HiLL PUBLISHING

Advances in Computer and Communication, 2026, 7(3), 187-191
https://www.hillpublisher.com/journals/acc/
ISSN Online: 2767-2875

Ensemble-learning-based Predictor for Anticancer Peptide Identification Using Multi-view Sequence Features

Xiaoqin Fan*, Xin Zhang, Mingxing Xu, Kexu Wu

Suqian University, Suqian 223800, Jiangsu, China.

How to cite this paper: Xiaoqin Fan, Xin Zhang, Mingxing Xu, Kexu Wu. (2026) Ensemble-learning-based Predictor for Anticancer Peptide Identification Using Multi-view Sequence Features. *Advances in Computer and Communication,* 7(3), 187-191.
DOI: 10.26855/acc.2026.09.015

Received: July 27, 2026
Accepted: August 25, 2026
Published: September 21, 2026

***Corresponding author:** Xiaoqin Fan, Suqian University, Suqian 223800, Jiangsu, China.

© 2026 by the author(s).
This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial-NoDerivatives (CC BY-NC-ND) license, which permits non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited and is not modified or adapted.
https://creativecommons.org/licenses/by-nc-nd/4.0/

Abstract

Anticancer peptides (ACPs) hold considerable promise as therapeutic agents, but screening candidate peptides in the laboratory is both expensive and slow, so reliable computational tools are in demand. A number of machine-learning and deep-learning predictors already exist, yet their accuracy on small benchmark datasets is still modest, and the benefit of merging different sequence descriptors has not been fully exploited. This paper describes an ensemble-learning workflow for identifying ACPs that draws on multiple complementary views of each sequence. Each peptide is first represented by amino acid composition (AAC) and composition of k-spaced amino acid pairs (CKSAAP), which jointly capture overall composition and local order; two physicochemical attributes (average charge and average hydrophobicity) are appended to these features. The resulting 822 dimensions are Z-score standardized and then pruned to 262 informative features by an XGBoost-based selector. Prediction is made by a soft-voting ensemble that pools a Random Forest and a tuned XGBoost classifier. Evaluated on the ACP-2.0 benchmark, this ensemble reaches 77.5% accuracy, a Matthews correlation coefficient of 0.551, and an area under the ROC curve (AUC) of 0.850, beating both base learners and the AntiCP 2.0 baseline while staying competitive with deep-learning methods that are far more demanding computationally. The results point to a practical, accurate tool for large-scale ACP screening.

Keywords

Anticancer peptides; ensemble learning; multi-view feature; XGBoost; random forest; soft voting

1. Introduction

Cancer ranks among the leading causes of death worldwide, and standard chemotherapy and radiotherapy often trigger drug resistance and severe off-target toxicity [1]. Anticancer peptides (ACPs) have drawn increasing interest as a treatment alternative owing to their high selectivity, low toxicity, and ability to cross membranes quickly [2]. Finding new ACPs through wet-lab experiments is nonetheless costly, slow, and labor-intensive [3]. As a result, computational approaches that can sift candidate peptides from large sequence databases now play a central role in drug-discovery workflows [4]. Building predictors that screen ACPs both efficiently and accurately is therefore a pressing practical need.

Over the last ten years, numerous machine learning-based models have been developed to predict anticancer peptides (ACPs). Early research adopted basic sequence composition features, including amino acid composition (AAC) and pseudo amino acid composition (PseAAC), and combined these descriptors with support vector machines for ACP prediction [5]. With the advancement of deep learning techniques, subsequent studies have

leveraged advanced neural network architectures, such as convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, which have achieved superior prediction performance across multiple benchmark datasets [6, 7]. Physicochemical properties and functional aspects of ACPs have also been reviewed in detail [8]. Despite these promising advances, two key limitations still persist in existing methods. First, most existing models rely on a single feature representation of peptide sequences, failing to integrate complementary information from diverse feature descriptors and thus limiting their feature utilization efficiency. Second, high-performance deep learning models usually require massive training data and substantial computational resources for model training and inference, which greatly restricts their practical application in routine ACP screening tasks with limited dataset scales. Interpretability of such models has also been recognized as a practical concern [9]. To address the aforementioned research limitations, this study aims to construct a lightweight and high-accuracy ACP prediction model capable of fusing multi-view sequence feature information.

