2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-3748-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/BIBM58861.2023.10385565

AIPPT: Predicts anti-inflammatory peptides using the most characteristic subset of bases and sequences by stacking ensemble learning strategies

Ruiqi Liu
School of Computer Science and Technology
Hainan University
Haikou, China
l1983127203@163.com

Xiuhao Fu
School of Computer Science and Technology
Hainan University
Haikou, China
fxh1001@hainanu.edu.cn

Shankai Yan
School of Computer Science and Technology
Hainan University
Haikou, China
skyan@hainanu.edu.cn

Zilong Zhang
School of Computer Science and Technology
Hainan University
Haikou, China
zhangzilong@hainanu.edu.cn

Feifei Cui*
School of Computer Science and Technology
Hainan University
Haikou, China
feifeicui@hainanu.edu.cn

***Abstract*—Therapeutic peptides play a vital role in developing peptide-based drugs. Recently, they have been applied as anti-inflammatory agents for a range of inflammatory conditions, including Alzheimer's disease and rheumatoid arthritis. Laboratory-based identification of peptides with anti-inflammatory properties is a highly time-consuming and labor-intensive endeavor. To tackle this issue, researchers have developed computational methods, primarily centered on machine learning, to streamline the procedure. This paper presents AIPPT, an intelligent and computationally efficient prediction tool that introduces a novel stacking framework for the reliable identification of anti-inflammatory peptides (AIP). The study specifically employs a combination of four feature encodings, where their importance is assessed using the LightGBM method to create an optimal feature subset, which is then input to the three classifiers. The output probabilities from the three classifiers are further fed into a meta-classifier, constructing a two-layer stacking model. Subsequently, the output probabilities from the three classifiers are incorporated into a meta-classifier, establishing a two-layer stacking model. Subsequently, the output probabilities from the three classifiers are incorporated into a meta-classifier, establishing a two-layer stacking model.**

***Index Terms*—anti-inflammatory peptides, machine learning, stacking, feature encodings**

I. Introduction

Inflammation constitutes a natural defensive response of the body to harmful stimuli, such as injuries and infections, which pose potential harm to affected tissues [1]. Inflammation serves as the common origin of numerous chronic diseases, such as cardiovascular disease, autoimmune diseases, rheumatoid arthritis and varicose veins. Treatment of inflammatory disorders typically involves the use of biologic agents and non-specific small molecule drugs, potentially resulting in undesirable side effects [2]. Peptide therapy, under normal circumstances, is more specific and less toxic than small molecule drugs. Peptide therapies represent a superior choice for the treatment of inflammatory and autoimmune diseases, and they have been widely embraced in medical practice [3]. Anti-inflammatory peptides (AIPs) are an essential constituent of the immune system in multicellular organisms, including humans [4]. Anti-inflammatory peptides play a pivotal role in various infectious diseases, such as systemic inflammatory response syndrome, bronchial asthma, Hepatitis B virus (HBV) infection, and acute pancreatitis. In conclusion, Anti-inflammatory peptides exert anti-inflammatory and protective effects on the body in numerous infectious diseases, offering an appealing avenue for disease treatment [5, 6].

Researchers in the field have developed several intelligent predictive models in the literature to distinguish AIPs [7, 8]. Numerous researchers strive to utilize machine learning algorithms in the pursuit of identifying AIP [9]. For instance, Gupta et al. introduced AntiInFlam [10], a model constructed by combining TPC and motif features as inputs for SVM [11]. Manavalan et al. proposed a Random Forest (RF)-based method [12, 13] called AIPpred for AIP prediction [14], as DPC-based models demonstrated notably better performance than other combination-based models [15]. Forest-based random feature selection was employed on the DPC to identify the most effective features for predicting AIPs. Khatun et al. also introduced PreAIP using RF and incorporated three distinct types of features [16]. Zhang et al. introduced AIEpred, which employs Pseudo (PSE) and Discrete Wavelet Transform (DWT) to convert the variable-length coding matrix into equidimensional features [17]. The method

979-8-3503-3748-8/23/$31.00 ©2023 IEEE

selects significant features using the t-test. Six distinct feature datasets are employed to construct various base classifiers, which are subsequently integrated. Zhao encoded samples with multiple features, including AAC [18], DDE [15], and GDC [19]. The optimal feature subset is generated via a two-step feature selection process. This optimal subset was employed as the input for developing iAIPs with the RF classifier. In the recently developed AIPstack [20], peptide sequences were represented using hybrid features that amalgamated two amino acid composition descriptors. Subsequently, a stacking ensemble model was constructed, utilizing random forests and extremely randomized trees as base-classifiers, with logistic regression [21] serving as the meta-classifier to aggregate the outputs from the base-classifiers.

