IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 30, 2020, accepted July 7, 2020, date of publication July 14, 2020, date of current version July 29, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3009125

cACP-2LFS: Classification of Anticancer Peptides Using Sequential Discriminative Model of KSAAP and Two-Level Feature Selection Approach

SHAHID AKBAR[1], MAQSOOD HAYAT[1], MUHAMMAD TAHIR [1], AND KIL TO CHONG[2]

[1]Department of Computer Science, Abdul Wali Khan University Mardan, Mardan 23200, Pakistan
[2]Department of Electronic and Information Engineering, Chonbuk National University, Jeonju 54896, South Korea

Corresponding authors: Maqsood Hayat (m.hayat@awkum.edu.pk) and Kil To Chong (kitchong@jbnu.ac.kr)

This work was supported by the Brain Research Program of the National Research Foundation (NRF) through the Korean Government (MSIT) under Grant NRF-2017M3C7A1044815.

ABSTRACT Cancer is a leading killer disease globally, it occurs when the cellular changes cause the abnormal growth and division of the cells. Conventional treatment such as therapies and wet experimental methods are deemed unsatisfactory and worthless because of its huge cost and laborious nature. However, the recent innovation of anticancer peptides (ACPs) offers an effective way to treat cancer affected cells. Due to the rapid growth of biological sequences, truly identification of ACPs has become a difficult task for scientists. Therefore, measuring the importance of ACPs, an efficient and reliable intelligent model is highly essential to accurately identify its pattern. In this study, three distinct nature encoding schemes are employed to obtain features from peptide sequences. However, K-space amino acid pair (KSAAP) is used to extract highly correlated and effective descriptors. Apart from the sequential features, composite physiochemical properties are applied to gather local structure descriptors. Furthermore, to represent the intrinsic residue information of amino acids, autocovariance is also used. Additionally, a novel two-level feature selection (2LFS) method is utilized to select high discriminative features and to minimize the dimensionality of the proposed descriptors. At last, to examine the performance of the proposed model, several learning hypotheses are investigated to select a superior operational engine. To measure the generalization capability, two diverse benchmark datasets are used. After evaluating the empirical outcomes, KSAAP using 2LFS reported high classification results on both datasets. Whereas, the classification outcomes reveal that our proposed cACP-2LFS achieved ∼11% improved performance accuracy than present models in the literature so far. It is expected that our proposed model might be useful in the area of medicine, proteomics, and research academia. The source code and all datasets are publicly available at https://github.com/shahidawkum/cACP-2LFS.

INDEX TERMS Anticancer peptides, support vector machine, K-space amino acid pair, two-level feature selection, composite physiochemical properties, classification.

I. INTRODUCTION

Cancer is the major health concern globally [1], and every year millions of death occur due to this devastating disease worldwide [2]. According to the recent statistics in 2018, about 9.6 million deaths occur due to cancer [3]. Additionally, the ratio of cancer deaths reported in developing countries is relatively high. To treat cancer patients various numbers of traditional therapies i.e., radiation, hormonal, and chemotherapy has been applied so far. However, these methods are ineffective because of their expensive cost and severe side impacts on the normal cells [4], [5]. Keeping these issues, the discovery of anticancer peptides (ACPs) are deemed as an alternative for the accurate identification of cancer affected cells. ACPs are usually the small fragment of a protein sequence called peptide containing less than 50 amino acids [6]. whereas, ACPs are mostly collected from antimicrobial peptides (AMPs) and have similar basic

The associate editor coordinating the review of this manuscript and approving it for publication was Fan-Hsun Tseng.

characteristics with ACPs [7]. Currently, various peptide-based therapies have been applied to treat various tumor types using clinical and preclinical trials [8], [9]. However, the identification of ACPs from the peptide sequences using experimental approaches is time-consuming, costly, and difficult to be applied in a high-throughput manner [10]. Therefore, the recent efforts have mainly focused on the development of computational methods, especially machine learning-based methods in order to expedite the identification of ACPs. Over the last decade, various intelligent statistical-based models have been proposed in the literature to accurately identify ACPs [11]–[14]. Hajisharifi *et al.*, generated a non-redundant training dataset, the Hajisharifi-Chen (HC) dataset, which contained 138 ACPs and 206 non-ACPs, whereas the biological sequences were formulated using pseudo amino acid composition (PseAAC) and local-alignment based kernel for the identification of ACPs [15]. The proposed model was evaluated using the support vector machine (SVM) and achieved an accuracy of 89.70%. In continuation, Chen *et al.*, developed a sequence-based predictor namely, ‘iACP’ using a similar HC dataset [16]. Whereas, iACP proposed an optimized g-gap dipeptide feature vector in combination with SVM and reported a better performance accuracy of 94.70%. In a sequel, Akbar *et al.*, proposed a hybrid model, namely, “iACP-GAEnsC” using a similar HC dataset for the accurate classification of ACPs [17]. Where the integrated feature space is calculated using three formulation techniques such as Reduce amino acid alphabet, gapped dipeptide, and amphiphilic-PseAAC. Finally, the ensemble learning approach using a genetic algorithm was applied by combining the prediction rates of the five individual classifiers and reported high accuracy of 96.45%. Furthermore, Kabir *et al.* proposed a “TargetACP” model using the HC dataset, that extracts sequential and evolutionary descriptors from peptide sequences [18]. Besides, an oversampling method was also applied to reduce the biases of the majority class. The proposed model achieved a remarkable result of 98.78% using the training dataset. Moreover, Khan *et al.* formulated the ACPs sequences of the HC dataset by employing the Split Amino acid composition based sequential approach [19]. The extracted sequential features achieved the performance accuracy of 93.31% using the SVM classifier. Similarly, Xu et. al., used g-gap dipeptide and maximum relevance-maximum distance-based statistical model to identify ACPs [20]. The proposed sequential model was trained using the similar HC dataset and reported an accuracy of 91.86%. Besides, Manavalan *et al.* developed SVM and RF-based predictors for the identification of ACPs [21], where features are calculated from peptide sequences using amino acid composition, dipeptide composition, and composite physiochemical properties. The encoding features were examined using two different predictors and reported the performance accuracy of 87.20% using the training dataset. However, among all the computational models that were trained using the HC dataset, only TargetAC and iACP-GAEnsC achieved encouraging performances with high prediction accuracies, but the overall utility of these two methods have certain limitations in terms of high computational cost, Interpretability and practical utility thereby require further improvements. On the other hand, several other computational models were also developed using different datasets, such as Vijayakumar *et al.*, Presented a webserver namely, “ACPP” for the accurate identification of ACPs [22]. ACPP model was trained using their own generated dataset, which consists of 4276 sequences, among which 257 are categorized as ACPs, and 4019 were non-ACPs. To formulate the ACP sequence, various protein-relatedness measure (PRM) parameters measures, such as combining compositional descriptors, distribution of amino acids, and particular regions of the amino acid were targeted. Finally, the performance of the extracted features was examined using SVM and obtained an accuracy of 97%. However, apart from the accuracy, the other performance evaluation parameters such as sensitivity and specificity were unbalanced due to biases of ACPs in the training dataset. More recently, Akbar *et. al.*, [23] proposed another intelligent model for ACPs using the two different nature benchmark datasets, selected from the AntiCP predictor [24]. Whereas, highly effective correlated residue information’s of amino acids are extracted using quasi-sequence order. In addition, a principal component analysis was also utilized to minimize redundancy and irrelevant features. The proposed cACP model achieved an improved accuracy of 96.91% and 89.54% using main and alternate dataset. Derived from recent publications [25]–[29], investigators have extensively emphasized the following Chou steps for developing an effective prediction model. These are: (i) to select or develop a valid benchmark dataset to train and test a model (ii) to represent the peptides sequence to effectively reflect their intrinsic correlation with targeted class, (iii) to propose or select an effective classification learner (iv) to assess the success rates of the classifier using cross-validation test.

