EAI Endorsed Transactions
on AI and Robotics
Research Article EAI.EU

A Stacking Based Ensemble Learning Approach for Accurate Identification of Tumor Homing Peptides in Precision Cancer Therapeutics

Jahid Hassan Akash[1], Rajib Mia[1], Abu Kowshir Bitto[1], Abdul Kadar Muhammad Masum[2*], Jobaer Ahmed[3], Fokrul Islam Khan[4]

[1]Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh
[2]Department of Computer Science and Engineering, Southeast University, Dhaka, Bangladesh
[3]College of Technology & Engineering, Westcliff University, California, United States
[4]College of Business, Westcliff University, California, United States

Abstract

The identification of tumor-homing peptides (THPs) plays a pivotal role in the development of targeted cancer therapies and precision medicine. Current THP identification methods still suffer from limited feature representation, moderate predictive performance, and insufficient generalization, highlighting the need for more robust ensemble frameworks. In this study, we propose STHPP, an innovative stacking-based ensemble machine learning approach designed to improve the accuracy and reliability of THP discovery. Two benchmark datasets, referred to as the "main" and "small" datasets of Shoombuatong were collected, merged, and pre-processed in preparation to create a large dataset and then split for training and testing. The STHPP model applies a two-layer ensemble architecture: first layer that aggregates three heterogenous baseline classifiers, Random Forest (RF), Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), and then second layer applies CatBoost as a meta-classifier for post-processing predictive results of the base models. The two-layer architecture uses model diversity and concepts in ensemble learning to enhance generalization performance. The STHPP framework proposed got outstanding performance with accuracy 0.98, precision 0.97, sensitivity 0.99, specificity 0.97, and a Matthews Correlation Coefficient (MCC) of 0.98. These are better than the performances of current state-of-the-art approaches, which illustrates the effectiveness of using the stacking strategy in complicated peptide classification problems. The finding showcases the potential of STHPP as a strong and scalable computational platform for propelling peptide-based drug discovery research and targeted oncology.

Keywords: Tumor-Homing Peptides, Stacking Ensemble Learning, Feature Extraction, Cancer Therapies, Precision Medicine

Received on 12 September 2025, accepted on 25 January 2026, published on 11 February 2026

doi: 10.4108/airo.10265

1. Introduction

Cancer is a complex and life-threatening disease characterized by the uncontrolled growth and proliferation of abnormal cells that deviate from the normal regulatory mechanisms governing cell division. As these malignant cells multiply, they invade and destroy surrounding healthy tissues, ultimately impairing organ function and disrupting physiological systems. A particularly dangerous feature of cancer is its ability to metastasize—spreading from its original site to distant organs via the lymphatic system or

*Corresponding author. Email: akmmasum@yahoo.com

bloodstream—making treatment increasingly difficult. Globally, cancer is one of the leading causes of mortality, second only to cardiovascular disease, and it imposes significant emotional, physical, and financial burdens on patients, families, and healthcare systems [1].

According to WHO/IARC 2022, cancer caused approximately 9.7 million deaths and 20 million new cases globally. These numbers are projected to increase dramatically by 2040, with an estimated 16 million deaths and 29.4 million new cases anticipated [2] [3]. Despite substantial advancements in early detection and treatment, current therapeutic strategies, particularly chemotherapy, remain limited by significant challenges. Chemotherapeutic agents often lack specificity, targeting both cancerous and healthy cells alike. This non-selectivity results in severe side effects, limits treatment duration and dosage, and ultimately compromises therapeutic efficacy. Therefore, developing methods that improve tumor-specific targeting while minimizing off-target toxicity is a critical objective in cancer research [4].

In this context, peptides have emerged as promising candidates for targeted drug delivery. Various peptide-based systems—such as cell-penetrating peptides (CPPs), homing peptides (HPs), and cell-penetrating homing peptides—have been proposed to enhance delivery precision [5]. Among these, tumor-homing peptides (THPs), a subclass of HPs, have shown significant potential as selective delivery agents [6]. THPs are short peptide sequences, typically comprising 3 to 30 amino acid residues, engineered to bind specifically to tumor cells or associated vasculature. They often feature motifs such as Arg-Gly-Asp (RGD) and Asn-Gly-Arg (NGR), which enable selective binding to tumor-associated antigens and blood vessels with poor antigenicity [7]. These characteristics make THPs a powerful tool for the selective delivery of therapeutic agents to various tumor types, including melanoma, colon, breast, lung, and prostate cancers.

Accurate identification of tumor-homing peptides is vital for advancing cancer diagnosis, prognosis, and personalized treatment. Traditional laboratory-based identification methods are time-consuming, resource-intensive, and costly. These challenges emphasize the urgency of developing computational systems capable of accurately identifying tumor-specific peptides while reducing false predictions. To overcome these limitations, researchers have increasingly adopted computational approaches, particularly machine learning (ML) and deep learning (DL) techniques, which offer rapid, scalable, and highly accurate alternatives. These models can process complex biological datasets to predict tumor-specific biomarkers and identify THPs with significantly improved efficiency and precision.

This study aims to contribute to this growing body of research by introducing STHPP, a stacking-based ensemble machine learning model specifically designed for tumor-homing peptide identification. The major contributions of this work are a comprehensive dataset was created by curating and merging two benchmark datasets, ensuring a richer and more reliable foundation for robust model training and evaluation. To further strengthen the dataset, diverse feature engineering techniques were applied, enabling the extraction of informative attributes that not only enhanced predictive performance but also improved interpretability. Building on this, a novel stacking-based ensemble framework was developed, which integrated multiple baseline classifiers with a meta-classifier. This innovative approach consistently outperformed existing models across all key evaluation metrics, demonstrating its effectiveness and superiority.

2. Related Work

In recent years, the identification of tumor-homing peptides (THPs) has gained significant interest due to their utility in targeted cancer treatment and site-specific drug delivery. Because conventional experimental techniques for peptide identification are labor-intensive, time-consuming, and costly, numerous computational approaches have been developed to simplify the prediction process. Several machine learning-based models have been suggested with different feature extraction methods and classification approaches with the aim of enhancing prediction reliability and accuracy. This section presents key dominant models, their strategies, feature extraction methods, performance measures, and key contributions and limitations with the aim of motivating better ensemble-based methods like the presented STHPP model.

