frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 28 April 2022
doi: 10.3389/fgene.2022.883766

Ensemble-AHTPpred: A Robust Ensemble Machine Learning Model Integrated With a New Composite Feature for Identifying Antihypertensive Peptides

Supatcha Lertampaiporn[1], *Apiradee Hongsthong*[1], *Warin Wattanapornprom*[2] *and Chinae Thammarongtham*[1]*

[1]Biochemical Engineering and Systems Biology Research Group, National Center for Genetic Engineering and Biotechnology, National Science and Technology Development Agency at King Mongkut's University of Technology Thonburi, Bangkok, Thailand, [2]Applied Computer Science Program, Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi, Bangkok, Thailand

OPEN ACCESS

Edited by:
Yanjie Wei,
Shenzhen Institutes of Advanced Technology (CAS), China

Reviewed by:
Deepika Mathur,
Icahn School of Medicine at Mount Sinai, United States
Piyush Agrawal,
National Cancer Institute, United States

***Correspondence:**
Chinae Thammarongtham
chinae@biotec.or.th

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: 25 February 2022
Accepted: 04 April 2022
Published: 28 April 2022

Citation:
Lertampaiporn S, Hongsthong A, Wattanapornprom W and Thammarongtham C (2022) Ensemble-AHTPpred: A Robust Ensemble Machine Learning Model Integrated With a New Composite Feature for Identifying Antihypertensive Peptides. Front. Genet. 13:883766. doi: 10.3389/fgene.2022.883766

Hypertension or elevated blood pressure is a serious medical condition that significantly increases the risks of cardiovascular disease, heart disease, diabetes, stroke, kidney disease, and other health problems, that affect people worldwide. Thus, hypertension is one of the major global causes of premature death. Regarding the prevention and treatment of hypertension with no or few side effects, antihypertensive peptides (AHTPs) obtained from natural sources might be useful as nutraceuticals. Therefore, the search for alternative/novel AHTPs in food or natural sources has received much attention, as AHTPs may be functional agents for human health. AHTPs have been observed in diverse organisms, although many of them remain underinvestigated. The identification of peptides with antihypertensive activity in the laboratory is time- and resource-consuming. Alternatively, computational methods based on robust machine learning can identify or screen potential AHTP candidates prior to experimental verification. In this paper, we propose Ensemble-AHTPpred, an ensemble machine learning algorithm composed of a random forest (RF), a support vector machine (SVM), and extreme gradient boosting (XGB), with the aim of integrating diverse heterogeneous algorithms to enhance the robustness of the final predictive model. The selected feature set includes various computed features, such as various physicochemical properties, amino acid compositions (AACs), transitions, n-grams, and secondary structure-related information; these features are able to learn more information in terms of analyzing or explaining the characteristics of the predicted peptide. In addition, the tool is integrated with a newly proposed composite feature (generated based on a logistic regression function) that combines various feature aspects to enable improved AHTP characterization. Our tool, Ensemble-AHTPpred, achieved an overall accuracy above 90% on independent test data. Additionally, the approach was applied to novel experimentally validated AHTPs, obtained from recent studies, which did not overlap with the training and test datasets, and the tool could precisely predict these AHTPs.

Keywords: antihypertensive, prediction, classification, ACE inhibitor, ACE inhibitory peptide, ensemble machine learning

INTRODUCTION

Hypertension is a global health issue due to its worldwide incidence and association with increased mortality and morbidity (Mills et al., 2020). Chronic hypertension is a substantial risk factor for heart diseases, stroke, cardiovascular diseases, congestive heart failure, glomerulonephritis, arteriosclerosis, and other diseases (Zhou et al., 2021).

The renin-angiotensin system (RAS) or the renin-angiotensin-aldosterone system (RAAS) is responsible for blood pressure regulation. The RAS regulates blood pressure and cardiac output by controlling the flow of blood through the heart (Wu et al., 2018).

One of the most important enzymes in the RAS system, angiotensin-converting enzyme (ACE), regulates blood pressure and fluid/salt homeostasis (He et al., 2014; Balgir and Sharma 2017). In the RAS, renin transforms angiotensinogen into angiotensin-I (ANG I), and subsequently, ACE transforms the inactive decapeptide angiotensin-I (ANG I) into the vasoconstrictor octapeptide angiotensin-II (ANG II). Excessive ACE activity results in the production of excessive amounts of angiotensin II and, as a result, an increase in blood pressure (i.e., it upregulates blood pressure) (Zhu et al., 2021).

ACE inhibition is a well-established technique for developing pharmaceuticals for the treatment of hypertension. Synthetic ACE inhibitors such as captopril, enalapril, cilazapril, benazepril, and lisinopril are typically used in clinical hypertension treatments (Daskaya-Dikmen, et al., 2017). However, the long-term treatment of hypertension with these drugs is accompanied by severe or mild adverse effects, such as cough, headache, diarrhea, dizziness, fatigue, angioedema, hyperkalemia, hypotension, or, in rare cases, renal impairment (De Leo et al., 2009; Nguyen et al., 2010; Norris and FitzGerald, 2013; Daskaya-Dikmen et al., 2017; Abachi et al., 2019; Festa et al., 2020).

Antihypertensive peptides (AHTPs) are bioactive peptides obtained from natural foods that have the effects/activities of ACE inhibitors against hypertension and are considered safe for consumption, with fewer adverse side effects than synthetic ACE inhibitor drugs or even no side effects. These natural ACE inhibitory bioactive peptides are highly desired for the development of functional foods, nutraceuticals and pharmaceuticals for the prevention and treatment of hypertension (Norris and FitzGerald, 2013; de Castro and Sato, 2015; Kumar et al., 2015; Abachi et al., 2019; Pujiastuti et al., 2019; Jiang et al., 2021; Zaky et al., 2022). Peptides are often multifunctional and may exhibit several health-promoting bioactivities, such as antioxidative, antihypertensive, anti-inflammatory, cytoprotective, and antimicrobial effects (He et al., 2019; Jakubczyk et al., 2020). Emerging evidence indicates that AHTPs may mediate antihypertensive effects by interacting with RAS-related renin, AT-II receptors, arginine–nitric oxide pathway, endothelin system, or Ca^{2+} channels in addition to ACE inhibition (Udenigwe and Mohan, 2014; Aluko 2015). AHTPs have major potential as functional ingredients (dietary compounds) in a daily diet aimed at helping prevent and safely manage hypertension and enhancing human health (Norris and FitzGerald, 2013; Jakubczyk et al., 2020). Therefore, the identification of new, nontoxic bioactive peptides derived from food or natural sources has received significant attention. As a consequence, an increasing number of food-derived antihypertensive peptides have been studied and reported (Martínez-Maqueda et al., 2012; Kumar et al., 2014; Abachi et al., 2019; Lee and Hur 2019; Pujiastuti et al., 2019; Lu et al., 2021). Finding new AHTPs in various organisms is currently a significant research topic. However, large-scale identification through wet laboratory experiments is a costly, time consuming, and labor-intensive approach (Li-Chan 2015; Pujiastuti et al., 2019; Festa et al., 2020). The use of bioinformatics and *in silico* methods for the identification of potential candidate AHTPs for subsequent experimental assays is necessary to shorten the process. The development of efficient computational approaches will facilitate the processes of discovery and screening, allowing potential novel AHTP candidates to be identified in a cost-, resource- and time-effective manner.

A few existing machine learning-based computational approaches are available for predicting AHTPs. mAHTPred is a meta-predictor that employs a two-step feature selection methodology (Manavalan et al., 2019). PAAP is an RF classification model approach based on varied combinations of amino acids, dipeptides, and pseudo amino acid composition descriptors (Win et al., 2018). AHTpin was developed to screen, predict, and design AHTPs by using an SVM-based regression model for tiny peptides and SVM-based classification models for small, medium and large peptides (Kumar et al., 2015). Additionally, an SVM prediction tool was recently built by using convolutional neural network (CNN) deep learning-based encoding features derived from amino acid compositions (AACs) and dipeptide composition features (Rauf et al., 2021).

Although certain tools for AHTP prediction are available, the development of our ensemble method is different from that of the existing approaches in several ways. First, we developed a weighted voting method for integrating the strengths of three independent machine learning models, each of which has high levels of performance in different aspects. Second, a new composite feature called comF2 was developed based on a logistic regression statistical framework. In both the RF and extreme gradient boosting (XGB) feature importance plots, this feature was ranked as the most significant. In addition, a Shapley additive explanations (SHAP) analysis revealed consistent results, showing that comF2 was the top-ranked feature and was capable of explaining large samples in the model; therefore, it could capture characteristics for most of the AHTPs in the training data. Third, our ensemble method

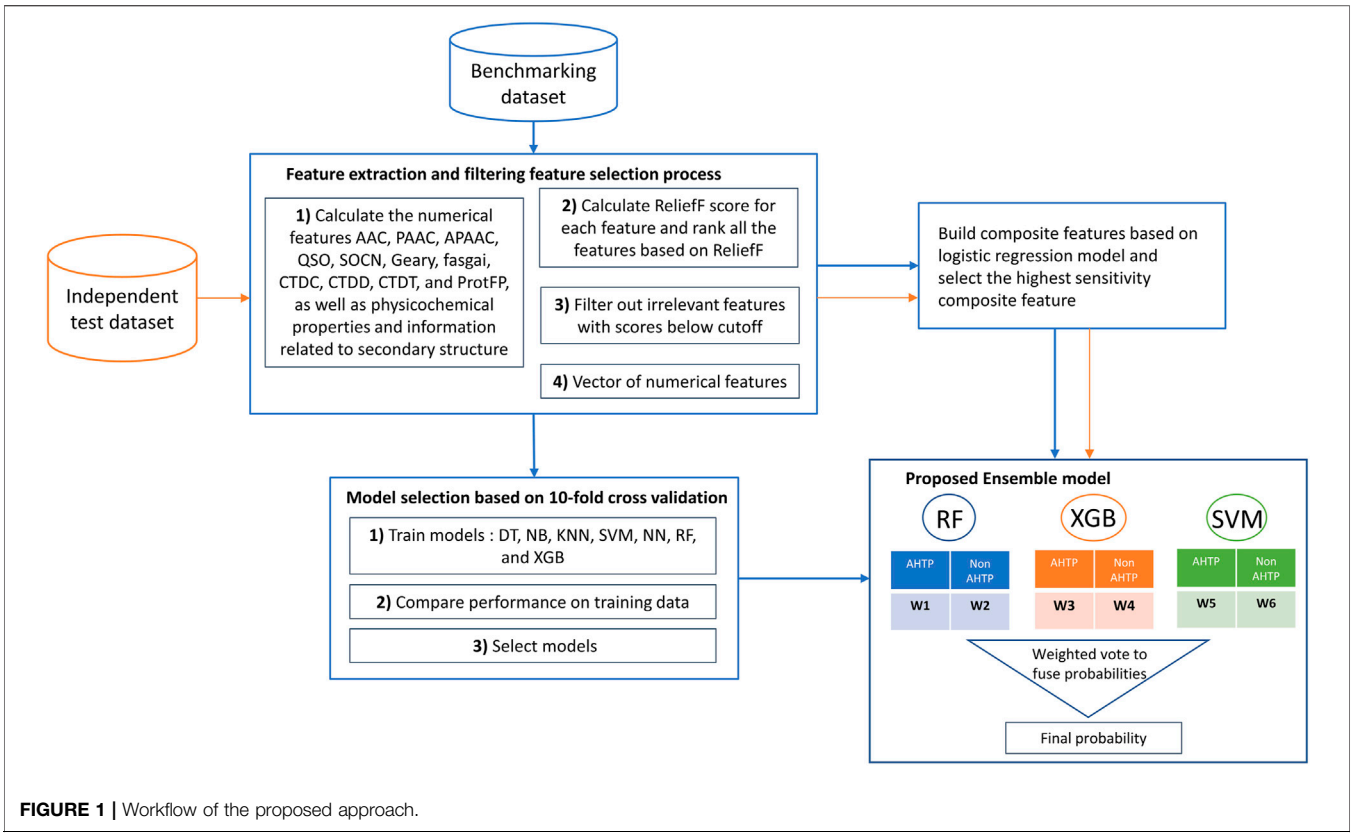

FIGURE 1 | Workflow of the proposed approach.

outperformed previously developed methods in terms of robustness and accuracy when predicting independent testing datasets, with an enhanced accuracy of 90.4%. Last, the technique could also correctly classify many novel unseen, and experimental AHTPs collected from recent studies.