In our analysis of the AIP method, we observed that existing predictors, such as AIPpred, exclusively employ a single feature encoding, potentially overlooking valuable information. These methods, including AntiInflam, AIPpred, PreAIP, and IAIPs, all rely on a single machine learning classifier. Nonetheless, ensemble learning offers the advantage of seamlessly integrating multiple individual models into a unified entity, resulting in enhanced prediction accuracy, stability, and robustness. However, if the dimensionality of the features used in the fusion process is excessively high, it can lead to a "dimensional disaster". Merging multiple features without discretion can introduce an excess of information, thereby impacting the prediction's effectiveness. The hybrid encoding utilized in the AIPstack method lacks a feature selection mechanism, potentially leading to unwarranted dimensionality. To address the aforementioned issues, this study developed an intelligent and computationally efficient predictive tool for anti-inflammatory peptides. We conducted an extensive exploration of feature encoders and classifiers, ultimately selecting three machine learning methods and four distinct feature encodings: AAC, ASDC [22, 23], DDE, and CTDD [24, 25, 26]. In the context of various biological problems, it is well-established that multiple features can effectively discriminate between positive and negative instances. Subsequently, by employing feature selection strategies, we identify the most relevant features. Feature selection serves to eliminate redundancy, decrease training time, and enhance model accuracy. Ultimately, we constructed a stacked ensemble model for predictive purposes. Extensive research has consistently demonstrated the superior performance of ensemble learning models compared to single algorithm-based models. Our experimental results unambiguously illustrate the superior performance of the proposed method in comparison to existing approaches.

II. Materials and methods

Figure 1 illustrates the comprehensive framework of AIPPT as introduced in this study. Step A involves establishing a baseline model for assessing the performance of the algorithms and encodings. During Step B, LightGBM [27] is employed to assess the importance of features obtained from the combination of the four encodings, facilitating the selection of the optimal feature subset. Step C involves feeding the output probabilities from the first layer into the second-layer meta-classifier, which is LR, for prediction purposes. Step D entails the evaluation of the model's prediction results.

A. Dataset preparation

Obtaining a benchmark dataset is essential for training our intelligent computational model. This acquisition ensures accurate evaluation of our proposed method's predictive performance and enables equitable comparisons with existing models. We employed the dataset from the recently proposed method "AIPstack". The dataset was initially sourced from the IEDB database [28]. Sequences with a similarity exceeding 80 percent are excluded. The training dataset comprises 1511 AIPs and an equal number of non-AIPs samples. The test dataset consists of 168 AIPs and an equivalent number of non-AIPs instances. The independent dataset encompasses 187 AIPs and an equal number of non-AIPs samples. Additionally, we utilized the test dataset of the Manavalan2018 dataset. The Manavalan2018 dataset enjoys widespread use, and we employ its test dataset to compare performance against some prior methods. Table I displays the sample distribution for each dataset.

TABLE I
THE NUMBER OF SAMPLES OF THE DATASET USED FOR THIS EXPERIMENT

	Positive	Negative
BD1-training	1511	1511
BD1-testing	168	168
BD1-ind	187	187
Manavalan2018-test	420	629

B. Feature encoding scheme

We employed thirteen sequence-based feature codes in this study to represent peptides. These feature encodings encompass AAC, ASDC, APAAC, CTDC, CTDT, CTDD, DDE, DPC, GAAC, GDPC, GTPC, KSCird, and CTirad. The descriptions and dimensions for each of these feature encodings are presented in Table II. We generated these descriptors using both the iFeature and Pfeature toolkits [29].

C. Machine learning method

In machine learning, predictive performance hinges not only on feature encoding but also on classifier selection. In this study, we utilized eight distinct machine learning classifiers: Adaptive Boosting (AdaBoost) [35], Random Forest (RF), Gradient Boosting Decision Tree (GBDT) [36], K-Nearest Neighbor (KNN) [37], Support Vector Machines (SVM), Logistic Regression (LR), Light Gradient Boosting Machine (LightGBM), and eXtreme Gradient Boosting (XGBoost) [38]. We implemented all classifier models using Scikit-learn version 1.1.1. For hyperparameter optimization, we utilized both grid search and Bayesian search techniques. Additionally, we employed 10-fold cross-validation. In each cross-validation fold, diverse hyperparameter combinations were explored for each machine learning classifier, and we selected the best hyperparameters based on the highest accuracy (ACC).

