IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 24, 2019, accepted December 19, 2019, date of publication December 24, 2019,
date of current version January 16, 2020.

*Digital Object Identifier 10.1109/ACCESS.2019.2962002*

# CanLect-Pred: A Cancer Therapeutics Tool for Prediction of Target Cancerlectins Using Experiential Annotated Proteomic Sequences

**AHMAD HASSAN BUTT AND YASER DAANIAL KHAN**
Department of Computer Science, School of Systems and Technology, University of Management and Technology, Lahore 54770, Pakistan
Corresponding author: Yaser Daanial Khan (yaser.khan@umt.edu.pk)

**ABSTRACT** Cancerlectins are significantly important group of lectins that have an inhibitory effect on cancer cells with respect to their growth. They have a vital role in various tumor cell interactions like adhesion, growth, metastasis, differentiation and mainly in cellular infection. The investigations associated with cancerlectins are applicable to relevant studies in laboratories, diagnostics and therapy in clinical applications, and drug discoveries in targeting cancers. Prediction of cancerlectins is considered a helpful task due to the fact that they are specifically useful in dissecting cancers. Although, several Bioinformatics tools have been developed to predict cancerlectins, however, the need for improvement in the quality of its prediction model requires enhancements in the annotation and determination process of cancerlectins. In this study, a new model is proposed that builds on statistical moments based features to distinguish cancerlectins from non-cancerlectins. The currently proposed model achieved an accuracy of 88.36% using jackknife test which is better than current state-of-the-art models. These outcomes suggest that the use of statistical moments could bear more effective and efficient results. For the accessibility of the scientific community, a user-friendly web server has been developed which will associate the researchers in medical science. Web server is freely accessible at https://www.biopred.org/canlect.

**INDEX TERMS** Cancerlectins, Hahn moments, lectins, moment invariants, PRIM.

## I. INTRODUCTION

In cellular biology, cells can be agglutinated by a kind of glycoprotein known as lectin. Diverse sugar structures can be specifically recognized through lectins, but somehow lectins lack catalytic movement. Lectins are proteins that bind to carbohydrate molecule and are distributed ubiquitously in nature [1]–[4]. They play a significant role in recognition of specific sugar structures and in a variety of cellular processes that involve cells, proteins and carbohydrates. They are able to reversibly bind carbohydrates [5]–[7]. Furthermore, Lectins greatly differ from antibodies. They also play important role in the innate immune system. They are often considered to be helpful in mediating against the invading microorganisms as a first line of defense. As compared to antibodies, they are not an outcome of response from the immune system. However, some antibodies also cause agglutinations by binding themselves to antigens and produce similar effects as lectins. Lectins are synthesized and secreted by almost all organisms, mainly including bacteria, viruses, vertebrates, plants and invertebrates [8]–[10]. Lectins are also involved in various biological activities such as growth in cells, development and differentiation of cells, cell migration and adhesion, interaction between extracellular and cell, apoptosis and inflammatory response. Many researchers in the field of molecular biology and immunology often consider lectins as therapeutics and diagnostics tool [11]–[13].

Leading the cause of death, cancer is the outcome of abnormal growth of cells which cannot be regulated. Lectins that are related to cancers are known as cancerlectins. These lectins are protective against the cancer cell growth mechanisms. The lectins are suggested to be used in the anti-tumor drugs development as they have minimum side effects. According to few recent researches [14]–[17], for

The associate editor coordinating the review of this manuscript and approving it for publication was Vincenzo Conti.