MATERIALS AND METHODS

Workflow

The workflow of Ensemble-AHTPpred is shown in **Figure 1**.

Datasets

In this study, we employed two nonredundant datasets from mAHTPred (Manavalan et al., 2019): a benchmarking dataset and an independent testing dataset. The balanced benchmarking dataset contained 913 unique AHTPs and 913 unique non-AHTPs. The 913 AHTPs were experimentally validated on the publicly available AHTPDB (Kumar et al., 2015) and BIOPEP (Minkiewicz et al., 2008; Iwaniak et al., 2016) databases. Note that experimentally validated non-AHTPs were not available as a public non-AHTP database. Therefore, the non-AHTPs were random peptides generated from Swiss-Prot proteins. Considering random sequences as a negative dataset is a routinely used standard procedure in many peptide-based prediction methods (Sharma et al., 2013; Kumar et al., 2015; Chen et al., 2016; Usmani et al., 2018; Manavalan et al., 2019) with the assumption that the probability of finding a random sequence to be positive is very low. Positive and negative training datasets have similar length distributions. The AHTPs in the benchmarking dataset have a length between 5 and 81 amino acids, with an average length of 7.7 amino acids. The non-AHTPs in the benchmarking dataset have a length between 5 and 45, with an average length of eight amino acids.

Another dataset, an independent dataset, was composed of 386 nonredundant, experimentally validated AHTPs (Win et al., 2018; Yi et al., 2018) and 386 random peptides generated from Swiss-Prot as negative samples. The AHTPs in the independent testing dataset have a length between 5 and 24 amino acids, with an average length of 6.48 amino acids. The non-AHTPs in the independent testing dataset have a length between 5 and 29, with an average length of 15.42 amino acids.

Features

The peptide properties that were relevant for predicting AHTPs were determined and encoded as a vector of 431 numerical features. The features can be grouped into seven main types as follows.

1) AAC descriptors: These descriptors were used as the fractions of each amino acid type within a protein sequence. The fractions of all 20 natural amino acids {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}, were calculated. (**AAC1-AAC20: 20 dimensions**).
2) Chou's pseudo amino acid composition (PseAAC) was generated in various modes: Chou's PseAAC (Chou, 2005)

has been widely used to convert complicated protein sequences with various lengths to fixed-length numerical feature vectors that incorporate sequence-order information. In comparison with an AAC, a PseAAC is more informative and capable of representing a protein sequence and incorporating information about its sequence order. Hence, it has been widely used for diverse amino acid sequence-based prediction problems (Chou, 2011). The PseAACs were calculated by using parameters of lambda = 3 and weight = 0.05 (**PAAC1-PAAC23: 23 dimensions**). PseAACs in parallel correlations (**Pse_PC1-Pse_PC22: 22 dimensions**), PseAACs in series correlations (**Pse_SC1-Pse_SC26: 26 dimensions**), and amphiphilic pseudo AACs with hydrophobicity correlation functions (**APAAC1_1-APAAC1_23: 23 dimensions**) and hydrophilicity correlation functions (**APAAC2_1-APAAC2_23: 23 dimensions**) were also calculated.

3) Composition/transition/distribution (C/T/D): The three descriptors based on the grouped AACs (Dubchak et al., 1995) [composition (**CTDC1-CTDC21: 21 dimensions**), transition (**CTDT1-CTDT21: 21 dimensions**) and distribution (**CTDD1-CTDD105: 105 dimensions**) descriptors] were calculated. C/T/D was calculated using the protr R package (Xiao et al., 2015). All amino acid residues were divided into three groups according to seven types of physicochemical properties, as defined in Dubchak et al. (1999). The seven physicochemical properties used for calculating these features were hydrophobicity, normalized van der Waals volume, polarity, polarizability, charge, secondary structures, and solvent accessibility.
4) Quasi-sequence-order descriptors: The quasi-sequence-order descriptors were derived from the distance matrix of the 20 amino acids (Chou 2000). Quasi-sequence-order descriptors (**QSO1-QSO46: 46 dimensions**) and sequence-order-coupling numbers (**SOCN1-SOCN6: 6 dimensions**) (lag = 3, w = 0.1) were calculated.
5) Various physicochemical and topological property-based features: The Crucian properties covariance index (**Crucian1-Crucian3: 3 dimensions**) (Cruciani et al., 2004), Z-scales based on physicochemical properties (**zscales1-zscales5: 5 dimensions**) (Sandberg et al., 1998), factor analysis scales of generalized amino acid information (**fasgai1-fasgai6: 6 dimensions**) (Liang and Li 2007), T-scales based on physicochemical properties (**tScales1-tScales5: 5 dimensions**) (Tian et al., 2007), VHSE-scales (principal component score vectors of hydrophobic, steric, and electronic properties) (**vhsescales1-vhsescales8: 8 dimensions**) (Mei et al., 2005), protFPs (**protFP1-protFP8: 8 dimensions**) (van Westen et al., 2013), ST-scales based on physicochemical properties (**stscales1-stscales8: 8 dimensions**) (Yang et al., 2010), MS-WHIM scores (**mswhimscore1-mswhimscore3: 3 dimensions**) (Zaliani and Gancia 1999), aliphatic indices of proteins (**aIndex: 1 dimension**) (Ikai, 1980), Geary autocorrelations (**geary1-geary12: 12 dimensions**), the autocovariance index (**autocov: 1 dimensions**) (Ikai, 1980), the potential protein interaction index (**Boman: 1 dimension**) (Boman, 2003), the net charge (**Charge: 1 dimension**), cross-covariance indices (**Crosscov1-Crosscov2: 2 dimensions**), instability indices (**Instaindex: 1 dimension**) (Guruprasad et al., 1990), the hmoment alpha helix (**Hmoment1: 1 dimensions**), the hmoment beta sheet (**Hmoment2: 1 dimensions**), BLOSUM matrix-derived descriptors (**Blosum1-8: 8 dimensions**), and the isoelectric point (**pI: 1 dimension**) were calculated by using the peptide R package (Osorio et al., 2015).
6) Occurrence of selected k-mer motifs: The YP, HLP, IYP, LHL, LPP, LRP, VPP, PEV, PFP, QTP, VLP, VYP, and YPF motifs (**13 dimensions**) were determined. First, we generated all 2-mers (400 dimensions) and all 3-mers (8000 dimensions). Then, we searched for the k-mer that was overrepresented in the positive and underrepresented in the negative datasets by calculating the log odds ratio score of the frequency of each k-mers in the positive versus negative datasets. Next, we ranked the discriminant k-mers based on the calculated log-odds score. Finally, we retained the top 2-mer and the top 12 3-mers as selected k-mer motif features that still need to be determined (the heatmap of log odds scores of 2-mers is shown in **Figure 5**).
7) Secondary structure conformation-related features: The aggregation, amyloid, turn, alpha-helix, helical aggregation, and beta-strand conformation secondary structure propensities were calculated using the Tango program (**tango1-tango6: 6 dimensions**) (Fernandez-Escamilla et al., 2004).

To further improve the prediction process with new informative features, we proposed a composite feature generation method *via* the fusion of the various selected features by using a logistic regression model. Various composite features based on various combinations of informative selected features were built by using logistic regression based on the benchmarking data and then compared through a 10-fold cross-validation process. The detailed process of building composite features is described in the hybrid feature section of ensemble-AMPPred (Lertampaiporn et al., 2021). A combination of features was used to fit a logistic regression model, which is represented by the following equation:

$$Prob.\ (\mathrm{Y} = AHTPs|x) = logistic(\mathrm{x}) = \left(e^{\beta 0+\beta 1\ \mathrm{X1}+\beta 2\mathrm{X2}+\beta 3\mathrm{X3}+\cdots+\beta \mathrm{nXn}} / \left(1 + e^{\beta 0+\beta 1\mathrm{X1}+\beta 2\mathrm{X2}+\beta 3\mathrm{X3}+\cdots+\beta \mathrm{nXn}}\right)\right)$$

Logit transformation (the logarithm of the odds ratio that Y is in the AHTP category) was applied to link a function with the logistic regression. The logit function is defined as

$$\mathrm{Logit}(\mathrm{x}) = \log\left(\frac{\mathrm{P}(Y = AHTP|X = X)}{P(Y = nonAHTP|X = X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_\mathrm{n} X_\mathrm{n}$$

Therefore, the composite feature was defined by the following equation:

$$\text{Composite feature} = \beta 0 + \beta 1\ \text{feature\_1} + \beta 2\text{feature\_2} + \beta 3\text{feature\_3} + \ldots + \beta n\ \text{feature\_n}$$

where β0 is the intercept; β1, β2, β3, and βn represent the regression coefficients for each selected feature in the equation; and feature_1, feature_2, . . ., and feature_n are the component features in the composite feature.

Feature Selection

A feature selection procedure based on ReliefF (Kononenko, 1994) scores was used as a preprocessing step to filter irrelevant features with a cutoff score. The ReliefF score for a feature was calculated based on how well the feature could distinguish between instances that were near each other. The ReliefF evaluation criterion selected features that aided in the separation of the samples from different classes and gave higher weights to the features that discriminated the samples from the neighborhoods of different classes.

Recursive feature elimination (RFE) (Tolosi and Lengauer, 2011) is a wrapper-type feature selection algorithm. RFE starts with all features in the training dataset and then searches for a subset of features by removing features through recursive elimination to eliminate the least relevant features one by one and refitting the model. This process is repeated until the optimal number of features is reached, ensuring that the classifier can achieve high performance.

Models

To select base classifiers for constructing an ensemble, seven machine learning algorithms were considered in our algorithm selection experiment—a naïve Bayes (NB) model, a neural network (NN), a support vector machine (SVM), k-nearest neighbors (kNN), a decision tree (DT), a random forest (RF) and an extreme gradient boost (XGB). Each algorithm has a different inductive bias and different learning hypotheses that can provide a potentially more independent and diverse set of predictions through the ensemble. For the hyperparameters, we used a grid search to find the optimal parameters.

The NB classifier is a simple probabilistic classifier based on Bayes' theorem and substantial independence assumptions between the features.

