HybAVPnet: A Novel Hybrid Network Architecture for Antiviral Peptides Prediction

Ruiquan Ge, Yixiao Xia, Minchao Jiang, Gangyong Jia, Xiaoyang Jing, Ye Li, and Yunpeng Cai

***Abstract*—Viruses pose a great threat to human production and life, thus the research and development of antiviral drugs is urgently needed. Antiviral peptides play an important role in drug design and development. Compared with the time-consuming and laborious wet chemical experiment methods, it is critical to use computational methods to predict antiviral peptides accurately and rapidly. However, due to limited data, accurate prediction of antiviral peptides is still challenging and extracting effective feature representations from sequences is crucial for creating accurate models. This study introduces a novel two-step approach, named HybAVPnet, to predict antiviral peptides with a hybrid network architecture based on neural networks and traditional machine learning methods. We adopted a stacking-like structure to capture both the long-term dependencies and local evolution information to achieve a comprehensive and diverse prediction using the predicted labels and probabilities. Using an ensemble technique with the different kinds of features can reduce the variance without increasing the bias. The experimental result shows HybAVPnet can achieve better and more robust performance compared with the state-of-the-art methods, which makes it useful for the research and development of antiviral drugs. Meanwhile, it can also be extended to other peptide recognition problems because of its generalization ability.**

***Index Terms*—Antiviral peptides, deep learning, machine learning, sequence analysis.**

I. Introduction

VIRUSES have become a great threat to humans and animals because of their high rates of infection and mortality [1]. Viruses can affect all species for long periods of time due to their genetic variation, diversity of transmission, and efficient survival within host cells [2], [3], [4]. Especially in recent years, the emergence and re-emergence of the current coronavirus disease 2019 (COVID-19) and severe acute respiratory syndrome (SARS) viruses have posed a serious threat to human life and society[5], [6], [7]. Therefore, it is urgent to develop effective antiviral drugs against various viral pathogens [8], [9]. However, some of the current antiviral agents often have severe side effects and can not kept pace with the evolution of more and more drug-resistant strains [10], [11], [12]. Meanwhile, antiviral drug development is time-consuming and laborious which is not effective enough to address the problem [13], [14].

In recent years, drug development based on peptides has attracted wide attention in the industry due to its highly selective, relatively safe, well tolerated and low production costs [15]. Antiviral peptides (AVPs), with 8 to 40 amino acids typically [16], [17], are a promising resource for the treatment of viral diseases [18]. Antiviral peptides can prevent the virus from attaching to or invading the host cell or interfering with viral replication and are easy to synthesis [4], [19], [20], [21]. Nowadays, there are some collected, experimentally validated AVP databases[22], such as AVPdb [23], HIPdb[24], APD3[25], CAMP[26] etc. AVPdb is a comprehensive resource of peptides that have been experimentally validated for antiviral activities. HIPdb is a specific database of experimentally validated HIV inhibiting peptides. Parts of AVPs are collected in the antimicrobial peptide database APD3 and CAMP.

Many computational tools have been developed to predict AVPs by using machine learning methods. AVPpred is the first AVP prediction tool developed using support vector machine (SVM) based on physiochemical properties [27]. Chang KY et al. employed four peptide features and used random forest (RF) classifier to identify AVPs [28]. Zare1 M et al. employed pseudo-amino acid composition (PseAAC) and adaboost with J48 as base classifier to identify AVPs [29]. AntiVPP 1.0 selected RF as the final classifier with the new two features relative frequency (Rfre) of all 20 natural amino acids and residues composition of peptides (PEP) to assess the antiviral peptides candidates [30]. PEPred-Suite employed an adaptive feature representation strategy to achieve better and robust performance using a two-step feature optimization strategy and eight RF models for eight types of functional peptides, respectively [31]. FIRM-AVP achieved a higher accuracy than other models using the informative filtered features from the physicochemical and structural properties of their amino acid sequences [32]. Charoenkwan P et al. also

Manuscript received 26 October 2022; revised 28 June 2023; accepted 2 April 2024. Date of publication 8 April 2024; date of current version 9 October 2024. This work was supported in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LY21F020017 and Grant 2022C03043, in part by the Joint Funds of the Zhejiang Provincial Natural Science Foundation of China under Grant U20A20386, in part by the National key research and development program of China under Grant 2019YFC0118404, and in part by the National Natural Science Foundation of China under Grant 61702146. *(Corresponding authors: Yunpeng Cai.)*

Ruiquan Ge is with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: gespring@hdu.edu.cn).

Yixiao Xia, Minchao Jiang, and Gangyong Jia are with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: xiayixiao0323@gmail.com; jamchaos666@gmail.com; gangyong@hdu.edu.cn).

Xiaoyang Jing is with the Toyota Technological Institute at Chicago, Chicago IL 60637 USA (e-mail: andersjing@gmail.com).

Ye Li is with the Key Laboratory for Health Informatics of the Chinese Academy of Sciences, Shenzhen Institute of advanced technology, Shenzhen 518055, China, and also with the Hangzhou Institute of Advanced Technology, Hangzhou 310024, China (e-mail: ye.li@siat.ac.cn).

Yunpeng Cai is with the Key Laboratory for Health Informatics of the Chinese Academy of Sciences, Shenzhen Institute of advanced technology, Shenzhen 518055, China (e-mail: yp.cai@siat.ac.cn).

Digital Object Identifier 10.1109/TCBB.2024.3385635

comprehensively summarized the above identified tools of AVPs from the feature encoding, classifiers, cross-validation and performance [33]. In addition, deep neural network methods also were employed to extract the high dimensional features for the identification of AVPs from the primary sequence. DeepPhy and DeepEvo were two different dual-channel deep neural network ensemble models of DeepAvp method[34].

Although the existing methods achieved good performance [35], [36], they are not satisfactory for drug development. Effective feature extraction posed a challenge to further improving of model performances [37], [38]. However, the number of underlying mechanisms relevant to AVP are diversified and manual extracted features, which is adopted by most current methods, are difficult to fully discover their functional characteristics [39]. Neural networks have the potential of identifying novel features but also pose the risk of biases on small data [40]. In this work, we proposed a novel hybrid stacked network architecture for antiviral peptides prediction, named HybAVPnet.