After investigating the existing models in the literature, it was found that most of the models used traditional methods such as AAC, g-gap DPC, and PseAAC to formulate peptide sequences. However, there were several flaws while processing short peptide sequences. i.e., in small peptide sequences, the performance of g-gap DPC may affect for the high value of g because some of the useful amino acid residues may be skipped. Similarly, PseAAC may also lose residue information by increasing tier values. Moreover, TargetACP used an oversampling approach to increase the instances of the minority class, which may cause overfitting issues and may discard useful data. Nonetheless, the performance of the aforementioned existing models is not adequate and needs more improvement. Therefore, to effectively deals with such issues, an accurate and computationally efficient classification model is proposed for anticancer peptides. Three distinct nature feature formulation methods, such as; K-space amino acid pair (KSAAP), Composite physiochemical properties (CPP), and auto-covariance (AC) is applied to collect valuable descriptors from ACP sequences. In contrast with other encoding schemes, KSAAP achieved improved prediction results. Whereas, KSAAP reflects the short-range interactions of amino acids among peptides sequences. On the other hand, to collect the high discriminative features, a novel two-level feature selection (2LFS) is utilized. 2LFS is an ensemble approach of the wrapper (SVM-RFE) and filter (mRMR) based feature selection methods. However, SVM-RFE collects the optimal features by computing the weights among extracted feature vectors. SVM-RFE uses the SVM training model to rank the features, which may increase the computational cost of the proposed model. Therefore, to make the model computationally effective, mRMR is utilized to deals with overfitting issues by eliminating irrelevant and redundant features. At last, the predictive results of our proposed model are examined using several classification algorithms, such as FKNN, SVM, and RF.

The rest of the paper is prepared as follows; section 2 explains methods and performance evaluation matrices; results and discussions are presented in section 3; and finally, the conclusion is represented at the end of the paper.

II. MATERIALS AND METHODS

A. DATASET

In the area of machine learning, the selection of a valid benchmark dataset is a rudimentary part for developing an intelligent predictor. However, choosing a suitable dataset has a high impact on performance rates. Measuring the effectiveness of the dataset on a computational model, we used two diverse nature benchmark datasets i.e., LEE dataset (S1) and independent dataset (S2). Whereas, both the datasets are divided into binary classes i.e., anticancer peptides (ACP) and non-anticancer peptides (non-ACP). LEE dataset is constructed using the screening procedure from various databases such as CancerPPDB [6] and APD3 [30]. However; most non-ACPs are selected from Tyagi independent datasets [24] and random selection from Swiss-Prot [31]. Thus, in total 844 unique sequences are used that are equally categorized into ACP and non-ACPs [21]. Whereas, the sequence length of most ACPs and non-ACPs are less than 25. On the other hand, to examine the generalization power of our proposed model, dataset S2 is also utilized, which consists of 150 ACPs (positive samples) and 150 non-ACPs (negative samples). The sequences of dataset S2 are selected from the Tyagi *et al.* [24] datasets and CancerPPDB [6]. Whereas, none of the sequences of the LEE dataset S1 is selected/repeated in independent dataset S2. Furthermore, the illegal amino acids such as, 'B', 'U', 'X', and 'Z', are eliminated from peptide sequences.

B. FEATURE EXTRACTION TECHNIQUES

1) K-SPACED AMINO ACID PAIRS (KSAAP)

K-spaced amino acid pairs (KSAAP) is an effective feature formulation scheme that highlights the valuable motif of protein fragments or sequences [32]–[34]. KSAAP is useful for identifying protein flexible or rigid regions and has been successfully applied for various post-translational modification sites [35], [36]. Consequently, measuring the significance of KSAAP descriptors over other formulation methods, KSAAP obtains valuable descriptors from the peptide sequences for the accurate classification of ACPs and non-ACPs. The detailed procedure of KSAAP is described as follows [34], [37]. For a protein fragment, it computes the occurrence frequency of amino acid pairs separated by K (j=0, 1, 2, …k) number of residue [38]. Whereas, the representation of the features is based on the frequency of k-spaced amino acid pairs in a local sequence window. For k=2, k-spaced pairs for j= 0, 1, and 2 are calculated. For each value of j, the corresponding feature spaces F_j i.e., F_0, F_1, and F$_2$ as shown in Eqs. (1), (2), and (3), respectively, are computed, each having a dimension of 441. The final feature space F can be calculated by combining the individual feature spaces as shown in Eq.(4). The value of each descriptor is computed by dividing the number of occurrences of that amino acid pair by the total number of j-spaced residue pairs $(N_0, N_1, \ldots .N_j)$ in the peptide. For j, $N_j = L-(j+1)$ where L represents is the length of the peptide sequence. In Figure 1, only a few windows have been illustrated to represent the mechanism of KSAAP for illustration.

$$F_0 = \left(\frac{M_{AA}}{N_0}, \frac{M_{AC}}{N_0}, \frac{M_{AD}}{N_0}, \ldots, \frac{M_{YY}}{N_0}\right)_{441} \tag{1}$$

$$F_1 = \left(\frac{M_{AxA}}{N_1}, \frac{M_{AxC}}{N_1}, \frac{M_{AxD}}{N_1}, \ldots, \frac{M_{YxY}}{N_1}\right)_{441} \tag{2}$$

$$F_2 = \left(\frac{M_{AxxA}}{N_2}, \frac{M_{AxxC}}{N_2}, \frac{M_{AxxD}}{N_2}, \ldots, \frac{M_{YxxY}}{N_2}\right)_{441} \tag{3}$$