Guan et al. [8] have suggested a stacking-based model known as StackTHPred. They used AAC, PAAC, PCP, BLOSUM62, and Z-Scale as their feature extraction methods. The model achieved an accuracy of 0.915 and an MCC of 0.831. Shoombuatong et al. [9] proposed a model named THPep, developed using an interpretable random forest classifier that integrates features such as AAC, DPC, and PAAC. This model achieved an overall accuracy of 0.901 and an MCC of 0.76. Charoenkwan et al. [10] developed a model using SCM and the propensity scores of 20 amino acids, named SCMTHP. It achieved an accuracy of 0.827, indicating moderate but not highly competitive performance. Sharma et al. [11] proposed an SVM-based model utilizing AAC and binary profiles of peptides. The maximum accuracies were 86.56% for AAC-based features, 82.03% for dipeptide composition, and 84.19% for binary profiles. Zou et al. [12] proposed a method to distinguish tumor-homing peptides (THPs) from non-THPs using PseRECM and PsePC for encoding, LASSO for feature selection, and SVM for classification. Their model achieved classification accuracies of 0.8902, 0.8849, and 0.9458 on the Main, Small, and Main90 datasets, respectively. Another stacking-based model, NEPTUNE, was proposed by Charoenkwan et al. [13] This model employed twelve different feature extractors and six machine learning classifiers, with SVM serving as the meta-model in the stacking framework. It achieved an accuracy of 0.888 and an MCC of 0.777.

Table 1 shows the taxonomy of existing research conducted on tumor-homing peptide prediction. While each of the approaches demonstrates specific strengths, they also come with certain limitations. This study addresses those limitations by introducing an improved stacking-based

ensemble model named STHPP, designed to more accurately and effectively identify tumor-homing peptides compared to existing prediction models.

3. Methodology

In this research we adopted a systematic and structured multi-stage methodology, beginning with data collection and consolidation. The two benchmark datasets were initially prepared as one main dataset and one secondary (small) dataset, and subsequently merged to create a rich and comprehensive data source. Merging was done to ensure enhanced data heterogeneity and robustness so that there could be a more solid foundation for model training and testing. Following the consolidation of datasets, the next important step was the application of various feature extraction techniques. These played an important role in identifying and isolating the most informative and useful features from the consolidated dataset. Based on considerations of useful statistical and biological characteristics, the process of feature extraction was optimized to use the learning capacity of the following machine learning models and improve overall prediction accuracy. Following this, different machine learning algorithms were deployed on the cleaned data. The models were evaluated based on how they predicted tumor-homing peptides (THPs), with particular emphasis on generalizability and reliability. The final step of the methodology was comprised of a comprehensive performance analysis. Several performance measurements, including accuracy, precision, recall, specificity, and Matthews Correlation Coefficient, were computed to assess each model's performance in an unbiased way. The comparison allowed us to determine the best-performing structure of the models and inform the design of our proposed stacking-based ensemble model, STHPP. Figure 1 displays the overall structure and workflow of the suggested STHPP model, with every step from preprocessing to final prediction indicated.

Table 1. Taxonomy of Existing Research in Tumor-Homing Peptide Prediction

Author	Method	Feature Extractor	Accuracy
Guan [8]	StackTHPred	AAC, PAAC, PCP, BLOSUM62, and Z-Scale	0.92
Shoombuatong [9]	THPep	AAC, DPC, and PAAC	0.90
Charoenkwan [10]	SCMTHP	AAC, DPC, PAAC	0.82
Sharma [11]	SVM	AAC, DPC, Binary Profile	0.86
Zou [12]	SVM	PseRECM, PsePC, LASSO	0.94
Charoenkwan [13]	NEPTUNE	AAC, DPC, AAI, APAAC, CTD, PAAC, PCP, RSs, RScharge, RSDHP, RSpolar, RSsecond	0.88

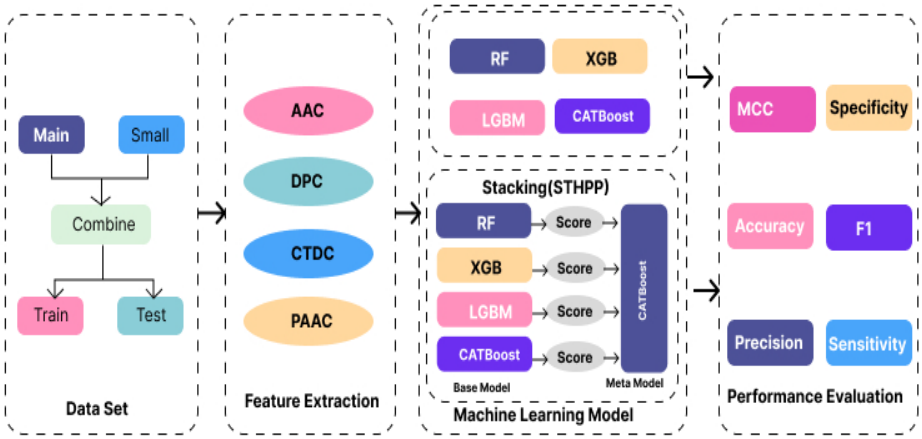

Figure 1. Working Procedure to Construct STHPP

Table 2. Description of Feature Encoding Method

Feature Extraction Method	Full Name of Descriptor	Feature Dimension
AAC	Amino Acid Composition	21

DPC	Di-Peptide Composition	400
CTDC	Composition, Transition, Distribution	39
PAAC	Pseudo-Amino Acid Composition	22

3.1. Dataset

During this investigation, two publicly available benchmark datasets referred to as the "main" and "small" datasets were compiled from the repository maintained at https://github.com/Shoombuatong/THPep [9]. The main dataset consists of 651 tumor-homing peptides (THPs) and an equal number of non-THPs, offering a balanced and extensive dataset for comprehensive model development. In contrast, the small dataset, derived as a representative subset of the main dataset, contains 469 THPs and 469 non-THPs. To enrich the training process and enhance model generalizability, both datasets were merged to construct a unified, high-quality dataset. To prevent data leakage , prior to merging, both datasets were checked for duplicate peptide sequences, no duplication is found. This consolidated dataset was then partitioned into distinct training and testing sets. A stratified sampling technique was applied during the train–test partition to ensure proportional representation of THPs and non-THPs. The training set comprises 959 THPs and 959 non-THPs, ensuring balanced class representation. The testing set includes 161 THPs and 161 non-THPs, facilitating robust model evaluation and unbiased performance assessment.