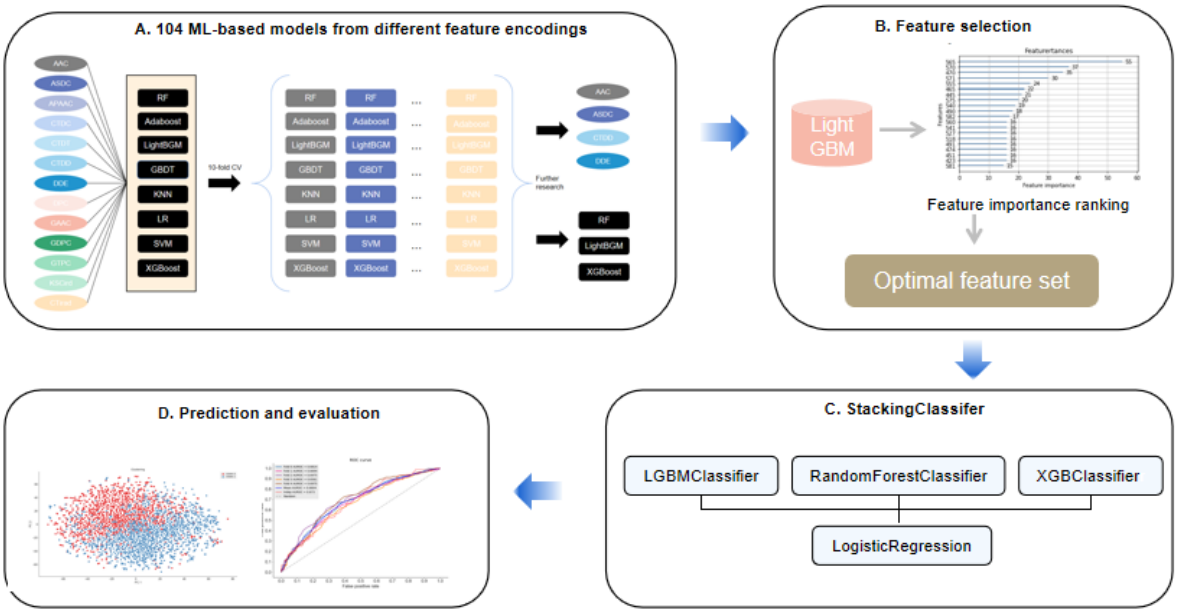

Fig. 1. The overall framework of our proposed approach.

TABLE II
SUMMARY OF 13 SEQUENCE-BASED FEATURE DESCRIPTORS

Feature descriptors	*Description*	*Dimension*	*References*
Amino acid composition (AAC)	Appearance number of 20 different amino acids	20-dimensional	[18]
Adaptive skip dipeptide composition (ASDC)	Adaptive skip dipeptide	composition400-dimensional	[22, 23]
Amphiphilic Pseudo-Amino Acid Composition(series) (APAAC)	Pseudo-amino acid composition	25-dimensional	[30, 31]
Composition (CTDC)	Percentage of particular amino acid propert groups	39-dimensional	[24, 25, 26]
Transition (CTDT)	Percentage of mutual conversion in amino	39-dimensional	[24, 25, 26]
Distribution (CTDD)	Distribution of amino acid properties	195-dimensional	[24, 25, 26]
Dipeptide composition (DPC)	Dipeptide composition	400-dimensional	[15]
Grouped amino acid composition (GAAC)	Physicochemical properties of amino acid	5-dimensional	[18, 32, 33]
Dipeptide deviation from expected mean(DDE)	Dipeptide deviation from expected mean	400-dimensional	[15]
Grouped dipeptide composition (GDPC)	Physicochemical properties of dipeptides	25-dimensional	[32]
Grouped tripeptide composition (GTPC)	Physicochemical properties of tripeptides	125-dimensional	[18, 32]
K-spaced conjoint triad (KSCTriad)	K-spaced conjoint triad	343-dimensional	[32]
Conjoint triad (CTriad)	Conjoint triad	343-dimensional	[34]

D. Performance evaluation metrics

Machine learning utilizes a range of performance metrics to evaluate the predictive accuracy of intelligent models. In this study, we employ specific performance evaluation metrics to accurately assess the proposed predictive factors.