$$F = F_0 + F_1 + \ldots, F_k \tag{4}$$

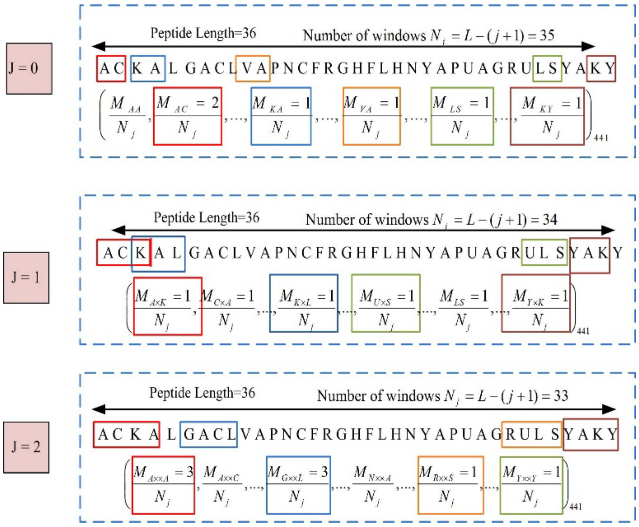
J = 0
Peptide Length=36
Number of windows $N_j = L-(j+1) = 35$
A C K A L G A C L V A P N C F R G H F L H N Y A P U A G R U L S Y A K Y
$\left(\frac{M_{AA}}{N_j}, \frac{M_{AC}=2}{N_j}, \ldots, \frac{M_{KA}=1}{N_j}, \ldots, \frac{M_{VA}=1}{N_j}, \ldots, \frac{M_{LS}=1}{N_j}, \ldots, \frac{M_{KY}=1}{N_j}\right)_{441}$
J = 1
Peptide Length=36
Number of windows $N_j = L-(j+1) = 34$
A C K A L G A C L V A P N C F R G H F L H N Y A P U A G R U L S Y A K Y
$\left(\frac{M_{A\times K}=1}{N_j}, \frac{M_{C\times A}=1}{N_j}, \ldots, \frac{M_{K\times L}=1}{N_j}, \ldots, \frac{M_{U\times S}=1}{N_j}, \ldots, \frac{M_{LS}=1}{N_j}, \ldots, \frac{M_{Y\times K}=1}{N_j}\right)_{441}$
J = 2
Peptide Length=36
Number of windows $N_j = L-(j+1) = 33$
A C K A L G A C L V A P N C F R G H F L H N Y A P U A G R U L S Y A K Y
$\left(\frac{M_{A\times\times A}=3}{N_j}, \frac{M_{A\times\times C}}{N_j}, \ldots, \frac{M_{G\times\times L}=3}{N_j}, \ldots, \frac{M_{N\times\times A}}{N_j}, \ldots, \frac{M_{R\times\times S}=1}{N_j}, \ldots, \frac{M_{Y\times\times Y}=1}{N_j}\right)_{441}$

FIGURE 1. Representation of KSAAP descriptor for K = 2 [38].

In this work, KSAAP feature vectors are collected using different K values, i.e. k =1, 2, 3, and 4.

2) COMPOSITE PHYSIOCHEMICAL PROPERTIES (CPP)

A peptide sequence consists of twenty unique amino acids, whereas each amino acid residue has various specific biological and physicochemical properties [39]. Physiochemical properties play a crucial role in identifying the structure and behavior of amino acids [40]. Moreover; these properties have direct or indirect influences on the functions and structures of proteins. It was also found that by combining the physicochemical properties effectively provides high discriminative features for the identification of protein types [41]. Hence, the twenty native amino acids are divided into several groups/clusters based on their nature [42]. In this work, the composite physiochemical properties based valuable and informative features are extracted from peptide sequences using eight different properties that are: charge, aliphatic, aromatic, acidic, hydrophilic, hydrophobic, small and tiny, as shown in Table 1. In addition, 20 features of amino acid composition are also combined with the final feature space.

TABLE 1. Composite physiochemical properties (CPP) groups of amino acids.

Amino Acid Property group	No; of Amino acids	No; of Features
Charge	Asp, Glu, His, Arg, Lys	5
Aromatic	Phe, His, Trp, and Tyr	4
Aliphatic	Ile, Leu, and Val	3
Acidic	Asp and Glu	2
Hydrophilic	Asp, Glu, Lys, Asn, Gln	5
Hydrophobic	Ala, Cys, Phe, Ile, Leu, Met, Val, Trp, Tyr	9
Small	Ala, Cys, Asp, Gly, Asn, Pro, Ser, Thr, and Val	9
Tiny	Ala, Cys, Gly, Ser, Thr	5
Amino acid composition	A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y	20

3) AUTO COVARIANCE (AC)

Auto covariance (AC) is a statistical encoding scheme that has been widely used in the area of bioinformatics to efficiently formulate the amino acid sequences [43]–[45]. AC is also considered more effective to minimize the loss of sequence order information as well as to represent the amino acid descriptor into a specific length [46]. In statistics, AC is the covariance of amino acid residues against a certain distance apart from the whole sequence [47]. In this work, AC is used to represent the average correlation factor among positions with a series of lag apart from the whole protein sequence P. whereas, AC can be computed by:

$$AC_{lag,j} = \frac{1}{n-lag}\sum_{i=1}^{n-lag}\left(p_{i,j}-\frac{1}{n}\sum_{i=1}^{n}p_{i,j}\right) \times \left(p_{(i+lag),j}-\frac{1}{n}\sum_{i=1}^{n}p_{i,j}\right) \quad (5)$$

where n represents the length of a sequence, lag denotes the distance between one residue and its neighbors, i is the position, and j represents one descriptor. The lag value should be smaller than the sequence length.

C. TWO LEVEL FEATURE SELECTION (2LFS)

In bioinformatics and machine learning, the feature extraction phase is highly valuable and important for the accurate classification of biological sequences. Although, the feature vectors with high dimensions require high training time that leads to producing low classification accuracy [48], [49]. Therefore, an effective feature selection is highly indispensable to select reliable features with minimum redundant and noisy features. In this work, a novel two-level feature selection (2LFS) approach is presented. Whereas, at level-1, SVM-RFE [50] based wrapper method is applied to select optimal features using SVM learner. However, in some cases, the computational time of SVM-RFE is high and may lead to overfitting because of its involvement in SVM training. Therefore, to deals with such issues, mRMR based filter feature selection method is applied in level-2 [51]. mRMR eradicates irrelevant and noisy features and considered much faster than the wrapper technique. Moreover, it is considered less prone to over-fitting. The detail explanation of 2LFS is described below:

1) SVM-RFE

SVM-RFE is a wrapper feature selection approach that selects optimal features by computing the weights among the extracted feature vector [52]. The evaluating criteria of SVM-RFE are almost similar to the general SVM learning model [53]. At first, SVM-RFE examines the weights against all the features using SVM parameters and then built a model. Whereas, in each turn, the weight of each feature is ordered and updated based on its closeness to the class label and the performance of the SVM model [54]. Consequently, the features with the smallest weight 'w' are eradicated and an optimal feature set is kept until vector space becomes empty. The same procedure is repeated for all no; of features and finally, the optimal feature set is formed.

In this work, SVM-linear kernel is applied and evaluation measure C_j of SVM is computed by weight 'w' using the following equation:

$$C_j = (w_j)^2 \quad (6)$$

whereas w_j represents the weight vector of the jth feature. while for the smallest value C_j, the jth element is eradicated.

The class interval of SVM-RFE can be calculated as:

$$Y = \frac{1}{2}|w|^2 \quad (7)$$

If the jth feature is removed, then the following Taylor expression is applied to calculate the variation ΔY as below:

$$\Delta Y = \frac{\partial Y}{\partial w}\Delta w + \frac{\partial^2 Y}{\partial w^2}(\Delta w)^2 \quad (8)$$

On the other hand, SVM-RFE is considered ineffective due to the ‘‘correlation bias’’ problem [55], where the importance of highly correlated features are underestimated. To deals with such a problem, correlation bias reduction (CBR) is also incorporated with SVM-RFE to select optimal features.

2) MINIMUM REDUNDANCY MAXIMUM RELEVANCE (mRMR)

Maximum Relevance and Minimum Redundancy (mRMR) is an effective feature selection that is applied to choose valuable, unique, and relevant features from extracted feature spaces [56]. Sometimes the feature encoding phase extract features that are highly correlated and are not capable to accurately predict the target class. Additionally, the Replicative space of feature highly affects the process of learning hypotheses [57]. Therefore, to deals with such an issue, an optimal feature set needs to be found with minimum correlated features. Whereas, such features can be effectively obtained using mRMR. mRMR ranked feature deals with larger dimensions of data. It also ensures the reduced dimension features with minimal loss of useful features. The relevancy between two vectors (a, b) can be calculated using the following mutual information (MI) formula:

$$MI(a, b) = \sum_{i,j \in N} P(a_i, b_j) \log \frac{P(a_i, b_j)}{P(a_i)P(b_j)} \tag{9}$$

where $P(a_i, b_j)$ represents the joint probability density function and $P(a_i)P(b_j)$ shows the marginal probability function.