To learn the effective features, HybAVPnet is consisted of a two-layer prediction models which are mixed of traditional machine learning models and deep learning models. In the first layer, two neural network and one group of LightGBM classifiers were employed to extract the different aspects of features using one-hot coding, composition, autocorrelation, and profile for amino acid sequences [41]. For the second layer, all the probability and label outputs of the first layer were fed into SVM classifier to obtain the final prediction [42]. The experimental results showed that HybAVPnet could achieve competitive advantages compared with the existing methods.

II. Material and Methods

A. Datasets

In order to compare our model with others, we use two groups of datasets from AVPpred. One dataset contains 604 AVPs with experimentally validated antiviral activities and 452 non-AVPs proved to be invalid, which is divided into training and testing subsets, named training set $T^{544P + 407N}$ (544 positive and 407 negative samples) and validation dataset $V^{60P + 45N}$ (60 positive and 45 negative samples). The another dataset consists of 604 effective AVPs and 604 non-experimental negative peptides from AntiBP2 [43], which is also divided into training and testing subsets, named training set $T^{544P + 544N}$ (544 positive and 544 negative samples) and validation dataset $V^{60P + 60N}$ (60 positive and 60 negative samples). After residues are excluded which do not include in the canonical 20 amino acids, the sequences of AVPs and non-AVPs were statistically analyzed and the amino acid frequency distribution of the datasets was shown in Fig. 1. It clearly showed that the frequency of amino acid “W” in the positive samples was high. However, there are no obvious rules for the distribution of other amino acids.

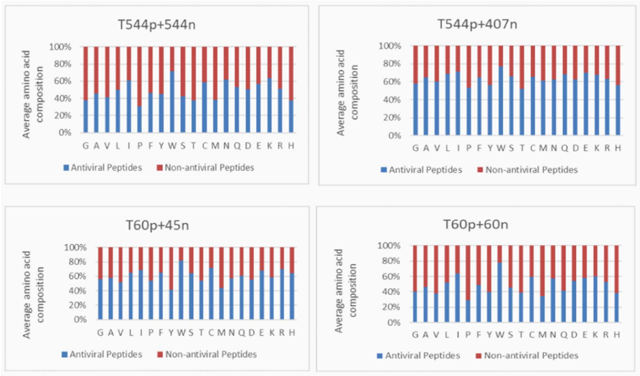


Fig. 1. Amino acid frequency distribution of AVPs and non-AVPs. The blue and red bars represent the amino acid frequency distribution of antiviral peptides and nonantiviral peptides respectively.

TABLE I
18 Kinds of Feature Representation Methods Based on Protein Primary Sequences

Category	Feature
Amino acid composition	Basic Kmer (kmer)
	Distance-based Residue(DR)
	Distance Pair(DP)
Autocorrelation	Auto covariance(feature-AC)
	Auto-cross covariance(ACC)
	Cross covariance(feature-CC)
	Physicochemical distance transformation(PDT)
Pseudo amino acid composition	Parallel correlation pseudo amino acid composition(PC-PseAAC)
	Series correlation pseudo amino acid composition(SC-PseAAC)
	General parallel correlation pseudo amino acid composition(PC-PseAAC-General)
	General series correlation pseudo amino acid composition(SC-PseAAC-General)
Profile-based features	Select and combine the n most frequent amino acids according to their Frequencies(Top-n-gram)
	Profile-based Physicochemical distance transformation(PDT-Profile)
	Distance-based Top-n-gram(DT)
	Profile-based Auto covariance(AC-PSSM)
	Profile-based Cross covariance(CC-PSSM)
	Profile-based Distance-based Top-n-gram(PSSM-DT)
	Profile-based Auto-cross covariance(ACC-PSSM)

B. Feature Representation

Considering the composition, frequency, physical and chemical properties of the sequence and other information, many features were extracted from the amino acid sequence [44]. Among of them, three kinds of features were extracted based on amino acid composition: Basic Kmer (kmer), Distance-based Residue (DR) and Distance Pair (DP) [45]. Four kinds of features were generated according to autocorrelation: auto covariance (feature-AC), auto-cross covariance (ACC), cross covariance (feature-CC), and physicochemical distance transformation (PDT) [46]. Based on pseudo amino acid composition (PseAAC) and frequency profile, we extracted four and seven kinds of features respectively [47]. In total, there are 18 kinds of features which were listed in Table I. Furthermore, all the features were also input into the neural network to explore the potential relationships between them. In addition, one-hot encoding method in natural language processing was employed

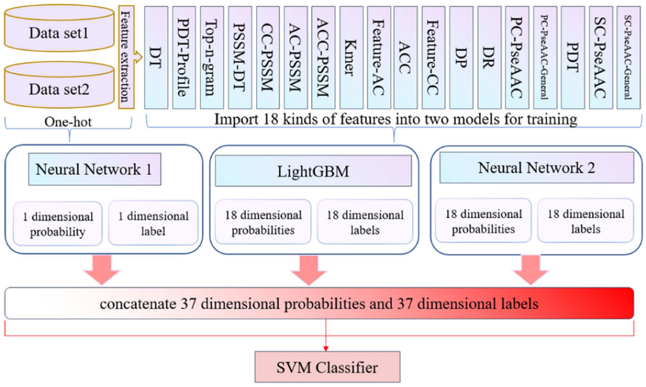


Fig. 2. HybAVPnet model architecture.

to extract the high dimensional features into the neural network structure [48].

C. Machine Learning Approaches

HybAVPnet identifies antiviral peptides by integrating several machine learning methods, i.e., Light Gradient Boosting Machine (LightGBM), SVM, Convolutional Neural Networks (CNN), and Bidirectional Long Short Term Memory (Bi-LSTM)[49].