Here, TP stands for True Positive, FP for False Positive, TN for True Negative, and FN for False Negative. We evaluate the classifier's performance using six distinct metrics, each offering a unique perspective. Precision and Sensitivity (SE) predominantly evaluate the classifier's capability to predict positive samples, while Specificity (SP) emphasizes its ability to predict negative samples. The Matthews Correlation Coefficient (MCC) quantifies the correlation between actual class labels and predicted labels. AUC represents the Area Under the ROC Curve, illustrating the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR) of a model [39, 40].

$$ACC = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

$$SE = \frac{TP}{TP + FN} \tag{2}$$

$$SP = \frac{TN}{TN + FP} \tag{3}$$

$$Pre = \frac{TP}{TP + FP} \tag{4}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

III. Results

A. Baseline model

During model construction, we utilized a total of 8 machine learning (ML) algorithms and 13 peptide sequence descriptors. For assessing the discriminative ability of these algorithms and descriptors in AIP classification, we created 104 baseline models on the training dataset, without hyperparameter tuning, and assessed their performance on the testing dataset. Figures 2 and 3 depict the AUC scores acquired from both the training and testing datasets. Tree-based models, such as RF, LightGBM, and XGBoost, exhibit strong performance. Particularly noteworthy is the baseline model, which employed the RF algorithm in conjunction with the CTDD descriptor, achieving the highest performance on the training dataset with an AUC of 0.8264. The second-best performance on the training dataset was accomplished by the baseline model utilizing XGBoost, which also incorporated the CTDD descriptor, resulting in an AUC of 0.821. Furthermore, with the inclusion of the CTDD descriptor, LightGBM and XGBoost exhibited superior performance compared to other descriptors. Moreover, within the ensemble model utilizing LightGBM, XGBoost, and RF as the base classifiers, we identified the most effective encodings, namely ASDC, AAC, and DDE, in addition to CTDD. Consequently, we selected DDE, AAC, CTDD, and ASDC for further investigation.

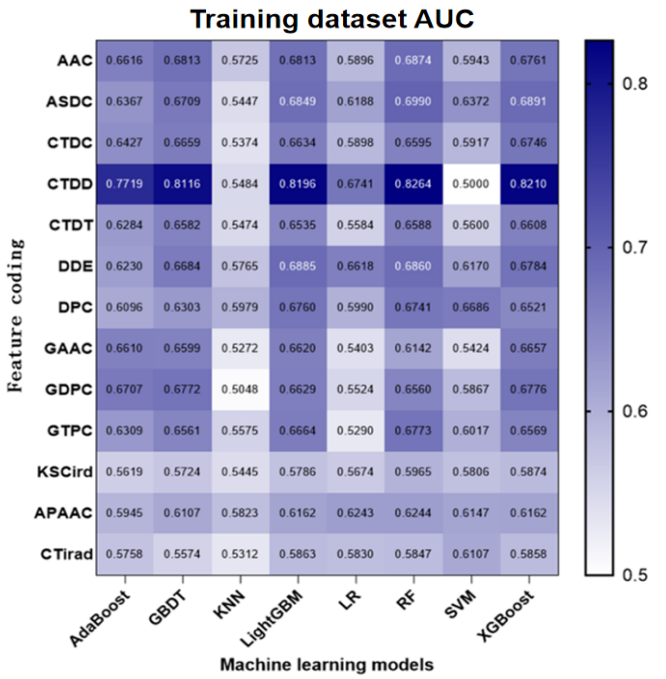

Fig. 2. The performance results of the baseline model on the training dataset. In the figure, the horizontal axis represents eight machine learning models, and the vertical axis represents 13 feature encodings, collectively constituting 104 (8 × 13) baseline models.

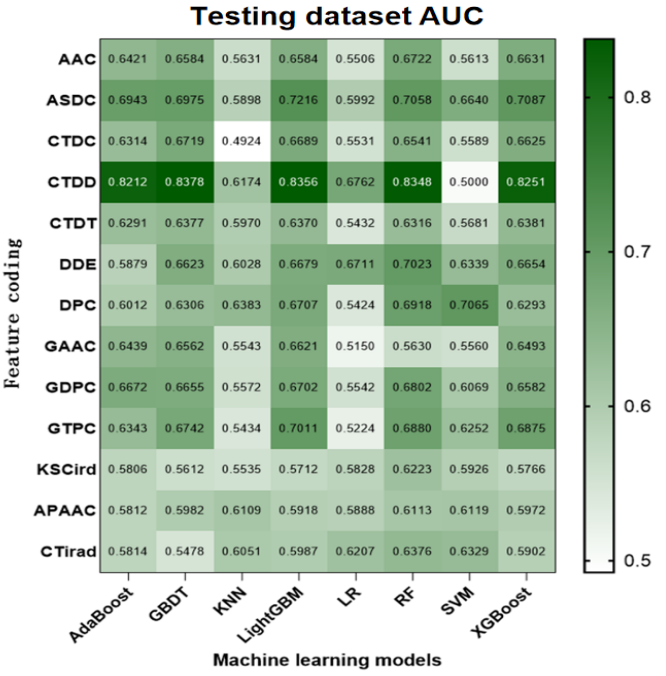

Fig. 3. The performance results of the baseline model on the testing dataset.In the figure, the horizontal axis represents eight machine learning models, and the vertical axis represents 13 feature encodings, collectively constituting 104 (8 × 13) baseline models.

