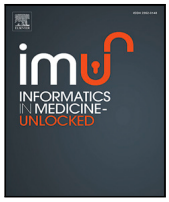
Accurately predicting anticancer peptide using an ensemble of heterogeneously trained classifiers

Informatics in Medicine Unlocked 42 (2023) 101348

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu



Accurately predicting anticancer peptide using an ensemble of heterogeneously trained classifiers

Sayed Mehedi Azim [a], Noor Hossain Nuri Sabab [b], Iman Noshadi [c], Hamid Alinejad-Rokny [d,e,f], Alok Sharma [g,h], Swakkhar Shatabda [b], Iman Dehzangi [a,i,*]

[a] Center for Computational and Integrative Biology, Rutgers University, Camden, NJ 08102, USA
[b] Department of Computer Science and Engineering, United International University, Plot 2, United City, Madani Avenue, Badda, Dhaka, 1212, Bangladesh
[c] Department of Bioengineering, University of California, Riverside, CA, 92507, USA
[d] BioMedical Machine Learning Lab, The Graduate School of Biomedical Engineering, The University of New South Wales (UNSW Sydney), Sydney, NSW, 2052, Australia
[e] UNSW Data Science Hub, UNSW Sydney, Sydney, NSW, 2052, Australia
[f] Health Data Analytics Program, AI-enabled Processes Research Centre, Macquarie University, Sydney, 2109, Australia
[g] Institute for Integrated and Intelligent Systems, Griffith University, Brisbane, Australia
[h] Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan
[i] Department of Computer Science, Rutgers University, Camden, NJ, 08102, USA

ARTICLE INFO

Keywords:
Anticancer peptides
Random Forest
Machine learning
Ensemble learning
Feature extraction
Heterogeneous classifiers

ABSTRACT

The use of therapeutic peptides for the treatment of cancer has received tremendous attention in recent years. Anticancer peptides (ACPs) are considered new anticancer drugs which have several advantages over chemistry-based drugs including high specificity, strong tumor penetration capacity, and low toxicity level for normal cells. Due to the rise of experimentally verified bioactive peptides, several in silico approaches became imperative for the investigation of the characteristics of ACPs. In this paper, we proposed a new machine learning tool named iACP-RF that uses a combination of several sequence-based features and an ensemble of three heterogeneously trained Random Forest classifiers to accurately predict anticancer peptides. Experimental results show that our proposed model achieves an accuracy of 75.9% which outperforms other state-of-the-art methods by a significant margin. We also achieve 0.52, 75.6%, and 76.2% in terms of Matthews Correlation Coefficient (MCC), Sensitivity, and Specificity, respectively. iACP-RF as a standalone tool and its source code are publicly available at: https://github.com/MLBC-lab/iACP-RF.

1. Introduction

Cancer is considered as a genetic disease since it is developed due to changes in genes that control cell function, especially how they grow and divide. According to the World Health Organization (WHO), in 2020 alone, 10 million people died prematurely due to cancer worldwide, which accounts for nearly one in six deaths. Due to the inadequacy of accurate and non-invasive markers, the detection of cancer is usually biased and not always correct [1,2]. Advancements in the field of proteomics and genomics have recently led to the discovery of peptide-based biomarkers, which have enhanced the detection of cancer at its early stage [3]. After diagnosing cancer, the next step is its treatment.

As of yet, chemotherapy, radiation therapy, hormonal therapy, and surgery are the conventional treatments available for treating cancer, which still come with adverse side effects and high expenses. In addition, there are still chances of recurrences of cancer after remission, even if the treatment shows promising results. This indicates the necessity to find more effective and improved treatment [4–6]. In recent years, peptide-based therapy has emerged as a novel and advanced strategy for the treatment of cancer. It has several advantages like good efficacy, high target specificity, low toxicity, easily synthesized and modified, and less immunogenic when combined with recombinant antibodies [7–9].

In the past few years, therapeutic peptides have been used as a diagnostic tool and proved to be successful in treating many diseases. So far, more than 7000 natural peptides have been reported to exhibit multiple bioactivities (antiviral, antifungal, antibacterial, anticancer, tumor-homing, etc.). So far, more than 60 drugs have been approved by

* Corresponding author at: Center for Computational and Integrative Biology, Rutgers University, Camden, NJ 08102, USA.
E-mail address: i.dehzangi@rutgers.edu (I. Dehzangi).

https://doi.org/10.1016/j.imu.2023.101348
Received 29 May 2023; Received in revised form 31 August 2023; Accepted 2 September 2023
Available online 6 September 2023

the Federal Drug Administration (FDA) and more than 500 are under clinical trials [10–17].

The term anticancer peptides (ACPs) refer to small peptides that exert selective and toxic properties toward cancer cells and represent a promising class of therapeutic agents as synthetic peptide-based drugs and vaccines due to their inherent high penetration and selectivity, as well as ease of modification. It was shown that affinity, stability, and selectivity for the elimination of cancer cells can be improved by designing ACPs [18]. Amino acid residues influence the anticancer properties by relying on cationic, hydrophobic, and amphiphilic properties with helical structures to drive cell permeability. Cationic amino acid residues like lysine, arginine, and histidine can particularly penetrate and disrupt cancer cell membranes to induce cytotoxicity. On the other hand, anionic amino acids like glutamic and aspartic acids provide antiproliferative activity against cancer cells [19]. Hydrophobic amino residues like phenylalanine, tryptophan, and tyrosine also exert their cytotoxic activities [20]. Also, cationic and hydrophobic amino acids that form the secondary structure of ACPs, plays a vital role in peptide-cancer cell membrane interaction that leads to cancer cell disruption and necrosis [21].

ACPs are small peptides (5–50 amino acids) and are cationic by nature. In general, they possess mostly α-helix based secondary structures (e.g. LL37, BMAP-27, BMAP-28, and Cecropin A). Some also fold into β-sheet (e.g. Lactoferrin, Defensins, etc.) and demonstrate extended linear structure like Tritrpticin, and Indolicidin [22,23]. Cancer cells display different properties in contrast to normal cells and possess a larger surface area due to the presence of a higher number of microvilli, negatively charged cell membrane, and higher fluidity of the membrane. Mitochondrial membrane lysis (apoptosis) is another means for ACPs to exhibit their function, recruiting other immune cells, or inhibiting angiogenesis pathway for attacking cancer cells and activating essential proteins which ultimately lyse cancer cells [24–28].

Accurate prediction of ACPs is essential to explore the novel therapeutic ACPs mechanism of action and development. Experimental processes to conduct different tasks in biology are time-consuming, labor-intensive, and expensive. Hence, there is a demand for developing fast and accurate computational tools. Many machine learning-based models have been developed to tackle different biological problems. Studies for predicting miRNA-disease associations have also benefited from computational approaches like machine learning by outperforming existing works [29–31]. Previously, various sequence-based computational methods were proposed for the prediction of ACPs. Among them, the most notables are AntiCP, iACP, ACPP, iACP-GAEnsC, MLACP, SAP, TargetACP, ACPred, ACP-DL, ACPred-FL, PTPD, Hajisharifi et al.'s method, Li and Wang's method, ACPred-Fuse and PEPred-Suite [29,30,32–43].

