Traditional Machine Learning Fusion Model: An Efficient and Accurate Machine Learning Model for Antiviral Peptide Identification

Zejun Lang[1], Yang Zhao[2]

[1]School of Mathematics, Hohai University, Nanjing, China, 211100
[2]College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, China, 266590

Abstract: *The rapid mutation and spread of viral diseases have intensified the challenge of drug resistance to traditional antivirals, making the development of new antiviral agents crucial. Antiviral peptides (AVPs) have emerged as promising candidates due to their unique membrane penetration mechanisms and low resistance risk. However, conventional experimental screening methods are time-consuming and costly, while existing machine learning approaches suffer from limitations in feature representation and generalization capabilities. This study proposes an ensemble machine learning model, AVP, designed to identify antiviral peptides efficiently and accurately. The model integrates Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT) classifiers using a soft-voting architecture with probability-based weighting. Regularization strategies, including L2 regularization for SVM and depth constraints for DT, are applied to enhance model stability. The model's performance is evaluated using five-fold cross-validation and ROC analysis. The AVP model achieves a training set AUC of 0.9980 and a test set AUC of 0.9784, demonstrating superior classification capability and generalization performance compared to traditional machine learning models. This study highlights the effectiveness of ensemble learning in fusing diverse feature-response patterns and provides a robust tool for antiviral peptide identification, accelerating the development of next-generation antiviral agents.*

Keywords: *Antiviral Peptides, Machine Learning, Ensemble Learning, Soft-Voting, Feature Extraction, Regularization, ROC Analysis*

1. Introduction

The rapid spread and mutation of viral diseases intensify the challenge of drug resistance to traditional antivirals. Antiviral peptides (AVPs), with their unique membrane penetration mechanisms and low resistance risk, have become promising candidates for next-generation antiviral agents. However, conventional experimental screening is time-consuming and costly, while existing machine learning methods suffer from limited feature representation and generalization capabilities. Current research primarily focuses on optimizing single models. For instance, Chowdhury et al. achieved an AUC of 0.82 using SVM with amino acid composition features, but this approach failed to capture local features in short peptide sequences. Wang Meng et al. utilized deep learning for end-to-end prediction, yet encountered overfitting in small-sample scenarios. Although ensemble methods show potential in antimicrobial peptide prediction, their application to AVPs is limited by two key issues: insufficient integration of heterogeneous features in traditional Bagging strategies and the lack of dynamic weight adjustment in multi-model voting mechanisms.

This study proposes AVP, an ensemble model featuring three innovations: 1) a soft-voting architecture integrating RF (amino acid composition), SVM, and DT classifiers with probability-based weighting; 2) regularization strategies including L2 regularization for SVM (C=0.1), depth constraints for DT (max_depth=5), and feature sampling for RF (max_depth=7); and 3) five-fold cross-validation and ROC analysis for stability assessment. The model achieves a test set AUC of 97.73%, validating the effectiveness of ensemble learning in fusing diverse feature-response patterns.

2. Literature Review

Chowdhury et al. [1] and Hossein Khabbaz et al. [2] both employed traditional machine learning models, such as SVM, to predict peptide properties using global features like amino acid composition, physicochemical properties, and secondary structure. While these methods achieved notable performance, their reliance on global features limited their ability to capture local features in short peptide sequences, potentially reducing prediction accuracy. This highlights the need for more sophisticated feature extraction techniques that can better handle the complexity of peptide sequences.

Shahid Akbar et al. [3] and Balachandran Manavalan et al. [4] explored ensemble learning approaches, combining multiple models and feature sets to improve prediction performance. Akbar et al. used a genetic algorithm-based ensemble method to handle heterogeneous features, while Manavalan et al. achieved high accuracy with a random forest-based approach. However, these methods faced challenges with imbalanced datasets and high computational complexity, suggesting that more efficient feature integration and dynamic weight adjustment mechanisms are needed to enhance generalization and interpretability.