Given the aforementioned research shortcomings, this study develops a novel ensemble learning framework for accurate ACP detection, which integrates multi-perspective sequence encoding strategies with a two-stage model construction pipeline. Specifically, the prediction workflow consists of three core steps: (i) extracting peptide sequence features from four dimensions, including AAC, CKSAAP, and two types of physicochemical properties for comprehensive sequence encoding; (ii) standardizing feature values via the Z-score method and eliminating redundant and noisy features through XGBoost-based feature importance screening to realize effective dimensionality reduction; (iii) inputting the optimized feature subsets into a soft voting ensemble model, which synthesizes the probability prediction results of optimized Random Forest and XGBoost classifiers to generate final identification outcomes. The major innovations and contributions of this work are summarized as follows:

• A multi-dimensional feature encoding strategy that unifies global component characteristics, local short-range sequence arrangement rules, and peptide physicochemical attributes to form a complete and complementary feature system.

• A refined XGBoost-driven feature selection module, which achieves substantial feature dimension compression while retaining nearly all valid feature information, balancing model efficiency and feature integrity.

• A high-efficiency soft voting ensemble model that obtains an AUC value of 0.850 on the ACP-2.0 benchmark dataset. Its predictive performance surpasses both individual base models and existing mainstream baseline algorithms.

2. Materials and Methods

2.1 Benchmark Dataset

We evaluated the predictor on the ACP-2.0 dataset, a widely used benchmark. It consists of 689 experimentally validated anticancer peptides (the positive class) and 689 non-anticancer peptides (the negative class). Stratified sampling split the data into a training set (80%, 1102 sequences) and an independent test set (20%, 276 sequences), keeping the class proportions identical in both parts.

2.2 Multi-View Sequence Feature Encoding

For every peptide, we computed three complementary groups of descriptors [10]:

(1) Amino acid composition (AAC): the occurrence frequency of the 20 standard amino acids, giving a 20-dimensional vector that summarizes overall residue usage.

(2) Composition of k-spaced amino acid pairs (CKSAAP): pair composition with gaps k = 0 and k = 1 encodes short-range order. Since each k contributes 400 dimensions, this yields an 800-dimensional order-sensitive representation.

(3) Physicochemical properties: average charge and average hydrophobicity were appended to enrich the descriptor set (2 further dimensions).

Taken together, the three views give an initial feature vector of 822 dimensions.

2.3 Data Standardization

The feature groups span very different numerical ranges, so we applied Z-score standardization (StandardScaler) to every feature to keep any one descriptor from dominating the training objective simply because of its scale.

2.4 Feature Selection

The 822-dimensional space is likely to contain redundant and noisy entries. We first trained an XGBoost-based

selector, which assigns an importance score to each feature. Any feature whose importance dropped below 1.2 times the mean was removed; the factor 1.2 was chosen empirically to trade off dimensionality reduction against information retention. This step left 262 informative features for model construction.

2.5 Ensemble Model Construction

We built a soft-voting ensemble from two complementary base classifiers:

Random Forest (RF): a bagging ensemble of decision trees that lowers prediction variance and resists overfitting; we used 500 estimators.

Extreme Gradient Boosting (XGBoost): a boosting algorithm that repeatedly corrects misclassified samples to reduce bias. A grid search with 5-fold cross-validation set its hyperparameters to n_estimators = 500, max_depth = 3, learning_rate = 0.05, and subsample = 0.8.