3.2. Feature Extraction

Feature extraction plays a pivotal role in machine learning (ML), as it converts raw biological sequences into structured numerical representations that are suitable for computational modeling. By isolating the most informative and discriminative features from the input data, ML models become not only more accurate but also more interpretable and generalizable [14]. In this study, we employed the iLearnPlus platform for feature extraction, leveraging four well-established encoding techniques: Amino Acid Composition (AAC), Di-Peptide Composition (DPC), Composition/Transition/Distribution Composition (CTDC), and Pseudo-Amino Acid Composition (PAAC). Each technique encapsulates distinct biological properties and is briefly described in the subsequent subsections.

Table 2 provides a simple summary of the feature encoding schemes applied in this study. Each of these schemes captures varied biochemical or structural information of the peptide sequences and varies in the dimension of the resulting feature vectors. AAC documents the overall frequency of amino acids, while DPC considers local pairwise residue patterns. CTDC encodes physicochemical properties via statistical measurements, and PAAC incorporates sequence-order information, improving the representation of peptide structures for improved machine learning performance.

3.3. Amino Acid Composition

The Amino Acid Composition (AAC) encoding process involves determining the frequency of each amino acid type within a protein or peptide sequence. This calculation encompasses the frequencies of all 20 naturally occurring amino acids.

$$f(t) = \frac{N(t)}{N}, \quad t \in \{A, C, D, \ldots, Y\} \qquad (1)$$

Here, N is the length of a protein or peptide sequence, and N(t) denotes the count of amino acid type *t*. The AAC descriptor has demonstrated efficacy in various applications, including the classification of nuclear receptors and the anticipation of anticancer peptides [15].

3.4. Di-Peptide Composition

The Dipeptide Composition (DPC) creates 400 descriptors and is formulated as follows:

$$D(r, s) = \frac{N_{rs}}{N-1}, \quad r, s \in \{A, C, D, \ldots, Y\} \qquad (2)$$

In this context, N_{rs} represents the count of dipeptides formed by amino acid types r and s [16].

3.5. Composition, Transition, Distribution, Composition

Using Composition, Transition, Distribution, Composition (CTDC), bioinformatics techniques are applied to extract characteristics from protein sequences and convert them into numerical representations. It includes the recording of physical and chemical characteristics, which helps with assignments such as protein categorization and structural forecasting.

$$C(r) = \frac{N_r}{N}, \quad r \in \{\text{polar, neutral, hydrophobic}\} \quad (3)$$

Here, N stands for the length of the sequence and N(r) for the count of amino acid type *r* in the encoded sequence [17].

3.6. Pseudo-Amino Acid Composition

The PAAC technique is utilized to transform protein sequences into numerical representations. This method captures a more extensive set of information than traditional amino acid composition, leading to improved performance in applications.

$$X_c = \frac{f_c}{\sum_{r=1}^{20}\left(f_r + w\sum_{j=1}^{\lambda}\theta_j\right)}, \quad (1 < c < 20) \qquad (4)$$

$$X_c = \frac{w\theta_{c-20}}{\sum_{r=1}^{20}\left(f_r + w\sum_{j=1}^{\lambda}\theta_j\right)}, \quad (21 < c < 20 + \lambda) \qquad (5)$$

The weight assigned to the sequence-order effect, represented as w, is commonly established at 0.05 in iLearnPlus, following the recommendation in [18].

3.7. Machine Learning Classifiers

In this research, we constructed identification models employing diverse ensemble learning classification methods, such as Random Forest, XGBoost, LGBM, CatBoost and Stacking (STHPP). The subsequent section provides an overview of each technique.

1) **Random Forest (RF):** The ensemble classifier known as Random Forest (RF) averages decision tree model outcomes across various dataset sub-samples to increase accuracy and reduce overfitting. The eventual result is predicted by majority vote. [19].
2) **Light Gradient Boosting Method (LGBM):** Light Gradient Boosting Method, or LightGBM, is a distributed architecture that uses tree-based training techniques to achieve gradient boosting efficiency. Through advanced techniques such as Exclusive Feature Bundling (EFB) and Gradient-based One-Side Sampling (GOSS), LightGBM overcomes the limitations of histogram-based strategies used in other gradient boosting decision tree (GBDT) frameworks. The model's efficiency is significantly enhanced by these combined techniques, giving it an edge over other GBDT architectures [20].
3) **Extreme Gradient Boosting (XGB):** XGBoost, or eXtreme Gradient Boosting, is a scalable, highly accurate, and robust gradient boosting algorithm built on decision trees. It excels in terms of faster training times and lower memory consumption. XGBoost provides feature importance scores for better model understanding and effectively handles large datasets through parallel processing. XGBoost rapidly and easily applied to various tasks. [21].
4) **Categorical Boosting (CatBoost):** CatBoost, short for Categorical Boosting, is a gradient boosting algorithm designed specifically for handling categorical features efficiently. It tackles the limitations of decision tree models by employing techniques like Ordered Target Statistics (OTSS), which eliminate the need for one-hot encoding—reducing memory usage and speeding up training, particularly on large datasets. CatBoost achieves competitive accuracy with fewer trees, and automatic hyperparameter tuning reduces manual effort. While decision tree models are generally challenging to interpret, CatBoost provides feature importance scores and partial dependence plots to offer valuable insights. Although it may not be optimal for every task, its categorical data handling capabilities make it useful for a wide range of machine learning applications. As a relatively new algorithm, CatBoost continues to improve through active development [22].
5) **Proposed Stacking Model (STHPP):** Our proposed ensemble classifier, denoted as STHPP, adopts a stacking-oriented methodology shows in figure 2. Stacking involves the amalgamation of classification or regression algorithms through a dual-layered estimation approach. In the initial layer, referred to as baseline models, specific classification techniques are employed. We employed RF, LGBM, XGB and CatBoost ML models as baseline classifier. These baseline models make predictions based on test datasets. The subsequent layer introduces the ultimate classifier method, termed the Meta-Classifier. We employed CatBoost ML algorithm as Meta-Classifier. Utilizing the predictions generated by all baseline models as input, the Meta-Classifier generates novel predictions. When compared to traditional machine learning models, the stacking-based prediction model yields higher accuracy. The combination of different algorithms and the addition of a two-stage prediction procedure is credited with this improvement. CatBoost was chosen as the meta-classifier due to its robustness in handling small to medium-sized datasets and its built-in support for overfitting control via Ordered Boosting and Oblivious Trees. CatBoost can learn complex non-linear relationships between base learners without requiring extensive manual feature engineering. Its implicit handling of categorical variables makes it possible to use regularization techniques, which makes it less susceptible to overfitting, which in the meta-learning phase can result in degraded performance because of over-aggregation of nearby predictions.