In the first layer of HybAVPnet, LightGBM is chosen as the predictor, which is a gradient boosting framework. The LightGBM is based on decision tree algorithms and supports efficient parallel training, with the advantages of faster training speed, lower memory consumption, better accuracy, distributed support, and rapid processing of massive data. SVM is a binary classifier, widely used in the supervised machine learning tasks. It is trying to find the best separated hyperplane in the feature spaces, and maximizes the interval between positive and negative samples on the training set, which makes it different from the perceptron. SVM performs effective in high dimensional spaces. And its kernel can be specified to solve the different problems. CNN (CNN1D) is a kind of feed forward neural network with convolution calculation. It is one of the representative algorithms of deep learning. CNN1D is widely used in sequence models. LSTM is a form of Recurrent Neural Network (RNN), which can take into account the relationship between front and back. So it is often used in sequence model. Bi-LSTM is a combination of forward LSTM and backward LSTM.

D. Computational Model

The framework of the whole model HybAVPnet is shown in Fig. 2, which is composed of three sub models: Neural Network1, LightGBM and Neural Network2. Considering that amino acid sequence has its related characteristics, we adopt a series of feature extraction methods to obtain a total of 18 kinds of features. Each kind of features are trained and predicted through LightGBM predictor and Neural Network2 to obtain the initial predicted results.

HybAVPnet consists of two parts, of which the first part includes three sub-models. The first sub-model unifies the protein sequences with different lengths into the same length. Then, the sequences are vectorized according to the specific one-hot coding form. The coded vectors are inputted into Neural Network1 to obtain its classification probabilities and classification labels as shown in Fig. 3. Through the embedding layer, each vector is converted into 64 dimensions (embedding (input_ dimensions = 26, output_ dimensions = 64, input_length = 1000)). Then the vectors are inputted into Conv1d (filters = 32, kernel_size = 1, activation = ‘relu’, strings = 1)). Finally, the outputs of Conv1d are imported into Bi-LSTM layer (bidirectional (LSTM (64, return_sequences = true)). Then, the network obtains the predicted labels and probabilities through two fully connected layers.

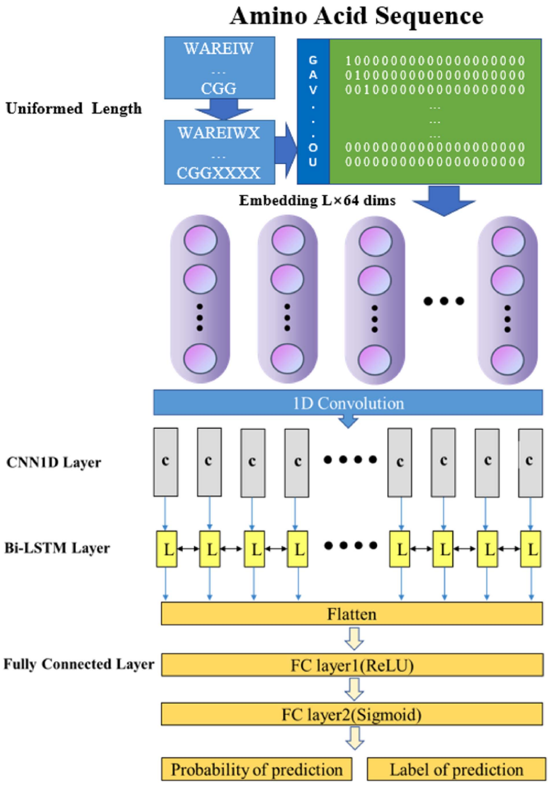


Fig. 3. Neural Network 1 framework.

The second sub model inputs the extracted 18 kinds of features into the LightGBM classifier for training and classification, and achieves the 18 dimensional classification probabilities and classification labels respectively. The third sub model Neural Network 2 also inputs 18 kinds of features into network similar to Neural Network1 without the embedding layer for training and classification. Through the above steps, we obtained a series of initial prediction results. Considering that the factors of predicted probability may have a great impact on the final results, both the probabilities and labels are inputted into the next layer of the network architecture. Finally, a total of 74 dimensional data from the classification probability and classification labels of three sub models is used as the training set for the next classifier.

In the last layer, some machine learning classifiers are evaluated to find the optimal solution, here we focuses on SVM, LightGBM, Bayes, Decision tree, KNN. Through the comparative experiments, SVM is selected as the final classifier.

III. Performance Evaluation

In the experiments, the following metrics were employed to verify the prediction performance of HybAVPnet, including Receiver Operating Characteristic curve (ROC), Sensitivity (Sn), Specificity (Sp), Accuracy (Acc), and the Matthews correlation coefficient (MCC)[50]. Five-fold cross-validation and independent test were conducted to evaluate the model on different datasets.

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

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

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

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

Where TP, FP, TN and FN indicate the number of true positives, false positives, true negatives, and false negatives respectively.

IV. Results and Discussion

A. Comparison With the Base Classifiers

Five-fold cross-validation was involved to evaluate the model in the training datasets $T^{544P + 407N}$, $T^{544P + 544N}$ and the independent testing datasets $V^{60P + 45N}$, $V^{60P + 60N}$ compared with a few base classifiers such as Random Forest, K-NN, LightGBM, Naïve Bayes, and Logistic Regression. As shown in Table II, HybAVPnet almost outperformes all the base classifiers on the four datasets except the sensitivity on $V^{60P + 60N}$ dataset compared with Logistic Regression.

B. Comparison With the State-of-Art Methods

To verify the effectiveness of HybAVPnet, five state-of-art predicted methods are employed for comparative analysis on the four datasets. The experimental results show that HybAVPnet performs significantly better than other models in $T^{544P + 407N}$, $T^{544P + 544N}$, $V^{60P + 45N}$ and $V^{60P + 60N}$ datasets. In the dataset $V^{60P + 60N}$, HybAVPnet is slightly lower than DeepEvo by 1.7% on sensitivity. To sum up, HybAVPnet achieves the best performance compared with other existing models in terms of evaluation on cross-validation and testing datasets as shown in Table III. Compared with other direct classification models, the classification method combining initial prediction maybe obtain better performance, such as DeepAvp and HybAVPnet. On the datasets $T^{544P + 407N}$ and $V^{60P + 45N}$, the performances of most predicting models are not good except for our method HybAVPnet.