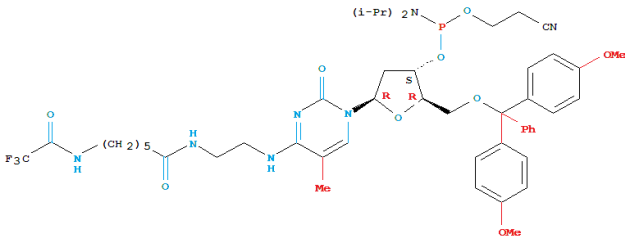

**FIGURE 1.** The chemical structure of a lectin found in bananas (http://www.lookchem.com/cas-117/117675-52-2.html).

impediment of the progression of tumors, many lectins are used as therapeutic agents which outcomes apoptosis and agglutination of cancer cells. Nagaimo lectin is believed to be helpful in the treatment of breast cancer [18]. It has shown in recent studies that HIV replication is inhibited by lectins in bananas (Figure.1) and is utilized in the investigations of treatment of AIDS [19]. Deficiency of mannose binding lectins may result in certain forms of skin infections and inflammatory skin diseases [13], [20]. Molecular changes triggered through mistletoe lectins can induct apoptotic cancer cells and inhibit cancer cell growth [21]. Galectins are highly potential in contributions to proliferations, tumorigenesis, metastasis and angiogenesis and thus are useful in the treatment and diagnosis of specific cancers, [11], [13], [22]–[24].

Most research studies suggest that lectins possess antitumor characteristics, but the knowledge of lectin biological properties and its protein interactions is still not sufficient for developing lectin based drugs [25], [26]. Furthermore, the currently available cancerlectins in natural form are limited and cannot fulfil the most requirements of cancerlectins based drug discovery. Hence, the need for the cancerlectin identification process exists significantly to provide affiliation in understanding the molecular biology of cancer mechanisms. Due to this fact that the availability of cancerlectins is limited, the newly discovered cancerlectins are of prime importance and are considered to be significant targets for advanced research in several applications of immunology and cancer research [27].

Many cancerlectins have been identified and annotated functionally by experimental assays. CancerLectinDB [27] includes cancerlectins integrated and archived with an overwhelming majority. The cancerlectins detected and annotated using experiments in CancerLectinDB [27] are extremely reliable and accurate. With the sequencing technology advancing continuously, new cancerlectins are increasing and being stored rapidly. Similarly, this new continuous experimental detection and storage of cancerlectins has led many computation prediction models to emerge naturally.

## II. RELATED WORK

In the past, several computational models have been developed for rapid and cost effective prediction of cancerlectins based on their evolutionary information, amino acid composition (AAC) and dipeptide composition (DPC). In [28] they proposed the pioneering work and developed a prediction model for cancerlectins using AAC, DPC, split-amino-acid compositions, domain and evolutionary information. Based on Support Vector Machines (SVM) with integration of position specific scores and domain information from PROSITE has shown better results in comparison with other features. In [29] they utilized g-gap dipeptides and developed a model for cancerlectin prediction with accuracy highest among other computational models. Furthermore, several machine learning models including Decision Trees, Random Forests, SVMs and Artificial Neural Networks (ANNs) have been used in classification of cancerlectins. However, the methodologies used in the aforementioned models do not utilize the most pertinent and obscure features from cancerlectin sequences and thus lack in predictive power and particularly accuracies not high enough to support reliable and efficient predictions. The currently proposed model produced most relevant and crucial features which were utilized with Random Forest classification and achieved the highest accuracy in comparison to the other state-of-the-art models.