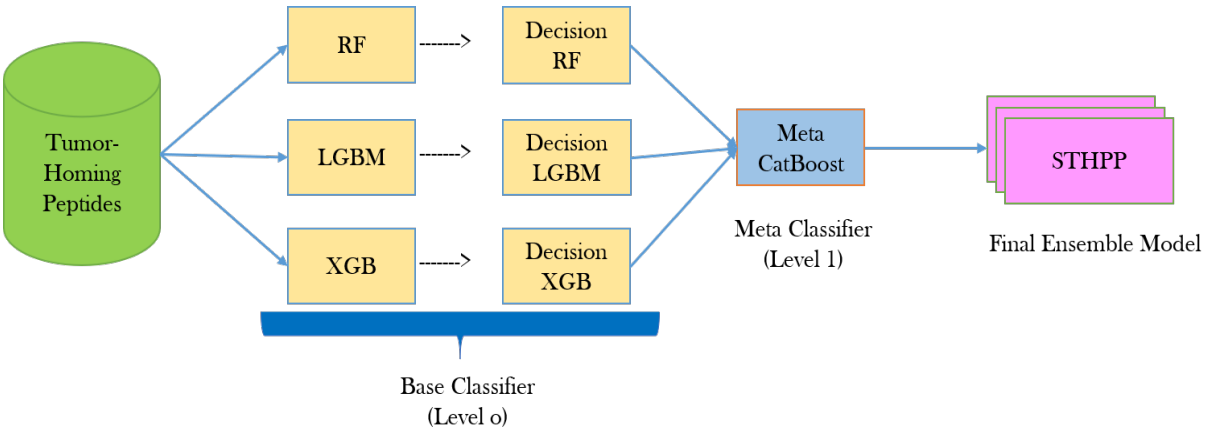

Figure 2. Proposed STHPP Model Architecture Diagram

3.8. Performance Evaluation

To evaluate the effectiveness of the applied machine learning classifiers, a variety of statistical assessment metrics were employed [23-30]. Chief among these are accuracy, sensitivity (recall), and specificity, all of which are derived from the confusion matrix. These metrics are particularly useful when the target classes in the dataset are relatively balanced. Accuracy evaluates the overall correctness of the model by measuring the proportion of true results (both true positives and true negatives) among the total number of cases examined. It is calculated using Equation (6). Sensitivity (Recall) refers to the proportion of actual positive cases that are correctly identified by the model. This metric highlights the model's ability to capture positive instances and is crucial in scenarios where false negatives are costly (Equation (7)). Specificity, on the other hand, represents the proportion of actual negatives correctly classified as such. It indicates the model's capability in recognizing negative instances (Equation (8)). In addition to these fundamental metrics, we incorporated Matthews Correlation Coefficient (MCC), it is a robust metric that considers true and false positives and negatives and produces a score between -1 and +1. A coefficient near +1 indicates a perfect prediction, 0 indicates a random prediction, and -1 signifies total disagreement between prediction and observation. It is beneficial when dealing with imbalanced datasets [25]. The formulas for the above metrics are provided below:

$$Accuracy = \frac{TP + TN}{TP+TN+FN+FP} \times 100\% \quad (6)$$

$$Sensitivity = \frac{TP}{TP+FN} \times 100\% \quad (7)$$

$$Specificity = \frac{TN}{TN + FP} \times 100\% \quad (8)$$

$$MCC = \frac{(TP \times TN)-(FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \times 100\% \quad (9)$$

Here, TP = True Positive, TN = True Negative, FP = False Positive, and FN = False Negative.

4. Result and Discussion

To test the effectiveness of the classifiers, the model's performance was evaluated using a variety of performance metrics. The comprehensive methodology is outlined as follows. We assessed five different models: RF, LGBM, XGB, CAT, and STHPP, each applied to features extracted using iLearnPlus methods—AAC, DPC, CTDC, and PAAC. A 5-fold cross-validation was conducted on the training dataset, followed by evaluation on an independent test dataset. To verify the significance of the observed performance differences, we conducted Wilcoxon signed-rank test between STHPP and the best-performing baseline. The results showed that the improvements in accuracy and MCC were statistically significant ($p < 0.01$), confirming the robustness of the ensemble approach.

Table 3 displays outcomes for various performance metrics assessed on the training dataset. Within the five machine learning algorithms, STHPP attains the highest accuracy score of 0.9432 in AAC, while CAT records the lowest accuracy at 0.8725 in DPC. Additionally, STHPP outperforms in other performance metrics, achieving the maximum scores for precision (0.9434), f1-score (0.9365), sensitivity (0.9314), specificity (0.9563), and MCC value (0.8899).

Table 4 displays the results for several performance measures on the testing dataset. STHPP demonstrates the maximum accuracy scores of 0.9884 in PAAC and 0.9807 in AAC. Moreover, STHPP achieves the highest scores across other performance metrics in PAAC, including precision (0.9434), f1-score (0.9365), sensitivity (0.9314), specificity (0.9563), and MCC value (0.8899). Conversely, AAC records scores of 0.9699, 0.9809, 0.9923, 0.9692, and 0.9771 in the respective metrics for AAC. Despite AAC's strong performance in the training dataset, we choose AAC as our proposed feature extractor method.