C. Ablation Experiments

In the selection of sub-model combinations, ablation experiments were conducted to determine the best combination. The SVM classifier is adopted as the last layer for each model. Seven different models are analyzed with LightGBM, Neural Netwok 1, Neural Network 2, and their different fused models. The average values of the evaluated indicators for the four experiments are taken as the final experimental results. The final results of the ablation experiments are shown in Table IV. It can be found that the results of the fused model HybAVPnet are better than other models whether on the testing set $V^{60P+45N}$ or $V^{60P + 60N}$ after optimizing the parameters of SVM.

TABLE II
Comparison of HybAVPnet with Base Classifiers on Five-fold Cross-validation and Independent Test Datasets

Dataset	Model	Acc(%)	Sn(%)	Sp(%)	MCC
$V^{544P+407N}$	Random Forest	84.34	85.09	83.15	0.68
	KNN	66.03	78.65	48.61	0.29
	LightGBM	85.09	86.95	82.53	0.70
	Naïve Bayes	78.17	77.07	80.51	0.57
	Logistic Regression	76.89	81.97	69.87	0.52
	HybAVPnet	**93.08**	**90.82**	**96.2**	**0.86**
$V^{544P+544N}$	Random Forest	88.89	85.41	92.52	0.78
	KNN	82.69	87.11	78.63	0.66
	LightGBM	91.30	88.09	94.53	0.83
	Naïve Bayes	86.76	82.24	91.24	0.74
	Logistic Regression	86.02	87.31	84.80	0.72
	HybAVPnet	**95.83**	**94.17**	**97.34**	**0.92**
$V^{60P+45N}$	Random Forest	80.77	88.33	70.45	0.60
	KNN	76.93	88.33	61.36	0.52
	LightGBM	82.69	88.33	75.00	0.64
	Naïve Bayes	78.85	81.67	75.00	0.57
	Logistic Regression	79.80	88.33	68.18	0.58
	HybAVPnet	**93.27**	**95.00**	**90.91**	**0.86**
$V^{60P+60N}$	Random Forest	92.37	95.00	89.65	0.84
	KNN	83.90	88.33	84.48	0.68
	LightGBM	92.37	93.33	91.38	0.84
	Naïve Bayes	84.74	81.67	87.93	0.70
	Logistic Regression	94.91	**98.33**	91.38	0.90
	HybAVPnet	**96.61**	95.00	**98.28**	**0.93**

The bold fonts indicate the best results.

It can be seen from Table IV that in the testing set $V^{60P + 45N}$, HybAVPnet performs better than other models in all evaluated indicators. However, in the testing set $V^{60P + 60N}$, compared with the LightGBM model, HybAVPnet leads it by 1.69% in accuracy, 3.45% in terms of specificity, 0.03% in terms of MCC, and the same in terms of sensitivity. Therefore, the combined output results of the three sub-models are chosen as the final experimental results for the input of the next SVM classifier after compared with the different fused models. Furthermore, the final experiments prove that the predicted probability has an important impact on the final classified evaluation. So we choose to integrate the predicted probabilities and the predicted labels into the final model. The results prove the fused model may have a strong sense of discrimination in the identification of antiviral peptides compared with a single ensemble model or neural network model.

TABLE III
COMPARISON OF HYBAVPNET WITH EXISTING METHODS ON FIVE-FOLD CROSS-VALIDATION AND INDEPENDENT TEST DATASETS

Data set	Model	Acc(%)	Sn(%)	Sp(%)	MCC
$T^{544P+407N}$	AVPpred	85.0	82.2	88.2	0.70
	Chang et al's method	85.1	86.6	83.0	0.70
	AntiVPP 1.0	-	-	-	-
	DeepPhy	88.0	85.5	79.7	0.65
	DeepEvo	83.5	84.6	82.1	0.66
	HybAVPnet	**93.08**	**90.82**	**96.2**	**0.86**
$T^{544P+544N}$	AVPpred	90.0	89.7	90.3	0.80
	Chang et al's method	91.5	89.0	94.1	0.83
	AntiVPP 1.0	-	-	-	-
	DeepPhy	88.5	88.0	89.0	0.77
	DeepEvo	90.1	89.3	90.8	0.80
	HybAVPnet	**95.83**	**94.17**	**97.34**	**0.92**
$V^{60P+45N}$	AVPpred	85.7	88.3	82.2	0.71
	Chang et al's method	89.5	91.7	86.7	0.79
	AntiVPP 1.0	-	-	-	-
	DeepPhy	80	83.3	75.6	0.59
	DeepEvo	87.60	90.00	84.40	0.75
	HybAVPnet	**93.27**	**95.00**	**90.91**	**0.86**
$V^{60P+60N}$	AVPpred	92.5	93.3	91.7	0.85
	Chang et al's method	93.0	91.7	95.0	0.87
	AntiVPP 1.0	93	87	97	0.87
	DeepPhy	89.2	88.3	90	0.78
	DeepEvo	93.30	**96.70**	90.00	0.87
	HybAVPnet	**96.61**	95.00	**98.28**	**0.93**

The bold fonts indicate the best results.

TABLE IV
COMPARISON OF THREE SUB MODELS AND JOINT MODEL IN TWO INDEPENDENT TEST DATASETS

Dataset	Model	Acc(%)	Sn(%)	Sp(%)	MCC
$V^{60P+45N}$	LG	89.42	91.67	86.36	0.78
	NN2	88.27	94.67	83.48	0.78
	NN1	75.58	86.00	61.36	0.49
	LG+NN1	90.38	91.67	88.64	0.80
	LG+NN2	90.38	93.33	86.36	0.80
	NN1+NN2	90.38	**95.00**	84.09	0.80
	HybAVPnet	**93.27**	**95.00**	**90.91**	**0.86**
$V^{60P+60N}$	LightGBM	94.92	**95.00**	94.83	0.90
	Neural Network 2	94.75	93.33	96.21	0.90
	Neural Network 1	85.59	85.67	85.52	0.71
	LG+NN1	93.22	93.33	93.10	0.86
	LG+NN2	94.92	95	94.83	0.90
	NN1+NN2	**96.61**	93.33	**1**	**0.93**
	HybAVPnet	**96.61**	**95.00**	98.28	**0.93**

The bold fonts indicate the best results. LG: LightGBM; NN1: Neural Network 1; NN2: Neural Network 2; HybAVPnet: LightGBM+ Neural Network 1+ Neural Network 2.