In one of the early studies, Tyagi et al. [32] used Support Vector Machine (SVM) to predict ACPs. Although they reported high specificity, their result was poor in terms of sensitivity. To develop iACP, Chen et al. used rigorous cross-validation by optimizing the g-gap dipeptide components to predict ACPs [33], whereas Akbar et al. used genetic algorithm-based ensemble classifiers to tackle this problem [35]. Later on, Manavalan et al. investigated the performance of SVM compared to the Random Forest (RF) classifier on Tyagi-B dataset. They showed that RF demonstrates better results compared to SVM for this problem [29]. In a similar study, Schaduangrat et al. used SVM and RF together to tackle this problem and achieved promising results [16]. Akbar et al. introduced cACP for ACP prediction, applying features like Quasi-sequence order, conjoint triad, and Geary autocorrelation descriptor, along with traditional ML methods such as SVM, RF, and KNN. They also utilized SVM for developing cACP-2LFS to predict ACPs and later proposed cACP-DeepGram, a Deep Neural Network approach using word embedding features, for accurate ACP classification [44–46].

More recently, Yi et al. used a long short-term memory (LSTM) model to predict ACPs and demonstrated promising results. In a different study, Wei et al. proposed a new adaptive feature representation strategy that learns the most representative features for different peptide types and used RF as their classification technique to solve this problem [30]. Although they were able to predict different peptide types simultaneously, their results were not satisfactory compared to other methods. More recently, Rao et al. fused the class and probabilistic features with handcrafted sequential features and showed that combinations of diverse and heterogeneous features have a more discriminative ability to predict ACPs [43]. The comprehensive review of computational approaches proposed to predict ACPs is presented in [16].

Despite all the efforts, the ACP prediction performance still remains limited. The main challenge of existing work is their limited ability in accurately classifying the ACPs. Although the existing works show high accuracy, they lack behind in terms of sensitivity. In addition, ensemble machine learning models combined with heterogeneous sets of features have not been explored adequately to tackle this problem. Although sequence-based feature extraction techniques like K-mer, k-Gapped K-mer, and Binary Profile features showed promising results as standalone techniques, there has not been any study to combine all three feature groups together for the prediction of ACPs.

To mitigate this gap, we propose a new ensemble of heterogeneously trained classifiers called iACP-RF to accurately predict anticancer peptides. To build this model, we use three effective feature extraction techniques namely K-mer, Binary profile feature, and k-Gapped K-mer. We utilize two variants of Gapped K-mer, which are 1-Gapped Di-Mono, and 1-Gapped Mono-Di. We then feed these three feature sets into three different Random Forest (RF) classifiers and then use majority voting to combine them and predict anticancer peptides. iACP-RF, as an ensemble of heterogeneous RF classifiers that are trained using different sets of features, demonstrates better results compared to the state-of-the-art methods found in the literature for predicting anticancer peptides. The key contributions of this research are as follows:

- Proposing a novel architecture to classify anticancer peptides (ACP).
- Outperforming existing models for predicting ACPs.
- Proposing a new ensemble of heterogeneous classifiers using Random Forest as the base classifier.
- Investigating different sets of attributes for feature extraction to build our proposed machine learning model iACP-RF.
- Building our model as a standalone tool namely iACP-RF, which is publicly available at https://github.com/MLBC-lab/iACP-RF.

2. Materials and methods

2.1. Dataset

Here we utilize a dataset obtained from iACP-FSCM study which is available at: iACP-FSCM. They accumulated the datasets used in their previous studies of anticancer peptides such as ACP-DL, ACPP, ACPred-FL, AntiCP, and iACP to build their own dataset [33,34,38,39,47,48]. This dataset contains ACPs with a length between 4 and 50 residues. They divided the dataset into two, namely main and alternate datasets, and divided both further into train and independent test sets. In each case, we train our model on the main and alternate datasets and test our model on their corresponding independent test sets.

The alternate dataset consists of 776 experimentally validated positive peptides and 776 negative peptides as training sets. It also contains 194 positive peptides and 194 negative peptides in the validation set used to verify the model's performance. The main dataset, contains ACPs with both anticancer and antimicrobial properties. This dataset consists of 689 positive anticancer peptides and 689 negative peptides. It also contains 172 positive peptides and 172 negative peptides in its validation set. Using these datasets enables us to directly compare our results with the state-of-the-art models found in the literature.

Table 1
Depiction of the features extracted using our employed feature extraction techniques.

Name of features	Number of features	Feature structure	Description
MonoMer Composition	20	X	When X = 1, 20 features of peptide
DiMer Composition	400	XX	When X = 2, 400 features of peptide
TriMer Composition	8000	XXX	When X = 3, 8000 features of peptide
1-Gapped Di-Mono Composition	8000	XX_X	When k-Gap = 1, 8000 features of peptide
1-Gapped Mono-Di Composition	8000	X_XX	When k-Gap = 1, 8000 features of peptide
Binary Profile Feature	1000	[0,0, …..0,1]	One hot vector is 20, peptide sequence has 50 residues, 20 x 50 = 1000

Table 2
Performance of the different ML methods for both main and alternate datasets using K-mer as a single feature.

Models	Main dataset				Alternate dataset			
	Ac	Sn	Sp	MCC	Ac	Sn	Sp	MCC
LR + LR + LR	68.9	59.3	78.5	0.39	89.0	85.6	92.3	0.70
DT + DT + DT	74.4	76.2	72.7	0.49	87.1	85.6	88.7	0.74
NB + NB + NB	71.5	72.1	70.9	0.43	86.6	75.8	97.4	0.75
KNN + KNN + KNN	69.2	95.3	43.0	0.45	66.5	97.4	35.6	0.42
SVM + SVM + SVM	72.7	73.3	72.1	0.45	90.0	82.0	98.0	0.81
RF + RF + RF	**75.9**	**75.6**	**76.2**	**0.52**	**93.1**	**89.2**	**96.9**	**0.86**

Table 3
Results achieved by different classifiers using Binary profile feature group.

Methods	Main dataset				Alternative dataset			
	Ac	Sn	Sp	MCC	Ac	Sn	Sp	MCC
LR	74.1	77.3	70.9	0.48	88.7	89.2	88.1	0.77
DT	74.4	75.6	73.3	0.49	82.0	89.2	74.7	0.64
KNN	67.4	94.2	40.7	0.41	64.9	95.9	34.0	0.38
SVM	74.4	79.1	69.8	0.49	89.0	82.0	95.9	0.79
NB	58.7	93.0	24.4	0.24	65.7	93.8	37.6	0.38
RF	76.2	79.1	73.3	0.52	76.2	79.1	73.3	0.52

2.2. Feature encoding

Extracting features is a major aspect of any research in order to implement an ML model. To accurately distinguish ACPs from non-ACPs and to develop an effective computational tool, extracting informative features with significant discriminatory information to present peptide sequences is crucial. In this paper, we utilized sequence-based k-mer composition and gapped k-mer composition for representing the sequences. These features are explained in more detail in the following sections.