Similarly, MI between feature and targeted classes can be computed as below:

$$MI(a, c) = \sum_{i,k \in N} P(a_i, c_k) \log \frac{P(a_i, c_k)}{P(a_i)P(c_k)} \tag{10}$$

where the variable ‘‘a’’ denotes the feature and ‘‘c’’ represents the targeted class.

Now the feature vector with minimal redundancy can be calculated using the following expression,

$$\min(mR) = \frac{1}{|v|^2} MI(a, b) \tag{11}$$

where a symbol $|v|$ represents the total number of features in vector v.

Furthermore, maximum relevance can be found using the following formula,

$$\max(MR) = \frac{1}{|v|} \sum_{a \in v} MI(a, c) \tag{12}$$

Finally, we have

$$Max(\nabla MI) = MR - mR \tag{13}$$

D. CLASSIFICATION APPROACHES

1) SUPPORT VECTOR MACHINE (SVM)

SVM has been effectively utilized for the prediction of various biological applications [58]. Initially, SVM was presented by Vapnik for binary problems, but subsequently, it was also extended to multi-class problems. In contrast, with other classification learners, SVM is considered more objective and efficient due to its accurate classification rates. SVM measures the predictive ability using different kernel functions that maps the input samples into a high dimensional feature vector [59]. However, the observations of the different classes are linearly classified using an optimal hyperplane. Whereas, the optimal hyperplane computes the maximum margin lines among the samples of different classes, which can effectively reduce the error rates [60]. The cost function of SVM is also convex like logistic regression. Therefore, while dealing with large datasets or features quadratic programming (QP) problems may occur but such a problem can be effectively solved using Sequential Minimal Optimization (SMO) based optimization method [61]. This divides the large QP problem into smaller sub-problems that can be analytically solved to avoid time complexity to some degree. SMO can be easily implemented using the ‘libsvm’ package. In this work, the radial base function (RBF) is utilized to train the peptide sequences. RBF as compare to other kernel functions is considered more effective due to its best hyperplane selection. Whereas, RBF uses two parameters such as; kernel width γ and regularization parameter C. furthermore, the values of these parameters are adjusted via grid search method.

The optimal hyperplane obtained using RBF kernel can be represented as below:

$$K(x_i, x_j) = \exp(-\gamma \, |x_i - x_j|^2) \tag{14}$$

The regularization parameter C and kernel parameter γ for search space can be, respectively [62], represented as

$$\begin{cases} 2^{-3} \leq C \leq 2^5 \\ 2^{-6} \leq \gamma \leq 2^4 \end{cases} \quad \text{with step of 1} \tag{15}$$

2) FUZZY K-NEAREST NEIGHBOR (FKNN)

FKNN an improved classification learner that overcome the drawbacks of the standard KNN algorithm [63]. Compare to other fuzzy methods, it has been considered more competitive in terms of accuracy and also provides a low error rate in classifying the objects. The main searching concept of both KNN and FKNN is similar, but in the case of standard KNN, every sample belongs to one majority class only [64]. While using FKNN the memberships of the data samples associated with more than one class.

In FKNN, the fuzzy class membership $u_i(P)$ is assigned to the test instance ‘‘P'' according to the following equation.

$$u_i(P) = \left[\frac{\sum_{j=1}^{K} u_i(P_j) D_j^{-2/(m-1)}}{\sum_{j=1}^{K} D_j^{-2/(m-1)}} \right] \quad i = 1, 2, 3, \ldots .C \tag{16}$$

where ‘m’ is a *Fuzzy* strength parameter, which controls the magnitude of the distance of the neighbors from the test instance, 'K' represents the number of nearest neighbors, and i=1, 2, 3,..., C denotes the number of classes. $D\_j = \|P - P_j\|$ represents the Euclidean distance between the test instance 'P' and its jth nearest references data P_j. Several distance metrics can be applied to compute the distance among instances, however, Euclidean distance is utilized in this study. finally, after examining the memberships of a query sample, it is then assigned to a class with the highest membership value.

3) RANDOM FOREST (RF)

Random forest is a widely used ensemble learner, which was initially developed by Breiman to effectively deals with classification and regression problems [65]. RF, as compared to other classification methods, is considered more efficient due to its simple training, fast prediction, and interpretability [66]. RF is an ensemble approach that involves multiple numbers of decision trees, wherein each tree the "n" number of features is randomly selected from the whole feature vector. It has been found that the random selection nature of RF is unbias that reduces the correlation among unpruned trees. In the next step, a bagging algorithm is used to produce a training feature set with resample instances [67]. In the third step, the decision tree is built using a randomly selected feature vector and resampled training set [68]. Finally, the number of decision trees is summarized and the final prediction is generated using a majority voting procedure.

E. PERFORMANCE EVALUATION MEASURES

In machine learning, the strength and efficiency of a computational predictor are examined through various criteria [69]. Whereas, a confusion matrix is maintained to keep the predicted results of a hypothesis learner. Generally, in prediction models, accuracy is employed to evaluate the strength and capability of classification learners, while in most of the cases the only accuracy is unable to predict the overall effectiveness of a model [70]. Therefore, the following performance parameters are utilized to accurately examine our proposed model.

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{17}$$

$$Sen = \frac{TP}{TP + FN} \tag{18}$$

$$Spe = \frac{TN}{TN + FP} \tag{19}$$

$$MCC = \frac{(TN \times TP) - (FN \times FP)}{\sqrt{(TP + FP)\,(TP + FN)\,(TN + FP)(TN + FN)}} \tag{20}$$

Here TP, FP, TN and FN represents the true positive, false positive, true negative, and false negative, respectively.

III. RESULTS AND DISCUSSIONS

In this section, we will explain the predicted outcomes of our proposed classification methods. The proposed work extracts features from the ACP sequence using three distinct nature encoding schemes. Furthermore, the extracted feature set is then processed using a 2LFS approach to select highly relevant and optimum features. Whereas, at first level SVM-RFE is applied to select optimal features. Moreover, to reduce the biases of highly correlated features. a CBR approach is also incorporated with SVM-RFE. Whereas, RFE uses the SVM training model to gather the optimal features that may lead to high computational time. Therefore, to speed up the proposed model, mRMR is adopted in level-2 to eradicate irrelevant and redundant descriptors. At last, the effectiveness of our model is examined using different nature hypotheses learners. The detailed graphical abstract of our proposed intelligent model is illustrated in Figure 2. Generally, in the area of machine learning several cross-validation (CV) tests are applied to enhance the success rate of a computational model. Among these tests; k-fold subsampling, independent, and jackknife tests are widely utilized to boost the prediction of hypothesis learners. In this paper, a 10-fold CV test is used to examine the success rates of our proposed model. Whereas, 9-folds are used to train the model and the remaining fold is used to test the model. In the below sub-sections, the predicted outcomes of our proposed formulation methods using training and independent datasets are briefly described.