B. Feature selection

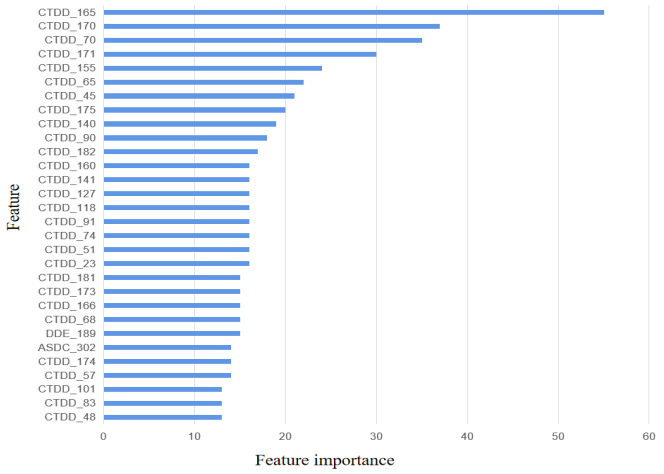

Fig. 4. Top 30 dimensions features for feature importance.

Hybrid Encoding may encompass high-dimensional features, leading to information redundancy that can adversely impact the model's performance and increase computational complexity. Consequently, a two-step feature optimization methodology is utilized to identify the most appropriate subset of features. Initially, the feature sorting process employs the LightGBM method to rank feature importance. Subsequently, we apply the Incremental Feature Selection (IFS) strategy, guided by accuracy, to determine the optimal feature subset. We calculated the feature importance of the 1015-dimensional features using the LightGBM algorithm. Figure 4 presents the importance scores for the top 30 dimensions features. The optimal feature dataset consists of 194 dimensions of features, used for training XGBoost, LightGBM, and RF classifiers. The optimal feature dataset consists of 194 dimensions of features, used for training XGBoost, LightGBM, and RF classifiers. Table III illustrates a performance comparison before and after feature selection.

TABLE III
COMPARISON OF THE PERFORMANCE OF THE THREE MODELS BEFORE AND AFTER FEATURE SELECTION ON THE TEST DATASET

	Methods	*Se*	*Sp*	*Pre*	*ACC*	*AUC*	*MCC*
Before feature selection	RF	0.726	0.773	0.762	0.750	0.818	0.500
	LightGBM	0.732	0.767	0.759	0.750	0.826	0.500
	XGBoost	0.750	0.761	0.759	0.755	0.816	0.511
After feature selection	RF	**0.761**	0.755	0.757	0.758	0.833	0.517
	LightGBM	**0.761**	0.773	0.771	**0.767**	0.830	0.535
	XGBoost	**0.761**	**0.785**	**0.777**	**0.767**	**0.834**	**0.536**

TABLE IV
COMPARISON OF THE OPTIMAL SINGLE CLASSIFIER XGBOOST AND THE ENSEMBLE MODEL AIPPT BEFORE AND AFTER

Methods	*Se*	*Sp*	*Pre*	*Acc*	*MCC*	*AUC*
XGBoost	0.761	0.785	0.777	0.767	0.834	0.536
AIPPT	0.761	0.785	0.780	0.773	0.838	0.547

C. Stack model

Following the acquisition of the optimal feature dataset, we utilized the stacking method to construct the final model[37]. Concretely, we concatenated the output probabilities of the three classifiers (RF, XGBoost, and LightGBM) that were based on the optimal feature dataset, creating a new 3-dimensional feature vector that conveyed multi-view information. Subsequently, this 3-dimensional vector was fed into an LR classifier to establish the ultimate predictor named AIPPT. We compared the ensemble model AIPPT with the best-performing individual model XGBoost, as shown in Table IV. It is evident that the ensemble model demonstrates improvements in Precision, ACC, MCC, and AUC.

D. Comparison of current study with previous predictors

The training dataset is identical to AIPstack. Firstly, let's compare with existing methods on the test dataset. This is depicted in Figure 6, and in line with our findings, our predictive factors surpassed previous methods across all evaluation metrics, achieving SE of 0.761, SP of 0.785, Precision of 0.780, ACC of 0.773, AUC of 0.838, and MCC of 0.547. This is depicted in Figure 6, and in line with our findings, our predictive factors surpassed previous methods across all evaluation metrics, achieving SE of 0.761, SP of 0.785, Precision of 0.780, ACC of 0.773, AUC of 0.838, and MCC of 0.547.