2.2.1. K-mer composition

K-mer is all the possible consecutive subsequences of length k obtained from peptide sequences, which denote the number of times each combination of k-mer exists in the sequence. With a sequence of size n, the number of k-mer possibilities is $n - k + 1$. To figure out the k-mer composition, the frequency of each k-mer of a particular sequence is calculated and then divided by the whole sequence length to normalize the result. This can be formulated as:

$$composition(s) = \frac{1}{n}\sum_{i=0}^{n-k} match(peptide[i : i + k - 1], s) \quad (1)$$

where n denotes the summation of nucleotides in the sequence, s represents a k-mer with a length of k, and peptide[i:i+k-1] denotes the substring of k peptides starting from the i index. The function match can be presented as the following formula:

$$match(s_j, s_l) = \begin{cases} 1, & if\ s_j == s_l \\ 0, & else \end{cases} \quad (2)$$

For instance, let us consider a peptide sequence consisting of twenty amino acids A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y. For the value of k = 1, we get 20 k-mers {'A','C','D','E','F','G','H','I','K', 'L','M','N','P','Q', 'R','S','T','V','W','Y'}, and using the formula (1) we can represent the sequence "ACDEFGHIKLMNPQRSTVWY" as a feature vector [1/20, 1/20, 1/20, 1/20, 1/20, 1/20, 1/20, 1/20, 1/20, 1/20, 1/20, 1/20, 1/20, 1/20, 1/20, 1/20, 1/20, 1/20, 1/20, 1/20].

2.2.2. K-gapped Mono-Di and Di-Mono composition

A full k-mer refers to a letter subsequence of length k. For example, AAGT is a full 4-mer. By contrast, a k-gapped Mono-Di refers to a subsequence containing three letters with k-number of gaps after one amino acid, whereas Di-Mono refers to a subsequence containing three letters with k-number of gaps after two amino acids. The normalized frequency of 3-mers with a single gap between them are used to calculate these features. X_XX is the form for 1-gapped Mono-Di where X is the amino acids A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, and 1-gapped Di-Mono is in the form of XX_X.

When k-gap is equal to n, then 20 x (20 x 20) x n features will be generated for protein sequences. For 1-gapped Mono-Di having k-gap = 1, a total of 20 x (20 x 20) x 1 = 8000 features is extracted, and the features are the number of A_AA, A_AC, A_AD, A_AE, A_AF, A_AG, A_AI, A_AK, A_AL, A_AM, A_AN, A_AP, A_AQ, A_AR, A_AS, A_AT, A_AV, A_AW, A_AY, A_CA, A_CC, …, Y_YA, Y_YC, Y_YD, Y_YE, Y_YF, Y_YG, Y_YH, Y_YI, Y_YK, Y_YL, Y_YM, Y_YN, Y_YP, Y_YQ, Y_YR, Y_YS, Y_YT, Y_YV, Y_YW, and Y_YY that are present the whole peptide sequence.

For 1-gapped Di-Mono having k-gap = 1, a total of 20 x (20 x 20) x 1 = 8000 features is extracted, and the features are the number of AA_A, AA_C, AA_D, AA_E, AA_F, AA_G, AA_H, AA_I, AA_K, AA_L, AA_M, AA_N, AA_P, AA_Q, AA_R, AA_S, AA_T, AA_V, AA_W, AA_Y, …, YY_A, YY_C, YY_D, YY_E, YY_F, YY_G, YY_H, YY_I, YY_K, YY_L, YY_M, YY_N, YY_P, YY_Q, YY_R, YY_S, YY_T, YY_V, YY_W, and YY_Y that are present in the whole peptide sequence.

We extracted these features using PyFeat, a toolkit implemented in Python for extracting various features from proteins, DNAs, and RNAs [49]. Table 1 shows the depiction of the features extracted using the employed feature extraction technique.

2.2.3. Binary profile feature

Binary profile feature (BPF) is a straightforward technique yet proves to be quite effective in the prediction of different functionalities from multi-omics data [38,50]. We generate Binary profiles for each peptide, by representing each amino acid as a vector of 20 dimensions in terms of one hot encoding. For instance, Cytosine can be written as a 20-size one hot vector which is [0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]. A sequence of length M can be represented by a vector of dimensions $M \times 20$. As the maximum size of peptide sequences in our datasets is 50 residues, we get 50 × 20, i.e., 1000 features for each peptide sequence. We padded the peptides that are shorter than 50 amino acids with dummy amino acid "X". We encoded this dummy amino acid with the [0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0] vector. In this way, we make sure that the lengths of sequences are equal and no redundant information is added.

2.3. Classifier

2.3.1. Random forest

Random forest is a meta-classifier that consists of a number of decision tree classifiers (referred to as base learners) trained on various sub-samples of training data that are generated based on the concept of bagging to solve regression and classification. It was first proposed

Table 4
Results achieved by different classifiers using the K-mer feature group.

Methods	Main dataset				Alternative dataset			
	Ac	Sn	Sp	MCC	Ac	Sn	Sp	MCC
LR	68.9	59.3	78.5	0.39	88.9	85.6	92.3	0.78
DT	68.3	73.3	63.4	0.37	85.1	84.5	85.6	0.70
KNN	74.1	73.8	74.4	0.48	90.2	86.1	94.3	0.81
SVM	74.1	70.3	77.9	0.48	91.8	89.2	94.3	0.84
NB	70.3	69.8	71.0	0.41	87.4	76.8	97.9	0.77
RF	75.6	74.4	76.7	0.51	92.0	88.7	95.4	0.84

Table 5
Results achieved by different classifiers using k-gapped mono-Di and k-gapped Di-mono feature groups.

Methods	Main dataset				Alternative dataset			
	Ac	Sn	Sp	MCC	Ac	Sn	Sp	MCC
LR	73.8	76.7	70.9	0.48	89.7	84.5	94.8	0.80
DT	62.5	79.7	45.3	0.27	67.0	91.2	42.8	0.40
KNN	57.3	95.9	18.6	0.23	49.7	99.5	0.0	−0.10
SVM	70.9	70.9	70.9	0.42	85.8	72.2	99.5	0.74
NB	73.0	65.1	80.8	0.47	83.2	68.6	97.9	0.70
RF	71.2	84.3	58.1	0.44	72.2	97.4	46.9	0.51

Table 6
Performance of the ensemble of different ML methods for both main and alternate dataset using K-mers, BPF, K-mers & Gapped K-mers as features.

Models	Main dataset				Alternate dataset			
	Ac	Sn	Sp	MCC	Ac	Sn	Sp	MCC
LR + LR + LR	75.3	74.4	76.2	0.51	90.2	87.1	93.3	0.81
DT + DT + DT	74.4	76.2	72.7	0.49	87.1	85.6	88.7	0.74
NB + NB + NB	71.5	72.1	70.9	0.43	86.6	75.8	97.4	0.75
KNN + KNN + KNN	69.2	95.3	43.0	0.45	66.5	97.4	35.6	0.42
SVM + SVM + SVM	72.7	73.3	72.1	0.45	90.0	82.0	98.0	0.81
AB + AB + AB	75.0	73.3	76.7	0.51	88.7	89.7	87.6	0.77
RoF + RoF + RoF	71.8	72.1	71.5	0.44	85.8	86.1	85.6	0.72
GBT + GBT + GBT	74.1	73.8	74.4	0.48	92.3	93.8	90.7	0.85
RF + RF + RF	**75.9**	**75.6**	**76.2**	**0.52**	**93.1**	89.2	**96.9**	**0.86**

in [51]. In bagging, the available training data is randomly subsampled through a technique called baggining to generate different subsamples from the original data. Random Forest estimates the outcome based on averaging the predictions of its base learners. RF has been widely used in similar studies and obtained promising results [16,29–31].