D. Comparison With the Different Classifiers

After the pre-classification of the three sub-models in the first step, 74 dimensional initial predicted results were used as the new training set. In the selection of the classifier in the second step as shown in Table V, a few of traditional machine learning classifiers were adopted to analyze their performances using SVM, Random Forest, LightGBM, Bayes, Decision tree, and KNN classifiers on $V^{60P+45N}$ and $V^{60P+60N}$ datasets.

TABLE V
COMPARISON OF THE DIFFERENT CLASSIFIERS IN THE LAST LAYER OF THE MODEL ARCHITECTURE

Data set	Model	Acc	Sn	Sp	MCC
$V^{60P+45N}$	SVM	**93.27**	95.00	**90.91**	**0.86**
	Random Forest	85.58	93.33	75.00	0.70
	LightGBM	81.73	86.67	75.00	0.62
	Naive Bayes	80.96	**97.66**	58.19	0.63
	Decision Tree	81.15	87.00	73.18	0.61
	KNN	90.77	93.00	87.73	0.81
$V^{60P+60N}$	SVM	**96.61**	**95.00**	**98.28**	**0.93**
	Random Forest	95.25	94.67	95.86	0.91
	LightGBM	87.29	78.33	96.55	0.76
	Naive Bayes	96.27	**95.00**	97.59	**0.93**
	Decision Tree	87.97	89.00	86.89	0.76
	KNN	96.10	**95.00**	97.93	**0.93**

The bold fonts indicate the best results.

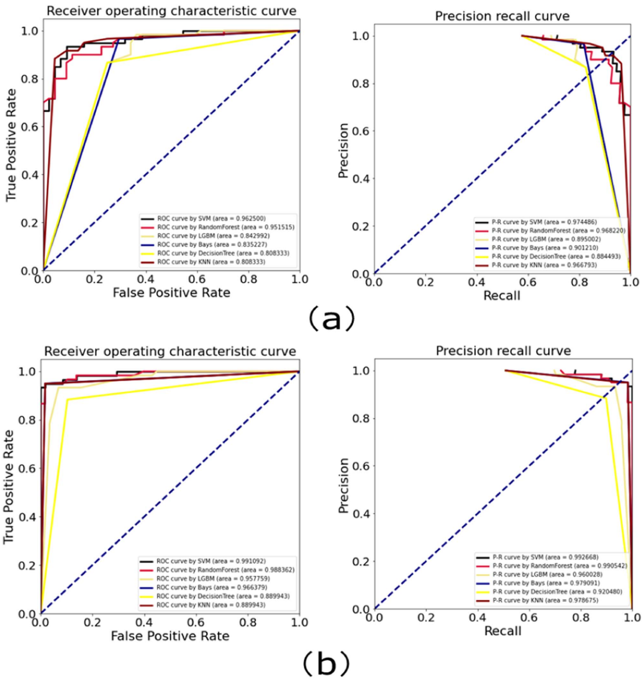


Fig. 4. Receiver operating characteristic (ROC) and precision recall (PR) curve of (a) $V^{60P+45N}$ and (b) $V^{60P+60N}$ datasets.

From Table V, we can see on testing set $V^{60p+45n}$, the performance of SVM and KNN is much better than other classifiers. Compared with KNN, SVM achieves 2.5%, 2%, 3.18% and 0.05 higher respectively in Acc, Sn, Sp and MCC. On testing set $V^{60p+60n}$, SVM performs better than other relatively good classifiers Bayes and KNN by 0.34% and 0.51% in Acc respectively, and similarly well in Sn and MCC. While in Sp, SVM achieves better performances than Bayes and KNN by 0.69% and 0.35% respectively.

Furthermore, Receiver operating characteristic (ROC) curve and Precision-Recall (PR) curve are drawn to evaluate the performance of each methods for intuitive comparison, as shown in Fig. 4. AUC represents the area under the ROC curve, which is plotted the true positive rate against false positive rate. AUPR stands for the area under PR curve that is plotted precision against recall. On the independent datasets $V^{60P+45N}$ and $V^{60P+60N}$, SVM can obtain the best balance in performances compared with Random Forest, LightGBM, Bayes, Decision

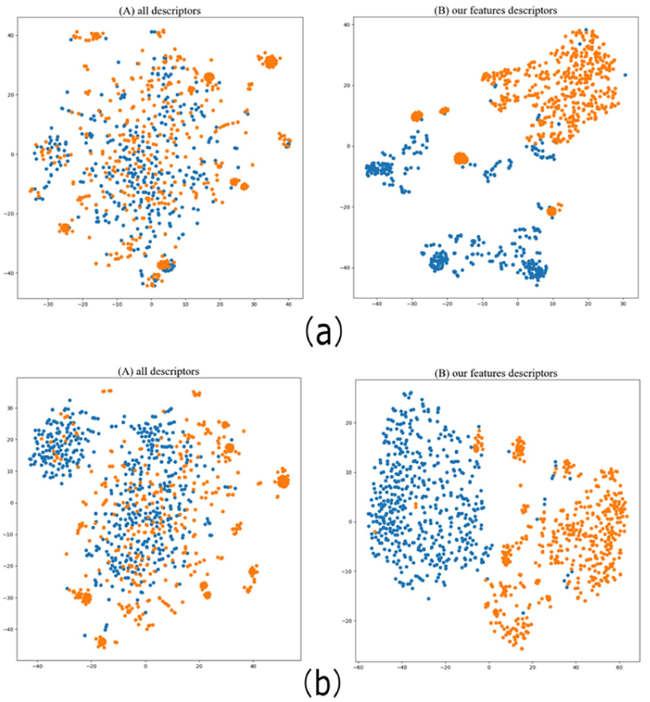


Fig. 5. t-distributed stochastic neighborhood embedding (t-SNE) visualization of all descriptors and our features descriptors in (a) $T^{544P + 407N}$ and (b) $T^{544P + 544N}$ datasets, respectively. The blue dot represents the distribution of nonantiviral peptides, and the orange dot represents the distribution of antiviral peptides.

tree and KNN classifiers. Therefore, SVM is selected as the last layer classifier.