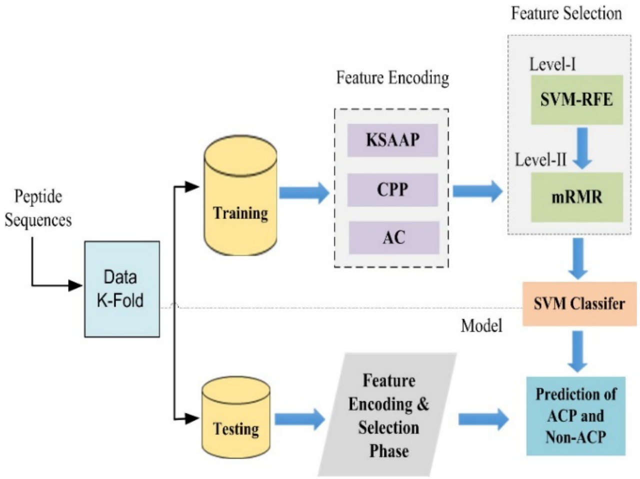

FIGURE 2. The framework of the proposed cACP-2LFS model.

A. PERFORMANCE ANALYSIS OF HYPOTHESIS LEARNERS USING LEE DATASET

The performance outcomes of the proposed three encoding schemes using different hypothesis learners were illustrated in Table.2. However, each of the encoding methods represents the feature vector of different dimensionality. Whereas, KSAAP extracts highly correlated sequential descriptors using different K-values i.e. (k = 0,1,2,3). In this paper, the highest performance was reported using K=3 having dimensions of (441 × 4) =1764D. Among all the hypothesis learners, SVM achieved the highest accuracy of 90.04% as

TABLE 2. Performance of proposed feature encoding schemes using LEE dataset.

Method	Classifier		Acc (%)	Sen (%)	Spe (%)	MCC
CPP	RF		87.51	83.46	91.47	0.75
	SVM		87.75	83.5	95.83	0.76
	FKNN		87.03	87.83	86.25	0.74
AC	RF		81.52	78.91	84.12	0.63
	SVM		81.99	75.12	88.86	0.65
	FKNN		82.12	72.99	91.47	0.66
KSAAP	K=0	RF	87.02	82.94	91.47	0.74
		SVM	87.2	82.69	91.72	0.75
		FKNN	87.44	78.19	96.68	0.76
	K=1	RF	87.68	85.31	90.05	0.75
		SVM	88.39	82.7	94.07	0.77
		FKNN	86.73	77.01	96.44	0.75
	K=2	RF	87.67	84.36	90.99	0.75
		SVM	89.34	82.46	96.21	0.79
		FKNN	86.73	77.49	95.97	0.75
	K=3	RF	88.51	86.02	90.99	0.77
		SVM	**90.04**	**84.36**	**95.73**	**0.81**
		FKNN	86.73	75.83	97.83	0.75

shown in Table.2. However, to develop a computationally efficient training model, a feature selection is highly essential to reduce the dimensional size of the extracted feature vector while retaining the significant features. Therefore, we proposed a 2LFS approach which comprises of two levels to collect optimal features. At level-1, SVM-RFE selects only 122 optimal features by examining the weights among all features using the SVM model. While training the SVM model, the linear kernel was utilized. However, SVM-RFE due to the SVM model may lead to high computational cost (time-complexity). Thus, mRMR based filter method was applied at the second level to choose only relevant features by eradicating irrelevant and redundant descriptors. mRMR further reduced the dimensions of KSAAP to 80D. After selecting the optimal subset of KSAAP performed remarkably and achieved an accuracy of 93.72%, with specificity, sensitivity, and MCC of 93.51%, 91.85%, and 0.86, respectively. The detailed classification results of the 2LFS approach using different classification algorithms are given in Table 3 and figure 3. On the other hand, CPP and AC descriptors reported an accuracy of 88.36% and 81.99%, respectively. Which was comparatively lower performance than KSAAP descriptor using all hypothesis learners.

B. PERFORMANCE ANALYSIS OF HYPOTHESIS LEARNERS USING INDEPENDENT DATASET

While discriminating ACP and non-ACP sequences, it may be possible to develop a prediction model whose prediction accuracy is over-estimated due to an overfitting problem. Therefore, to avoid overfitting, we used an independent dataset to examine the generalization capability of our model. The detailed empirical outcomes of the independent dataset using cACP-2LFS are given in Table.4. Among all hypothesis learners, SVM achieved the highest prediction outcomes with an accuracy of 94.15%, sensitivity of 91.33%, and specificity of 96.23% and MCC of 0.87. Whereas, these results were reported using the KSAAP descriptors (K=0) whose dimensions are reduced to 72D using 2LFS.

TABLE 3. Performance rates of LEE dataset after applying 2LFS approach.

Method	Classifier	Acc (%)	Sen (%)	Spe (%)	MCC
Level 1 (SVM-RFE)					
CPP	RF	87.04	82.73	91.22	0.74
	SVM	87.36	84.31	90.42	0.76
	FKNN	87.64	85.65	89.57	0.75
AC	RF	80.81	77.25	84.36	0.62
	SVM	81.99	72.28	91.71	0.65
	FKNN	82.35	75.83	99.86	0.65
KSAAP	RF	89.93	89.1	90.76	0.8
	SVM	**92.11**	**90.52**	**94.32**	**0.85**
	FKNN	83.41	93.6	73.33	0.68
Level 2 (SVM-RFE) + mRMR					
CPP	RF	87.15	83.7	90.52	0.75
	SVM	88.36	84.19	92.47	0.77
	FKNN	87.28	84.92	89.57	0.75
AC	RF	81.17	77.72	84.61	0.63
	SVM	81.99	71.8	92.19	0.65
	FKNN	81.87	71.57	92.19	0.65
KSAAP	RF	90.28	88.86	91.69	0.81
	SVM	**93.72**	**91.85**	**94.71**	**0.86**
	FKNN	85.54	91.47	79.62	0.72

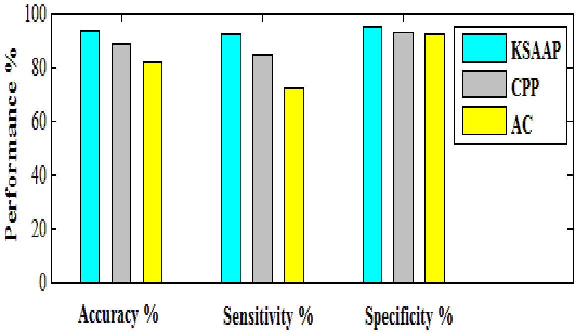

FIGURE 3. Performance analysis of LEE dataset after 2LFS feature selection.

TABLE 4. Performance rates of the proposed model using independent dataset.

Method	Classifier	Acc (%)	Sen (%)	Spe (%)	MCC
CPP	RF	90.81	89.33	91.67	0.81
	SVM	91.99	92.42	95.34	0.84
	FKNN	89.91	92.67	88.01	0.79
AC	RF	88.33	83.33	92.67	0.76
	SVM	87.67	82.67	92.71	0.76
	FKNN	86.77	78.11	95.33	0.75
KSAAP	RF	91.01	88	95	0.82
	SVM	**94.15**	**91.33**	**96.23**	**0.87**
	FKNN	88.01	77.33	98.67	0.78

TABLE 5. Performance comparisons of 'cACP-2LFS' model with existing models.