The NN was a multilayer perceptron (MLP). An MLP is a neural network with at least three layers: an input layer, a hidden layer, and an output layer (parameters: number of epochs: 500; learning rate: 0.3; and momentum for updating weights: 0.2).

The SVM model is a supervised learning model with associated learning algorithms for data classification and regression analysis. The SVM assigns training examples to coordinates in a high-dimensional space to widen the distance between the two classes and separates the two classes with a simple hyperplane (parameters: C = 36.0; kernel = 'Radial Basis Function'; and gamma = 0.119).

The KNN method is a well-known nonparametric technique used in statistical pattern classification due to its simplicity, intuitiveness, and effectiveness. The essential principle is that an unclassified object is assigned to the class to which the majority of its k nearest neighbors belong (parameters: k = 7 and distance = inverse weight).

The DT is another nonparametric supervised learning method used for classification and regression. It develops a model that accurately predicts the value of a target variable by inferring basic decision rules from data attributes. A tree can be thought of as an approximation to a piecewise constant (parameter: confidence factor = 0.25).

The RF algorithm is one of the most commonly used bagging ensemble algorithms because of its flexibility and ease of use. This algorithm can produce good results without hyperparameter tuning. The RF approach is an ensemble technique with the ability to achieve high accuracy and prevent overfitting by making use of voting with multiple decision trees (parameters: no. estimators = 350 and max_depth = 12).

The XGB algorithm is a gradient boosting ensemble algorithm. The boosting algorithm adjusts the model weights according to a differential loss function and then uses the adjusted weights in the next training iteration [parameters: no. estimators (nrounds) = 800; max_depth = 10; eta = 0.01; and subsample = 0.8].

The proposed method was implemented by using Perl, Python, and R scripts. The program was run on a Fedora Linux-based machine. All the data, the trained models and the standalone program are available to download at http://ncrna-pred.com/Ensemble_AHTPpred.htm.

We adopted 10-fold cross-validation to investigate the classification performance of the various models on the benchmarking dataset. Based on the 10-fold cross validation results, model selection processes were performed. Then, the best-performing models were selected based on their diverse measurements and later used as the base classifiers of the ensemble model. Thereafter, the individual base classifiers were iteratively trained to find the optimal weight for each class of each classifier. The probability weight set (w_1, w_2, w_3, w_4, w_5, w_6) was estimated by using the level of confidence in predicting each class (AHTP or non-AHTP), which fluctuated among the classes. The probabilities acquired from the base classifiers were aggregated through weighted voting to obtain the final prediction of the ensemble model.

Probability-weighted voting = (W_1*Prob. (RF $_{\text{class=AHTP}}$)) + (W_2*Prob. (RF $_{\text{class=non-AHTP}}$)) + (W_3*Prob. (XGB class = $_{\text{AHTP}}$)) + (W_4*Prob. (XGB $_{\text{class=non-AHTP}}$)) + (W_5*Prob (SVM $_{\text{class=AHTP}}$)) + (W_6*Prob (SVM $_{\text{class=non-AHTP}}$)).

To evaluate the classification performance of the model, the following metrics were used:

$$\text{ACC} = \frac{TP + TN}{(TP + TN + FP + FN)}$$

$$\text{Sn} = \frac{TP}{(TP + FN)}$$

$$\text{Sp} = \frac{TN}{(TN + FP)}$$

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

where ACC, Sn, Sp, and MCC are accuracy, sensitivity, specificity, and Matthew's coefficient correlation, respectively. These measurements were calculated based on the numbers of true

TABLE 1 | Physicochemical property-based composition of amino acids.

Physicochemical property-based composition of amino acids	Positive dataset (AHTPs)	Negative dataset (non-AHTPs)
Molecular weight of the peptide (Da)	888.2	**912.5**
Number of amino acids in the sequence	7.75	**8.05**
% Composition of charged residues (DEKHR)	19.91	**24.94**
% Composition of aliphatic residues (ILV)	**22.34**	22.04
% Composition of aromatic residues (FHWY)	**14.42**	10.32
% Composition of polar residues (DERKQN)	25.81	**31.49**
% Composition of neutral residues (AGHPSTY)	**43.44**	37.68
% Composition of hydrophobic residues (CVLIMFW)	30.75	**30.83**
% Composition of positively charged residues (HKR)	12.75	**12.96**
% Composition of negatively charged residues (DE)	7.16	**11.98**
% Composition of tiny residues (ACDGST)	22.65	**34.97**
% Composition of small residues (EHILKMNPQV)	**61.94**	51
% Composition of large residues (FRWY)	**15.41**	14.03

The higher values, between the two datasets, are shown in bold.

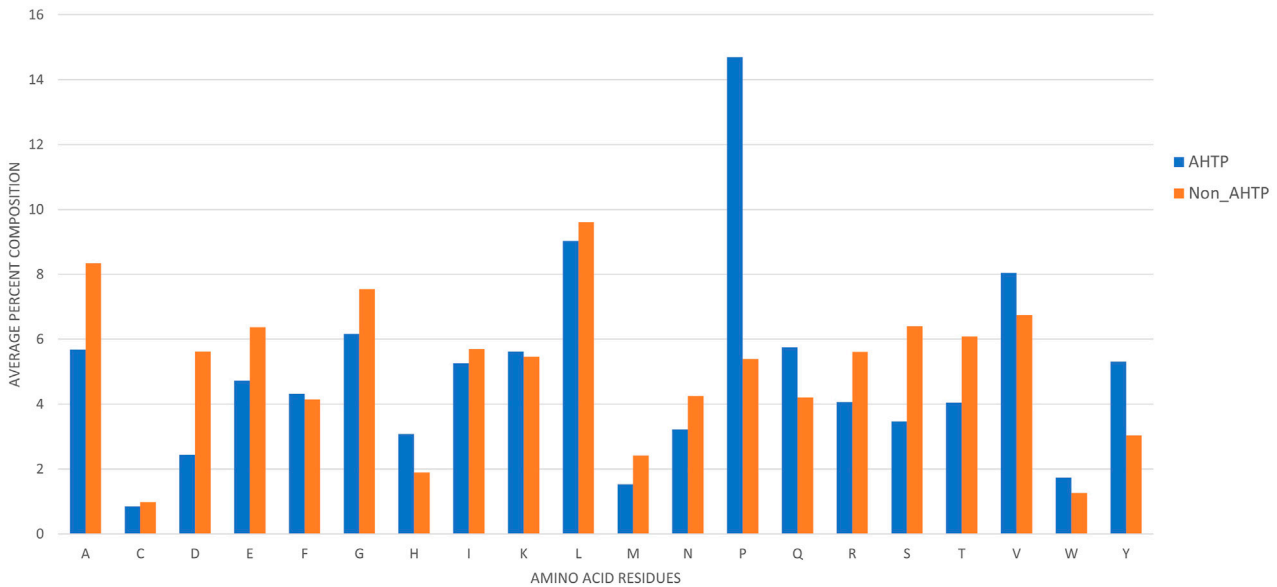

FIGURE 2 | Percent average composition of amino acid residues present in the positive and negative datasets.

positives (TPs), true negatives (TNs), false positives (FPs) and false negatives (FNs). The area under the receiver operating characteristic (ROC) curve (AUC) was calculated to assess the tradeoff between the sensitivity and specificity performance of the different methods. The ROC curve is a plot of the TP vs. FP rates at different thresholds. For a perfect predictor, the AUC is equal to 1.

RESULTS AND DISCUSSION

Amino Acid Composition and Positional Residue Analysis

The activity of peptides depends on their structure and amino acid composition. To understand the relation between the composition and antihypertensive function of a peptide, the composition of AHTPs and non-AHTPs were analyzed/investigated. Generally, most antihypertensive peptides are relatively short peptide residues with lengths that vary from 2 amino acids to 20 amino acids. The amino acid composition is a quantitative measure of the fraction of each amino acid type within a protein. The percent amino acid composition based on the physicochemical properties of amino acids (whole peptides) was computed and calculated using COPid (Kumar et al., 2008) and includes the composition of charged (DEKHR), aliphatic (ILV), aromatic (FHWY), polar (DERKQN), neutral (AGHPSTY), hydrophobic (CVLIMFW), positively charged (HKR), negatively charged (DE), tiny (ACDGST), small (EHILKMNPQV) and large (FRWY) residues, as summarized in **Table 1** (a category with higher composition is shown in bold).

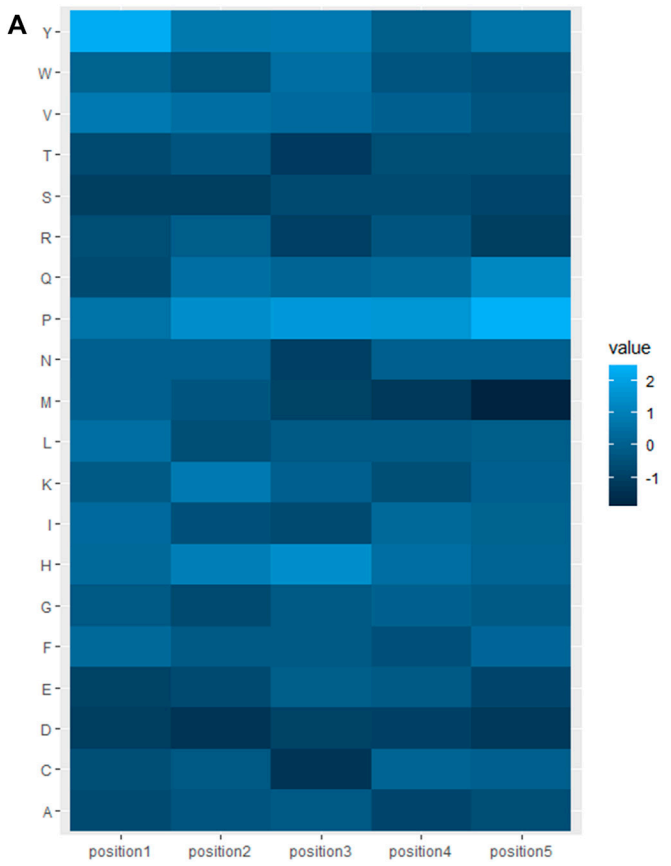

FIGURE 3 | N-terminal features of AHTP positive data and non-AHTP negative data: **(A)** Heatmap of log odds ratios, where a lighter color denotes overrepresented amino acid residues in AHTPs compared to non-AHTPs (positive log odds score) and, a darker color denotes underrepresented amino acid residues in AHTPs compared to non-AHTPs (negative log odds score). **(B)** Sequence logos of positions one to five of the AHTP positive dataset. **(C)** Sequence logos of positions one to five of the non-AHTP negative dataset.

When comparing positive and negative of benchmarking datasets, we can see that AHTPs include more aliphatic (ILV), aromatic (FHWY), and neutral (AGHPSTY) amino acid residues than non-AHTP sequences.

Amino acid residues present in AHTPs and non-AHTPs were compared, as shown in **Figure 2**. Histidine (H), proline (P), glutamine (Q), valine (V), tryptophan (W) and tyrosine (Y) more frequently occurred in AHTPs than in non-AHTPs, especially proline (P), which is highly abundant in AHTPs. In contrast, certain residues such as cysteine (C), aspartic acid (D), methionine (M), and tryptophan (W) occurred rarely in AHTPs. Certain types of residues occured frequently in both AHTPs and non-AHTPs, such as leucine (L) and valine (V). Amino acids such as alanine (A), aspartic Acid (D), and serine (S) were less frequent in AHTPs than in non-AHTPs.