E. Visual Analysis

To better interpret the feature representation between the sub models, we adopted t-distributed stochastic neighborhood embedding (t-SNE) to visualize and compare the feature space distribution on $T^{544P + 407N}$ and $T^{544P + 544N}$ datasets [51]. In the experiment of t-SNE, as shown in Fig. 5, the new feature distribution in HybAVPnet is the most efficient and effective compared with all features descriptors to discriminate AVPs from non-AVPs.

V. Conclusion

Due to their advantages and good performance, antiviral peptides have potential wide applications in the development of antiviral drugs. To this end, some computational models have been developed to quickly and accurately identify AVPs. In this work, we present a novel hybrid network tool named HybAVPnet to predict AVPs. HybAVPnet takes full advantage of traditional machine learning models and deep learning models to obtain the effective feature representation of amino acid sequences at sequential, structural, and evolutionary levels. HybAVPnet can capture the long-term dependencies and semantic dependence. As a new dimension, predictive probability increases the ability of feature expression to a certain extent. Experimental results demonstrated the proposed HybAVPnet model could achieve more discriminative power for the prediction of AVPs and could be easier to separate the positive samples and negative samples. Furthermore, a serial of comparative experiments showed the consistently stability and robustness of HybAVPnet from the five-fold cross-validation and independent test. We expect that HybAVPnet can help the development of antiviral peptide drugs and the treatment of related diseases for researchers [52]. In the future, we will strive to develop predictive models for various therapeutic peptides to better serve precision medicine [53].

References

[1] S. Calvignac-Spencer, A. Dux, J. F. Gogarten, F. H. Leendertz, and L. V. Patrono, “A great ape perspective on the origins and evolution of human viruses,” *Adv. Virus Res.*, vol. 110, pp. 1–26, 2021.

[2] M. M. Islam and D. Koirala, “Toward a next-generation diagnostic tool: A review on emerging isothermal nucleic acid amplification techniques for the detection of SARS-CoV-2 and other infectious viruses,” *Analytica Chimica Acta*, vol. 1209, May 29, 2022, Art. no. 339338.

[3] J. A. Jackman, “Antiviral peptide engineering for targeting membrane-enveloped viruses: Recent progress and future directions,” *Biochim Biophys Acta Biomembr*, vol. 1864, no. 2, Feb. 1, 2022, Art. no. 183821.

[4] F. Zarif, M. I. Anasir, J. X. Koh, M. F. Chew, and C. L. Poh, “Stability and antiviral activity of SP40 peptide in human serum,” *Virus Res.*, vol. 303, Oct. 2, 2021, Art. no. 198456.

[5] S. Mahmud et al., “Antiviral peptides against the main protease of SARS-CoV-2: A molecular docking and dynamics study,” *Arabian J. Chem.*, vol. 14, no. 9, Sep. 2021, Art. no. 103315.

[6] H. Heydari, R. Golmohammadi, R. Mirnejad, H. Tebyanian, M. Fasihi-Ramandi, and M. Moosazadeh Moghaddam, “Antiviral peptides against Coronaviridae family: A review,” *Peptides*, vol. 139, May 2021, Art. no. 170526.

[7] B. K. Maiti, “Potential role of peptide-based antiviral therapy against SARS-CoV-2 infection,” *ACS Pharmacol. Transl. Sci.*, vol. 3, no. 4, pp. 783–785, Aug. 14, 2020.

[8] M. Saito et al., “Macrocyclic peptides exhibit antiviral effects against influenza virus HA and prevent pneumonia in animal models,” *Nat. Commun.*, vol. 12, no. 1, May 11, 2021, Art. no. 2654.

[9] H. Shin, S. J. Park, J. Kim, J. S. Lee, and D. H. Min, “A graphene oxide-based fluorescent nanosensor to identify antiviral agents via a drug repurposing screen,” *Biosensors Bioelectron.*, vol. 183, Jul. 1, 2021, Art. no. 113208.

[10] A. N. Zelikin and F. Stellacci, “Broad-spectrum antiviral agents based on multivalent inhibitors of viral infectivity,” *Adv. Healthcare Mater.*, vol. 10, no. 6, Mar. 2021, Art. no. e2001433.

[11] T. Hu, O. Agazani, S. Nir, M. Cohen, S. Pan, and M. Reches, “Antiviral activity of peptide-based assemblies,” *ACS Appl. Mater. Interfaces*, vol. 13, no. 41, pp. 48469–48477, Oct. 20, 2021.

[12] N. A. Murugan, K. M. P. Raja, and N. T. Saraswathi, “Peptide-Based Antiviral Drugs,” *Adv. Exp. Med. Biol.*, vol. 1322, pp. 261–284, 2021.

[13] A. Hollmann, N. P. Cardoso, J. C. Espeche, and P. C. Maffia, “Review of antiviral peptides for use against zoonotic and selected non-zoonotic viruses,” *Peptides*, vol. 142, Aug. 2021, Art. no. 170570.

[14] S. Saikia, M. Bordoloi, R. Sarmah, and B. Kolita, “Antiviral compound screening, peptide designing, and protein network construction of influenza a virus (strain a/Puerto Rico/8/1934 H1N1),” *Drug Develop. Res.*, vol. 80, no. 1, pp. 106–124, Feb. 2019.

[15] K. Yan, H. Lv, Y. Guo, Y. Chen, H. Wu, and B. Liu, “TPpred-ATMV: Therapeutic peptide prediction by adaptive multi-view tensor learning model,” *Bioinformatics*, vol. 38, no. 10, pp. 2712–2718, May 13, 2022.