2.3.2. Ensemble classifier

Machine learning has been widely used to tackle different problems in biological science including genomics, proteomics, microarrays, systems biology, evolution, and text mining [52–55]. Among different ML approaches, ensemble classifiers are considered among the most effective ones. Ensemble learning is a concept to train multiple classifiers and combine their predictions as a single classifier. In general, it is expected that the output of the ensemble classifier to be better compared to any of its ensemble members with uncorrelated error on the target data sets [56]. Ensemble models were originally designed to reduce the variance which results in the improvement of the performance. Where variance indicates the performance change of a model when it fits with a different set of data. An ideal machine learning model is considered to have low variance and low bias and these two are affected by one another. From previous studies it is evident that some ensemble techniques reduce the error of both bias and variance parts, consequently, improving prediction performance [57,58]. Ensemble classifiers have been shown effective in enhancing prediction performance for different problems in bioinformatics as well [59–64].

To predict the ACP sites with precision, various types of models have been used. We investigate the effectiveness of several classifiers using single feature extraction methods (k-mer, Binary Profile Feature, and k-gapped Mono-Di and Di-Mono features, separately) to predict ACPs. We observed satisfactory results in some cases, with better sensitivity or specificity as shown in Tables 2–5. However, the achieved results are biased toward negative samples, which shows that a combination of models is the next best approach to consider.

In this paper, we use an ensemble of three Random Forest (RF) classifiers which are trained heterogeneously using different feature sets to predict the ACPs. We aggregate the final output of these classifiers using majority voting. We extract MonoMer, DiMer, TriMer and feed them to the first RF model, Binary profile feature into the second RF, and a combination of K-mers, 1-gapped Mono-Di, and 1-gapped Di-Mono to the third RF as input feature vectors. To build our model, we have also studied several popular classification techniques such as Linear Regression (LR), K-Nearest Neighbor (KNN), Naïve Bayes (NB), Decision Tree (DT), and Support Vector Machines (SVM) which are widely used for similar problems and attained promising results [65]. However, among all these classifiers, an ensemble of RF classifiers attained the best results and significantly outperformed other classifiers. We have investigated a different number of base learners for our employed RF classifiers. Of all the variations of base learners, using 150, 250, and 400 for our three RF models, we obtained the best results. Since our employed dataset is considerably small, we experimented with unpruned trees where the nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples, which is two in our experiment. The remaining hyperparameters were kept as default, which are:

criterion ='gini', max_depth = None, min_samples_split = 2, min_samples_leaf = 1, min_weight_fraction_leaf = 0.0, max_features = 'sqrt', max_leaf_nodes = None, min_impurity_decrease = 0.0, bootstrap = True, oob_score = False, n_jobs=None, random_state = None, verbose = 0, warm_start = False, class_weight = None, ccp_alpha = 0.0, max_samples = None. The general architecture of iACP-RF is shown in Fig. 1.

3. Result analysis

3.1. Evaluation metrics

Evaluating the performance of a prediction method is crucial to find its reliability and generality with respect to the experimental dataset. To evaluate the performance of our model and to compare our results with previous studies, we use different measurements including sensitivity (Sn), specificity (Sp), accuracy (Ac), and Matthews correlation coefficient (MCC). These measurements are calculated using the following equations:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} * 100 \quad (3)$$

$$Sensitivity = \frac{tp}{tp + fn} * 100 \quad (4)$$

$$Specificity = \frac{tn}{tn + fp} * 100 \quad (5)$$

$$MCC = \frac{(tp \times tn) - (fp \times fn)}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}} \quad (6)$$

Where *tp* represents the total true positive predictions, *tn* represents the total true negative predictions, *fp* represents the false positive predictions, and *fn* represents the false negative predictions.

3.2. Comparison with different ML approaches to build our ensemble classifier

First, to build our ensemble model, we compare different machine learning classifiers including SVM, LR, DT, NB, KNN, and RF. We further studied several other classifiers including the Adaboost (AB), Rotation Forest (RoF), and Gradient boosting trees (GBT). The results achieved for the best combinations of different classifiers are presented in Table 6. As shown in the table, using an ensemble of heterogeneous RF classifiers, we obtain the best results. Hence, we use this classifier to build iACP-RF.

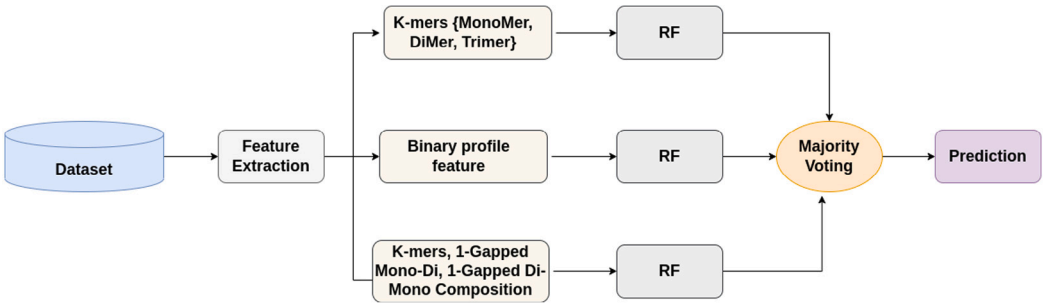


Fig. 1. The general architecture of iACP-RF as an ensemble of heterogeneously trained RF classifiers to predict ACP sites.

Table 7
Performance of individual Random Forest models using different feature sets.

Methods	Main dataset				Alternate dataset			
	Ac	Sn	Sp	MCC	Ac	Sn	Sp	MCC
K-mer, for k value of 1,2,3 (MonoMer, DiMer, TriMer) 8420 feature	75.6	74.4	76.7	0.51	92.0	88.7	95.4	0.84
K-gapped di-Mono 8000 feature	73.5	83.7	63.4	0.48	70.9	95.4	46.4	0.48
K-gapped mono-Di 8000 feature	**69.2**	**82.0**	56.4	0.40	74.7	96.4	53.1	0.55
K-mer and gapped k-mer 24420 feature	75.6	75.0	76.2	0.51	91.5	87.1	95.9	0.83
Binary Profile Feature 1000 feature	76.2	79.1	73.3	0.52	89.7	86.1	93.3	0.80

3.3. Investigating the impact of each classifier used to build iACP-RF

To investigate the performance of our proposed model, it was first crucial to find out how the individual Random Forest models perform using the given feature sets. The separate models showed impressive results using different types of features in terms of accuracy, sensitivity, specificity, and MCC scores, as shown in Table 7. However, the models are not consistent with their performance as the table depicts. Also, none of the individual RF models obtain better results than the combination of all three classifiers.