C-terminal and N-terminal positional residue analysis was also performed by calculating the average amino acid composition of position one to position five of the N- and C-termini in AHTPs (positive) and non-AHTPs (negative). The log odds ratios between positive and negative N- and C-termini were calculated. The log-odds ratios of positive versus negative termini were calculated as [log2 (P_a/N_a)], where P_a and N_a are the observed frequencies of amino acid a in the positive and negative training datasets, respectively. Heatmaps of log odds ratios were plotted for the N-terminal and C-terminal regions, as shown in **Figures 3A**, **4A**. The sequence logos of positions one to five of the N- or C-terminus were generated by using Seq2Logo (Thomsen and Nielsen, 2012). **Figures 3B,C** display N-terminal positional sequence logos of AHTPs and non-AHTPs, respectively. (In sequence logos, specific colors were assigned to amino acids as follows, purple represents nonpolar sidechains (G A V L I M F W P), blue represents basic amino acid (K R H), Red represents acidic amino acid (D E), and green represents polar sidechains (S T C Y N Q); the height of the amino acids is proportional to their frequency at that position.) The most abundant amino acids in the N-terminus of AHTPs were Leu (9.069%), Pro (14.896%), Tyr (5.214%) and Val (8.697%). The most abundant amino acids in the C-terminus of AHTPs were Leu (9.003%), Pro (16.605%), and Val (7.338%). The most abundant amino acids in the N- and C-termini of non-AHTPs were Leu, Ala, Gly, and Val. The most abundant 2-mers in the N-terminus of AHTPs were YP, LP, PF, PP, and VP, while the most abundant 2-mers in the C-terminus of AHTPs were IP,

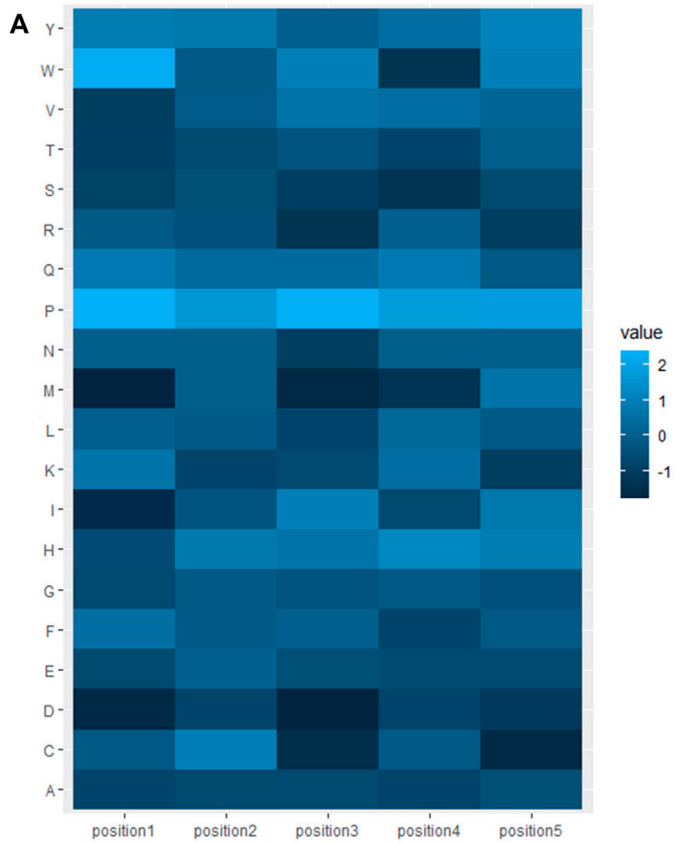

FIGURE 4 | C-terminal analysis of AHTP positive data and non-AHTP negative data: **(A)** Heatmap of log odds ratios, where a lighter color denotes overrepresented amino acid residues in AHTPs compared to non-AHTPs (positive log odds score) and a darker color denotes underrepresented amino acid residues in AHTPs compared to non-AHTPs (negative log odds score). **(B)** Sequence logos of positions one to five of the AHTP positive dataset. **(C)** Sequence logos of C-terminal positions one to five of the non-AHTP negative dataset.

FP, PL, PP, PV, QP and VP. The most abundant 2-mers in the N- and C-termini of non-AHTPs were AA, LA, AL, LG, LE, and AR.

In addition, a heatmap of the log odds score of occurrences of the 2-mer motif in the whole sequence of AHTPs vs. in the whole sequence of non-AHTPs was also plotted, as shown in **Figure 5**. TyrPro (log odds = 4.393), ProPhe (log odds = 3.896) and ProHis (log odds = 3.340) were overrepresented in AHTP positive data compared to non-AHTP negative data. In contrast, AspSer (log odds = −5.708), MetThr (log odds = −4.292) and CysLeu (log odds = −4.070) were overrepresented 2-mers in non-AHTP negative data relative to AHTP positive data.

Performance Evaluation Based on the Benchmarking Dataset to Select the Base Models for the Ensemble

Before training a prediction model, feature extraction and feature selection are two important steps for extracting various numerical features to represent biological sequences and then selecting relevant and discriminative features so that a machine learning model can further analyze and detect the generalized pattern of the data of interest. In this work, we extracted a total of 431 numerical features to represent peptide sequences.

Since we collected as many features that could explain the peptides as possible, these 431 extracted features may have contained irrelevant and noninformative features with respect to explaining the AHTPs. Feature selection is required to eliminate irrelevant and redundant features that do not explain the target class. Furthermore, feature selection mitigates the curse of dimensionality (by reducing the number of dimensions) and prevents overfitting. Filter, wrapper, and embedding techniques are the three primary feature selection methods. Both the wrapper and embedding methods are tightly coupled with specific classification algorithms. The wrapper requires one predetermined classification algorithm and relies on its performance to evaluate and select the feature subset. This approach seeks the features that are best suited to the predetermined algorithm. As a result, these methods first necessitated determining the classification algorithm to be used. However, we intended to create an ensemble consisting of multiple

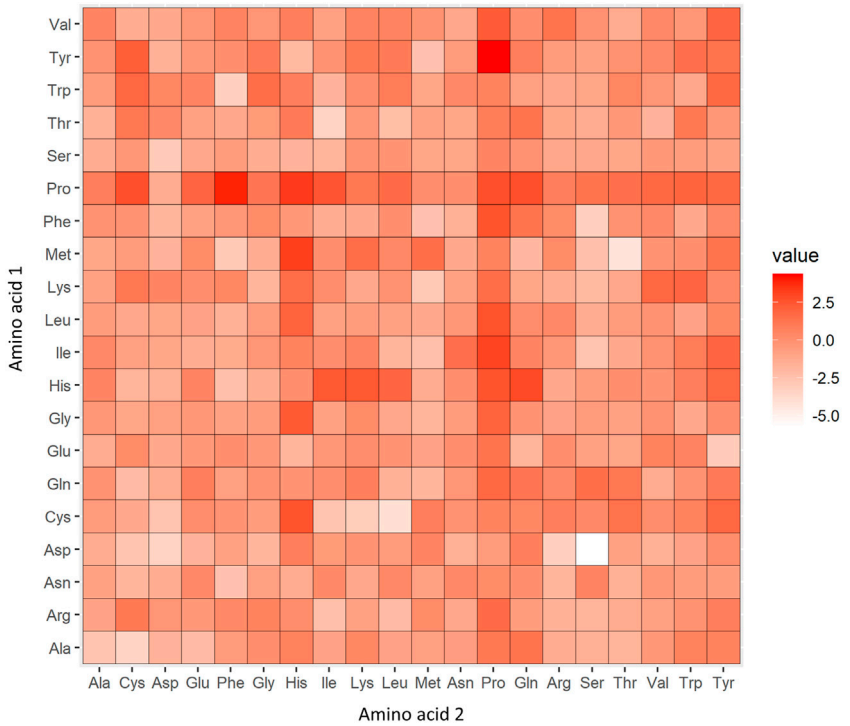

FIGURE 5 | Heatmap of the log odds scores of 2-mers abundant in the positive versus negative datasets. In the heatmap, a red color (high log odds score) denotes 2-mers overrepresented in AHTPs compared to non-AHTPs, and a white color (low log odds score) denotes 2-mers underrepresented in AHTPs compared to non-AHTPs.

TABLE 2 | Classification performance of different trained models.

	DT	NB	KNN	NN	SVM	XGB	RF
ACC (%)	73.494%	74.465%	74.918%	76.177%	80.504%	78.925%	**80.668%**
Sn	0.714	0.696	0.690	0.721	0.758	**0.789**	0.752
Sp	0.756	0.814	0.808	0.803	0.852	0.791	**0.861**
AUC	0.766	0.793	0.791	0.831	**0.878**	0.861	0.877

The highest values are in bold.

classification algorithms. Therefore, the filtering procedure was used initially to remove irrelevant features during this step. Note that the filtering method may not eliminate redundant features. We applied the filtering method based on ReliefF scores. After applying the filtering method, a total of 379 features had scores that were higher than the cutoff score. The vector containing these 379 numerical features was then used to train the 7 algorithms.

The training process was carried out *via* 10-fold cross-validation on a benchmarking dataset to investigate the classification performance of different trained models. **Table 2** shows the performance of the individual trained models. Different algorithms were able to take advantage of different characteristics and relationships contained in a given dataset. In this process, we detected and combined the strengths of distinct algorithms to form a resilient and stable ensemble. The findings support the "no free lunch" theorem, which states that there is no single best algorithm that is superior in terms of every metric. The ROC curves of individual classification model performance are plotted in **Figure 6**.

Based on the performance obtained during the training process, **Table 2** shows that XGB had the highest sensitivity (0.789), followed by the SVM (0.758). The AUC provides a measure for evaluating which models are better on average by weighing the tradeoff between sensitivity and specificity. For the AUC metric, the SVM model achieved the highest score of 0.878, followed by the RF model (0.877), indicating that these two models achieved a good balance between positive and negative prediction. The RF model had the highest classification accuracy of 80.668% among the seven

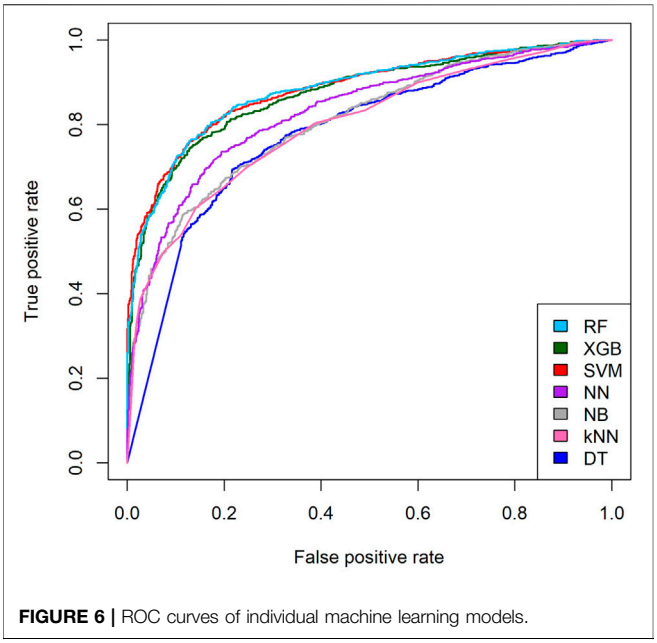

FIGURE 6 | ROC curves of individual machine learning models.

trained models. Accordingly, based on the evaluation, we chose the SVM, the RF, and XGB as the ensemble members because of their superior performance in terms of different metrics.