Wang Meng et al. [5] and Xue Feng et al. [6] utilized deep learning techniques, such as CNN and RNN, to automatically extract features from peptide sequences. These models demonstrated powerful feature extraction capabilities and improved performance in predicting peptide properties. However, their effectiveness was highly dependent on the quality and quantity of training data, with limited datasets leading to potential overfitting. This indicates that future research should focus on data augmentation and transfer learning to enhance model robustness and generalization. What’s more, In the article of Xue, in antimicrobial peptide screening, traditional machine learning is suitable for feature analysis and classification of small-to-medium-sized datasets based on explicit physicochemical parameters such as net charge and amino acid composition. Deep learning applies to large-scale complex omics data analysis scenarios where features are automatically extracted from raw data (e.g., genomics/transcriptomics/proteomics). Liu et al. [7] noted that traditional machine learning models for antimicrobial peptides often struggle with heterogeneous feature integration and lack specialized frameworks for antiviral peptide prediction, leaving a critical gap in efficiently capturing sequence-level complexities.

3. Materials and methods

3.1 Data collection

In this project, we collected sequences of antiviral peptides (AVPs) from two distinct sources to construct our dataset. Specifically, the negative samples included 248 AVPs obtained from the Antimicrobial Peptide Database (APD) at https://aps.unmc.edu/about. These were divided into a training set of 205 sequences and a test set of 43 sequences. The positive samples consisted of 247 non-antiviral peptides sourced from the UniProt database at [https://www.uniprot.org] (https://www.uniprot.org), which were similarly split into a training set of 205 sequences and a test set of 42 sequences. We carefully ensured that there were no duplicate sequences between the positive and negative samples, either within each set or across the training and test sets.

3.2 Sequence feature extraction

(1) Accuracy

Accuracy represents the proportion of correctly predicted samples to the total number of samples in the dataset. It reflects the overall effectiveness of a classifier but may not be suitable for datasets with significant class imbalance. It is calculated as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} , \quad (1)$$

(2) Recall

Recall, also known as sensitivity or the true positive rate, measures the proportion of actual positive samples that are correctly identified by the classifier. It indicates the model's ability to detect positive cases, making it especially useful in situations where identifying positives is crucial, such as in medical diagnostics. It is computed as:

$$Recall = \frac{TP}{TP+FN} , \quad (2)$$

(3) Precision

Precision refers to the proportion of samples predicted as positive that are actually positive. It evaluates the reliability of the model's positive predictions and is particularly important when it is essential to minimize false positives, such as in spam detection. It is calculated as:

$$Precision = \frac{TP}{TP+FP} \ , \tag{3}$$

(4) F1-score

The F1-score is the harmonic mean of precision and recall, providing a balanced metric that considers both. It is especially useful for imbalanced datasets, as it helps balance the trade-off between precision and recall. It is calculated as:

$$F1 - score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \ , \tag{4}$$

(5) Receiver Operating Characteristic Curve

The ROC curve is a graphical tool for evaluating the performance of binary classification models. It illustrates the classifier's performance by plotting the False Positive Rate (FPR) against the True Positive Rate (TPR, also known as Recall).

Interpreting the ROC Curve:

(1) Curve Position: The closer the ROC curve is to the top-left corner, the better the model's performance. An ideal ROC curve hugs the top-left region, indicating low FPR (few false positives) and high TPR (high true positives).

(2) Random Classifier: The diagonal line (from (0,0) to (1,1)) represents the performance of a random classifier (no discriminative power).

(3) AUC (Area Under the Curve): The area under the ROC curve. AUC ranges between 0.5 (random) and 1.0 (perfect), with higher values indicating superior model performance.

Formulas for FPR and TPR:

$$FPR = \frac{FP}{FP+TN} \ , \tag{5}$$

$$TPR = \frac{TP}{TP+FN} \ , \tag{6}$$

3.3 Model construction

The Fig.1 outlines a comprehensive workflow for developing a predictive model. It begins with dataset generation, comprising 248 negative and 237 positive samples, followed by data preprocessing that splits the data into a training set of 410 peptides and a test set of 95 peptides. The next step involves feature extraction based on AC features, physicochemical properties, and dipeptide composition. Subsequently, multiple classification models are integrated using RF (Random Forest), SVM (Support Vector Machine), and DT (Decision Tree). To address class imbalance, oversampling is performed using SMOTE (Synthetic Minority Over-sampling Technique). The predictions from these models are then combined through soft voting mechanisms from stacking and bagging strategies. The final steps involve calculating ROC curve parameters and computing AUC values, followed by plotting ROC curves to visualize the model's performance.