Figure 3 presents a detailed performance comparison conducted on the training dataset across several subplots. Subplot (A) illustrates the results for the Amino Acid Composition (AAC), while Subplot (B) focuses on the Di-Peptide Composition (DPC). Subplot (C) displays the

outcomes for the Composition, Transition, Distribution, Composition (CTDC), and Subplot (D) provides insights of Pseudo-Amino Acid Composition (PAAC). Each subplot offers a comprehensive view of the respective ML classifiers performances, facilitating a clearer understanding of the differences and strengths among with the feature extraction methods.

Figure 4 offers a comprehensive comparison of performance on the testing dataset, depicted across multiple subplots. Subplot (A) highlights the performance of Amino Acid Composition (AAC), while Subplot (B) focuses on Di-Peptide Composition (DPC). Subplot (C) presents the results for Composition, Transition, and Distribution (CTDC), and Subplot (D) showcases the performance of Pseudo-Amino Acid Composition (PAAC). Each subplot provides a detailed overview of the respective ML classifiers, enabling a clearer understanding of their individual differences and strengths with our applied feature extraction methods.

Figure 5 provides a comprehensive performance comparison of various feature extractors. The figure is divided into two distinct subplots to highlight the performance on different datasets. Subplot (A) illustrates the performance metrics obtained from the training dataset, providing insight into how well each feature extractor performs during the model's learning phase. On the other hand, Subplot (B) focuses on the performance of these feature extractors on the testing dataset, showcasing how effectively the model generalizes to unseen data. By examining both subplots, we can assess the consistency and reliability of each feature extraction method across both the training and testing stages.

Table 3. Performance Evaluation of the Applied Classifier on Four Feature Extraction Methods using Cross Validation

Extractor	Classifier	Accuracy	Precision	F1 Score	Sensitivity	Specificity	MCC
AAC	RF	0.9320	0.9018	0.9266	0.9460	0.9177	0.8660
	XGB	0.9304	0.8976	0.9252	0.9545	0.9106	0.8610
	LGBM	0.9308	0.8977	0.9256	0.9554	0.9106	0.8626
	CAT	0.9250	0.8945	0.9191	0.9451	0.9085	0.8505
	STHPP	0.9432	0.9434	0.9365	0.9314	0.9563	0.8899
DPC	RF	0.8879	0.8796	0.8751	0.8697	0.9008	0.7665
	XGB	0.9061	0.8660	0.8999	0.9365	0.8811	0.8133
	LGBM	0.8933	0.8595	0.8852	0.9125	0.8776	0.7871
	CAT	0.8725	0.8655	0.8572	0.8491	0.8917	0.7422
	STHPP	0.9069	0.8991	0.8964	0.8920	0.9177	0.8127
CTDC	RF	0.9258	0.8992	0.9195	0.9374	0.9099	0.8473
	XGB	0.9223	0.8933	0.9160	0.9400	0.9078	0.8463
	LGBM	0.9246	0.8913	0.9190	0.9485	0.9050	0.8501
	CAT	0.9165	0.8862	0.9099	0.9348	0.9015	0.8333
	STHPP	0.9312	0.9387	0.9224	0.8980	0.9437	0.8479
PAAC	RF	0.9258	0.8992	0.9195	0.9374	0.9099	0.8620
	XGB	0.9223	0.8933	0.9160	0.9400	0.9078	0.8371
	LGBM	0.9246	0.8913	0.9190	0.9485	0.9050	0.8363
	CAT	0.9165	0.8862	0.9099	0.9348	0.9015	0.8336
	STHPP	0.9312	0.9387	0.9224	0.8980	0.9437	0.8663

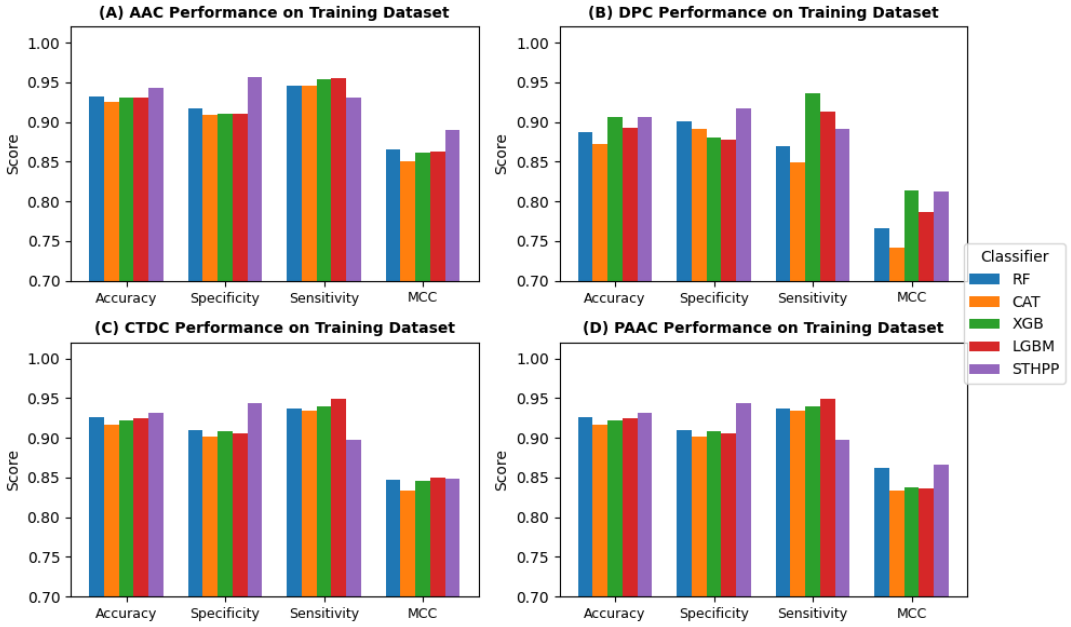

Figure 3. Performance comparison of ML classifiers on training dataset using different feature extraction methods: (A) Amino Acid Composition (AAC), (B) Di-Peptide Composition (DPC), (C) Composition, Transition, Distribution Composition (CTDC), and (D) Pseudo-Amino Acid Composition (PAAC)

