bioinformatics for neuropeptide analysis. *J Sep Sci* 2008;**31**(3):427–45.
4. Secher A, Kelstrup CD, Conde-Frieboes KW, *et al.* Analytic framework for peptidomics applied to large-scale neuropeptide identification. *Nat Commun* 2016;**7**(1):11436.

5. Hayakawa E, Watanabe H, Menschaert G, *et al.* A combined strategy of neuropeptide prediction and tandem mass spectrometry identifies evolutionarily conserved ancient neuropeptides in the sea anemone Nematostella vectensis. *PLoS ONE* 2019;**14**(9):e0215185.
6. Fricker LD, Lim J, Pan H, *et al.* Peptidomics: identification and quantification of endogenous peptides in neuroendocrine tissues. *Mass Spectrom Rev* 2006;**25**(2):327–44.
7. Wang Y, Wang M, Yin S, *et al.* NeuroPep: a comprehensive resource of neuropeptides. *Database* 2015;**2015**.
8. Burbach JPH. Neuropeptides from concept to online database www.Neuropeptides.Nl. *Eur J Pharmacol* 2010; **626**(1):27–48.
9. Kim Y, Bark S, Hook V, *et al.* NeuroPedia: neuropeptide database and spectral library. *Bioinformatics* 2011;**27**(19): 2772–3.
10. Altschul S, Madden TL, Schäffer AA, *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *FASEB J* 1998;**12**(8):A1326–6.
11. Ofer D, Linial M. NeuroPID: a predictor for identifying neuropeptide precursors from metazoan proteomes. *Bioinformatics* 2014;**30**(7):931–40.
12. Kang JJ, Fang Y, Yao P, *et al.* NeuroPP: a tool for the prediction of neuropeptide precursors based on optimal sequence composition. *Interdisciplinary Sciences-Computational Life Sciences* 2019;**11**(1):108–14.
13. Agrawal P, *et al.* NeuroPIpred: a tool to predict, design and scan insect neuropeptides. *Scientific Reports* 2019;**9**.
14. Bin Y, Zhang W, Tang W, *et al.* Prediction of neuropeptides from sequence information using ensemble classifier and hybrid features. *J Proteome Res* 2020;**19**(9):3732–40.
15. Wei LY, Tang JJ, Zou Q, *et al.* SkipCPP-Pred: an improved and promising sequence-based predictor for predicting cell-penetrating peptides. *BMC Genomics* 2017;**18**(7):1–11.
16. Zhang, L., Chengjin Zhang, Rui Gao, *et al.*, *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.
17. Wang YJ, Zhang Q, Sun MA, *et al.* High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles. *Bioinformatics* 2011;**27**(6):777–84.
18. Cai L, Wang L, Fu X, *et al.* ITP-Pred: an interpretable method for predicting, therapeutic peptides with fused features low-dimension representation. *Brief Bioinform* 2021;**22**(4):bbaa367.
19. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting - rejoinder. *Ann Stat* 2000;**28**(2):400–7.
20. Zarayeneh N, Hanifeloo Z. “Antimicrobial peptide prediction using ensemble learning algorithm.” 2020. arXiv preprint arXiv:2005.01714.
21. Liu J, Shang W and Lin W . Improved stacking model fusion based on weak classifier and word2vec. In: *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*. IEEE, 2018, p. 820–4.
22. Gromiha MM, Suwa M. A simple statistical method for discriminating outer membrane proteins with better accuracy. *Bioinformatics* 2005;**21**(7):961–8.
23. Lin H, Ding H. Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition. *J Theor Biol* 2011;**269**(1):64–9.
24. Ding H, Guo SH, Deng EZ, *et al.* Prediction of Golgi-resident protein types by using feature selection technique. *Chemom Intell Lab Syst* 2013;**124**:9–13.
25. Lee TY, Lin ZQ, Hsieh SJ, *et al.* Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences. *Bioinformatics* 2011;**27**(13):1780–7.
26. Guthrie D, Allison B, Liu W, *et al.* A closer look at skip-gram modelling. *LREC*. Citeseer. 2006, p. 1222–25.
27. Chung CR, Kuo TR, Wu LC, *et al.* Characterization and identification of antimicrobial peptides with different functional activities. *Brief Bioinform* 2020;**21**(3):1098–114.
28. Chaudhary K, Kumar R, Singh S, *et al.* A web server and mobile app for computing Hemolytic potency of peptides. *Sci Rep* 2016;**6**(1):22843.
29. Dash M, Liu H. Feature selection for classification. *Intell Data Anal* 1997;**1**:131–56.
30. Kira K, Rendell LA. The feature selection problem: traditional methods and a new algorithm. *Aaai* 1992, p. 129–34.
31. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw* 2010;**36**(11):1–13.
32. Song Q, Jiang H, Liu J. Feature selection based on FDA and F-score for multi-class classification. *Expert Syst Appl* 2017;**81**:22–7.