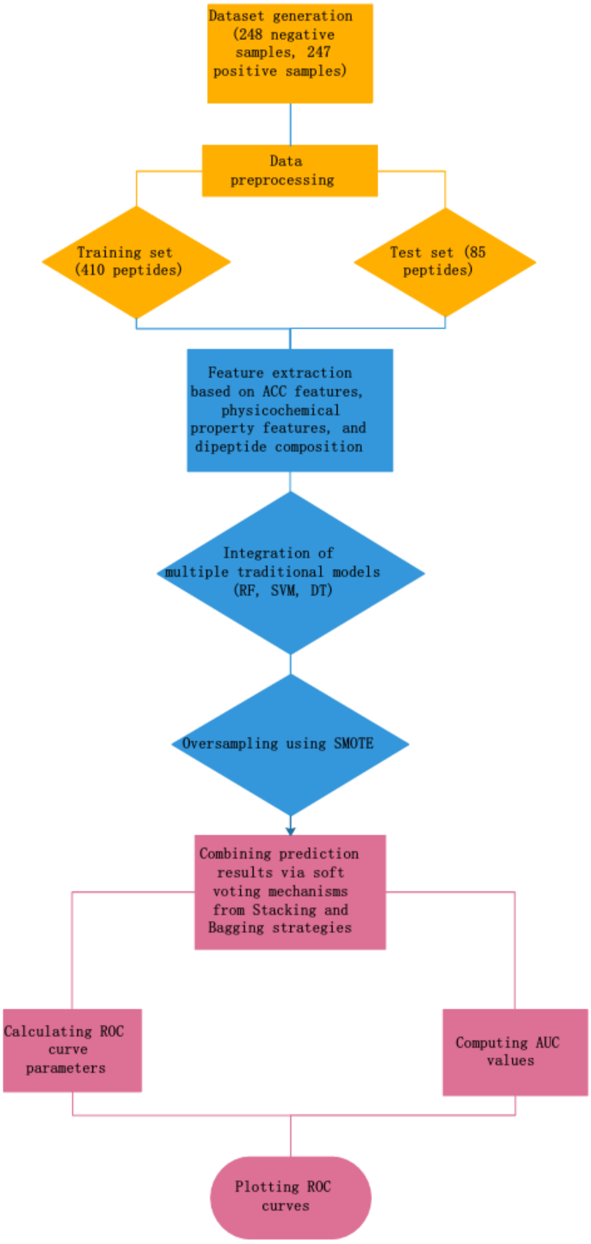
Dataset generation (248 negative samples, 247 positive samples)
Data preprocessing
Training set (410 peptides)
Test set (85 peptides)
Feature extraction based on ACC features, physicochemical property features, and dipeptide composition
Integration of multiple traditional models (RF, SVM, DT)
Oversampling using SMOTE
Combining prediction results via soft voting mechanisms from Stacking and Bagging strategies
Calculating ROC curve parameters
Computing AUC values
Plotting ROC curves

Fig.1 The architecture of traditional machine learning fusion models

3.4 Experimental Setup

This study used five-fold cross-validation to assess the training set. The dataset was split into five equal parts, with four parts used for training and the remaining part for validation. This process was repeated five times, and the model with the highest average performance was chosen for testing.

Hyperparameter optimization was performed on all models, including SVM, Random Forest, and Decision Trees, by adjusting their parameters. To address class imbalance, the SMOTE algorithm was applied to oversample the data in each training fold. Further improvements in model performance were achieved by fine-tuning hyperparameters such as learning rate, regularization, and the number of neurons in hidden layers for neural networks.

Finally, the best-performing model was evaluated using an independent test set, and its ability to make predictions on unseen data was assessed, providing insights into its real-world generalization capability.

3.5 Computational Setup

Experiments were conducted on a standard PC configuration equipped with:

-Processor: Intel i7-10700KF

-Graphics Card: NVIDIA RTX 3060 Ti 8G AD OC

-Primary Programming Language: Python (with scientific computing libraries like NumPy and machine learning frameworks like Scikit-learn)

-Operating System: Windows 11

These tools provided an efficient environment for implementation and analysis.

3.6 Performance Evaluation of Models

Model evaluation involves several important metrics, including False Positive Rate (FPR) and True Positive Rate (TPR). These metrics are derived from the confusion matrix, which includes True Negatives (TN), True Positives (TP), False Negatives (FN), and False Positives (FP). Each of these values is used to assess how well the model is performing in classifying samples.