[16] N. Schaduangrat, C. Nantasenamat, V. Prachayasittikul, and W. Shoombuatong, “Meta-iAVP: A sequence-based meta-predictor for improving the prediction of antiviral peptides using effective feature representation,” *Int. J. Mol. Sci.*, vol. 20, no. 22, Nov. 15, 2019, Art. no. 5743.

[17] X. Liang et al., “Antiviral effects of Bovine antimicrobial peptide against TGEV in vivo and in vitro,” *J. Vet. Sci.*, vol. 21, no. 5, Sep. 2020, Art. no. e80.

[18] I. Sadeghian, R. Heidari, S. Sadeghian, M. J. Raee, and M. Negahdaripour, “Potential of cell-penetrating peptides (CPPs) in delivery of antiviral therapeutics and vaccines,” *Eur. J. Pharmaceut. Sci.*, vol. 169, Feb. 1, 2022, Art. no. 106094.

[19] S. Basith, B. Manavalan, T. H. Shin, and G. Lee, “Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening,” *Med. Res. Rev.*, vol. 40, no. 4, pp. 1276–1314, Jul. 2020.

[20] S. Mahmud et al., “Prospective role of peptide-based antiviral therapy against the main protease of SARS-CoV-2,” *Front. Mol. Biosciences*, vol. 8, 2021, Art. no. 628585.

[21] S. Nakkeeran et al., “Mining the genome of bacillus velezensis VB7 (CP047587) for MAMP genes and non-ribosomal peptide synthetase gene clusters conferring antiviral and antifungal activity,” *Microorganisms*, vol. 9, no. 12, Dec. 3, 2021, Art. no. 2511.

[22] A. Qureshi, H. Tandon, and M. Kumar, “AVP-IC50 Pred: Multiple machine learning techniques-based prediction of peptide antiviral activity in terms of half maximal inhibitory concentration (IC50),” *Biopolymers*, vol. 104, no. 6, pp. 753–763, Nov. 2015.

[23] A. Qureshi, N. Thakur, H. Tandon, and M. Kumar, “AVPdb: A database of experimentally validated antiviral peptides targeting medically important viruses,” *Nucleic Acids Res.*, vol. 42, pp. D1147–D1153, Jan. 2014.

[24] A. Qureshi, N. Thakur, and M. Kumar, “HIPdb: A database of experimentally validated HIV inhibiting peptides,” *PLoS One*, vol. 8, no. 1, 2013, Art. no. e54908.

[25] G. Wang, X. Li, and Z. Wang, “APD3: The antimicrobial peptide database as a tool for research and education,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1087–D1093, Jan. 4, 2016.

[26] S. Thomas, S. Karnik, R. S. Barai, V. K. Jayaraman, and S. Idicula-Thomas, “CAMP: A useful resource for research on antimicrobial peptides,” *Nucleic Acids Res.*, vol. 38, pp. D774–D780, Jan. 2010.

[27] N. Thakur, A. Qureshi, and M. Kumar, “AVPpred: Collection and prediction of highly effective antiviral peptides,” *Nucleic Acids Res.*, vol. 40, pp. W199–W204, Jul. 2012.

[28] K. Y. Chang and J. R. Yang, “Analysis and prediction of highly effective antiviral peptides based on random forests,” *PLoS One*, vol. 8, no. 8, 2013, Art. no. e70166.

[29] M. Zare, H. Mohabatkar, F. K. Faramarzi, M. M. Beigi, and M. Behbahani, “Using chou’s pseudo amino acid composition and machine learning method to predict the antiviral peptides,” *Open Bioinf. Journa*, vol. 9, pp. 13–19, 2015.

[30] J. F. Beltran Lissabet, L. H. Belen, and J. G. Farias, “AntiVPP 1.0: A portable tool for prediction of antiviral peptides,” *Comput Biol. Med.*, vol. 107, pp. 127–130, Apr. 2019.

[31] L. Wei, C. Zhou, R. Su, and Q. Zou, “PEPred-Suite: Improved and robust prediction of therapeutic peptides using adaptive feature representation learning,” *Bioinformatics*, vol. 35, no. 21, pp. 4272–4280, Nov. 1, 2019.

[32] A. S. Chowdhury, S. M. Reehl, K. Kehn-Hall, B. Bishop, and B. M. Webb-Robertson, “Better understanding and prediction of antiviral peptides through primary and secondary structure feature importance,” *Sci. Rep.*, vol. 10, no. 1, Nov. 6, 2020, Art. no. 19260.

[33] P. Charoenkwan, N. Anuwongcharoen, C. Nantasenamat, M. M. Hasan, and W. Shoombuatong, “In silico approaches for the prediction and analysis of antiviral peptides: A review,” *Curr. Pharmaceut. Des.*, vol. 27, no. 18, pp. 2180–2188, 2021.

[34] J. Li, Y. Pu, J. Tang, Q. Zou, and F. Guo, “DeepAVP: A dual-channel deep neural network for identifying variable-length antiviral peptides,” *IEEE J. Biomed. Health Inform.*, vol. 24, no. 10, pp. 3012–3019, Oct. 2020.

[35] Y. Pang, L. Yao, J. H. Jhong, Z. Wang, and T. Y. Lee, “AVPIden: A new scheme for identification and functional prediction of antiviral peptides based on machine learning approaches,” *Brief. Bioinf.*, vol. 22, no. 6, Nov. 5, 2021, Art. no. bbab263.

[36] P. B. Timmons and C. M. Hewage, “ENNAVIA is a novel method which employs neural networks for antiviral and anti-coronavirus activity prediction for therapeutic peptides,” *Brief. Bioinf.*, vol. 22, Jul. 23, 2021, Art. no. bbab258.

[37] G. Agarwal and R. Gabrani, “Antiviral Peptides: Identification and Validation,” *Int. J. Peptide Res. Therapeutics*, vol. 27, no. 1, pp. 149–168, 2021.