Note that the input vectors for the SVM model were drawn from a separate collection of features. Because the RF and XGB have built-in feature selection, we used the complete 379-feature vector as the input feed. However, for the SVM-based model, we used RFE as an additional wrapper feature selection step to remove redundant features and reduce the computational time and memory. As a result, the feature subset used as the input vector for the SVM model was reduced from 379 to 256 attributes.

Each model was assigned a weight, which was proportional to the model classification accuracy across all classes. In addition, the capacities for classification and prediction on different classes may have been unequal. Therefore, the classifier with the highest prediction confidence was given greater weight for that class. Subsequently, the training process was conducted *via* 10-fold cross validation to find the optimal class weights for each classifier/predictor in the ensemble. Thereafter, the individual classifiers (SVM, RF, and XGB) were aggregated through weighted voting to obtain the final probability and prediction.

The New Composite Feature is Significant for Improving the Sensitivity of the Method

We propose utilizing a logistic regression equation to create additional composite features, based on the fusion of two or more existing features. In contrast to sophisticated black-box classification models, regression is a powerful way to determine the unique relationships between a large number of features and a target class. In this work, we created a number of composite features and selected the two with the highest sensitivity, which we refer to as comF and comF2. These features were merged into the feature vector as the input of the ensemble model.

The comF feature is defined as

$$comF = 0.8634 - 0.157tscales4 - 0.154CTDC19 - 0.135protFP6 + 0.133CTDC21 - 0.132fasgai4 + 0.122mswhimscore1 - 0.12hydrophobicity$$

The comF2 feature is defined as

$$comF2 = 0.1786 + 0.1522APAAC1\_15 - 2.2951CTDC10 - 0.6069CTDC19 - 0.0065CTDD49 + 0.2176QSO19 + 0.9747fasgai4 + 0.3691ProtFP3 + 2.0823Pse\_PC13$$

where APAAC1_15 denotes the amphiphilic PseAAC of amino acid R (the sequence-order coupling mode was used along a protein sequence *via* a hydrophobicity correlation function; the hydrophobic properties of amino acids were taken into account) and CTDC10 denotes the percentage of a particular amino acid in the polarizability group 1 (polarization between 0 and 1.08: amino acids G, A, S, D, and T) relative to protein length. CTDC19 is the percentage of a particular amino acid in solvent access group 1 (buried: amino acids A, L, F, C, G, I, V, and W) relative to the protein length. CTDD49 is the percentage of a particular amino acid in polarization group 1 (polarization between 0 and 1.08: amino acids G, A, S, D, and T) located in 75% of the residues of the protein chain. QSO19 is the quasi-sequence order of the normalized occurrence of amino acid Y, fasgai4 is a descriptor that reflects compositional characteristics, ProtFP3 is the scales-based descriptor derived from the amino acid properties of all AA indices (protein fingerprint 3), and Pse_PC13 is the parallel correlation PseAAC of amino acid P.

Interestingly, we discovered some intriguing aspects within the comF2 composite feature. Particular component properties of comF2, such as the distant locations of certain amino acids Y, R, and P, had beneficial impacts on the equation; this is consistent with the results of many research papers demonstrating that certain residues are dominant in the C-termini or N-termini of potent AHTPs. Hydrophobic residues with aliphatic side chains at the C-terminus promoted ACE inhibitory activity (Nimalaratne et al., 2015; Asoodeh et al., 2016; Jiang et al., 2021; Wang et al., 2021). Other studies have demonstrated that the positively charged lysine and arginine amino acids (K and R) contribute to the strong potency of ACE inhibitory peptides (Wei et al., 2019; Maky and Zendo, 2021). The richness of proline (P) and its number of occurrences in a sequence positively influenced the potency of ACE inhibition (Abachi et al., 2019; Festa et al., 2020; Pavlicevic et al., 2020). The presence of a polar amino acid at the C-terminus along with hydrophobic amino acids at the N-terminus may have contributed to the activity (Ryan et al., 2011; Udenigwe et al., 2012; de Castro and Sato, 2015). Moreover, the equation was adversely affected (according to the minus sign) by component

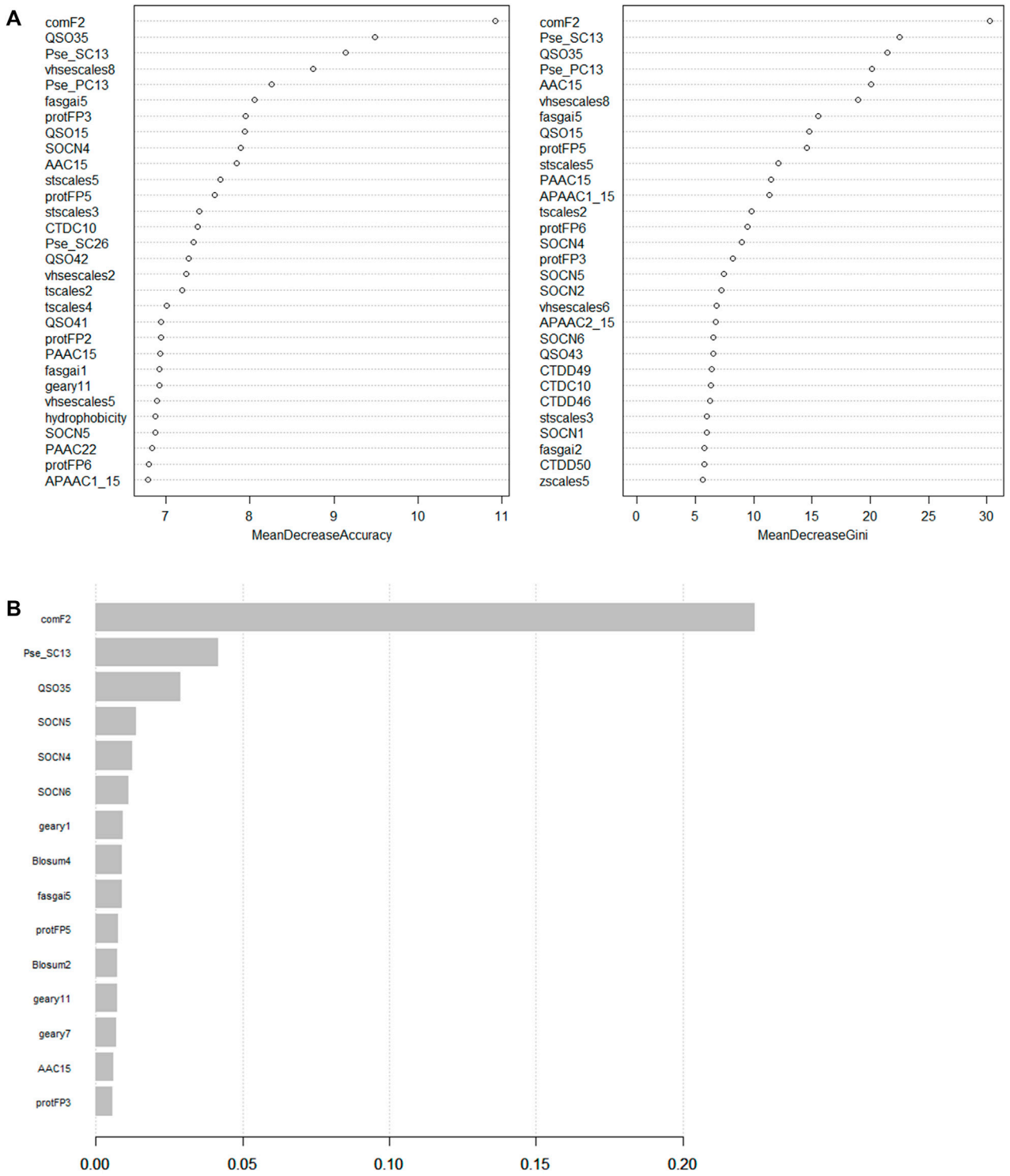

FIGURE 7 | Importance plots and SHAP plot: **(A)** Importance plots yield by the RF (left: permutation importance; right: Gini importance). **(B)** Importance plot yielded by the XGB model. **(C)** SHAP summary plot of the top 15 features; **(D)** dependence plot of composite feature comF2 for the AHTP class. **(E)** Density distribution of the SHAP plot's top six features (sample) by class in the training data.

properties involving low-polarization amino acids (CTDC10 and CTDD49) and those with restricted solvent access (CTDC19; buried structure).

Because the RF and XGB have built-in feature importance analysis mechanisms, we discovered that the composite feature comF2 was the highest-ranking feature in both models based on

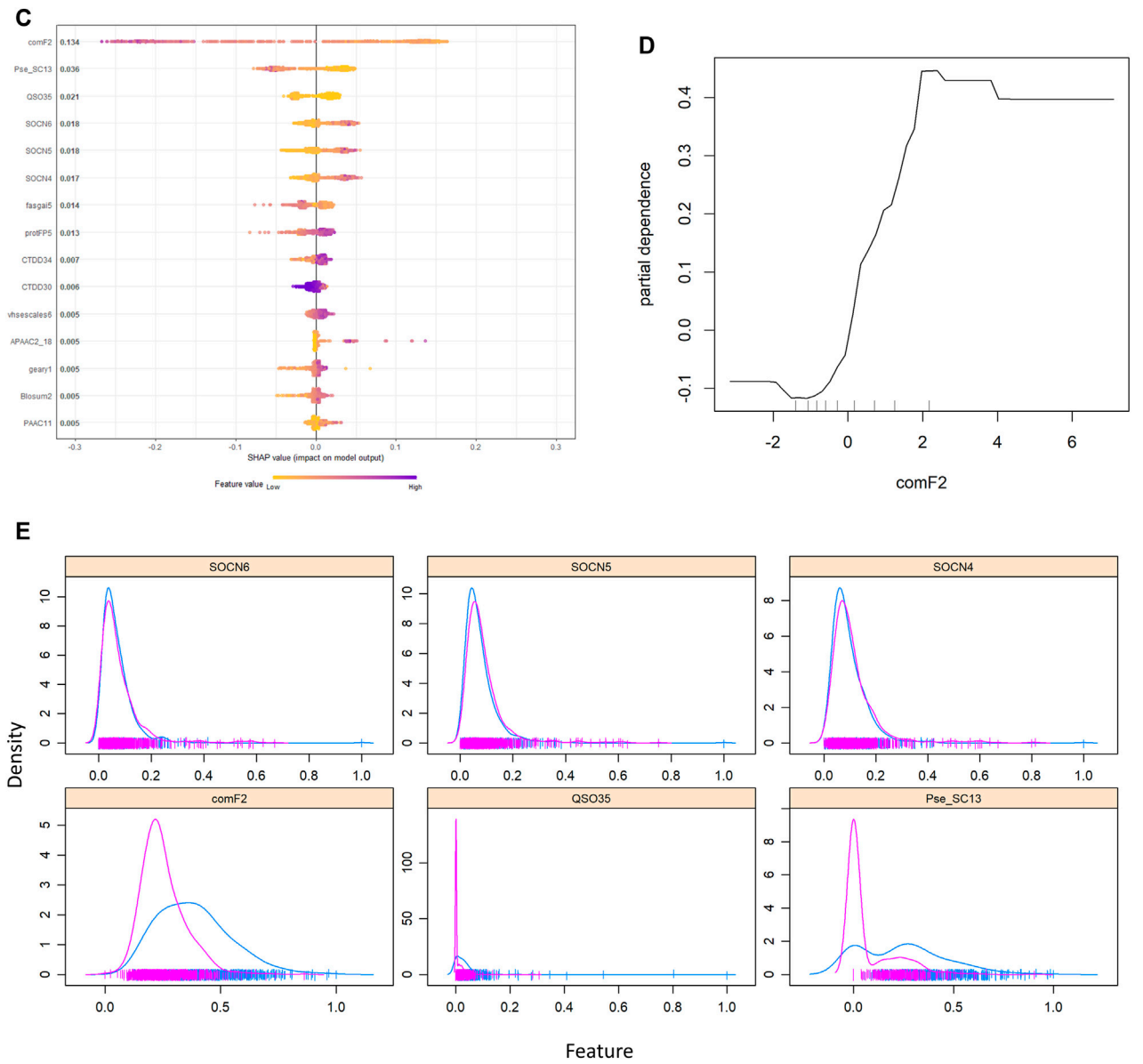

FIGURE 7 | (Continued).