Dataset	Predictors	Acc (%)	Sen (%)	Spe (%)	MCC
LEE Dataset	SVMACP[21]	81.40	77.50	85.30	0.63
	RFACP [21]	82.70	70.60	94.80	0.67
	cACP-2LFS	**93.72**	**91.85**	**94.71**	**0.86**
Independent Dataset	iACP [16]	92.67	93.33	92.00	0.85
	Fm-Li [71]	93.61	89.86	96.12	0.87
	cACP-2LFS	**94.15**	**91.33**	**96.23**	**0.87**

C. PERFORMANCE COMPARISON OF cACP-2LFS MODEL WITH EXISTING METHODS

The comparison analysis of the cACP-2LFS model with the existing state of art methods is provided in Table 5. In the case of LEE dataset, previously, Manavalan *et al.* proposed two predictors namely, 'SVMACP' and 'RFACP' to predict peptides sequences using physicochemical properties and AAC & DPC based sequential features [21]. Whereas, SVMACP achieved an accuracy of 81.40 % with specificity, sensitivity, and MCC of 77.50%, 85.30 and 0.63, respectively. Similarly, RFACP performed better and reported an accuracy of 82.70%, a specificity of 70.60 %, and a sensitivity of 94.80 % and MCC of 0.67. In contrast, our cACP-2LFS model using LEE dataset achieved a remarkable prediction accuracy of 93.72%, the sensitivity of 91.85%, the specificity of 94.71%, and MCC of 0.86. On the other hand, cACP-2LFS independent dataset S2, achieved an accuracy of 94.15%, with 91.33% specificity and 96.23% sensitivity which was ∼2% higher than the existing prediction models available in the literature [16], [71]. Furthermore, the detailed comparison of the cACP-2LFS model with existing studies is illustrated in Table 5, Figures 4 & 5.

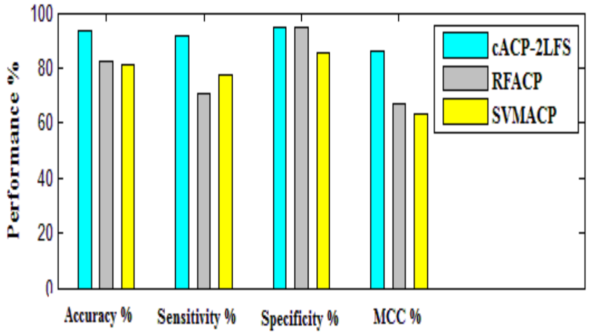

FIGURE 4. Performance comparison of cACP-2LFS and existing models using LEE dataset.

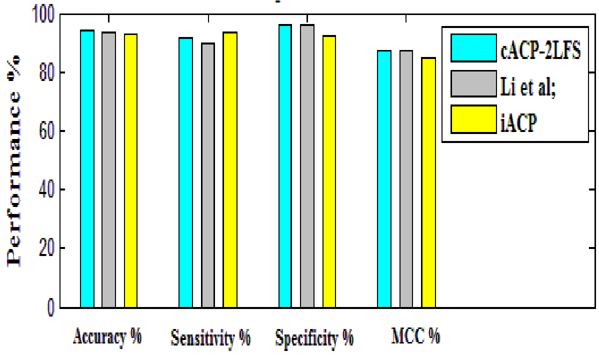

FIGURE 5. Performance comparison of cACP-2LFS and existing models using independent dataset.

IV. CONCLUSION

In this study, we proposed an effective and reliable computational model for the identification of ACPs. To extract high discriminative features, three distinct nature formulation schemes are employed. Whereas, KSAAP and CPP are utilized to extract highly correlated sequential and local structure features. Besides, AC is also employed to gather neighboring residue information from peptide sequences. Furthermore, to select optimal features and to eradicate irrelevant and noisy features, a novel 2LFS is applied. Various hypothesis learners are utilized to investigate the classification rates of our proposed model. It is observed that the KSAAP feature set in combination with 2LFS outperformed than existing studies in literature and reported the highest accuracy of 93.72% and 94.11% using LEE and independent datasets, respectively. Consequently, it is expected that our proposed work will be considered a useful tool for research academia and drug discovery.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, "Global cancer statistics," *CA, Cancer J. Clinicians*, vol. 61, no. 2, pp. 69–90, Mar. 2011.

[2] J. Ferlay, H.-R. Shin, F. Bray, D. Forman, C. Mathers, and D. M. Parkin, "Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008," *Int. J. Cancer*, vol. 127, no. 12, pp. 2893–2917, Dec. 2010.

[3] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, ‘‘Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,’’ *CA, Cancer J. Clinicians*, vol. 68, no. 6, pp. 394–424, Nov. 2018.

[4] J. Thundimadathil, ‘‘Cancer treatment using peptides: Current therapies and future prospects,’’ *J. Amino Acids*, vol. 2012, pp. 1–13, Dec. 2012.

[5] M. Maliepaard, G. L. Scheffer, I. F. Faneyte, M. A. van Gastelen, A. C. L. M. Pijnenborg, A. H. Schinkel, M. J. van de Vijver, R. J. Scheper, and J. H. M. Schellens, ‘‘Subcellular localization and distribution of the breast cancer resistance protein transporter in normal human tissues,’’ *Cancer Res.*, vol. 61, no. 8, pp. 3458–3464, 2001.

[6] A. Tyagi, A. Tuknait, P. Anand, S. Gupta, M. Sharma, D. Mathur, A. Joshi, S. Singh, A. Gautam, and G. P. S. Raghava, ‘‘CancerPPD: A database of anticancer peptides and proteins,’’ *Nucleic Acids Res.*, vol. 43, no. D1, pp. D837–D843, Jan. 2015.

[7] D. Gaspar, A. S. Veiga, and M. A. R. B. Castanho, ‘‘From antimicrobial to anticancer peptides. A review,’’ *Frontiers Microbiol.*, vol. 4, p. 294, Oct. 2013.

[8] V. Gregorc *et al.*, ‘‘Phase I study of NGR-hTNF, a selective vascular targeting agent, in combination with cisplatin in refractory solid tumors,’’ *Clin. Cancer Res.*, vol. 17, no. 7, pp. 1964–1972, 2011.

[9] P. Khalili, ‘‘A non-RGD-based integrin binding peptide (ATN-161) blocks breast cancer growth and metastasis *in vivo*,’’ *Mol. Cancer Therapeutics*, vol. 5, no. 9, pp. 2271–2280, Sep. 2006.

[10] L. Wei, C. Zhou, H. Chen, J. Song, and R. Su, ‘‘ACPred-FL: A sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides,’’ *Bioinformatics*, vol. 34, no. 23, pp. 4007–4016, 2018.

[11] P. Feng and Z. Wang, ‘‘Recent advances in computational methods for identifying anticancer peptides,’’ *Current Drug Targets*, vol. 20, no. 5, pp. 481–487, Mar. 2019.

[12] B. Rao, C. Zhou, G. Zhang, R. Su, and L. Wei, ‘‘ACPred-Fuse: Fusing multi-view information improves the prediction of anticancer peptides,’’ *Briefings Bioinf.*, pp. 1–10, 2019.

[13] N. Schaduangrat, C. Nantasenamat, V. Prachayasittikul, and W. Shoombuatong, ‘‘ACPred: A computational tool for the prediction and analysis of anticancer peptides,’’ *Molecules*, vol. 24, no. 10, p. 1973, May 2019.