Table 4. Performance Evaluation of the Applied Classifier on Four Feature Extraction Methods using Independent Test

Extractor	Classifier	Accuracy	Precision	F1 Score	Sensitivity	Specificity	MCC
AAC	RF	0.9730	0.9489	0.9737	1.0000	0.9461	0.9730
	XGB	0.9692	0.9420	0.9701	1.0000	0.9384	0.9402
	LGBM	0.9769	0.9558	0.9774	1.0000	0.9538	0.9548
	CAT	0.9692	0.9420	0.9701	1.0000	0.9384	0.9402
	STHPP	0.9807	0.9699	0.9809	0.9923	0.9692	0.9771
DPC	RF	0.9307	0.9375	0.9302	0.9384	0.9384	0.8773
	XGB	0.9653	0.9416	0.9662	0.9923	0.9384	0.9321
	LGBM	0.9538	0.9402	0.9545	0.9692	0.9384	0.9081
	CAT	0.9153	0.9218	0.9147	0.9076	0.9230	0.8308
	STHPP	0.9615	0.9615	0.9615	0.9615	0.9692	0.9307
CTDC	RF	0.9769	0.9558	0.9774	1.0000	0.9384	0.9475
	XGB	0.9769	0.9558	0.9774	1.0000	0.9538	0.9548
	LGBM	0.9692	0.9420	0.9701	1.0000	0.9384	0.9402
	CAT	0.9730	0.9489	0.9737	1.0000	0.9461	0.9475
	STHPP	0.9884	0.9774	0.9885	1.0000	0.9769	0.9771
PAAC	RF	0.9692	0.9420	0.9701	1.0000	0.9538	0.9475
	XGB	0.9730	0.9489	0.9737	1.0000	0.9461	0.9548
	LGBM	0.9769	0.9558	0.9774	1.0000	0.9538	0.9548
	CAT	0.9769	0.9558	0.9774	1.0000	0.9538	0.9548
	STHPP	0.9884	0.9847	0.9885	0.9923	0.9846	0.9769

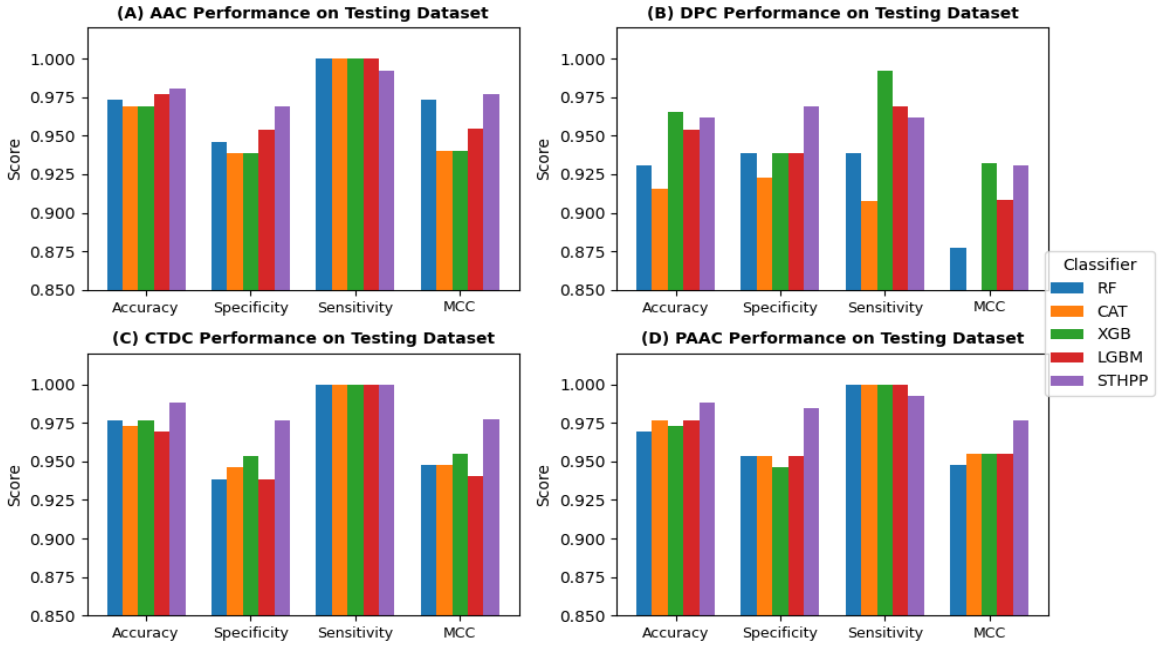

Figure 4. Performance comparison of ML classifiers on testing dataset using different feature extraction methods: (A) Amino Acid Composition (AAC), (B) Di-Peptide Composition (DPC), (C) Composition, Transition, Distribution Composition (CTDC), and (D) Pseudo-Amino Acid Composition (PAAC).

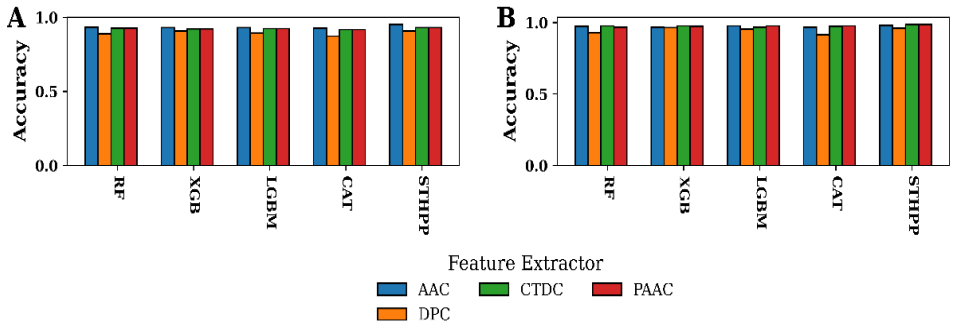

Figure 5. Comparative Performance of Different Feature Extractors on (A) Training Dataset and (B) Testing Dataset.