## III. MATERIALS AND METHODS

### A. BENCHMARK DATASET

In order to In order to compare performance of the proposed model with the existing state-of-the-art, originally constructed benchmark dataset [28] was employed in the proposed model. This dataset was originally extracted from CancerLectinDB [27] with 509 cancerlectin protein sequences. To remove similar sequences CD-HIT [30] tool was used with 100% similarity ratio. This resulted in 385 positive cancerlectin sequences. To obtain the negative dataset, UniProtKB (http://www.uniprot.org) was searched with keyword ''Lectin'' and a total of 1550 sequences was obtained. Furthermore, sequences annotated with ambiguous terms like ''by similarity'', ''fragment'', ''probably'', ''probable'' and ''putative'' were excluded which resulted in 891 lectin sequences. 71 sequences were found to be common among the 385 cancerlectins and 891 lectins which were removed and lectins were reduced to 820. Random 385 lectin sequences were selected from 820 lectin sequences to balance the dataset. Moreover, to remove redundancy bias and homology, CD-HIT [30] software was used to exclude sequences having a 50% cutoff ratio for sequence similarity. Finally, the benchmark dataset of 404 sequences were constructed from which 178 are cancerlectins and 226 are non-cancerlectin protein sequences. In order to further examine the performance and efficiency of our proposed model, an independent dataset was also constructed by collecting 40 cancerlectin and 40 non-cancerlectin sequences manually from UniProtKB (http://www.uniprot.org) and NCBI (https://www.ncbi.nlm.nih.gov). The independent dataset is available at http://www.biopred.org/canlect/supl.html. Table.1 includes the breakdown of the benchmark dataset.

**TABLE 1. Breakdown of the benchmark datasets.**

| Lectins | Benchmark Dataset [27] | Independent Dataset |
|---|---|---|
| Cancerlectins | 178 | 40 |
| Non-Cancerlectins | 226 | 40 |
| Overall | 404 | 80 |

Prediction models based on statistical measures mostly have training and testing datasets. But in case of jackknife tests, the construction of training and testing datasets is not required. The benchmark datasets used in jackknife tests can be defined using eqs.(1):

$$X = X^+ \cup X^- \tag{1}$$

where $X^+$ represents 178 cancerlectins $X^-$ represents 226 non-cancerlectins and U denotes the symbol of ''union'' in the set theory.

### B. FEATURE EXTRACTION

The biological sequence formulation is the core requirement in developing an effective Bioinformatics prediction model. The sequence is formulated, without losing any sequence-pattern information or key-order characteristics, with a vector or a discrete model. The reason for this fact, as explained in a comprehensive state-of-the-art review [31], that the formulations of a vector requires to be computed as sequences cannot be handled directly by the existing machine learning algorithms. However, during this whole process there might be possible that some chance of the sequence-pattern information to be lost during a discrete model formulation. To overcome this loss of crucial information from the sequence of proteins, (PseAAC) pseudo amino acid composition was proposed by [32]. In almost all areas of systems biology and Bioinformatics [31], the concept of Chou's PseAAC has been extensively utilized and has become an integral part of many research studies. In the near past, an efficient and a very useful web-server called 'Pse-in-One2.0' [33] was developed, an updated version of 'Pse-in-One' [34], which enabled researchers to generate DNA/RNA and protein/peptide sequence pseudo components for any preferred feature vector as required by the research community.

Two kinds of model construction are usually used to present protein samples. Both discrete and sequential modeling is mostly utilized to represent proteins in vector formulations. The sequential model expresses the protein sequence as its amino acid sequence by the using the following equation (2):

$$\mathbf{X} = Z_1 Z_2 Z_3 Z_4 Z_5 Z_6 \ldots Z_n \tag{2}$$

where $Z_1$ is the first amino acid representation in protein X and $Z_L$ is the last amino acid. The total length of the sequence is represented as 'n'.

In the second model, discrete model representation of a protein sample is represented using its amino acid composition (AAC). The protein $\mathbf{X}$ representation using a discrete model is defined using equation (3):

$$\mathbf{X} = [d_1 \quad d_2 \quad d_3 \quad \ldots \quad d_{20}]^T \tag{3}$$

where $d_a$ $(a = 1, 2, 3, \ldots 20)$ represents the useful component features defined by the extraction methods using relevant amino acids in protein $\mathbf{X}$. These components are further utilized in the feature extraction methods based on the statistical moments.

### C. STATISTICAL MOMENTS

In statistics and probability distributions, quantitative measures of certain types are useful in concentration study of unique configurations. The studies related to these configurations in data collections of pattern recognition problems are known as moments [35]. Moments are beneficial in many pattern recognition relevant problems for producing features that are not reliant on parameters from the given pattern or sequence [36]–[40].