As shown in Table 7, RF trained on K-mers shows consistent results. However, it comes short of specificity on the alternate dataset. Whereas RF trained on K- gapped di-Mono achieves an underwhelming result with respect to specificity on both main and alternate datasets despite achieving outstanding results in terms of sensitivity. They scored 83.7% and 95.4% in both the main and alternate datasets, respectively which was an increase of 11.1% and 7.8% in sensitivity compared to iACP-FSCM which is considered the state-of-the-art ACP predictor. Similarly, RF trained using K-gapped Mono-Di is also capable of achieving high true positives rate of 82.0% and 96.4%, respectively on the main and alternate datasets, which are an increase of 9.4% and 8.8%, respectively compared to iACP-FSCM. However, it achieves poor specificity scores. Promising results used for each individual classifier demonstrate the effectiveness of our proposed features and employed classifiers to tackle this problem. Using K-mer and Gapped K-mer (24420 feature) achieves an increase of 2.4% on the main dataset in terms of sensitivity compared to iACP-FSCM. In addition, using Binary Profile Feature (1000 feature) we obtained an increase of 6.5% in sensitivity on the main dataset compared to iACP-FSCM.

By studying different feature sets for separate models, we demonstrate that single independent models are not able to achieve consistent results in terms of all the metrics used in this study for evaluation measurements. By using an ensemble of heterogeneously trained Random Forest methods, we achieve consistent performance for both the main and alternate datasets, with 75.6% and 89.2% in terms of sensitivity and 76.2% and 96.9% in terms of specificity, respectively. Our MCC scores also outperform iACP-FSCM by 0.04, 0.03, 0.06, and 0.09 respectively as shown in Table 8.

These results show that not only our proposed features and employed classifiers are able to achieve promising results to tackle this problem, but also our proposed ensemble of heterogeneously trained classifiers can enhance the prediction performance with respect to all metrics reported in this study compared to previous studies found in the literature.

3.4. Comparison with other state-of-the-art approaches

We then compare the results achieved by iACP-RF on both main and alternative datasets to the state-of-the-art methods found in the literature to predict anticancer peptides. The results for this comparison are presented in Table 9. As shown in this table, iACP-RF significantly outperforms iACP-FSCM which is the most recent and accurate ACP predictor on the alternative dataset. iACP-RF achieves promising results especially in terms of sensitivity on the main dataset compared to iACP-FSCM. iACP-RF demonstrates 4.2%, 1.6%, 6.7%, and 0.09 enhancements in terms of accuracy, sensitivity, specificity, and MCC, respectively over iACP-FSCM on the alternative dataset.

Although, in general, iACP-FSCM demonstrates better results on the main dataset compared to our proposed model, as shown in Table 9, iACP-RF achieved 75.6% in terms of Sensitivity compared to 72.6% for iACP-FSCM. It shows that our model is better than determining actual ACP sites. Considering that the main aim of this study is to have better performance in predicting positive samples, iACP-RF can be considered a model with better precision. Our Receiver operating characteristic (ROC) curves in Fig. 3 show that our model predicts the positive instances, with the Area Under the Curve (AUC) of 0.85 on the main dataset, and 0.96 on the alternate dataset. Fig. 2 shows the confusion matrix for the testing data. As shown in this figure, iACP-RF can be recognized as a model of good precision.

Note that although iACP and ACPred achieve better sensitivity than our model, they perform very poorly on negative samples which in turn, results in low specificity and consequently, very poor MCC. This result is mainly related to the dataset that they used to build their model and how they trained with a significant bias toward positive samples. In general, considering the significantly better MCC for our model compared to these two models, we can infer that iACP-RF is more accurate than these two models for predicting ACPs. Although ACPred-Fuse showed their performance to exceed the other existing models [43], we are able to outperform their result. iACP-RF also outperforms ACPred-Fuse in all the metrics for alternate and main datasets by a significant margin.

3.5. Performance of proposed model on external dataset

To further investigate the effectiveness of our proposed method for predicting ACPs, we tested our model on three external datasets used in recent studies [66–68]. Table 10 shows the experimental results using the datasets collected from various studies. Our proposed method shows stable prediction performance in all the datasets including the main and alternate datasets used in this study using 5-fold cross-validation.

Table 8
Performance of selected individual and ensemble model.

Methods	Main dataset				Alternate dataset			
	Ac	Sn	Sp	MCC	Ac	Sn	Sp	MCC
RF1 (K-mer-8420)	75.6	74.4	76.7	0.51	92.0	88.7	95.4	0.84
RF2 (BPF-1000)	76.2	79.1	73.3	0.52	89.7	86.1	93.3	0.80
RF3 (K-mer and gapped k-mer- 24420)	75.6	75.0	76.2	0.51	91.5	87.1	95.9	0.83
iACP-RF (ensemble of RF1, RF2, RF3)	75.9	**75.6**	76.2	0.52	**93.1**	**89.2**	**96.9**	**0.86**

Table 9
Performance comparison of pre-existing models with our proposed method using both main and alternate datasets.

Methods	Main dataset				Alternative dataset			
	Ac	Sn	Sp	MCC	Ac	Sn	Sp	MCC
AntiCP	50.6	100	1.2	0.07	90.0	89.7	90.2	0.80
iACP	55.1	77.9	32.2	0.11	77.6	78.4	76.8	0.55
ACPred	53.5	85.6	21.4	0.09	85.3	87.1	83.5	0.71
PEPred-Suite	53.5	33.1	73.8	0.08	57.5	40.2	74.7	0.16
ACPred-FL	44.8	67.1	22.5	−0.12	43.8	60.2	25.6	−0.15
ACPred-Fuse	68.9	69.2	68.6	0.38	78.9	64.4	93.3	0.60
iACP-FSCM	82.5	72.6	90.3	0.64	88.9	87.6	90.2	0.77
K-mer + BPF + Gapped K-mer (8420, 1000, 16000)	76.5	**81.4**	71.5	0.52	**92.0**	**91.8**	92.3	**0.84**
K-mer + BPF + K-mer (8420, 1000, 8420)	75.6	75	76.2	0.51	91.8	88.2	95.4	0.83
iACP-RF (K-mer + BPF + K-mer & Gapped K-mer,) (8420, 1000, and 24420 respectively)	75.9	**75.6**	76.2	0.52	**93.1**	**89.2**	**96.9**	**0.86**

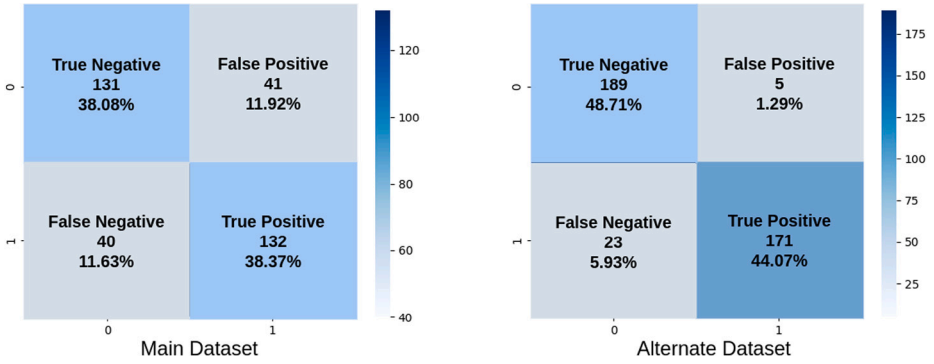


Fig. 2. The confusion matrix for iACP-RF for main and alternate datasets.