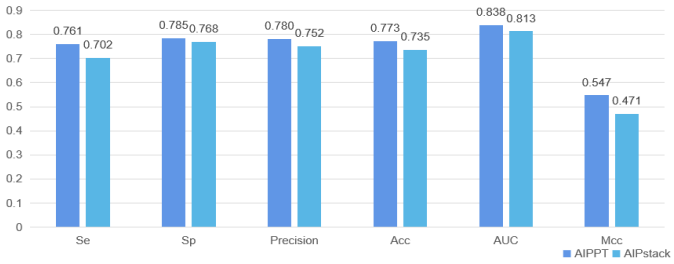

Fig. 5. Comparison of performance on AIPstack and AIPPT on the test dataset.

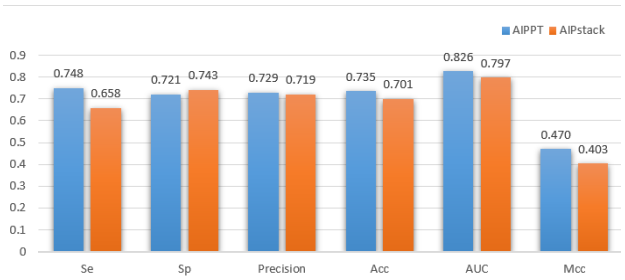

Fig. 6. Comparison of performance on AIPstack and AIPPT on the independent dataset.

Employing an independent dataset for evaluation enables us to validate the model's predictive performance on novel and unseen data while assessing its real-world applicability. To fulfill this objective, we used a test dataset to compare

its generalization ability with that of existing methods. This is illustrated in Figure 7, where our approach demonstrated strong performance, achieving SE of 0.748, SP of 0.721, Precision of 0.729, ACC of 0.725, AUC of 0.826, and MCC of 0.470. Remarkably, our method consistently outperformed existing models in five of the six evaluation metrics, except for SP. These results strongly indicate that our model exhibits robust generalization capability and effectively discriminates anti-inflammatory peptides in practical predictions.

In this study, we also employed the test dataset from Manavalan2018 to assess our model's performance. The figure below provides a comparison of different models using this dataset. The results depicted in the figure indicate that our approach achieved Se=0.800, Sp=0.801, ACC=0.800, MCC=0.593, and AUC=0.787 on this dataset, attaining the highest scores in SE, ACC, and MCC. This demonstrates that the proposed method can perform effectively on various datasets.

TABLE V
COMPARISON OF AIPPT AND EXISTING METHODS ON THE MANAVALAN2018 TEST DATASET

Methods	*Se*	*Sp*	*Acc*	*MCC*	*AUC*
Antilnflam (LA)	0.258	**0.892**	0.638	0.197	0.647
Antilnflam(MA)	0.786	0.417	0.565	0.210	0.706
AIEpred	0.555	0.899	0.762	0.495	0.767
AIPpred	0.741	0.746	0.744	0.479	0.813
iAIPs	0.567	0.874	0.751	0.471	**0.822**
AIPPT	**0.800**	0.801	**0.800**	**0.593**	0.787

E. Limitations of the study

A limitation of this study is the exclusive use of machine learning classifiers, without incorporating deep learning models. Deep learning requires extensive datasets, but the available data for AIP is rather constrained [41]. Another limitation is the sole exploration of a single feature selection method, without conducting comparisons with multiple methods. The third limitation is the lack of a web server development for practical deployment.

IV. DISCUSSION

Extensive research is conducted on anti-inflammatory peptides for the development of innovative bioactive compounds and pharmaceuticals. Nevertheless, the precise identification of potential anti-inflammatory peptides remains a challenging task. We introduce a novel sequence-based approach called AIPPT. Comprehensive experiments have shown that our proposed feature representation model effectively distinguishes between AIP and non-AIP within the feature space.

We investigated eight distinct ML algorithms and thirteen feature encoding methods. Among these, the combination of AAC, ASDC, DDE, and CTDD encoding techniques exhibited excellent performance. To handle the increased feature dimensions arising from the hybrid features, we utilized the LightGBM algorithm for feature selection to identify the most significant ones. In contrast to the prevailing reliance on RF in the construction of models in existing methods, we adopted a stacking ensemble approach. This approach utilizes LightGBM, RF, and XGBoost as base classifiers, with LR serving as the meta-classifier. Based on the experimental results, the ensemble approach outperformed individual machine learning models.