Many different moment orders are used to describe several data properties. Some moments are used for estimation of the size of data and some reveal data orientation and data eccentricity. Various statisticians have formed several moments based on polynomials and distribution functions. Raw, Central and Hahn moments are further utilized to explicate the problem in the current research study [41].

Raw moments are the moments that are used in mean, variance and asymmetry calculation of probability distribution. These raw moments are neither scale-invariant nor location-invariant. The same process is followed in case of Central moments, but the calculations are performed using the data centroid. The central moments are scale-variant, but they remain location-invariant with respect to centroid as they are calculated along the data centroid. The Hahn moments are based on Hahn polynomials. These Hahn moments are neither scale-invariant nor location-variant [42]–[45]. These moments are able to extract obscure features from protein sequences are primarily significant because they are sensitive to ordered biological sequence information. In the proposed study, the linear structured protein sequence is utilized which used 2D version of the aforementioned moments and hence expressed by eqs.(2). This linear structured sequence information is further transformed into a 2D notation. A row major scheme, defined in eqs. (4), is used to transform a linear protein structure to a 2D structure:

$$m = \lceil \sqrt{p} \rceil \tag{4}$$

where 'p' is the sequence length of a sample protein and 'm' represents the 2D square matrix dimensions. The N' matrix, in eqs.(5), is formed using the ordering obtained from

equation (4) having 'k x k' rows and columns respectively.

$$N' = \begin{bmatrix} O_{1\to1} & O_{1\to2} & \cdots & O_{1\to j} & \cdots & O_{1\to k} \\ O_{2\to1} & O_{2\to2} & \cdots & O_{2\to j} & \cdots & O_{2\to k} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ O_{i\to1} & O_{i\to2} & \cdots & O_{i\to j} & \cdots & O_{i-k} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ O_{k\to1} & O_{k\to2} & \cdots & O_{k\to j} & \cdots & O_{k\to k} \end{bmatrix} \quad (5)$$

The Raw moments are computed using the values of N'. The Raw moments of $R(a, b)$, a 2D continuous function having order (a + b), are computed from eqs. (6):

$$R_{ab} = \sum_p \sum_q p^a q^b N'(p, q) \quad (6)$$

The raw moments from the above equation are computed up to 3rd order. These raw moments assume the origin of data as the reference point from where these moments are computed and the origin is utilized as the distance between the components [46]–[49]. The Raw moment unique features computed up to 3rd order are labeled as $R_{00}, R_{01}, R_{10}, R_{11}, R_{02}, R_{20}, R_{12}, R_{21}, R_{30}$ & $R_{03}$.

The center of gravity of any data is also considered as its centroid. A data point from where all the data is uniformly distributed in all directions. These directions are relations of its weighted average [46], [49]–[51]. The central moments unique features computed up to 3rd order, using the centroid of the data as their reference point, are computed from the following eqs.(7):

$$C_{ab} = \sum_p \sum_q (p - \bar{p})^a (q - \bar{q})^b N'(p, q) \quad (7)$$

The unique features from Central moments, up to 3rd order, are labeled as $C_{00}, C_{01}, C_{10}, C_{11}, C_{02}, C_{20}, C_{12}, C_{21}, C_{30}$ & $C_{03}$. Here the centroids are calculated as $\bar{p}$ and $\bar{q}$ from eqs. (8) and eqs. (9):