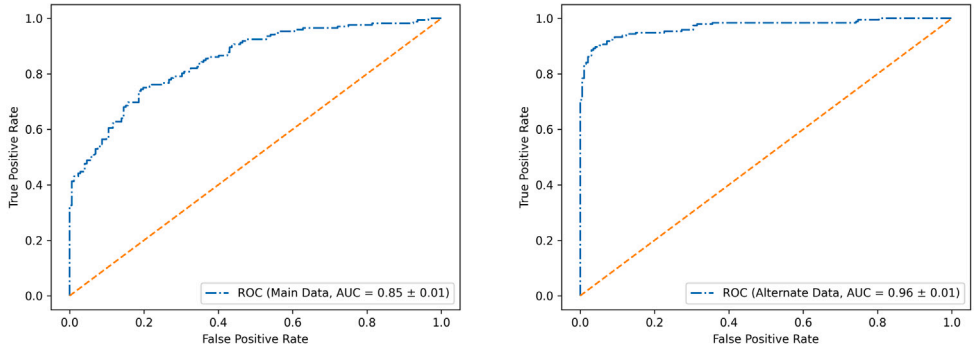


Fig. 3. Receiver operating characteristic (ROC) curves for independent datasets (main and alternate).

4. Discussion

In recent years, peptide-based therapy has emerged as a novel and promising strategy for the treatment of cancer. It has several advantages like high target specificity, low toxicity, good efficacy, easily synthesized and modified, and less immunogenic when combined with recombinant antibodies compared to conventional approaches. As it is challenging to discover ACP from protein sequence data using experimental methods, which emphasizes on the rapid advancement of computational methods due to its efficient nature.

In this paper, we proposed a novel prediction method named iACP-RF to accurately predict anticancer peptides. Our model demonstrates

Table 10
Results achieved by iACP-RF on external datasets used in previous studies using 5-fold cross-validation.

Dataset		Ac	Sn	Sp	MCC	F-1 score
Main (iACP-RF)		85.4	83.0	88.2	0.72	0.86
Alternate (iACP-RF)		97.0	95.7	98.1	0.93	0.97
ACPNet	iACP-RF	**91.0**	**88.1**	**94.4**	**0.82**	**0.91**
	ACPNet	89.6	87.8	91.4	0.79	89.4
Alsanea et al.	iACP-RF	95.2	95.0	**100**	0.66	**97.4**
	Alsanea et al.	97.1	97.4	96.9	0.93	96.2
StackACPred	iACP-RF	**91.9**	**90.2**	**93.6**	**0.83**	**0.92**
	StackACPred	84.5	84.1	84.9	0.70	–

promising performance on the main and alternate datasets. Furthermore, it shows its effectiveness in distinguishing ACPs from non-ACPs on external datasets compared to previous studies.

Although the performance of individual features showed promising results with single classifiers, the results were imbalanced. However, empirical studies show that using ensemble models reduces both bias and variance to improve prediction accuracy. Thus, we experimented using a combination of several sequence-based features namely K-mer, BPF, 1-Gapped Di-Mono, and 1-Gapped Mono-Di, and achieved better outcomes compared to existing methods. Among different combinations of features being studied to build our model, using K-mer, BPF, K-mer + 1-Gapped Di-Mono + 1-Gapped Mono-Di respectively, feeding into our heterogeneously trained base classifiers RF1, RF2, and RF3 models combined using majority voting, the best performance was achieved.

Even so, a critical challenge in the machine learning pipeline when working with a small amount of data is that the model can overfit on the training data and be biased toward the dominant class. In this study, we used a balanced dataset consisting of the same number of samples in the positive and negative classes, which helps in getting a balanced prediction for both classes. Testing the model's performance using two independent datasets and three external datasets, along with a 5-fold CV with high and consistent performance, proves the model is performing in an optimal manner avoiding overfitting.

Despite the merits of our proposed method, it has several limitations. First, tuning the parameters to get optimal performance requires more data. Since the dataset we worked with contains a handful of samples, tuning the parameters optimally was not feasible. Second, finding the optimal number of classifiers to ensemble is critical and there is no conventional way to find the optimal number of base learners. Finally, the commonly used evaluation metrics used to evaluate the performance of binary classifiers can be too specific. To address these limitations and mitigate these issues we aim to build an explainable machine learning pipeline in the future for predicting anticancer peptides.

5. Conclusion

Anticancer peptides play a crucial role in the study of anticancer drugs and the treatment of cancer. Targeting cancer cells is essential in the treatment of cancer. However, a lack of "guiding missiles" to target such cells leads to less effective treatment progress. Peptide properties can be used both in molecularly targeted drugs and 'guiding missiles' to inhibit cell proliferation or eradicate cancer cells completely. In this paper, we proposed an ensemble of heterogeneously trained Random Forest models for predicting ACPs using a combination of several sequence-based features namely K-mer, Binary profile feature, 1-Gapped Di-Mono, and 1-Gapped Mono-Di. iACP-RF tool outperforms existing methods by a significant margin for all the metrics in the alternate dataset and shows an enhancement of 3% in terms of sensitivity for the main dataset. On the alternate dataset, we outperform iACP-FSCM in all counts of accuracy, sensitivity, specificity, and MCC score by a margin of 5.5%, 1.6%, 6.7%, and 0.09, respectively. Our results demonstrate the effectiveness of iACP-RF in predicting anticancer peptides compared to previously proposed models found in the literature. iACP-RF as a standalone predictor and all its source code are publicly available at: https://github.com/MLBC-lab/iACP-RF.

Funding sources

This material is based upon work supported by the National Science Foundation under Grant No. 2152059.

Declaration of competing interest

None.