their importance plots (as shown in **Figures 7A,B**). It is well known that the value of a feature (as measured by information gain) varies depending on how frequently it is employed at the leaf nodes. We also conducted SHAP (Shapley Additive exPlanations) analysis as a follow-up to our initial investigation. SHAP is a game-theoretic framework for

TABLE 3 | Performance evaluation of the proposed method using benchmarking dataset.

Method	ACC	Sn	Sp	MCC	AUC
CNN + SVM (Rauf et al., 2021)	0.958	0.996	0.920	0.920	0.958
mAHTPred (Manavalan et al., 2019)	0.848	0.821	0.874	0.697	0.903
PAAP (Win et al., 2018)	0.791	0.865	0.780	0.585	NA
AHTpin_AAC (Kumar et al., 2015)	0.785	0.777	0.793	0.567	NA
AHTpin_ATC (Kumar et al., 2015)	0.785	0.783	0.787	0.573	NA
Our ensemble	0.858	0.832	0.885	0.718	0.926

TABLE 4 | Performance evaluation of the proposed method using independence testing dataset.

Method	ACC	Sn	Sp	MCC	AUC
CNN + SVM (Rauf et al., 2021)	0.895	0.948	0.841	0.795	0.895
mAHTPred (Manavalan et al., 2019)	0.883	0.894	0.873	0.767	0.951
PAAP (Win et al., 2018)	NA	NA	NA	NA	NA
AHTpin_AAC (Kumar et al., 2015)	0.800	0.821	0.780	0.601	0.852
AHTpin_ATC (Kumar et al., 2015)	0.820	0.798	0.842	0.641	0.888
Our ensemble	0.904	0.920	0.889	0.809	0.965

TABLE 5 | Performance evaluation of the proposed method using recently reported novel AHTPs.

Peptide sequence	IC_{50}	Source	References	Correctly identify by our method (Yes/No)
YLYELR	9.37 μM	Scorpion venom	Setayesh-Mehr et al. (2021)	Yes
AFPYYGHHLG	17.22 μM	Scorpion venom	Setayesh-Mehr et al. (2021)	Yes
LVLPGE	13.5 μM	Broccoli protein	Pei et al. (2021)	Yes
IPPAYTK	23.5 μM	Broccoli protein	Dang et al. (2019)	Yes
LVLPGELAK	184 μM	Broccoli protein	Dang et al. (2019)	Yes
TFQGPPHGIQVER	3.4 μM	Broccoli protein	Dang et al. (2019)	Yes
LIIPQH	120.1 μM	Rice wine lees	He et al. (2021)	Yes
LIPPEH	60.49 μM	Rice wine lees	He et al. (2021)	Yes
QTDEYGNPPR	210.03 μM	Black tea	Lu et al. (2021)	Yes
AGFAGDDAPR	178.91 μM	Black tea	Lu et al. (2021)	No
IDESLR	196.31 μM	Black tea	Lu et al. (2021)	No
IQDKEGIPPDQQR	121.11 μM	Black tea	Lu et al. (2021)	Yes
DAFGSFLYEYSE	-	Ricotta cheese	Pontonio et al. (2021)	No
RHPYFYAPELLYYANK	-	Ricotta cheese	Pontonio et al. (2021)	Yes
VERGRRITSV	6.82 μM	Walnut Glutelin-1	Wang et al. (2021)	No
FVIEPNITPA	6.36 μM	Walnut Glutelin-1	Wang et al. (2021)	Yes
LSGYGP	2.57 μM	Tilapia	Chen et al. (2020)	Yes
LVPPHA	414.88 μM	Radix Astragali	Wu et al. (2020)	Yes
SAGGYIW	0.002 μM	Wheat gluten	Zhang et al. (2020)	Yes
APATPSFW	0.875 μM	Wheat gluten	Zhang et al. (2020)	Yes
PPNNNPASPDFSSS	-	Soy protein	Daliri et al. (2019)	Yes
GPKALPII	-	Soy Protein	Daliri et al. (2019)	Yes
IIRCTGC	-	Soy protein	Daliri et al. (2019)	No
IGPGPFSR	47.22 μM	Mussel lamellidens	Ankhi et al. (2022)	Yes
FHAPWK	16.83 μM	Cassia obtusifolia seeds	Shih et al. (2019)	Yes

explaining the output of any machine learning model. It correlates optimal credit allocation with local explanations by using classic Shapley values (Lundberg and Lee 2017). Since it averages the marginal contributions across all permutations, the performance of SHAP is notably more consistent than that of the information gain technique. The SHAP summary plot in **Figure 7C** is somewhat consistent with the information gain-based importance plot, which shows that comF2 was the most significant feature, followed by Pse_SC13 and QSO35. According to the SHAP plot, the comF2 feature had an effect on the likelihoods for a larger model sample. Every dot in the SHAP plot represents a sample from the data. For each sample, the color of the corresponding dot refers to the value of the associated feature. The x-axis represents the feature's influence on the model's prediction. The high spread of comF2 indicates that it could capture and provide more useful information to the model to predict/identify the classes. Moreover, the partial dependence plot (PDP) of comF2 presents the impact of this feature on the predicted outcome, as shown in **Figure 7D**, allowing for a better understanding of the feature's interdependence with the target class (AHTP). According to the comF2 PDP illustrated in **Figure 7D**, the higher the value of the comF2 feature is, the higher the chance of the sample being classified into the AHTP class by the model (comF2 greater than two likelihoods of being in the AHTP class). Additionally, **Figure 7E** depicts the distribution of the top six features. A substantial distribution difference was observed between the AHTP and non-AHTP classes in the histogram of the comF2 feature. However, some overlap occurred between the two classes' territories. The functionality of comF2 can be enhanced, resulting in an increase in prediction performance.

Comparison With Existing Prediction Methods

To evaluate the performance of the proposed method, we used the benchmarking dataset and the independence testing dataset (as shown in **Tables 3**, **4**, respectively), and then we compared and evaluated our ensemble method with the available prediction tools based on the results reported in (Manavalan et al., 2019; Rauf et al., 2021). As shown in **Table 3**, our technique achieved 85.8% accuracy on the benchmarking dataset or training dataset, outperforming most of the other methods. However, while the CNN + SVM technique surpassed our ensemble for the training dataset, our ensemble performed substantially better on the independent dataset.

When testing was performed on the independent data, accuracies of 90.4% were achieved, as shown in **Table 4**, and our method significantly outperformed the other methods.

Performance Evaluation of Our Model With Novel Antihypertensive Peptides From Recent Studies

Novel AHTPs derived from food or natural sources are receiving significant attention. Therefore, an increasing number of food-derived or natural sources AHTPs have been researched and reported. To further assess the

generalization performance and robustness of the proposed method on new unseen data, we collected various experimental AHTPs from recent studies. These published AHTPs have been validated by *in vitro* or *in vivo* experimental assays. The results are summarized in **Table 5**. Note that these peptides did not overlap with our training data. Our ensemble model correctly classified these novel AHTPs from different sources with an accuracy of 80%.

CONCLUSION

In this work, an ensemble model with a combination of XGB, RF, and SVM machine learning algorithms integrated by weighted voting was developed to achieve improved sensitivity and reduce the false positive rate in terms of predicting AHTPs. A new composite feature for AHTPs, comF2, was proposed and incorporated to improve the sensitivity of the developed method. The components of the comF2 feature were selected by a machine learning process based solely on a single training dataset (benchmarking dataset). However, we hypothesize that this new feature can be improved and adjusted to be more sensitive by combining novel knowledge or the information contained in the structure-function relationships (structure-activity relationships) of AHTPs reported in recent studies or by experts/biologists in the field. This knowledge can be expanded by incorporating more recent information or new significant features found in the future to further improve the proposed approach.

Currently, deep learning (DL) has become very prominence because of its ability to identify patterns in large volumes of raw data (scalability) and its ability in perform automatic feature extraction from raw data (feature encoding/learning). However, DL does not have an explicit feature engineering step because it has automated feature extraction. We are interested in feature engineering, extraction, and selection; therefore, we apply machine learning, including DL-related algorithms so called neural nets. We exploited various features that are more explainable in terms of biological meaning, and we tried to capture an explainable relationship in the hybrid feature that may be an advantage in AHTP design in the future. We used the ensemble method, which is well-known to ensure generalization and to reduce the problem of overfitting of individual models. For precision of classification tools, both positive and negative dataset are important for model training. Availability of experimentally validated negative datasets, particularly sequences with similar amino acid compositions to those of AHTPs, will be beneficial for further improvement. Moreover, additional negative datasets containing other classes of peptides, for example, antioxidant, antimicrobial, and anticancer peptides and neuropeptides, which have been experimentally confirmed for their activities and do not show any antihypertensive activity will be more advantageous. To make this tool more useful, implementation as a webserver will be more accessible to bioactive peptide research communities.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

Conceptualization: AH, CT, and SL. Formal analysis: SL. Methodology: SL. Writing—original draft: SL. Writing—review and editing: AH, CT, WW, and SL. Funding acquisition: WW.

FUNDING

This research was funded by King Mongkut's University of Technology Thonburi, Thailand.