$$\bar{p} = \frac{R_{10}}{R_{00}}, \quad (8)$$

$$\bar{q} = \frac{R_{01}}{R_{00}} \quad (9)$$

For computing Hahn moments, transformation into square matrix notations from 1D notation is required. Discrete Hahn moments or orthogonal moments, also known as moments of 2D, require a square matrix input data in 2D structure [52]. These moments possess inverse properties as they are orthogonal in nature. The reconstruction of the original data can be performed using the discrete Hahn moments inverse functions. After calculating moments, it is further observed that the positional and compositional features of a protein sequence are somehow conserved within the computed moments [35], [41]–[45], [52], [53]. To compute the Orthogonal Hahn moments, 2D input data in the form of N' matrix was utilized. The Hahn polynomial of order 'm' can be computed from eqs.(10):

$$h_m^{x,y}(r, M) = (M + y - 1)_m (M - 1)_m \sum_{s=0}^{m} (-1)^s \times \frac{(-m)_s (-r)_s (2M + x + y - m - 1)_s}{(M + y - 1)_s (M - 1)_s} \cdot \frac{1}{s!} \quad (10)$$

Here 'p' and 'q' (p > -1, q > -1) are controlling the shape of polynomials using adjustable parameters. The aforementioned Pochhammer symbol was defined as follows in eqs.(11):

$$(\mathbb{P})_s = \mathbb{P}(\mathbb{P} + 1) \ldots (\mathbb{P} + s - 1) \quad (11)$$

And was simplified further by the Gamma operator in eqs.(12):

$$(\mathbb{P})_s = \frac{\Gamma(\mathbb{P} + s)}{\Gamma(\mathbb{P})} \quad (12)$$

A weighting function and square norm are usually used to scale the raw values of Hahn moments given as in eqs.(13):

$$h_m^{\tilde{x},y}(r, M) = h_m^{x,y}(r, M) \sqrt{\frac{\mathbb{P}(r)}{s_m^2}}, \quad m = 0, 1, \ldots, M - 1 \quad (13)$$

Meanwhile, in eqs.(14),

$$\mathbb{P}(r) = \frac{\Gamma(p + r + q)\Gamma(q + r + 1)(p + q + r + 1)_M}{(p + q + 2r + 1)m!(M - r - 1)!} \quad (14)$$

For the 2D discrete data, the Hahn moments are computed up to 3rd order as follows in eqs.(15):

$$H_{pq} = \sum_{j=0}^{M-1} \sum_{i=0}^{M-1} N'_{i,j}\, h_p^{\tilde{x},y}(j, M)\, h_q^{\tilde{x},y}(i, M), \quad p, q = 0, 1, \ldots, M - 1 \quad (15)$$

The Hahn moments based unique features are represented by $H_{00}, H_{01}, H_{10}, H_{11}, H_{02}, H_{20}, H_{12}, H_{21}, H_{30}$ & $H_{03}$. 10 Raw, 10 Central and 10 Hahn moments for every protein sequence are computed which are up to 3rd order and are further unified into the miscellany Super Feature Vector (SFV).

### D. POSITION-RELATIVE-INCIDENT-MATRIX(PRIM)

For identifying the protein characteristics, the ordered location of the amino-acids in the protein sequences are of pivotal significance [36], [48], [51], [54]. The relative position of an amino acid, in any protein sequence, is considered a core pattern that utilizes the physical features of the protein sequence. The PRIM is used to represent the protein sequence in (20 x 20) order. The relative position of every amino-acid in the given protein sequence is extracted in the form of the following matrix using eqs.(16):

$$V_{PRM} = \begin{bmatrix} O_{1\to1} & O_{1\to2} & \cdots & O_{1\to j} & \ldots & O_{1\to20} \\ O_{2\to1} & O_{2\to2} & \cdots & O_{2\to j} & \cdots & O_{2\to20} \\ \vdots & \vdots & \ldots & \vdots & \ldots & \vdots \\ O_{i\to1} & O_{i\to2} & \cdots & O_{i-j} & \cdots & O_{i-20} \\ \vdots & \vdots & \ldots & \vdots & \ldots & \vdots \\ O_{k\to1} & O_{k\to2} & \cdots & O_{k\to j} & \cdots & O_{k\to20} \end{bmatrix} \quad (16)$$