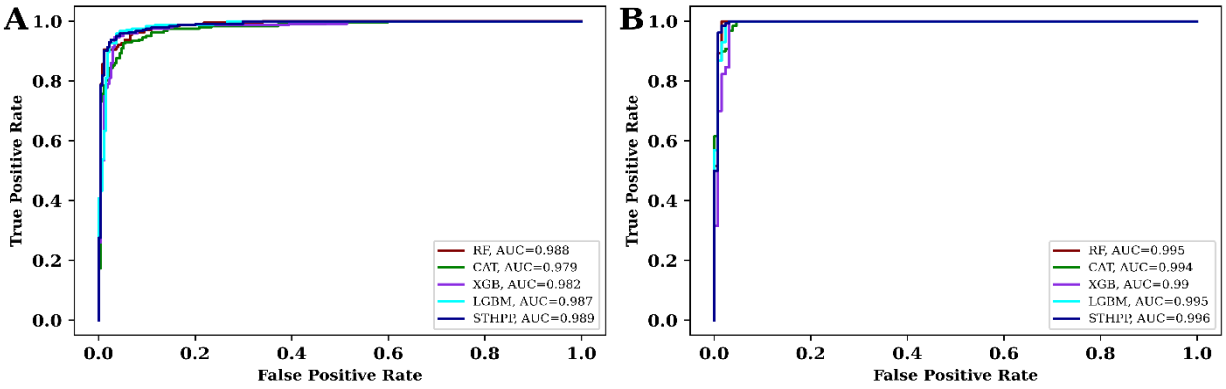

Figure 6. Receiver Operating Characteristic (ROC) Curves with Area Under the Curve (AUC) scores of various classifiers on (A) Training Dataset and (B) Testing Dataset.

Figure 6 presents the Receiver Operating Characteristic (ROC) curves, along with the corresponding Area Under the Curve (AUC) scores, for a variety of classifiers. The figure is organized into two separate subplots, each serving a specific purpose in evaluating the classifiers' performance. Subplot (A) depicts the ROC curves generated using the training dataset, which illustrates how well each classifier performs during the model training process. This provides valuable insights into the ability of the classifiers to distinguish between classes during training. Subplot (B), on the other hand, displays the ROC curves for the testing dataset, highlighting the classifiers' performance when applied to unseen data. By comparing both subplots, one can assess not only the classifiers' ability to learn from the training data but also their generalization performance on new, unseen test data. The inclusion of AUC scores further enhances this evaluation by providing a quantitative measure of each classifier's overall performance.

Extensive research has been conducted on predicting THPs; however, there remains scope for further advancements in this area. In our study, we utilize a THP sequence dataset for THP prediction. Following dataset collection, we amalgamated the main and small datasets to create a unified dataset. Subsequently, we employed four feature extraction methods—AAC, DPC, CTDC, and PAAC. Five supervised ML algorithms were applied for precise THP prediction. Post-application of the ML approaches, we evaluated the results using various performance metrics, including accuracy, precision, f1-score, sensitivity, specificity, and MCC. The models previously implemented comprised StackTHPred [8], THPep [9], SCMTHP [10], SVM [11], and NEPTUNE [13]. Among these, our STHPP demonstrated superior performance in accuracy, specificity, and MCC. An example, StackTHPred [8], which relies on gradient boosting-based feature selection before stacking, STHPP introduces a two-layer heterogeneous ensemble combining four diverse classifiers (RF, LGBM, XGB, CatBoost) at the base level and CatBoost at the meta-level. Additionally, STHPP uses raw fused features from multiple encoding schemes (AAC, DPC, etc.) without feature filtering, thereby preserving informative signals that may be removed in feature selection. The training strategy in STHPP incorporates a broader diversity of learners and a stratified k-fold CV at both levels to reduce overfitting.

In Table 5, it is observed that our proposed STHPP model achieved the highest accuracy score of 0.9807 among the five machine learning algorithms evaluated. Additionally, STHPP demonstrated superior performance across other performance metrics, with scores of 0.9699, 0.9809, 0.9923, 0.9692, and 0.9771 for precision, f1-score, sensitivity, specificity, and MCC value, respectively. However, this study has some limitations. First, we used secondary dataset. Secondly, we applied only four feature extractors and five ML models. In the future, we will try to collect a primary dataset and apply more feature extraction methods. In addition, we will apply some Deep Learning algorithms.

Table 5. Comparison of the proposed model with existing methods.

Method	Accuracy	Precision	F1-Score	Sensitivity	Specificity	MCC
StackTHPred[8]	0.915	N/A	N/A	0.915	0.915	0.831
THPep[9]	0.908	N/A	N/A	0.918	0.8797	0.77
SCMTHP[10]	0.827	N/A	N/A	0.869	0.785	0.656
SVM[11]	0.8656	N/A	N/A	0.8063	0.8971	0.70
NEPTUNE[13]	0.888	N/A	N/A	N/A	N/A	0.777
STHPP	0.9807	0.9699	0.9809	0.9923	0.9692	0.9771

5. Conclusion

The research presented in this article showcases the efficacy of the AAC feature encoding method in the identification of THPs. The utilization of Stacking Classifiers STHPP approach has demonstrated promising results, offering a robust and accurate framework for the prediction of tumor homing peptides. Incorporating diverse evaluation metrics such as accuracy, precision, f1-score, sensitivity, specificity, and MCC offers a thorough evaluation of the model's performance. The findings suggest that the proposed methodology enhances the accuracy and reliability of identifying tumor homing peptides, contributing valuable insights to peptide-based drug development and targeted therapies. The combination of feature encoding techniques with advanced machine learning strategies, as demonstrated in this research, paves the way for further advancements in bioinformatics and computational biology. This work lays a foundation for future studies aiming to improve the precision and specificity of peptide identification, ultimately advancing our understanding of tumor-targeting mechanisms and their therapeutic.

References

[1] Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. cell. 2011 Mar 4;144(5):646-74.