The final model demonstrated noteworthy enhancements, with a 3.8-point increase in ACC and a 2.5-point increase in AUC on the test dataset, surpassing existing methods across all six evaluation metrics. Similar improvements were noted during the independent dataset evaluation. Additionally, AIP surpassed existing methods on the Manavalan2018 dataset. Additionally, AIP surpassed existing methods on the Manavalan2018 dataset. The exceptional performance on unfamiliar datasets demonstrated its robustness and reliability.

The theoretical basis for the enhanced AIP performance can be summarized as follows:

- Exploring a wider range of feature encoding methods, including those not previously employed.
- Employing a broader range of machine learning methods for comparison, facilitating the identification of more suitable approaches.
- Implementing feature selection techniques to choose optimal feature subsets and enhance model performance.
- Employing the stacking ensemble method for further performance enhancement.

The results confirm the enhanced capability of our novel model in identifying anti-inflammatory peptides, as it surpasses other methods in discrimination. Further research and work are necessary in the future. We plan to develop a web server for this method and integrate more promising algorithms, including advanced deep learning methods, to enhance anti-inflammatory peptide prediction.

ACKNOWLEDGMENT

The work was supported by the National Natural Science Foundation of China (No. 62101100, 62262015)

REFERENCES

[1] Y. Wu, S. Antony, J. L. Meitzler, and J. H. Doroshow, "Molecular mechanisms underlying chronic inflammation-associated cancers," *Cancer Letters*, vol. 345, no. 2, pp. 164–173, 2014. Inflammation and Gastrointestinal and Liver Cancers.

[2] S. La Manna, C. Di Natale, D. Florio, and D. Marasco, "Peptides as therapeutic agents for inflammatory-related diseases," *International journal of molecular sciences*, vol. 19, no. 9, p. 2714, 2018.

[3] C. de la Fuente-Núñez, O. N. Silva, T. K. Lu, and O. L. Franco, "Antimicrobial peptides: Role in human disease and potential as immunotherapies," *Pharmacology Therapeutics*, vol. 178, pp. 132–140, 2017.

[4] M. Dadar, Y. Shahali, S. Chakraborty, M. Prasad, F. Tahoori, R. Tiwari, and K. Dhama, "Antiinflammatory peptides: current knowledge and promising prospects," *Inflammation Research*, vol. 68, pp. 125–145, 2019.

[5] I. J. Elenkov and G. P. Chrousos, "Stress hormones, proinflammatory and antiinflammatory cytokines, and autoimmunity,"