Here the indication score of the $j^{th}$ position amino-acid is determined by the $N_{i\to j}$ with respect to the first occurrence of the $i^{th}$ amino-acid. This score is substitution of biological evolutionary process is performed by amino-acid type '$j$'. The positional values of 20 native amino-acid occurrences are represented in alphabetical order. 400 coefficients are obtained from successful computations from position relative incidences in the form of $N_{PRIM}$ matrix. (Figure.2) shows the 2D-HeatMap of the summation of all PRIMs from cancerlectin benchmark dataset.

10 Hahn moments, 10 Central moments and 10 Raw moments were computed using the 2D $N_{PRIM}$ matrix up to 3rd order. 30 more unique features were further incorporated into the miscellany SFV.

### E. REVERSE-POSITION-RELATIVE-INCIDENT-MATRIX(R-PRIM)

In cell biology, there are often many cases where the biological sequences are homologous in nature. This usually happens when the same ancestor is part of the evolution process and more than one sequence is evolved from it. In such cases, the performance of the classifier is hugely affected using these homologous sequences. Hence, to produce accurate results, effective and reliable sequence similarity searching is performed during results processing. In machine learning, accuracy and efficiency is hugely dependent on the meticulousness and thoroughness of algorithms through which most pertinent features in the data are extracted. During the learning phase in machine learning algorithms, learning and adaptation, of the most embedded obscure patterns in the data, are performed to undercover the relevant and pertinent features [36], [48], [51], [54]. R-PRIM and PRIM computations have the same procedure but only R-PRIM works with the reverse protein sequence ordering. Computing R-PRIM uncovered hidden patterns which enabled alleviation of any ambiguities between homologous sequences. R-PRIM was also formed as 2D matrix of (20 x 20) order with 400 coefficients. It is defined by eqs.(17):

$$N_{R-PRM}=\begin{bmatrix} O_{1\to1} & O_{1\to2} & \cdots & O_{1\to j} & \cdots & O_{1\to20} \\ O_{2\to1} & O_{2\to2} & \cdots & O_{2\to j} & \cdots & O_{2\to20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ O_{i\to1} & O_{i\to2} & \cdots & O_{i\to j} & \cdots & O_{i\to20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ O_{k\to1} & O_{k\to2} & \cdots & O_{k\to j} & \cdots & O_{k\to20} \end{bmatrix} \quad (17)$$

10 Hahn moments, 10 Central moments and 10 Raw moments were computed using the 2D $N_{R-PRIM}$ matrix up to 3rd order. 30 more unique features were further incorporated into the miscellany SFV.

### F. FREQUENCY-DISTRIBUTION-VECTOR (FDV)

The distribution of occurrence in every amino-acid of a protein sequence was utilized to form a frequency distribution vector (FDV). The FDV is defined as in eqs.(18):

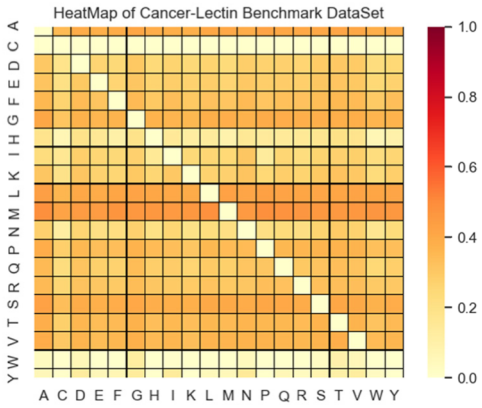


**FIGURE 2.** The HeatMap of PRIMs from Cancerlectins benchmark Dataset.

$$\mu = \{\alpha_i, \ldots, \alpha_{20}\} \quad (18)$$

Here the occurrence frequency of the $i^{th}$ $(1 \le i \le 20)$ relevant amino-acid is represented as $\alpha_i$. However, the alleviation of information about the position relevance of amino-acids in a sequence is performed using these measures. 20 features from FDV are also further incorporated into the miscellany SFV.