References

[1] Hazelton William D, Luebeck E Georg. Biomarker-based early cancer detection: Is it achievable? Sci Transl Med 2011;3(109):109fs9.
[2] Virnig Beth A, Baxter Nancy N, Habermann Elizabeth B, Feldman Roger D, Bradley Cathy J. A matter of race: Early-versus late-stage cancer diagnosis. Health Aff 2009;28(1):160–8.
[3] Omenn Gilbert S. Strategies for genomic and proteomic profiling of cancers. Stat Biosci 2016;8(1):1–7.
[4] Mahassni Sawsan Hassan, Al-Reemi Roaa Mahdi. Apoptosis and necrosis of human breast cancer cells by an aqueous extract of garden cress *(Lepidium sativum)* seeds. Saudi J Biol Sci 2013;20(2):131–9.
[5] Gerber Bernd, Freund Mathias, Reimer Toralf. Recurrent breast cancer: Treatment strategies for maintaining and prolonging good quality of life. Deutsches Arzteblatt Int 2010;107(6):85.
[6] Khan Samee Ullah, Baik Ran. MPPIF-net: Identification of plasmodium falciparum parasite mitochondrial proteins using deep features with multilayer bi-directional LSTM. Processes 2020;8(6):725.
[7] Thundimadathil Jyothi. Cancer treatment using peptides: Current therapies and future prospects. J Amino Acids 2012;2012.
[8] Marqus Susan, Pirogova Elena, Piva Terrence J. Evaluation of the use of therapeutic peptides for cancer treatment. J Biomed Sci 2017;24(1):1–15.
[9] McGregor Duncan Patrick. Discovering and improving novel peptide therapeutics. Curr Opin Pharmacol 2008;8(5):616–9.
[10] Schulte Imke, Tammen Harald, Selle Hartmut, Schulz-Knappe Peter. Peptides in body fluids and tissues as markers of disease. Exp Rev Mol Diagn 2005;5(2):145–57.
[11] Diamandis Eleftherios P. Peptidomics for cancer diagnosis: Present and future. J Proteome Res 2006;5(9):2079–82.
[12] Cicero Arrigo FG, Fogacci Federica, Colletti Alessandro. Potential role of bioactive peptides in prevention and treatment of chronic diseases: A narrative review. Br J Pharmacol 2017;174(11):1378–94.
[13] Agrawal Piyush, Bhalla Sherry, Usmani Salman Sadullah, Singh Sandeep, Chaudhary Kumardeep, Raghava Gajendra PS, et al. CPPsite 2.0: A repository of experimentally validated cell-penetrating peptides. Nucleic Acids Res 2016;44(D1):D1098–103.
[14] Mathur Deepika, Prakash Satya, Anand Priya, Kaur Harpreet, Agrawal Piyush, Mehta Ayesha, et al. PEPlife: A repository of the half-life of peptides. Sci Rep 2016;6(1):1–7.
[15] Agrawal Piyush, Singh Harinder, Srivastava Hemant Kumar, Singh Sandeep, Kishore Gaurav, Raghava Gajendra PS. Benchmarking of different molecular docking methods for protein-peptide docking. BMC Bioinform 2019;19(13):105–24.
[16] Schaduangrat Nalini, Nantasenamat Chanin, Prachayasittikul Virapong, Shoombuatong Watshara. ACPred: A computational tool for the prediction and analysis of anticancer peptides. Molecules 2019;24(10):1973.
[17] Usmani Salman Sadullah, Bedi Gursimran, Samuel Jesse S, Singh Sandeep, Kalra Sourav, Kumar Pawan, et al. THPdb: Database of FDA-approved peptide and protein therapeutics. PLoS One 2017;12(7):e0181748.
[18] Chiangjong Wararat, Chutipongtanate Somchai, Hongeng Suradej. Anticancer peptide: Physicochemical property, functional aspect and trend in clinical application. Int J Oncol 2020;57(3):678–96.
[19] Xie Jing, Bi Ye, Zhang Huan, Dong Shiyan, Teng Lesheng, Lee Robert J, et al. Cell-penetrating peptides in diagnosis and treatment of human diseases: From preclinical research to clinical application. Front Pharmacol 2020;11:697.
[20] Luan Chi-Hao, Parker Timothy M, Gowda D Channe, Urry Dan W. Hydrophobicity of amino acid residues: Differential scanning calorimetry and synthesis of the aromatic analogues of the polypentapeptide of elastin. Biopolymers: Orig Res Biomol 1992;32(9):1251–61.
[21] Hoskin David W, Ramamoorthy Ayyalusamy. Studies on anticancer activities of antimicrobial peptides. Biochim Biophys Acta (BBA)-Biomembranes 2008;1778(2):357–75.

[22] Gaspar Diana, Veiga A Salomé, Castanho Miguel ARB. From antimicrobial to anticancer peptides. A review. Front Microbiol 2013;4:294.

[23] Deslouches Berthony, Di Y Peter. Antimicrobial peptides with selective antitumor mechanisms: Prospect for anticancer applications. Oncotarget 2017;8(28):46635.

[24] Sok Miha, Šentjurc Marjeta, Schara Milan. Membrane fluidity characteristics of human lung cancer. Cancer Lett 1999;139(2):215–20.

[25] Yoon Wan-Hee, Park Hae-Duck, Lim Kyu, Hwang Byung-Doo. Effect of O-glycosylated mucin on invasion and metastasis of HM7 human colon cancer cells. Biochem Biophys Res Commun 1996;222(3):694–9.

[26] Ran Sophia, Downes Amber, Thorpe Philip E. Increased exposure of anionic phospholipids on the surface of tumor blood vessels. Cancer Res 2002;62(21):6132–40.

[27] Dobrzyńska Izabela, Szachowicz-Petelska Barbara, Sulkowski Stanisław, Figaszewski Zbigniew. Changes in electric charge and phospholipids composition in human colorectal cancer cells. Mol Cell Biochem 2005;276(1):113–9.

[28] Felício Mário R, Silva Osmar N, Gonçalves Sônia, Santos Nuno C, Franco Octávio L. Peptides with dual antimicrobial and anticancer activities. Front Chem 2017;5:5.

[29] Manavalan Balachandran, Basith Shaherin, Shin Tae Hwan, Choi Sun, Kim Myeong Ok, Lee Gwang. MLACP: Machine-learning-based prediction of anticancer peptides. Oncotarget 2017;8(44):77121.

[30] Wei Leyi, Zhou Chen, Su Ran, Zou Quan. PEPred-Suite: Improved and robust prediction of therapeutic peptides using adaptive feature representation learning. Bioinformatics 2019;35(21):4272–80.

[31] Dehzangi Abdollah, Phon-Amnuaisuk Somnuk, Dehzangi Omid. Using random forest for protein fold prediction problem: An empirical study. J Inf Sci Eng 2010;26(6):1941–56.

[32] Tyagi Atul, Kapoor Pallavi, Kumar Rahul, Chaudhary Kumardeep, Gautam Ankur, Raghava GPS. In silico models for designing and discovering novel anticancer peptides. Sci Rep 2013;3(1):1–8.

[33] Chen Wei, Ding Hui, Feng Pengmian, Lin Hao, Chou Kuo-Chen. iACP: A sequence-based tool for identifying anticancer peptides. Oncotarget 2016;7(13):16895.

[34] Vijayakumar Saravanan, Ptv Lakshmi. ACPP: A web server for prediction and design of anti-cancer peptides. Int J Pept Res Therapeutics 2015;21(1):99–106.

[35] Akbar Shahid, Hayat Maqsood, Iqbal Muhammad, Jan Mian Ahmad. iACP-GAEnsC: Evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space. Artif Intell Med 2017;79:62–70.

[36] Xu Lei, Liang Guangmin, Wang Longjie, Liao Changrui. A novel hybrid sequence-based model for identifying anticancer peptides. Genes 2018;9(3):158.

[37] Kabir Muhammad, Arif Muhammad, Ahmad Saeed, Ali Zakir, Swati Zar Nawab Khan, Yu Dong-Jun. Intelligent computational method for discrimination of anticancer peptides by incorporating sequential and evolutionary profiles information. Chemometr Intell Lab Syst 2018;182:158–65.