[14] H.-C. Yi, Z.-H. You, X. Zhou, L. Cheng, X. Li, T.-H. Jiang, and Z.-H. Chen, ‘‘ACP-DL: A deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation,’’ *Mol. Therapy-Nucleic Acids*, vol. 17, pp. 1–9, Sep. 2019.

[15] Z. Hajisharifi, M. Piryaiee, M. M. Beigi, M. Behbahani, and H. Mohabatkar, ‘‘Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via AMES test,’’ *J. Theor. Biol.*, vol. 341, pp. 34–40, Jan. 2014.

[16] W. Chen, H. Ding, P. Feng, H. Lin, and K.-C. Chou, ‘‘iACP: A sequence-based tool for identifying anticancer peptides,’’ *Oncotarget*, vol. 7, no. 13, p. 16895, 2016.

[17] S. Akbar, M. Hayat, M. Iqbal, and M. A. Jan, ‘‘IACP-GAEnsC: Evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space,’’ *Artif. Intell. Med.*, vol. 79, pp. 62–70, Jun. 2017.

[18] M. Kabir, M. Arif, S. Ahmad, Z. Ali, Z. N. K. Swati, and D.-J. Yu, ‘‘Intelligent computational method for discrimination of anticancer peptides by incorporating sequential and evolutionary profiles information,’’ *Chemometric Intell. Lab. Syst.*, vol. 182, pp. 158–165, Nov. 2018.

[19] M. A. Akmal, N. Rasool, and Y. D. Khan, ‘‘Prediction of N-linked glycosylation sites using position relative features and statistical moments,’’ *PLoS ONE*, vol. 12, no. 8, Aug. 2017, Art. no. e0181966.

[20] L. Xu, G. Liang, L. Wang, and C. Liao, ‘‘A novel hybrid sequence-based model for identifying anticancer peptides,’’ *Genes*, vol. 9, no. 3, p. 158, Mar. 2018.

[21] B. Manavalan, S. Basith, T. H. Shin, S. Choi, M. O. Kim, and G. Lee, ‘‘MLACP: Machine-learning-based prediction of anticancer peptides,’’ *Oncotarget*, vol. 8, no. 44, p. 77121, 2017.

[22] S. Vijayakumar and L. Ptv, ‘‘ACPP: A Web server for prediction and design of anti-cancer peptides,’’ *Int. J. Peptide Res. Therapeutics*, vol. 21, no. 1, pp. 99–106, Mar. 2015.

[23] S. Akbar, A. U. Rahman, M. Hayat, and M. Sohail, ‘‘CACP: Classifying anticancer peptides using discriminative intelligent model via Chou's 5-step rules and general pseudo components,’’ *Chemometric Intell. Lab. Syst.*, vol. 196, Jan. 2020, Art. no. 103912.

[24] A. Tyagi, P. Kapoor, R. Kumar, K. Chaudhary, A. Gautam, and G. P. S. Raghava, ‘‘In silico models for designing and discovering novel anticancer peptides,’’ *Sci. Rep.*, vol. 3, no. 1, p. 2984, Dec. 2013.

[25] P. Feng, H. Ding, H. Yang, W. Chen, H. Lin, and K.-C. Chou, ‘‘IRNA-PseColl: Identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC,’’ *Mol. Therapy-Nucleic Acids*, vol. 7, pp. 155–163, Jun. 2017.

[26] W. Chen, H. Tang, and H. Lin, ‘‘MethyRNA: A Web server for identification of N6-methyladenosine sites,’’ *J. Biomol. Struct. Dyn.*, vol. 35, no. 3, pp. 683–687, Feb. 2017.

[27] S. Ahmad, M. Kabir, and M. Hayat, ‘‘Identification of heat shock protein families and J-protein types by incorporating dipeptide composition into Chou's general PseAAC,’’ *Comput. Methods Programs Biomed.*, vol. 122, no. 2, pp. 165–174, Nov. 2015.

[28] W. Chen, P.-M. Feng, H. Lin, and K.-C. Chou, ‘‘IRSpot-PseDNC: Identify recombination spots with pseudo dinucleotide composition,’’ *Nucleic Acids Res.*, vol. 41, no. 6, p. e68, Apr. 2013.

[29] B. Liu, R. Long, and K.-C. Chou, ‘‘IDHS-EL: Identifying DNase I hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework,’’ *Bioinformatics*, vol. 32, no. 16, pp. 2411–2418, Aug. 2016.

[30] G. Wang, X. Li, and Z. Wang, ‘‘APD3: The antimicrobial peptide database as a tool for research and education,’’ *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1087–D1093, Jan. 2016.

[31] T. U. Consortium, ‘‘UniProt: The universal protein knowledgebase,’’ *Nucleic Acids Res.*, vol. 45, no. D1, pp. D158–D169, 2016.

[32] Z. Ju and S.-Y. Wang, ‘‘Prediction of lysine formylation sites using the composition of k-spaced amino acid pairs via Chou's 5-steps rule and general pseudo components,’’ *Genomics*, vol. 112, no. 1, pp. 859–866, Jan. 2020.

[33] M. Hasan, M. Khatun, M. Mollah, N. Haque, C. Yong, and G. Dianjing, ‘‘NTyroSite: Computational identification of protein nitrotyrosine sites using sequence evolutionary features,’’ *Molecules*, vol. 23, no. 7, p. 1667, Jul. 2018.

[34] H. Fu, Y. Yang, X. Wang, H. Wang, and Y. Xu, ‘‘DeepUbi: A deep learning framework for prediction of ubiquitination sites in proteins,’’ *BMC Bioinf.*, vol. 20, no. 1, p. 86, Dec. 2019.

[35] M. M. Hasan, Y. Zhou, X. Lu, J. Li, J. Song, and Z. Zhang, ‘‘Computational identification of protein pupylation sites by using profile-based composition of k-Spaced amino acid pairs,’’ *PLoS ONE*, vol. 10, no. 6, Jun. 2015, Art. no. e0129635.

[36] Z. Ju and J.-Z. Cao, ‘‘Prediction of protein N-formylation using the composition of k-spaced amino acid pairs,’’ *Anal. Biochem.*, vol. 534, pp. 40–45, Oct. 2017.

[37] J. Song, Y. Wang, F. Li, T. Akutsu, N. D. Rawlings, G. I. Webb, and K.-C. Chou, ‘‘IProt-sub: A comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites,’’ *Briefings Bioinf.*, vol. 20, no. 2, pp. 638–658, Mar. 2019.

[38] M. Usman, S. Khan, and J.-A. Lee, ‘‘AFP-LSE: Antifreeze proteins prediction using latent space encoding of composition of k-Spaced amino acid pairs,’’ *Sci. Rep.*, vol. 10, no. 1, pp. 1–13, Dec. 2020.

[39] M. Hayat and A. Khan, ‘‘Mem-PHybrid: Hybrid features-based prediction system for classifying membrane protein types,’’ *Anal. Biochem.*, vol. 424, no. 1, pp. 35–44, May 2012.

[40] J. Jia, Z. Liu, X. Xiao, B. Liu, and K.-C. Chou, ‘‘Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition,’’ *J. Biomol. Struct. Dyn.*, vol. 34, no. 9, pp. 1946–1961, Sep. 2016.

[41] M. Hayat and A. Khan, ‘‘WRF-TMH: Predicting transmembrane helix by fusing composition index and physicochemical properties of amino acids,’’ *Amino Acids*, vol. 44, no. 5, pp. 1317–1328, May 2013.