### G. AAPIV(ACCUMULATIVE-ABSOLUTE-POSITION-INCIDENCE-VECTOR)

The frequency distribution vector stores the distributional information of amino-acids but does not have any relevant amino-acid relative positional information. Using AAPIV the relative positional information was accommodated from 20 native amino-acids in a protein sequence with a length of 20 associated critical features [51], [54]. These 20 critical features from AAPIV (see eqs.(19)) are also incorporated into the miscellany SFV.

$$AAPIV = \{\Psi_i, \ldots, \Psi_{20}\} \quad (19)$$

Here $\Psi_i$ is from protein sequence $R_x$ having 'n' total amino-acids, which can be calculated using eqs.(20):

$$\Psi_i = \sum_{x=1}^{n} R_x \quad (20)$$

### H. RAAPIV(REVERSE-ACCUMULATIVE-ABSOLUTE-POSITION-INCIDENCE-VECTOR)

R-AAPIV and AAPIV computations have the same procedure but only R-AAPIV works with the reverse protein sequence ordering. Computing R-AAPIV utilized reverse relative positional information by under covering deep and hidden patterns of every sample features [51], [54]. R-AAPIV is formed as the following eqs.(21) and generates 20 valuable features. These 20 unique critical features from R-AAPIV are also incorporated into the miscellany SFV.

$$\mathrm{R}-AAPIV = \{\Psi_i, \ldots, \Psi_{20}\} \quad (21)$$

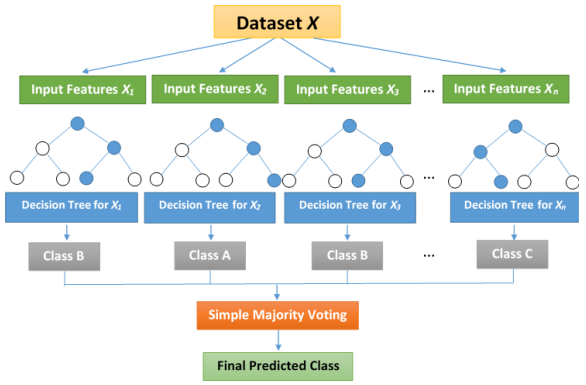


**FIGURE 3.** The structure of the random forest classifier.

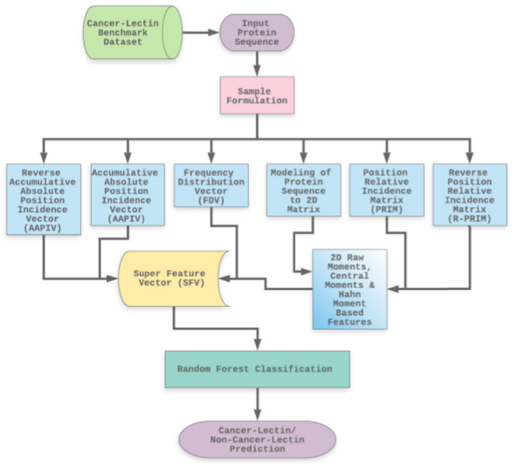


**FIGURE 4.** The flowchart of the overall proposed model.

Here $\alpha_i$ is any element of R-AAPIV, from protein sequence $R_x$ having 'n' total amino-acids, can be calculated using eqs.(22):

$$\Psi_i = \sum_{x=1}^{n} Reverse(R)_x \tag{22}$$

After extracting features using all the aforementioned methods, the SFV of 150-D features was constructed to be used for further processing in classification algorithm.

## IV. CLASSIFICATION ALGORITHMS

### A. RANDOM FORESTS

Many research studies, relevant to bioinformatics, have utilized ensemble learning methods in past. These studies have performance measures of highly efficient and accurate outcomes. The aggregation results of many classifiers are utilized in ensemble learning methods. Boosting [55], [56] and Bagging [57] are the two most commonly used methods which perform tree-based classifications.