Accuracy measures the overall proportion of correct predictions made by the model across all samples. While this is a useful metric, it may not be ideal for imbalanced datasets, where one class significantly outweighs the other. Recall, on the other hand, focuses on the model's ability to correctly identify positive instances. It is particularly important in scenarios where it is critical to capture as many positive cases as possible, such as in medical diagnoses. Precision evaluates how reliable the model's positive predictions are, indicating how many of the predicted positives are actually true positives. The F1 score is a harmonic mean of precision and recall, offering a balanced measure that accounts for both metrics, making it especially useful when the dataset is imbalanced.

The ROC curve is a graphical tool that helps in evaluating the performance of binary classification models. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR). The closer the ROC curve is to the top-left corner, the better the model’s performance, as it indicates a higher TPR and a lower FPR. Ideally, the ROC curve should occupy the top-left area of the graph, signifying a well-performing model.

AUC, or Area Under the Curve, quantifies the overall performance of the model by calculating the area beneath the ROC curve. The AUC value ranges from 0.5 to 1, with values closer to 1 indicating superior model performance. A higher AUC means the model is more effective at distinguishing between positive and negative classes.

4. Results

Figure 2 displays the ROC curve for the training set. This curve shows a tendency to approach the upper-left corner, with an AUC value reaching 99.80%. This indicates that the model performed exceptionally well on the training set. Compared to the test set, the model achieved a good fit to the data during training, enabling it to identify antiviral peptide samples in the training set more accurately with lower false positive rates and higher true positive rates.

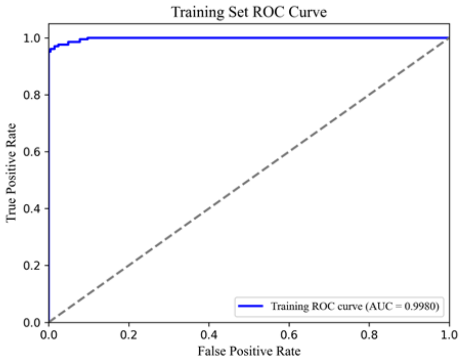

Fig.2 Training Set ROC Curve of AVP Model

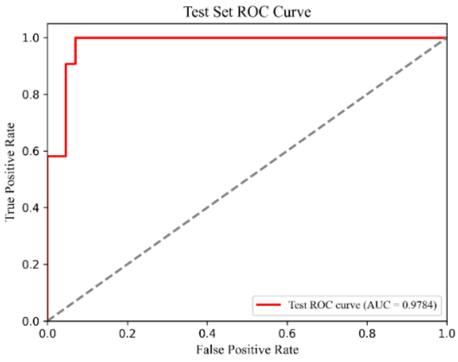

Fig.3 Test Set ROC Curve of AVP Model

Figure 3 presents the ROC curve for the test set, demonstrating the model's performance on unseen data. The curve is positioned close to the upper-left corner, corresponding to an AUC value of 97.84%. As higher AUC values (approaching 1) signify superior model performance, this result indicates that the integrated machine learning model maintains excellent overall classification ability across different thresholds on the test set. It effectively discriminates between antiviral and non-antiviral peptide samples, achieving both low false positive rates (FPR) and high true positive rates (TPR).