Annals of the New York Academy of Sciences, vol. 966, no. 1, pp. 290–303, 2002.
[6] T. J. Standiford, "Anti-inflammatory cytokines and cytokine antagonists," *Current Pharmaceutical Design*, vol. 6, no. 6, pp. 633–649, 2000.
[7] C. Ao, S. Jiao, Y. Wang, L. Yu, and Q. Zou, "Biological sequence classification: A review on data and general methods," *Research*, vol. 2022, p. 0011, 2022.
[8] F. Cui, Z. Zhang, C. Cao, Q. Zou, D. Chen, and X. Su, "Protein–dna/rna interactions: Machine intelligence tools and approaches in the era of artificial intelligence and big data," *Proteomics*, vol. 22, no. 8, p. 2100197, 2022.
[9] Z. Lv, M. Li, Y. Wang, and Q. Zou, "Editorial: Machine learning for biological sequence analysis," *Frontiers in Genetics*, vol. 14, 2023.
[10] S. Gupta, A. K. Sharma, V. Shastri, M. K. Madhu, and V. K. Sharma, "Prediction of anti-inflammatory proteins/peptides: an insilico approach," *Journal of translational medicine*, vol. 15, no. 1, pp. 1–11, 2017.
[11] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.
[12] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.
[13] S. Jiao, Q. Zou, H. Guo, and L. Shi, "ittca-rf: a random forest predictor for tumor t cell antigens," *Journal of translational medicine*, vol. 19, no. 1, pp. 1–11, 2021.
[14] B. Manavalan, T. H. Shin, M. O. Kim, and G. Lee, "Aippred: sequence-based prediction of anti-inflammatory peptides using random forest," *Frontiers in pharmacology*, vol. 9, p. 276, 2018.
[15] V. Saravanan and N. Gautham, "Harnessing computational biology for exact linear b-cell epitope prediction: a novel amino acid composition-based feature descriptor," *Omics: a journal of integrative biology*, vol. 19, no. 10, pp. 648–658, 2015.
[16] M. Khatun, M. M. Hasan, H. Kurata, *et al.*, "Preaip: computational prediction of anti-inflammatory peptides by integrating multiple complementary features," *Frontiers in genetics*, vol. 10, p. 129, 2019.
[17] J. Zhang, Z. Zhang, L. Pu, J. Tang, and F. Guo, "Aiepred: an ensemble predictive model of classifier chain to identify anti-inflammatory peptides," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 18, no. 5, pp. 1831–1840, 2020.
[18] M. Bhasin and G. P. Raghava, "Classification of nuclear receptors based on amino acid composition and dipeptide composition," *Journal of Biological Chemistry*, vol. 279, no. 22, pp. 23262–23266, 2004.
[19] D. Zhao, Z. Teng, Y. Li, and D. Chen, "iaips: identifying anti-inflammatory peptides using random forest," *Frontiers in Genetics*, vol. 12, p. 773202, 2021.
[20] H. Deng, C. Lou, Z. Wu, W. Li, G. Liu, and Y. Tang, "Prediction of anti-inflammatory peptides by a sequence-based stacking ensemble model named aipstack," *Iscience*, vol. 25, no. 9, 2022.
[21] M. P. LaValley, "Logistic regression," *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008.
[22] L. Wei, P. Xing, R. Su, G. Shi, Z. S. Ma, and Q. Zou, "Cppred-rf: a sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency," *Journal of Proteome Research*, vol. 16, no. 5, pp. 2044–2053, 2017.
[23] L. Wei, J. Tang, and Q. Zou, "Skipcpp-pred: an improved and promising sequence-based predictor for predicting cell-penetrating peptides," *BMC genomics*, vol. 18, no. 7, pp. 1–11, 2017.
[24] Z.-R. Li, H. H. Lin, L. Han, L. Jiang, X. Chen, and Y. Z. Chen, "Profeat: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence," *Nucleic acids research*, vol. 34, no. suppl_2, pp. W32–W37, 2006.
[25] I. Dubchak, I. Muchnik, S. R. Holbrook, and S.-H. Kim, "Prediction of protein folding class using global description of amino acid sequence.," *Proceedings of the National Academy of Sciences*, vol. 92, no. 19, pp. 8700–8704, 1995.
[26] K. Tomii and M. Kanehisa, "Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins," *Protein Engineering, Design and Selection*, vol. 9, no. 1, pp. 27–36, 1996.
[27] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.
[28] Y. Kim, J. Ponomarenko, Z. Zhu, D. Tamang, P. Wang, J. Greenbaum, C. Lundegaard, A. Sette, O. Lund, P. E. Bourne, *et al.*, "Immune epitope database analysis resource," *Nucleic acids research*, vol. 40, no. W1, pp. W525–W530, 2012.
[29] A. Pande, S. Patiyal, A. Lathwal, C. Arora, D. Kaur, A. Dhall, G. Mishra, H. Kaur, N. Sharma, S. Jain, *et al.*, "Pfeature: A tool for computing wide range of protein features and building prediction models," *Journal of Computational Biology*, vol. 30, no. 2, pp. 204–222, 2023.
[30] K.-C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, vol. 21, no. 1, pp. 10–19, 2005.
[31] K.-C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins: Structure, Function, and Bioinformatics*, vol. 43, no. 3, pp. 246–255, 2001.
[32] Z. Chen, P. Zhao, F. Li, A. Leier, T. T. Marquez-Lago, Y. Wang, G. I. Webb, A. I. Smith, R. J. Daly, K.-C. Chou, *et al.*, "ifeature: a python package and web server for features extraction and selection from protein and peptide sequences," *Bioinformatics*, vol. 34, no. 14, pp. 2499–2502, 2018.
[33] T.-Y. Lee, S.-A. Chen, H.-Y. Hung, and Y.-Y. Ou, "Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites," *PloS one*, vol. 6, no. 3, p. e17331, 2011.
[34] H.-B. Shen and K.-C. Chou, "Using ensemble classifier to identify membrane protein types," *Amino acids*, vol. 32, pp. 483–488, 2007.
[35] R. E. Schapire, "Explaining adaboost," in *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pp. 37–52, Springer, 2013.
[36] P. Xuan, C. Sun, T. Zhang, Y. Ye, T. Shen, and Y. Dong, "Gradient boosting decision tree-based method for predicting interactions between target genes and drugs," *Frontiers in genetics*, vol. 10, p. 459, 2019.
[37] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.
[38] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, *et al.*, "Xgboost: extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.
[39] F. Hsieh and B. W. Turnbull, "Nonparametric and semiparametric estimation of the receiver operating characteristic curve," *The annals of statistics*, vol. 24, no. 1, pp. 25–40, 1996.
[40] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve.," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
[41] F. Cui, Z. Zhang, and Q. Zou, "Sequence representation approaches for sequence-based protein prediction tasks that use deep learning," *Briefings in Functional Genomics*, vol. 20, no. 1, pp. 61–73, 2021.