33. Henseler J, Ringle C, Sarstedt M. A new criterion for assessing discriminant validity in variance-based structural equation modeling. *J Acad Mark Sci* 2015;**43**:115–35.
34. Xu Y, Lin C, Zhao W. Producing computationally efficient KPCA-based feature extraction for classification problems. *Electron Lett* 2010;**46**(6):452–U100.
35. Vapnik VN. An overview of statistical learning theory. *IEEE Trans Neural Netw* 1999;**10**(5):988–99.
36. Breiman L. Random forests. *Mach Learn* 2001;**45**(1):5–32.
37. Friedman JH. Stochastic gradient boosting. *Computational Statistics & Data Analysis* 2002;**38**(4):367–78.
38. Chen, T and Guestrin, C. Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, p. 785–94.
39. Weinberger KQ, Saul LK. Distance metric learning for large margin nearest neighbor classification. *J Mach Learn Res* 2009;**10**:207–44.
40. Balabin RM, Lomakina EI. Neural network approach to quantum-chemistry data: accurate prediction of density functional theory energies. *J Chem Phys* 2009;**131**(7):074104.
41. Wang B, Mei C, Wang Y, *et al.* Imbalance data processing strategy for protein interaction sites prediction. *IEEE/ACM transactions on computational biology and bioinformatics*, 2019.
42. Deng A, Zhang H, Wang W, *et al.* Developing computational model to predict protein-protein interaction sites based on the XGBoost algorithm. *Int J Mol Sci* 2020; **21**(7):2274.
43. Yue Z, Chu X, Xia J. PredCID: prediction of driver frameshift indels in human cancer. *Brief Bioinform* 2021;**22**(3):bbaa119.
44. Basith S, Manavalan B, Hwan Shin T, *et al.* Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening. *Med Res Rev* 2020;**40**(4):1276–314.
45. Shoombuatong W, Schaduangrat N, Pratiwi R, *et al.* THPep: a machine learning-based approach for predicting tumor homing peptides. *Comput Biol Chem* 2019;**80**:441–51.
46. Su R, Hu J, Zou Q, *et al.* Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools. *Brief Bioinform* 2020;**21**(2):408–20.

47. Manavalan B, Basith S, Shin TH, *et al.* 4mCpred-EL: an ensemble learning framework for identification of DNA N4-Methylcytosine sites in the mouse genome. *Cell* 2019;**8**(11):1332.
48. Manavalan B, Basith S, Shin TH, *et al.* Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. *Molecular Therapy-Nucleic Acids* 2019;**16**:733–44.
49. Boopathi V, Subramaniyam S, Malik A, *et al.* mACPpred: a support vector machine-based meta-predictor for identification of anticancer peptides. *Int J Mol Sci* 2019;**20**(8):1964.
50. Manavalan B, Basith S, Shin TH, *et al.* mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics* 2019;**35**(16):2757–65.
51. Qiang X, Zhou C, Ye X, *et al.* CPPred-FL: a sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learning. *Brief Bioinform* 2020;**21**(1):11–23.
52. Schaduangrat N, Nantasenamat C, Prachayasittikul V, *et al.* Meta-iAVP: a sequence-based meta-predictor for improving the prediction of antiviral peptides using effective feature representation. *Int J Mol Sci* 2019;**20**(22):5743.
53. Wei L, Zhou C, Chen H, *et al.* ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* 2018;**34**(400W):4016.
54. Rao B, Zhou C, Zhang G, *et al.* ACPred-fuse: fusing multi-view information improves the prediction of anticancer peptides. *Brief Bioinform* 2020;**21**(5):1846–55.
55. Hökfelt T, Barde S, Xu ZQD, *et al.* Neuropeptide and small transmitter coexistence: fundamental studies and relevance to mental illness. *Frontiers in Neural Circuits* 2018;**12**:106.
56. Chen Z, Zhao P, Li F, *et al.* iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 2018;**34**(14): 2499–502.
57. Chen Z, Zhao P, Li F, *et al.* iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief Bioinform* 2020;**21**(3):1047–57.
58. Cao RZ, Adhikari B, Bhattacharya D, *et al.* QAcon: single model quality assessment using protein structural and contact information with machine learning techniques. *Bioinformatics* 2017;**33**(4):586–8.
59. Chan, L., Ian Morgan, Hayden Simon, *et al.*, *Survey of AI in Cybersecurity for Information Technology Management. 2019 Ieee Technology & Engineering Management Conference (Temscon)*, 2019.
60. Hou J, Wu T, Cao R, *et al.* Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. *Proteins-Structure Function and Bioinformatics* 2019;**87**(12):1165–78.
61. Conover M, Staples M, Si D, *et al.* AngularQA: protein model quality assessment with LSTM networks. *Comput Math Biophys* 2019;**7**(1):1–9.
62. Hou J, Cao R, Cheng J. Deep convolutional neural networks for predicting the quality of single protein structural models. *bioRxiv* 2019. 10.1101/590620 preprint: not peer reviewed.