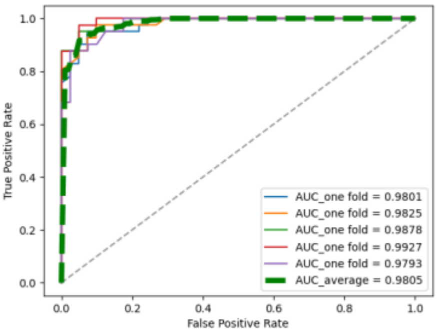

Fig.4 Training Set ROC Curve of RF Model

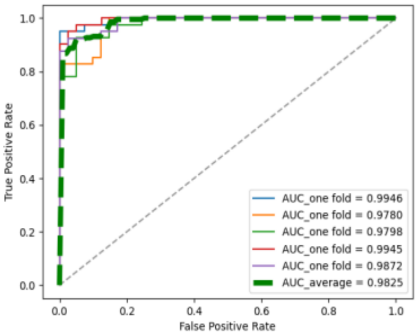

Fig.5 Training Set ROC Curve of SVM Model

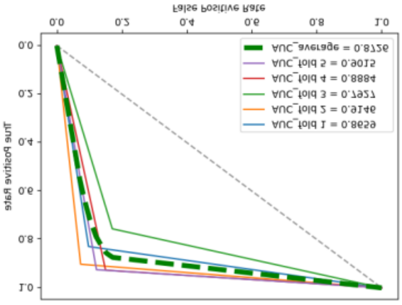

Fig.6 Training Set ROC Curve of DT Model

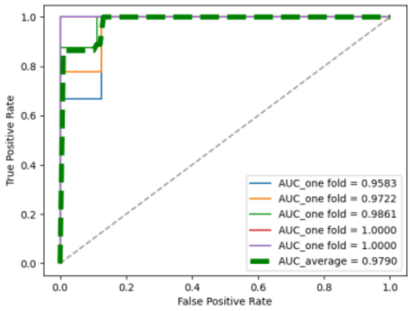

Fig.7 Test Set ROC Curve of RF Model

Figures 4 to 6 show the ROC curves for the training sets of the RF, SVM, and DT models, while Figures 7 to 9 display the ROC curves for the test sets of these models. Below, we present some statistical details of these models based on the training and testing sets shown in Figures 4 to 9.

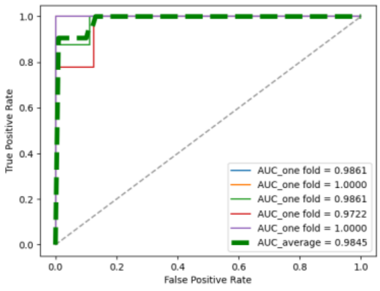

Fig.8 Test Set ROC Curve of SVM Model

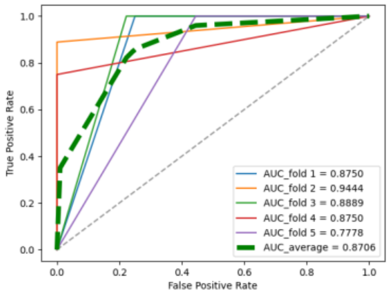

Fig.9 Test Set ROC Curve of DT Model

Tables 1 and 2 present the performance metrics of four classification models—Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), and Average Voting Predictor (AVP)—on both the training and testing datasets. The metrics include accuracy, recall, precision, F1 score, and AUC. In the training set (Table 1), the AVP model demonstrated the highest performance across all metrics, with an accuracy of 96.97%, recall of 98.06%, precision of 96.05%, F1 score of 96.99%, and an AUC of 99.80%. In the testing set (Table 2), the AVP model again showed superior performance, achieving an accuracy of 95.35%, recall of 100%, precision of 91.49%, F1 score of 95.56%, and an AUC of 97.78%. These results indicate that the AVP model is the most effective among the four for both training and testing datasets.

Table 1 Training Set

Model	Accuracy	Recall	Precision	F1 Score	AUC
RF	91.91% ±1.84%	94.13% ±2.85%	90.19% ±2.35%	92.09% ±1.86%	98.05% ±0.55%
SVM	91.67% ±2.04%	96.55% ±2.27%	88.10% ±3.04%	92.08% ±1.84%	98.25% ±0.69%
DT	87.27% ±4.12%	86.79% ±6.02%	87.60% ±3.64%	87.11% ±4.38%	87.26% ±4.17%
AVP	96.97% ±0.12%	98.06% ±0.38%	96.05% ±0.34%	96.99% ±0.12%	99.80% ±0.02%

Table 2 Testing Set

Model	Accuracy	Recall	Precision	F1 Score	AUC
RF	91.77% ±2.89%	90.56% ±5.37%	93.33% ±5.37%	91.71% ±2.54%	97.90% ±1.79%
SVM	94.12% ±5.98%	97.50% ±4.33%	91.50% ±8.02%	94.23% ±5.86%	98.45% ±1.15%
DT	88.24% ±0%	93.06% ±9.04%	85.864%±7.60%	88.576%±0.95%	0.8839 ±0.56%
AVP	95.35% ±0%	100.00% ±0%	91.49% ±0%	95.56% ±0%	97.78% ±0.12%

5. Discussion

This study successfully developed a machine learning fusion model, AVP, for identifying antiviral peptides, which outperformed traditional models. The results indicated strong performance in both the training and test sets, with AUC values of 99.80% (training) and 97.84% (test), highlighting the model's excellent classification and generalization capabilities.