REFERENCES

Abachi, S., Bazinet, L., and Beaulieu, L. (2019). Antihypertensive and Angiotensin-I-Converting Enzyme (ACE)-Inhibitory Peptides from Fish as Potential Cardioprotective Compounds. *Mar. Drugs* 17 (11), 613. doi:10.3390/md17110613

Aluko, R. E. (2015). Antihypertensive Peptides from Food Proteins. *Annu. Rev. Food Sci. Technol.* 6, 235–262. PMID: 25884281. doi:10.1146/annurev-food-022814-015520

Ankhi, H., Madhushrita, D., K, D. T. D., Pubali, D., and Jana, C. (2022). Isolation of an Antihypertensive Bioactive Peptide from the Freshwater Mussel Lamellidens Marginalis. *Int. J. Food Nutr. Sci.* 11, 1–8. doi:10.54876/ijfans_01-08

Asoodeh, A., Homayouni-Tabrizi, M., Shabestarian, H., Emtenani, S., and Emtenani, S. (2016). Biochemical Characterization of a Novel Antioxidant and Angiotensin I-Converting Enzyme Inhibitory Peptide from *Struthio camelus* Egg white Protein Hydrolysis. *J. Food Drug Anal.* 24 (2), 332–342. doi:10.1016/j.jfda.2015.11.010

Balgir, P. P., and Sharma, M. (2017). Biopharmaceutical Potential of ACE-Inhibitory Peptides. *J. Proteomics Bioinform.* 10, 171–177. doi:10.4172/jpb.1000437

Boman, H. G. (2003). Antibacterial Peptides: Basic Facts and Emerging Concepts. *J. Intern. Med.* 254 (3), 197–215. doi:10.1046/j.1365-2796.2003.01228.x

Chen, J., Ryu, B., Zhang, Y., Liang, P., Li, C., Zhou, C., et al. (2020). Comparison of an angiotensin-I-converting Enzyme Inhibitory Peptide from tilapia (*Oreochromis niloticus*) with Captopril: Inhibition Kinetics, *In Vivo* Effect, Simulated Gastrointestinal Digestion and a Molecular Docking Study. *J. Sci. Food Agric.* 100 (1), 315–324. doi:10.1002/jsfa.10041

Chen, W., Ding, H., Feng, P., Lin, H., and Chou, K.-C. (2016). iACP: a Sequence-Based Tool for Identifying Anticancer Peptides. *Oncotarget* 7 (13), 16895–16909. doi:10.18632/oncotarget.7815

Chou, K.-C. (2000). Prediction of Protein Subcellular Locations by Incorporating Quasi-Sequence-Order Effect. *Biochem. Biophysical Res. Commun.* 278, 477–483. doi:10.1006/bbrc.2000.3815

Chou, K.-C. (2011). Some Remarks on Protein Attribute Prediction and Pseudo Amino Acid Composition. *J. Theor. Biol.* 273, 236–247. doi:10.1016/j.jtbi.2010.12.024

Chou, K.-C. (2005). Using Amphiphilic Pseudo Amino Acid Composition to Predict Enzyme Subfamily Classes. *Bioinformatics* 21, 10–19. doi:10.1093/bioinformatics/bth466

Cruciani, G., Baroni, M., Carosati, E., Clementi, M., Valigi, R., and Clementi, S. (2004). Peptide Studies by Means of Principal Properties of Amino Acids

Derived from MIF Descriptors. *J. Chemometrics* 18, 146–155. doi:10.1002/cem.856

Daliri, E. B.-M., Ofosu, F. K., Chelliah, R., Kim, J.-H., Oh, D.-H., and Oh, D. H. (2019). Development of a Soy Protein Hydrolysate with an Antihypertensive Effect. *Ijms* 20 (6), 1496. doi:10.3390/ijms20061496

Dang, Y., Zhou, T., Hao, L., Cao, J., Sun, Y., and Pan, D. (2019). *In Vitro* and *In Vivo* Studies on the Angiotensin-Converting Enzyme Inhibitory Activity Peptides Isolated from Broccoli Protein Hydrolysate. *J. Agric. Food Chem.* 67, 6757–6764. doi:10.1021/acs.jafc.9b01137

Daskaya-Dikmen, C., Yucetepe, A., Karbancioglu-Guler, F., Daskaya, H., and Ozcelik, B. (2017). Angiotensin-I-Converting Enzyme (ACE)-Inhibitory Peptides from Plants. *Nutrients* 9 (4), 316. doi:10.3390/nu9040316

de Castro, R. J. S., and Sato, H. H. (2015). Biologically Active Peptides: Processes for Their Generation, Purification and Identification and Applications as Natural Additives in the Food and Pharmaceutical Industries. *Food Res. Int.* 74, 185–198. doi:10.1016/j.foodres.2015.05.013

De Leo, F., Panarese, S., Gallerani, R., and Ceci, L. (2009). Angiotensin Converting Enzyme (ACE) Inhibitory Peptides: Production and Implementation of Functional Food. *Cpd* 15 (31), 3622–3643. doi:10.2174/138161209789271834

Dubchak, I., Muchnik, I., Holbrook, S. R., and Kim, S. H. (1995). Prediction of Protein Folding Class Using Global Description of Amino Acid Sequence. *Proc. Natl. Acad. Sci. U.S.A.* 92, 8700–8704. doi:10.1073/pnas.92.19.8700

Dubchak, I., Muchnik, I., Mayor, C., Dralyuk, I., and Kim, S.-H. (1999). Recognition of a Protein Fold in the Context of the Scop Classification. *Proteins* 35, 401–407. doi:10.1002/(sici)1097-0134(19990601)35:4<401::aid-prot3>3.0.co;2-k

Fernandez-Escamilla, A.-M., Rousseau, F., Schymkowitz, J., and Serrano, L. (2004). Prediction of Sequence-dependent and Mutational Effects on the Aggregation of Peptides and Proteins. *Nat. Biotechnol.* 22, 1302–1306. doi:10.1038/nbt1012

Festa, M., Sansone, C., Brunet, C., Crocetta, F., Di Paola, L., Lombardo, M., et al. (2020). Cardiovascular Active Peptides of Marine Origin with ACE Inhibitory Activities: Potential Role as Anti-hypertensive Drugs and in Prevention of SARS-CoV-2 Infection. *Ijms* 21 (21), 8364PMC7664667. doi:10.3390/ijms21218364.PMID:33171852

Guruprasad, K., Reddy, B. V. B., and Pandit, M. W. (1990). Correlation between Stability of a Protein and its Dipeptide Composition: a Novel Approach for Predicting *In Vivo* Stability of a Protein from its Primary Sequence. *Protein Eng. Des. Sel* 4 (2), 155–161. doi:10.1093/protein/4.2.155

He, R., Aluko, R. E., and Ju, X.-R. (2014). Evaluating Molecular Mechanism of Hypotensive Peptides Interactions with Renin and Angiotensin Converting Enzyme. *PLoS ONE* 9 (3), e91051. doi:10.1371/journal.pone.0091051

He, R., Wang, Y., Yang, Y., Wang, Z., Ju, X., and Yuan, J. (2019). Rapeseed Protein-Derived ACE Inhibitory Peptides LY, RALP and GHS Show Antioxidant and Anti-inflammatory Effects on Spontaneously Hypertensive Rats. *J. Funct. Foods* 55, 211–219. doi:10.1016/j.jff.2019.02.031

He, Z., Liu, G., Qiao, Z., Cao, Y., and Song, M. (2021). Novel Angiotensin-I Converting Enzyme Inhibitory Peptides Isolated from Rice Wine Lees: Purification, Characterization, and Structure-Activity Relationship. *Front. Nutr.* 8, 746113. doi:10.3389/fnut.2021.746113

Ikai, A. (1980). Thermostability and Aliphatic index of Globular Proteins. *J. Biochem.* 88 (6), 1895–1898. doi:10.1093/oxfordjournals.jbchem.a133104

Iwaniak, A., Minkiewicz, P., Darewicz, M., Sieniawski, K., and Starowicz, P. (2016). BIOPEP Database of Sensory Peptides and Amino Acids. *Food Res. Int.* 85, 155–161. doi:10.1016/j.foodres.2016.04.031

Jakubczyk, A., Karaś, M., Rybczyńska-Tkaczyk, K., Zielińska, E., and Zieliński, D. (2020). Current Trends of Bioactive Peptides-New Sources and Therapeutic Effect. *Foods*, 9(7). PMCID, 846PMC7404774. doi:10.3390/foods9070846. PMID:32610520

Jiang, Q., Chen, Q., Zhang, T., Liu, M., Duan, S., and Sun, X. (2021). The Antihypertensive Effects and Potential Molecular Mechanism of Microalgal Angiotensin I-Converting Enzyme Inhibitor-like Peptides: A Mini Review. *Ijms* 22 (8), 4068. doi:10.3390/ijms22084068

Kononenko, I. (1994). "Estimating Attributes: Analysis and Extensions of RELIEF," in *Machine Learning: ECML-94. ECML 1994. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence)*. Editors F. Bergadano and L. De Raedt (Berlin, Heidelberg: Springer), 784. doi:10.1007/3-540-57868-4_57

Kumar, M., Thakur, V., and Raghava, G. P. (2008). COPid: Composition Based Protein Identification. *Silico Biol.* 8 (2), 121–128.

Kumar, R., Chaudhary, K., Singh Chauhan, J., Nagpal, G., Kumar, R., Sharma, M., et al. (2015). An In Silico Platform for Predicting, Screening and Designing of Antihypertensive Peptides. *Sci. Rep.* 5, 12512. doi:10.1038/srep12512

Kumar, R., Chaudhary, K., Sharma, M., Nagpal, G., Chauhan, J. S., Singh, S., et al. (2014). AHTPDB: A Comprehensive Platform for Analysis and Presentation of Antihypertensive Peptides. *Nucleic Acids Res.* 43, D956–D962. doi:10.1093/nar/gku1141

Lee, S. Y., and Hur, S. J. (2019). Purification of Novel Angiotensin Converting Enzyme Inhibitory Peptides from Beef Myofibrillar Proteins and Analysis of Their Effect in Spontaneously Hypertensive Rat Model. *Biomed. Pharmacother.* 116, 109046. doi:10.1016/j.biopha.2019.109046

Lertampaiporn, S., Vorapreeda, T., Hongsthong, A., and Thammarongtham, C. (2021). Ensemble-AMPPred: Robust AMP Prediction and Recognition Using the Ensemble Learning Method with a New Hybrid Feature for Differentiating AMPs. *Genes* 12 (2), 137. doi:10.3390/genes12020137

Li-Chan, E. C. (2015). Bioactive Peptides and Protein Hydrolysates: Research Trends and Challenges for Application as Nutraceuticals and Functional Food Ingredients. *Curr. Opin. Food Sci.* 1, 28–37. doi:10.1016/j.cofs.2014.09.005

Liang, G., and Li, Z. (2007). Factor Analysis Scale of Generalized Amino Acid Information as the Source of a New Set of Descriptors for Elucidating the Structure and Activity Relationships of Cationic Antimicrobial Peptides. *QSAR Comb. Sci.* 26 (6), 754–763. doi:10.1002/qsar.200630145

Lu, Y., Wang, Y., Huang, D., Bian, Z., Lu, P., Fan, D., et al. (2021). Inhibitory Mechanism of Angiotensin-Converting Enzyme Inhibitory Peptides from Black tea. *J. Zhejiang Univ. Sci. B* 22 (7), 575PMC8284085–589. PMCID. doi:10.1631/jzus.B2000520.PMID:34269010

Lundberg, S. M., and Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* 30. arXiv:1705.07874 [cs.AI].