The two models were combined by averaging their predicted class probabilities, P_final = (P_RF + P_XGB) / 2; a peptide is labeled an ACP when P_final > 0.5. Figure 1 summarizes the full workflow.

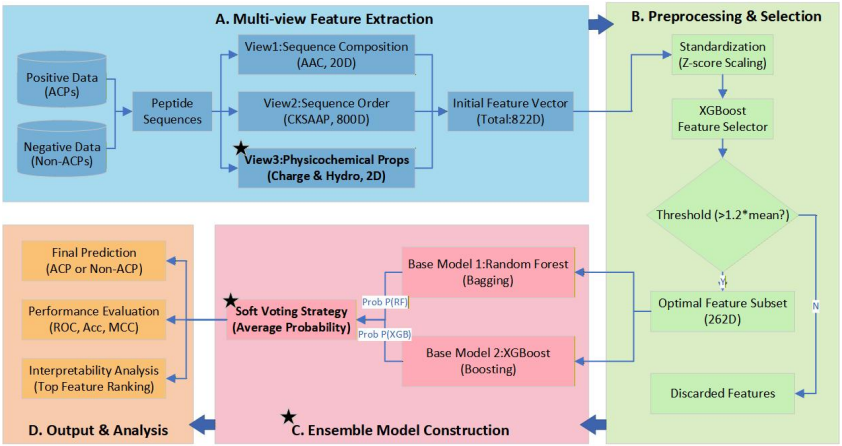

Figure 1. Overall workflow of the proposed ensemble-learning predictor.

2.6 Performance Evaluation

We assessed performance with five standard metrics: sensitivity (Sn), specificity (Sp), accuracy (Acc), the Matthews correlation coefficient (MCC), and the area under the ROC curve (AUC). The first four are defined below:

Sensitivity (Sn) / Recall:

$$Sn = \frac{TP}{TP+FN} \tag{1}$$

Specificity (Sp):

$$Sp = \frac{TN}{TN+FP} \tag{2}$$

Accuracy (Acc):

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

Matthews Correlation Coefficient (MCC):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{4}$$

Here TP, TN, FP, and FN stand for true positives, true negatives, false positives, and false negatives, respectively. AUC was taken as the area under the ROC curve of sensitivity against (1−specificity).

3. Results

3.1 Performance of Base Learners and the Ensemble

We evaluated the two base classifiers and the soft-voting ensemble on the independent test set. Random Forest

achieved an accuracy of 0.750 and an MCC of 0.502, while the tuned XGBoost reached 0.765 and 0.529. The ensemble beat both, scoring the best values on every metric: accuracy 0.775, specificity 0.797, MCC 0.551, and AUC 0.850. The steady gain in MCC suggests the ensemble trades off positive and negative predictions more evenly than either model alone.

3.2 ROC Analysis

The ensemble's ROC curve encloses the curves of both base learners, which confirms that pairing Random Forest's variance reduction with XGBoost's bias reduction improves generalization. Against XGBoost alone (AUC = 0.837), the ensemble delivers a clear gain in AUC (0.850).

3.3 Comparison with State-of-the-Art Predictors

We also compared our framework with four recent ACP predictors on the same ACP-2.0 test set: AntiCP 2.0 (the baseline), iACP-SEI, ENNAACT, and ACP-DPE. Our model reaches 77.54% accuracy, above AntiCP 2.0 (75.58%) and close to iACP-SEI (77.78%), without needing computationally expensive pre-trained protein language models. Deep models such as ACP-DPE do score higher (82.81%), but they behave as opaque black boxes; our approach holds competitive accuracy at a small fraction of the computational cost, a real advantage when screening at scale.

3.4 Confusion Matrix Analysis

The ensemble correctly labeled 104 ACPs (true positives) and 110 non-ACPs (true negatives). Next to the standalone Random Forest, it shows a more even balance between sensitivity and specificity and a lower false-positive rate, matching its higher MCC.