[38] Yi Hai-Cheng, You Zhu-Hong, Zhou Xi, Cheng Li, Li Xiao, Jiang Tong-Hai, et al. ACP-DL: A deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation. Mol Therapy-Nucleic Acids 2019;17:1–9.

[39] Wei Leyi, Zhou Chen, Chen Huangrong, Song Jiangning, Su Ran. ACPred-FL: A sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. Bioinformatics 2018;34(23):4007–16.

[40] Wu Chuanyan, Gao Rui, Zhang Yusen, De Marinis Yang. PTPD: Predicting therapeutic peptides by deep learning and word2vec. BMC Bioinformatics 2019;20(1):1–8.

[41] Hajisharifi Zohre, Piryaiee Moien, Beigi Majid Mohammad, Behbahani Mandana, Mohabatkar Hassan. Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. J Theoret Biol 2014;341:34–40.

[42] Li Feng-Min, Wang Xiao-Qian. Identifying anticancer peptides by using improved hybrid compositions. Sci Rep 2016;6(1):1–6.

[43] Rao Bing, Zhou Chen, Zhang Guoying, Su Ran, Wei Leyi. ACPred-Fuse: Fusing multi-view information improves the prediction of anticancer peptides. Brief Bioinform 2020;21(5):1846–55.

[44] Akbar Shahid, Rahman Ateeq Ur, Hayat Maqsood, Sohail Mohammad. cACP: Classifying anticancer peptides using discriminative intelligent model via Chou's 5-step rules and general pseudo components. Chemometr Intell Lab Syst 2020;196:103912.

[45] Akbar Shahid, Hayat Maqsood, Tahir Muhammad, Chong Kil To. cACP-2LFS: classification of anticancer peptides using sequential discriminative model of KSAAP and two-level feature selection approach. IEEE Access 2020;8:131939–48.

[46] Akbar Shahid, Hayat Maqsood, Tahir Muhammad, Khan Salman, Alarfaj Fawaz Khaled. cACP-DeepGram: Classification of anticancer peptides via deep neural network and skip-gram-based word embedding model. Artif Intell Med 2022;131:102349.

[47] Gautam Ankur, Chaudhary Kumardeep, Kumar Rahul, Sharma Arun, Kapoor Pallavi, Tyagi Atul, et al. In silico approaches for designing highly effective cell penetrating peptides. J Transl Med 2013;11(1):1–12.

[48] Charoenkwan Phasit, Chiangjong Wararat, Lee Vannajan Sanghiran, Nantasenamat Chanin, Hasan Md, Shoombuatong Watshara, et al. Improved prediction and characterization of anticancer activities of peptides using a novel flexible scoring card method. Sci Rep 2021;11(1):1–13.

[49] Muhammod Rafsanjani, Ahmed Sajid, Md Farid Dewan, Shatabda Swakkhar, Sharma Alok, Dehzangi Abdollah. PyFeat: A Python-based effective feature generation tool for DNA, RNA and protein sequences. Bioinformatics 2019;35(19):3831–3.

[50] Azim Sayed Mehedi, Sharma Alok, Noshadi Iman, Shatabda Swakkhar, Dehzangi Iman. A convolutional neural network based tool for predicting protein AMPylation sites from binary profile representation. Sci Rep 2022;12(1):1–7.

[51] Breiman Leo. Random forests. Mach Learn 2001;45(1):5–32.

[52] Larranaga Pedro, Calvo Borja, Santana Roberto, Bielza Concha, Galdiano Josu, Inza Inaki, et al. Machine learning in bioinformatics. Brief Bioinform 2006;7(1):86–112.

[53] Chen Xing, Li Tian-Hao, Zhao Yan, Wang Chun-Chun, Zhu Chi-Chi. Deep-belief network for predicting potential miRNA-disease associations. Brief Bioinform 2021;22(3):bbaa186.

[54] Ha Jihwan, Park Chihyun. MLMD: Metric learning for predicting miRNA-disease associations. IEEE Access 2021;9:78847–58.

[55] Ha Jihwan. MDMF: Predicting miRNA–Disease association based on matrix factorization with disease similarity constraint. J Personal Med 2022;12(6):885.

[56] Yang Yun. Temporal data mining via unsupervised ensemble learning. Elsevier; 2016.

[57] Witten Daniela, James Gareth. An introduction to statistical learning with applications in R. springer publication; 2013.

[58] Rokach Lior. Pattern classification using ensemble methods, vol. 75. World Scientific; 2010.

[59] Yang Pengyi, Hwa Yang Yee, B. Zhou Bing, Y. Zomaya Albert. A review of ensemble methods in bioinformatics. Curr Bioinform 2010;5(4):296–308.

[60] Miah Md Ochiuddin, Muhammod Rafsanjani, Al Mamun Khondaker Abdullah, Farid Dewan Md, Kumar Shiu, Sharma Alok, et al. CluSem: Accurate clustering-based ensemble method to predict motor imagery tasks from multi-channel EEG data. J Neurosci Methods 2021;364:109373.

[61] Dehzangi Abdollah, Paliwal Kuldip, Sharma Alok, Dehzangi Omid, Sattar Abdul. A combination of feature extraction methods with an ensemble of different classifiers for protein structural class prediction problem. IEEE/ACM Trans Comput Biol Bioinform 2013;10(3):564–75.

[62] Dehzangi Abdollah, Phon-Amnuaisuk Somnuk, Dehzangi Omid. Enhancing protein fold prediction accuracy by using ensemble of different classifiers. Aust J Intell Inf Process Syst 2010;26(4):32–40.

[63] Dehzangi Abdollah, Karamizadeh Sasan. Solving protein fold prediction problem using fusion of heterogeneous classifiers. Information, Int Interdiscip J 2011;14(11):3611–22.

[64] Azim Sayed Mehedi, Haque Md Rakibul, Shatabda Swakkhar. Oric-ens: A sequence-based ensemble classifier for predicting origin of replication in s. Cerevisiae. Comput Biol Chem 2021;92:107502.

[65] Dai Wei, Chang Qi, Peng Wei, Zhong Jiancheng, Li Yongjiang. Identifying human essential genes by network embedding protein-protein interaction network. In: International symposium on bioinformatics research and applications. Springer; 2019, p. 127–37.

[66] Arif Muhammad, Ahmed Saeed, Ge Fang, Kabir Muhammad, Khan Yaser Daanial, Yu Dong-Jun, et al. StackACPred: Prediction of anticancer peptides by integrating optimized multiple feature descriptors with stacked ensemble approach. Chemometr Intell Lab Syst 2022;220:104458.

[67] Sun Mingwei, Yang Sen, Hu Xuemei, Zhou You. ACPNet: A deep learning network to identify anticancer peptides by hybrid sequence information. Molecules 2022;27(5):1544.

[68] Alsanea Majed, Dukyil Abdulsalam S, Riaz Bushra, Alebeisat Farhan, Islam Muhammad, Habib Shabana. To assist oncologists: An efficient machine learning-based approach for anti-cancer peptides classification. Sensors 2022;22(11):4005.