Maky, M. A., and Zendo, T. (2021). Generation and Characterization of Novel Bioactive Peptides from Fish and Beef Hydrolysates. *Appl. Sci.* 11, 10452. doi:10.3390/app112110452

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019). mAHTPred: A Sequence-Based Meta-Predictor for Improving the Prediction of Anti-hypertensive Peptides Using Effective Feature Representation. *Bioinformatics* 35, 2757–2765. doi:10.1093/bioinformatics/bty1047

Martínez-Maqueda, D., Miralles, B., Recio, I., and Hernández-Ledesma, B. (2012). Antihypertensive Peptides from Food Proteins: a Review. *Food Funct.* 3 (4), 350–361. doi:10.1039/c2fo10192k

Mei, H., Liao, Z. H., Zhou, Y., and Li, S. Z. (2005). A New Set of Amino Acid Descriptors and its Application in Peptide QSARs. *Biopolymers* 80 (6), 775–786. doi:10.1002/bip.20296

Mills, K. T., Stefanescu, A., and He, J. (2020). The Global Epidemiology of Hypertension. *Nat. Rev. Nephrol.* 16 (4), 223–237. doi:10.1038/s41581-019-0244-2

Minkiewicz, P., Dziuba, J., Iwaniak, A., Dziuba, M., and Darewicz, M. (2008). BIOPEP Database and Other Programs for Processing Bioactive Peptide Sequences. *J. AOAC Int.* 91 (4), 965–980. doi:10.1093/jaoac/91.4.965

Nguyen, Q., Dominguez, J., Nguyen, L., and Gullapalli, N. (2010). Hypertension Management: An Update. *Am. Health Drug Benefits.* 3, 47–56.

Nimalaratne, C., Bandara, N., and Wu, J. (2015). Purification and Characterization of Antioxidant Peptides from Enzymatically Hydrolyzed Chicken Egg white. *Food Chem.* 188, 467–472. doi:10.1016/j.foodchem.2015.05.014

Norris, R., and J., R. (2013). "Antihypertensive Peptides from Food Proteins," in *Bioactive Food Peptides in Health and Disease*. Editors B. Hernandez-Ledesma and C. Hsieh (IntechOpen). doi:10.5772/51710

Osorio, D., Rondón-Villarreal, P., and Torres, R. (2015). Peptides: A Package for Data Mining of Antimicrobial Peptides. *R. J.* 7 (1), 4–14. doi:10.32614/rj-2015-001

Pavlicevic, M., Maestri, E., and Marmiroli, M. (2020). Marine Bioactive Peptides-An Overview of Generation, Structure and Application with a Focus on Food Sources. *Mar. Drugs* 18 (8), 424. doi:10.3390/md18080424

Pei, J., Hua, Y., Zhou, T., Gao, X., Dang, Y., and Wang, Y. (2021). Transport, *In Vivo* Antihypertensive Effect, and Pharmacokinetics of an Angiotensin-Converting Enzyme (ACE) Inhibitory Peptide LVLPGE. *J. Agric. Food Chem.* 69 (7), 2149–2156. doi:10.1021/acs.jafc.0c07048

Pontonio, E., Montemurro, M., De Gennaro, G. V., Miceli, V., and Rizzello, C. G. (2021). Antihypertensive Peptides from Ultrafiltration and Fermentation of the Ricotta Cheese Exhausted Whey: Design and Characterization of a Functional Ricotta Cheese. *Foods* 10 (11), 2573. doi:10.3390/foods10112573

Pujiastuti, D. Y., Ghoyatul Amin, M. N., Alamsjah, M. A., and Hsu, J.-L. (2019). Marine Organisms as Potential Sources of Bioactive Peptides that Inhibit the Activity of Angiotensin I-Converting Enzyme: A Review. *Molecules* 24 (14), 2541. doi:10.3390/molecules24142541

Rauf, A., Kiran, A., Hassan, M. T., Mahmood, S., Mustafa, G., and Jeon, M. (2021). Boosted Prediction of Antihypertensive Peptides Using Deep Learning. *Appl. Sci.* 11 (5), 2316. doi:10.3390/app11052316

Ryan, J. T., Ross, R. P., Bolton, D., Fitzgerald, G. F., and Stanton, C. (2011). Bioactive Peptides from Muscle Sources: Meat and Fish. *Nutrients* 3 (9), 765–791. doi:10.3390/nu3090765

Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M., and Wold, S. (1998). New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *J. Med. Chem.* 41, 2481–2491. doi:10.1021/jm9700575

Setayesh-Mehr, Z., Ghasemi, L. V., and Asoodeh, A. (2021). Evaluation of the *In Vivo* Antihypertensive Effect and Antioxidant Activity of HL-7 and HL-10 Peptide in Mice. *Mol. Biol. Rep.* 48 (7), 5571–5578. doi:10.1007/s11033-021-06576-7

Sharma, A., Kapoor, P., Gautam, A., Chaudhary, K., Kumar, R., Chauhan, J. S., et al. (2013). Computational Approach for Designing Tumor Homing Peptides. *Sci. Rep.* 3, 1607. doi:10.1038/srep01607

Shih, Y.-H., Chen, F.-A., Wang, L.-F., and Hsu, J.-L. (2019). Discovery and Study of Novel Antihypertensive Peptides Derived from Cassia Obtusifolia Seeds. *J. Agric. Food Chem.* 67 (28), 7810–7820. doi:10.1021/acs.jafc.9b01922

Thomsen, M. C. F., and Nielsen, M. (2012). Seq2Logo: a Method for Construction and Visualization of Amino Acid Binding Motifs and Sequence Profiles Including Sequence Weighting, Pseudo Counts and Two-Sided Representation of Amino Acid Enrichment and Depletion. *Nucleic Acids Res.* 40, W281–W287. Web Server issue)–W287. doi:10.1093/nar/gks469

Tian, F., Zhou, P., and Li, Z. (2007). T-scale as a Novel Vector of Topological Descriptors for Amino Acids and its Application in QSARs of Peptides. *J. Mol. Struct.* 830, 106–115. doi:10.1016/j.molstruc.2006.07.004

Toloşi, L., and Lengauer, T. (2011). Classification with Correlated Features: Unreliability of Feature Ranking and Solutions. *Bioinformatics* 27, 1986–1994. doi:10.1093/bioinformatics/btr300

Udenigwe, C. C., Li, H., and Aluko, R. E. (2012). Quantitative Structure-Activity Relationship Modeling of Renin-Inhibiting Dipeptides. *Amino Acids* 42, 1379–1386. doi:10.1007/s00726-011-0833-2

Udenigwe, C. C., and Mohan, A. (2014). Mechanisms of Food Protein-Derived Antihypertensive Peptides Other Than ACE Inhibition. *J. Funct. Foods* 8, 45–52. doi:10.1016/j.jff.2014.03.002

Usmani, S. S., Bhalla, S., and Raghava, G. P. S. (2018). Prediction of Antitubercular Peptides from Sequence Information Using Ensemble Classifier and Hybrid Features. *Front. Pharmacol.* 9, 954. doi:10.3389/fphar.2018.00954

van Westen, G. J., Swier, R. F., Wegner, J. K., IJzerman, A. P., van Vlijmen, H. W., and Bender, A. (2013). Benchmarking of Protein Descriptor Sets in Proteochemometric Modeling (Part 1): Comparative Study of 13 Amino Acid Descriptor Sets. *J. Cheminform* 5 (1), 41. doi:10.1186/1758-2946-5-41

Wang, J., Wang, G., Zhang, Y., Zhang, R., and Zhang, Y. (2021). Novel Angiotensin-Converting Enzyme Inhibitory Peptides Identified from Walnut Glutelin-1 Hydrolysates: Molecular Interaction, Stability, and Antihypertensive Effects. *Nutrients* 14 (1), 151. doi:10.3390/nu14010151

Wei, D., Fan, W., and Xu, Y. (2019). *In Vitro* Production and Identification of Angiotensin Converting Enzyme (ACE) Inhibitory Peptides Derived from Distilled Spent Grain Prolamin Isolate. *Foods* 8 (9), 390. doi:10.3390/foods8090390

Win, T. S., Schaduangrat, N., Prachayasittikul, V., Nantasenamat, C., and Shoombuatong, W. (2018). PAAP: A Web Server for Predicting Antihypertensive Activity of Peptides. *Future Med. Chem.* 10, 1749–1767. doi:10.4155/fmc-2017-0300

Wu, C. H., Mohammadmoradi, S., Chen, J. Z., Sawada, H., Daugherty, A., and Lu, H. S. (2018). Renin-Angiotensin System and Cardiovascular Functions. *Arterioscler Thromb. Vasc. Biol.* 38 (7), e108–e116. doi:10.1161/ATVBAHA.118.311282

Wu, J.-S., Li, J.-M., Lo, H.-Y., Hsiang, C.-Y., and Ho, T.-Y. (2020). Anti-hypertensive and Angiotensin-Converting Enzyme Inhibitory Effects of Radix Astragali and its Bioactive Peptide AM-1. *J. Ethnopharmacology* 254, 254112724. doi:10.1016/j.jep.2020.112724

Xiao, N., Cao, D.-S., Zhu, M.-F., and Xu, Q.-S. (2015). Protr/ProtrWeb: R Package and Web Server for Generating Various Numerical Representation Schemes of Protein Sequences. *Bioinformatics* 31, 1857–1859. doi:10.1093/bioinformatics/btv042

Yang, L., Shu, M., Ma, K., Mei, H., Jiang, Y., and Li, Z. (2010). ST-scale as a Novel Amino Acid Descriptor and its Application in QSAM of Peptides and Analogues. *Amino acids* 38 (3), 805–816. doi:10.1007/s00726-009-0287-y

Yi, Y., Lv, Y., Zhang, L., Yang, J., and Shi, Q. (2018). High Throughput Identification of Antihypertensive Peptides from Fish Proteome Datasets. *Mar. Drugs* 16, 365. doi:10.3390/md16100365

Zaky, A. A., Simal-Gandara, J., Eun, J.-B., Shim, J.-H., and Abd El-Aty, A. M. (2022). Bioactivities, Applications, Safety, and Health Benefits of Bioactive Peptides from Food and By-Products: A Review. *Front. Nutr.* 8, 815640. doi:10.3389/fnut.2021.815640

Zaliani, A., and Gancia, E. (1999). MS-WHIM Scores for Amino Acids: a New 3D-Description for Peptide QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* 39 (3), 525–533. doi:10.1021/ci980211b

Zhang, P., Chang, C., Liu, H., Li, B., Yan, Q., and Jiang, Z. (2020). Identification of Novel Angiotensin I-Converting Enzyme (ACE) Inhibitory Peptides from Wheat Gluten Hydrolysate by the Protease of *Pseudomonas aeruginosa*. *J. Funct. Foods* 65, 103751. doi:10.1016/j.jff.2019.103751

Zhou, B., Perel, P., Mensah, G. A., and Ezzati, M. (2021). Global Epidemiology, Health burden and Effective Interventions for Elevated Blood Pressure and Hypertension. *Nat. Rev. Cardiol.* 18 (11), 785–802. doi:10.1038/s41569-021-00559-8

Zhu, J., Li, J., Guo, Y., Quaisie, J., Hong, C., and Ma, H. (2021). Antihypertensive and Immunomodulatory Effects of Defatted Corn Germ Hydrolysates: An *In Vivo* Study. *Front. Nutr.* 8, 679583. doi:10.3389/fnut.2021.679583

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.