4. Discussion

4.1 Why the Ensemble Generalizes Better

The ensemble's edge (AUC = 0.850) comes from the complementary error patterns of its two base learners. As a bagging method, Random Forest averages many decorrelated trees, which dampens variance and sensitivity to noise. As a boosting method, XGBoost focuses on the instances that are hardest to classify, which lowers bias. Averaging their probabilities merges both mechanisms and produces predictions that are stable as well as well-calibrated—and that is exactly why the ensemble outperforms either classifier on its own.

4.2 Efficiency and Practical Value

Accuracy aside, the pipeline is deliberately light. Feature selection trims the input from 822 to 262 dimensions, and neither base learner needs pre-trained embeddings or large parameter sets. The predictor can therefore be trained and rolled out quickly—well matched to high-throughput screening, where many candidate sequences have to be ranked fast.

4.3 Toward Interpretability

This study foregrounds predictive performance, but the trained ensemble is also naturally interpretable: Random Forest and XGBoost both expose feature-importance scores that show which sequence and physicochemical attributes drive each classification. A full biological reading of these determinants—linking the learned feature rankings to electrostatic and hydrophobic membrane interactions—appears in a companion study, so we note it only in passing here.

5. Conclusion

We have presented an ensemble-learning predictor for anticancer peptide identification built on multi-view sequence features. The framework encodes peptides with AAC and CKSAAP plus two physicochemical descriptors, applies Z-score normalization and XGBoost-based feature selection, and then fuses Random Forest with XGBoost through soft voting. On the ACP-2.0 benchmark, it reaches 77.5% accuracy and an AUC of 0.850, ahead of both base learners and the established baseline, and stays highly competitive with recent deep-learning methods at a much lower computational cost. The source code and datasets are publicly available at https://github.com/xiaoqin-fan/Anticancer-Peptide-Prediction, and we expect the tool to support high-throughput screening of potential anticancer drugs.

Data Availability

The ACP-2.0 benchmark dataset is publicly available. The source code and implementation of the proposed predictor are provided at https://github.com/xiaoqinfan/Anticancer-Peptide-Prediction.

References

[1] Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229-63.

[2] Ghadiri N, Zali H, Vafaei M, et al. Bioactive peptides: An alternative therapeutic approach for cancer management. Front Immunol. 2024;15:1310443.

[3] Xu M, Li W, Wang X, et al. Integrating traditional machine learning and deep learning for precision screening of anticancer peptides: A novel approach for efficient drug discovery. ACS Omega. 2024;9(14):16820-31.

[4] Kim H, Kim J, Lee I. Artificial intelligence in drug discovery: A comprehensive review of data-driven and machine learning approaches. Biotechnol Bioprocess Eng. 2020;25(6):895-930.

[5] Schaduangrat N, Lohavanichbutr P, Saeaeung A, et al. Acpred: A computational tool for the prediction and analysis of anticancer peptides. Molecules. 2019;24(10):1973.

[6] Ahmed S, Mugal I, Ijaz MT, et al. Acp-Mhcnn: An accurate multi-headed deep-convolutional neural network to predict anticancer peptides. Sci Rep. 2021;11:23676.

[7] Yi HC, You ZH, Huang DS, et al. Acp-Dl: A deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation. Mol Ther Nucleic Acids. 2019;17:1-9.

[8] Chiangjong W, Chutipongtanate S, Hongeng S. Anticancer peptide: Physicochemical property, functional aspect and trend in clinical application. Int J Oncol. 2020;57(3):678-96.

[9] Van Hilten A, Chevyrev V, Maticzka D, et al. Designing interpretable deep learning applications for functional genomics: A quantitative analysis. Brief Bioinform. 2024;25(5):bbae344.

[10] Yuan Q, Chen J, Zhao H, et al. Prediction of anticancer peptides based on an ensemble model of deep learning and machine learning using ordinal positional encoding. Brief Bioinform. 2023;24(1):bbac630.