[42] A. Nath and S. Karthikeyan, ‘‘Enhanced prediction and characterization of CDK inhibitors using optimal class distribution,’’ *Interdiscipl. Sci., Comput. Life Sci.*, vol. 9, no. 2, pp. 292–303, Jun. 2017.

[43] I. A. Doytchinova and D. R. Flower, ‘‘VaxiJen: A server for prediction of protective antigens, tumour antigens and subunit vaccines,’’ *BMC Bioinf.*, vol. 8, no. 1, p. 4, Dec. 2007.

[44] Y. Guo, M. Li, M. Lu, Z. Wen, and Z. Huang, ‘‘Predicting G-protein coupled receptors-G-protein coupling specificity based on autocross-covariance transform,’’ *Proteins, Struct., Function, Bioinf.*, vol. 65, no. 1, pp. 55–60, Jul. 2006.

[45] Y. Fang, Y. Guo, Y. Feng, and M. Li, "Predicting DNA-binding proteins: Approached from Chou's pseudo amino acid composition and other specific sequence features," *Amino Acids*, vol. 34, no. 1, pp. 103–109, Jan. 2008.

[46] Y. Guo, L. Yu, Z. Wen, and M. Li, "Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences," *Nucleic Acids Res.*, vol. 36, no. 9, pp. 3025–3030, May 2008.

[47] L. Yang, Y. Li, R. Xiao, Y. Zeng, J. Xiao, F. Tan, and M. Li, "Using auto covariance method for functional discrimination of membrane proteins based on evolution information," *Amino Acids*, vol. 38, no. 5, pp. 1497–1503, May 2010.

[48] Z. Chen, C. Wu, Y. Zhang, Z. Huang, B. Ran, M. Zhong, and N. Lyu, "Feature selection with redundancy-complementariness dispersion," *Knowl.-Based Syst.*, vol. 89, pp. 203–217, Nov. 2015.

[49] I. M. Johnstone and D. M. Titterington, *Statistical Challenges of High-Dimensional Data*. London, U.K.: The Royal Society, 2009.

[50] G. Roffo, "Feature selection library (MATLAB toolbox)," 2016, *arXiv:1607.01327*. [Online]. Available: http://arxiv.org/abs/1607.01327

[51] S. Wang, Y.-H. Zhang, Y.-D. Cai, G. Huang, and L. Chen, "Analysis and prediction of myristoylation sites using the mRMR method, the IFS method and an extreme learning machine algorithm," *Combinat. Chem. High Throughput Screening*, vol. 20, no. 2, pp. 96–106, Jun. 2017.

[52] L. Xi, S. Li, Y. Wei, X. Wu, H. Liu, and X. Yao, "Recognition of protein folding kinetics pathways based on amino acid properties information derived from primary sequence," *Chemometric Intell. Lab. Syst.*, vol. 126, pp. 76–82, Jul. 2013.

[53] F. Ali, M. Kabir, M. Arif, Z. N. Khan Swati, Z. U. Khan, M. Ullah, and D.-J. Yu, "DBPPred-PDSD: Machine learning approach for prediction of DNA-binding proteins using discrete wavelet transform and optimized integrated features space," *Chemometric Intell. Lab. Syst.*, vol. 182, pp. 21–30, Nov. 2018.

[54] F. Ali, S. Ahmed, Z. N. K. Swati, and S. Akbar, "DP-BINDER: Machine learning model for prediction of DNA-binding proteins by fusing evolutionary and physicochemical information," *J. Comput.-Aided Mol. Des.*, vol. 33, no. 7, pp. 645–658, Jul. 2019.

[55] D. Tang, W. Jin, N. Qin, and H. Li, "Feature selection and analysis of single lateral damper fault based on SVM-RFE with correlation bias reduction," in *Proc. 35th Chin. Control Conf. (CCC)*, Jul. 2016, pp. 3840–3845.

[56] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[57] M. Khan, M. Hayat, S. A. Khan, and N. Iqbal, "Unb-DPC: Identify mycobacterial membrane protein types by incorporating un-biased dipeptide composition into Chou's general PseAAC," *J. Theor. Biol.*, vol. 415, pp. 13–19, Feb. 2017.

[58] S. Akbar and M. Hayat, "IMethyl-STTNC: Identification of N6-methyladenosine sites by extending the idea of SAAC into Chou's PseAAC to formulate RNA sequences," *J. Theor. Biol.*, vol. 455, pp. 205–211, Oct. 2018.

[59] S. Akbar, M. Hayat, M. Iqbal, and M. Tahir, "IRNA-PseTNC: Identification of RNA 5-methylcytosine sites using hybrid vector space of pseudo nucleotide composition," *Frontiers Comput. Sci.*, vol. 14, no. 2, pp. 451–460, Apr. 2020.

[60] X. Xiao, J.-L. Min, W.-Z. Lin, Z. Liu, X. Cheng, and K.-C. Chou, "IDrug-target: Predicting the interactions between drug compounds and target proteins in cellular networking via benchmark dataset optimization approach," *J. Biomol. Struct. Dyn.*, vol. 33, no. 10, pp. 2221–2233, Oct. 2015.

[61] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, Apr. 2011.

[62] F.-Y. Dao, H. Lv, F. Wang, C.-Q. Feng, H. Ding, W. Chen, and H. Lin, "Identify origin of replication in saccharomyces cerevisiae using two-step feature selection technique," *Bioinformatics*, vol. 35, no. 12, pp. 2075–2083, Jun. 2019.

[63] H.-B. Shen, J. Yang, and K.-C. Chou, "Fuzzy KNN for predicting membrane protein types from pseudo-amino acid composition," *J. Theor. Biol.*, vol. 240, no. 1, pp. 9–13, May 2006.

[64] M. Hayat and A. Khan, "Discriminating outer membrane proteins with fuzzy K-Nearest neighbor algorithms based on the general form of Chou's PseAAC," *Protein Peptide Lett.*, vol. 19, no. 4, pp. 411–421, Apr. 2012.

[65] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[66] T. Jo and J. Cheng, "Improving protein fold recognition by random forest," *BMC Bioinf.*, vol. 15, no. S11, p. S14, Dec. 2014.

[67] X. Ma, J. Guo, and X. Sun, "DNABP: Identification of DNA-binding proteins based on feature selection using a random forest and predicting binding residues," *PLoS ONE*, vol. 11, no. 12, Dec. 2016, Art. no. e0167345.

[68] M. Hayat, A. Khan, and M. Yeasin, "Prediction of membrane proteins using split amino acid and ensemble classification," *Amino Acids*, vol. 42, no. 6, pp. 2447–2460, Jun. 2012.

[69] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Comput. Appl.*, vol. 29, no. 10, pp. 685–693, May 2018.

[70] S. Akbar, M. Hayat, M. Kabir, and M. Iqbal, "IAFP-gap-SMOTE: An efficient feature extraction scheme gapped dipeptide composition is coupled with an oversampling technique for identification of antifreeze proteins," *Lett. Organic Chem.*, vol. 16, no. 4, pp. 294–302, Mar. 2019.

[71] F.-M. Li and X.-Q. Wang, "Identifying anticancer peptides by using improved hybrid compositions," *Sci. Rep.*, vol. 6, no. 1, p. 33910, Dec. 2016.