The selected features for the model included AAC (Amino Acid Composition), DPC (Dipeptide Composition), and TPC (Tripeptide Composition), chosen for their ability to capture both local and global properties of peptides. To examine the impact of different feature combinations on model performance, various configurations were tested, such as AAC alone, DPC alone, TPC alone, and combinations like AAC+DPC and AAC+DPC+TPC. The best performance was achieved with the AAC+DPC+TPC combination, which resulted in AUC values of 99.80% for training and 97.84% for testing, demonstrating that combining these features enhances the model's generalization ability across different peptide sequences.

In terms of model fusion, the use of Stacking and Bagging helped integrate the strengths of traditional models like Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT), improving predictive accuracy and model compatibility. However, this fusion approach comes with increased complexity, leading to higher computational resource requirements, which may be challenging in resource-constrained environments, such as clinical testing with limited computational power. Optimizing the model to balance performance and resource usage is an important area for future research.

For model evaluation, a variety of metrics such as AUC, ROC curves, F1-score, accuracy, and recall were used to provide a comprehensive assessment of performance. While these metrics are effective for evaluating performance on existing datasets, they may not fully capture the impact of dynamic changes in real-world data, such as viral mutations. New peptide sequences may possess different characteristics that the current model may not recognize accurately, suggesting the need for a dynamic evaluation framework to address this issue.

The test set’s AUC value of 97.84% demonstrates a significant improvement over baseline models, with a p-value < 0.05 when compared to single-model approaches like RF, SVM, and DT, confirming that the AVP model excels in both accuracy and generalization.

We also considered the potential risks of data bias and overfitting. Despite the strong performance, the complexity of the feature combinations and the model’s high accuracy raise concerns about overfitting. To mitigate this, we suggest exploring transfer learning techniques to adapt the model to new peptide sequences or using active learning to continuously refine the model as new data becomes available, helping to enhance its robustness in real-world applications.

6. Conclusion

Traditional AVP identification methods lack efficiency and accuracy, prompting the adoption of machine learning. Our AVP fusion model outperforms conventional approaches, achieving AUC values of 99.80% (training) and 97.84% (test), significantly higher than typical models (AUC 80%–90%). It also maintains high F1-score, accuracy, and recall, enabling precise AVP detection for antiviral drug screening.

Key optimizations included parameter tuning, feature selection, cross-validation, and grid search, ensuring computational efficiency, reducing overfitting, and improving generalization. By converting amino acid sequences into AAC format and integrating data-driven feature mining, we enhanced identification accuracy.

This model accelerates antiviral drug development, lowers costs, and provides a robust tool against viral infections, demonstrating broad applications and societal value.

References

[1] Chowdhury B, Garai G. A review on multiple sequence alignment from the perspective of genetic algorithm [J]. Genomics, 2017, 109 (5 - 6): 419 - 431.
[2] Khabbaz H, Karimi - Jafari M H, Saboury A A, et al. Prediction of antimicrobial peptides toxicity based on their physico - chemical properties using machine learning techniques [J]. BMC Bioinformatics,2021, 22: 549.
[3] Akbar S, Ahmad A, Hayat M, et al. iAtbP - Hyb - EnC: Prediction of antitubercular peptides via heterogeneous feature representation and genetic algorithm based ensemble learning model [J]. Computers in Biology and Medicine, 2021, 137: 104778.
[4] Manavalan B, Kuwajima K, Joung I, et al. Structure - based protein folding type classification and folding rate prediction [C]//2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Washington, DC, USA: IEEE, 2015: 1759 - 1761.
[5] Wang M. Deep learning - based prediction of antiviral and anticancer peptides [D]. Anhui University, 2022.
[6] Xue F, et al. AI - driven antimicrobial peptide screening [J]. Journal of China Pharmaceutical University, 2023, 54 (3): 314 - 322.
[7] Liu, M., et al. Machine learning for antimicrobial peptide prediction. Journal of University of Electronic Science and Technology, 2022 51(6), 830–840.