[2] Global cancer burden growing, amidst mounting need for services (1 February 2024) World Health Organization. Available at: https://www.iarc.who.int/news-events/global-cancer-burden-growing-amidst-mounting-need-for-services (Accessed: 07 December 2025)

[3] Ferlay J, Colombet M, Soerjomataram I, Parkin DM, Piñeros M, Znaor A, Bray F. Cancer statistics for the year 2020: An overview. International journal of cancer. 2021 Aug 15;149(4):778-89.

[4] Mäe M, Myrberg H, El-Andaloussi S, Langel Ü. Design of a tumor homing cell-penetrating peptide for drug delivery. International Journal of peptide research and therapeutics. 2009 Mar;15(1):11-5.

[5] Svensen N, Walton JG, Bradley M. Peptides for cell-selective drug delivery. Trends in pharmacological sciences. 2012 Apr 1;33(4):186-92.

[6] Khongorzul P, Ling CJ, Khan FU, Ihsan AU, Zhang J. Antibody–drug conjugates: a comprehensive review. Molecular Cancer Research. 2020 Jan 1;18(1):3-19.

[7] Kondo E, Iioka H, Saito K. Tumor-homing peptide and its utility for advanced cancer medicine. Cancer science. 2021 Jun;112(6):2118-25.

[8] Guan J, Yao L, Chung CR, Chiang YC, Lee TY. Stackthpred: identifying tumor-homing peptides through gbdt-based feature selection with stacking ensemble architecture. International Journal of Molecular Sciences. 2023 Jun 19;24(12):10348.

[9] Shoombuatong W, Schaduangrat N, Pratiwi R, Nantasenamat C. THPep: A machine learning-based approach for predicting tumor homing peptides. Computational biology and chemistry. 2019 Jun 1;80:441-51.

[10] Charoenkwan P, Chiangjong W, Nantasenamat C, Moni MA, Lio' P, Manavalan B, Shoombuatong W. SCMTHP: A new approach for identifying and characterizing of tumor-homing peptides using estimated propensity scores of amino acids. Pharmaceutics. 2022 Jan 4;14(1):122.

[11] Sharma A, Kapoor P, Gautam A, Chaudhary K, Kumar R, Chauhan JS, Tyagi A, Raghava GP. Computational approach for designing tumor homing peptides. Scientific reports. 2013 Apr 5;3(1):1607.

[12] Zou H, Yang F, Yin Z. Identification of tumor homing peptides by utilizing hybrid feature representation. Journal of Biomolecular Structure and Dynamics. 2023 May 24;41(8):3405-12.

[13] Charoenkwan P, Schaduangrat N, Moni MA, Manavalan B, Shoombuatong W. NEPTUNE: a novel computational approach for accurate and large-scale identification of tumor homing peptides. Computers in Biology and Medicine. 2022 Sep 1;148:105700.

[14] Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of machine learning research. 2003;3(Mar):1157-82.

[15] Wei L, Zhou C, Chen H, Song J, Su R. ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. Bioinformatics. 2018 Dec 1;34(23):4007-16.

[16] Bhasin M, Raghava GP. Classification of nuclear receptors based on amino acid composition and dipeptide composition. Journal of Biological Chemistry. 2004 May 1;279(22):23262-6.

[17] Saravanan V, Gautham N. BCIgEPRED—a dual-layer approach for predicting linear IgE epitopes. Molecular Biology. 2018 Mar;52(2):285-93.

[18] Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Structure, Function, and Bioinformatics. 2001 May 15;43(3):246-55.

[19] Tuj Jannat F, Biplob KB, Bitto AK. Predicting bangladesh life expectancy using multiple depend features and regression models. InInternational Conference on Machine Intelligence and Signal Processing 2022 Mar 12 (pp. 47-58). Singapore: Springer Nature Singapore.

[20] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems. 2017;30.

[21] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. InProceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 785-794).

[22] Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems. 2018;31.

[23] Sanni RR, Guruprasad HS. Analysis of performance metrics of heart failured patients using Python and machine learning algorithms. Global transitions proceedings. 2021 Nov 1;2(2):233-7.

[24] Chicco D, Tötsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData mining. 2021 Feb 4;14(1):13.

[25] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics. 2020 Jan 2;21(1):6.

[26] Abu Kowshir Bitto, Rezwana Karim, M. H. Begum, M. F. I. K. Khan, Dr. Md. Maruf Hassan, and Prof. Dr. Abdul kadar Muhammad Masum, "Explainable AI Based Deep Ensemble Convolutional Learning for Multi-Categorical Ocular Disease Prediction", *EAI Endorsed Transactions on AI and Robotics*, vol. 4, Jul. 2025.

[27] Prof. Dr. Abdul kadar Muhammad Masum, A. K. Bitto, S. I. Talukder, M. F. I. Khan, M. S. Alam, and K. M. Mohi Uddin, "An Explainable AI Based Deep Ensemble Transformer Framework for Gastrointestinal Disease Prediction from Endoscopic Images", *EAI Endorsed Transactions on AI and Robotics*, vol. 4, Aug. 2025.

[28] K. Kalaivani, P. Deepan, G. Ganesh, J. Ravichandran, and S. Dhiravidaselvi, "Hyperband-Optimized Convolutional Neural Network Model for Efficient Brain Tumor Classification and Prediction", *EAI Endorsed Transactions on AI and Robotics*, vol. 4, Sep. 2025.

[29] Badruzzaman Biplob KB, Sammak MH, Bitto AK, Mahmud I. COVID-19 and Suicide Tendency: Prediction and Risk Factor Analysis Using Machine Learning and Explainable AI. EAI Endorsed Transactions on Pervasive Health & Technology. 2024 Jan 1;10(1).

[30] Mia R, Hasan T, Bitto AK, Mahadi Hassan M, Shamsul Alam M, Abdul Kadar Muhammad Masum. Enhancing the Prediction of IL-4 Inducing Peptides Using Stacking Ensemble Model. EAI Endorsed Trans AI Robotics [Internet]. 2025 Oct. 27 [cited 2026 Feb. 9];4. Available

from: https://publications.eai.eu/index.php/airo/article/view/9867