[38] C. Zannella et al., “Broad-spectrum antiviral activity of the amphibian antimicrobial peptide temporin l and its analogs,” *Int. J. Mol. Sci.*, vol. 23, no. 4, Feb. 13, 2022, Art. no. 2060.

[39] S. Akbar, F. Ali, M. Hayat, A. Ahmad, S. Khan, and S. Gul, “Prediction of Antiviral peptides using transform evolutionary & SHAP analysis based descriptors by incorporation with ensemble learning strategy,” *Chemometrics Intell. Lab. Syst.*, vol. 230, Nov. 15, 2022, Art. no. 104682.

[40] H. Kurata, S. Tsukiyama, and B. Manavalan, “iACVP: Markedly enhanced identification of anti-coronavirus peptides using a dataset-specific word2vec model,” *Brief. Bioinf.*, vol. 23, no. 4, Jul. 18, 2022, Art. no. bbac265.

[41] J. Yan et al., “LightGBM: Accelerated genomically designed crop breeding through ensemble learning,” *Genome Biol.*, vol. 22, no. 1, Sep. 20, 2021, Art. no. 271.

[42] D. B. Vukovic, K. Romanyuk, S. Ivashchenko, and E. M. Grigorieva, “Are CDS spreads predictable during the COVID-19 pandemic? Forecasting based on SVM, GMDH, LSTM and Markov switching autoregression,” *Expert Syst. Appl.*, vol. 194, May 15, 2022, Art. no. 116553.

[43] S. Lata, N. K. Mishra, and G. P. Raghava, “AntiBP2: Improved version of antibacterial peptide prediction,” *BMC Bioinf.*, vol. 11, Jan. 18, 2010, Art. no. S19.

[44] B. Liu, “BioSeq-Analysis: A platform for DNA, RNA and protein sequence analysis based on machine learning approaches,” *Brief. Bioinf.*, vol. 20, no. 4, pp. 1280–1294, Jul. 19, 2019.

[45] B. Liu, H. Wu, and K.-C. Chou, “Pse-in-one 2.0: An improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences,” *Natural Sci.*, vol. 09, no. 04, pp. 67–91, 2017.

[46] X. Jing, Q. Dong, D. C. Hong, and R. Lu, “Amino acid encoding methods for protein sequences: A comprehensive review and assessment,” *EEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, pp. 1918–1931, Apr. 16, 2019.

[47] S. M. Krishnan, “Using Chou’s general PseAAC to analyze the evolutionary relationship of receptor associated proteins (RAP) with various folding patterns of protein domains,” *J Theor Biol*, vol. 445, pp. 62–74, May 14, 2018.

[48] S. Okada, M. Ohzeki, and S. Taguchi, “Efficient partition of integer optimization problems with one-hot encoding,” *Sci. Rep.*, vol. 9, no. 1, Sep. 10, 2019, Art. no. 13036.

[49] S. Dai, Y. Ding, Z. Zhang, W. Zuo, X. Huang, and S. Zhu, “GrantExtractor: Accurate grant support information extraction from biomedical fulltext based on Bi-LSTM-CRF,” *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 1, pp. 205–215, Jan./Feb. 2021.

[50] S. Mei et al., “A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction,” *Brief. Bioinf.*, vol. 21, no. 4, pp. 1119–1135, Jul. 15, 2020.

[51] D. Kobak and G. C. Linderman, “Initialization is critical for preserving global data structure in both t-SNE and UMAP,” *Nature Biotechnol.*, vol. 39, no. 2, pp. 156–157, Feb. 2021.

[52] T. Todorovski, D. Kalafatovic, and D. Andreu, “Antiviral peptide-based conjugates: State of the art and future perspectives,” *Pharmaceutics*, vol. 15, no. 2, p. 357, Feb. 2023.

[53] G. Wang, C. M. Zietz, A. Mudgapalli, S. Wang, and Z. Wang, “The evolution of the antimicrobial peptide database over 18 years: Milestones and new features,” *Protein Sci*, vol. 31, no. 1, pp. 92–106, Jan. 2022.

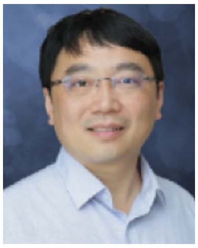
Ruiquan Ge (Member, IEEE) received the PhD degree in computer science from the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, in 2017. Currently, he is an associate professor with the School of Computer Science and Technology, Hangzhou Dianzi University. His research interests include bioinformatics, health information, machine learning, and medical imaging.

Yixiao Xia is currently working toward the PhD degree with the School of Computer Science and Technology, Hangzhou Dianzi University. His research interests include cloud computing and bioinformatics.

Minchao Jiang is currently working toward the master's degree with the School of Computer Science and Technology, Hangzhou Dianzi University. He is particularly interested in data mining and computer vision.

Gangyong Jia (Member, IEEE) received the PhD degree from the Department of Computer Science, University of Science and Technology of China, Hefei, China, in 2013. He is currently an associate professor with the Department of Computer Science, Hangzhou Dianzi University, China. He has served as a reviewer of Microprocessors and Microsystems. His current research interests include edge AI, IoT, cloud computing, and operating system. He is a member of CCF.

Xiaoyang Jing received the PhD degree from the School of Computer Science and Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China. He is a postdoctoral with Toyota Technological Institute at Chicago, Chicago, IL, USA. His main research interests include bioinformatics, Big Data, and machine learning.



Ye Li (Member, IEEE) received the BS and MS degrees in electrical engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 1999 and 2002, respectively, and the PhD degree in electrical engineering from Arizona State University, AZ. He is a professor with the Shenzhen Institute of Advanced Technology (SIAT), Chinese Academy of Sciences. In 2007, He worked in Cadence Design Systems, Inc., San Jose, CA, the U.S. Since 2008, he is the director of the Research Center for Biomedical Information Technology in SIAT. His research interests include body sensor networks, wearable computing, and health data mining.

Yunpeng Cai (Member, IEEE) received the PhD degree in computer science and technology from Tsinghua University, Beijing, China, in 2007. He is currently a professor with Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. His research interests include health Big Data, health informatics, bioinformatics, and machine learning.