In boosting method extra weights are propagated to points, through trees which are successive, and then later predicted incorrectly by the previous classifiers. The prediction is decided using the weighted vote in the end. In bagging method, from the data using a bootstrap sample, each tree is constructed independently and the successive trees do not rely on previous trees. The prediction is decided using the simple majority vote in the end.

Random Forests were introduced by [58] which added randomness as an additional layer to bagging. The classification trees construction changed after random forests. These changes are reflected by using a different bootstrap sample of data for adding the construction of each tree. In standard classification trees, the splitting of each node is performed among all the variables by dividing each node equally. However, random forests choose the best predictor among a subset of predictors for splitting of each node, which are chosen randomly at that node (Figure.3 shows the structure of the random forest classifier). This strategy is counterintuitive and performs very well against many other classifiers, such as discriminant analysis, support vector machine and neural networks and it is robust against overfitting as well [59].

### B. ALGORITHM: SUPERVISED LEARNING USING RANDOM FORESTS

The python library Scikit-Learn [60] was utilized for the implementation of random forest classifier for our model trainings and simulations. The default parameter value of 10 was increased to 25 in order to increase the number of processing trees in the classifier. It was observed in previous study [61], that the theoretical upper limit of trees in random forest classifier is 128 and will not improve the efficiency nor the accuracy of classifier further if there is any increase in upper limit of number of trees. During the experimentation process, minimal contribution to the accuracy of the classifier was observed if the forest was implemented using more than 25 trees, but the overall size of the proposed model was enhanced substantially. (Figure.4) illustrates a flowchart to show the overall process of the proposed model.

## V. EXPERIMENTS AND RESULTS

The overall performance of the proposed model is examined by some methods that will assess and verify how well the prediction model has performed. Several parameters based on estimates and assessments are used to measure the performance of classifiers.

### A. JACKKNIFE TEST

During the evaluation of statistical predictors, several cross-validation tests are applied commonly. Jackknife is considered a consistent and a reliable test among these tests. Therefore, many classifiers are assessed by extensive jackknife tests in most research studies. The accurate estimation of a model can be measured using the jackknife method which is calculated over the entire dataset. The testing of a model is performed on the left out items after successful training. After the training and testing are successfully completed,

a confusion matrix is obtained through which the true positive and negative values and false positive and negative values are utilized to estimate the accuracy associated with items in the data. The mean of all the accuracies is used at the end for final accuracy of the prediction model.

During jackknife, each protein sequence is tested and the rest of the protein sequences are used in calculation of the remaining parameters. This is performed by leaving out each protein sequence from the dataset and calculating its estimate and the average of all these estimations from the left out proteins. The performance of our prediction model on a benchmark dataset is listed in Table 4.

### B. K-FOLD CROSS-VALIDATION TEST

K-fold cross validation (KFCV) technique is most commonly used by practitioners for estimation of errors in classifications. Also known as rotation estimation, KFCV splits a dataset into 'K' folds which are randomly selected and are equal in size approximately. The prediction error of the fitted model is calculated by predicting the $k^{th}$ part of the data which is dependent on other K-1 parts to fit the model. The error estimates of K from the prediction are combined together using the same procedure for each k = 1, 2, . . . , K.

In the KFCV tests, the selection of 'K' is considered as a significant attribute. To testify errors in prediction models, cross validations (K = 10) tests have been used in many research studies. 10-Fold tests proved to have accurate results in our proposed model and proved to be much better than other classifiers. These results are listed in Table 6.

## VI. EVALUTAION PARAMETERS

To estimate the performance of the prediction model, three cross validation methods are often used. For performance evaluation of statistical based classifiers, sub-sampling (5-fold or 10-fold cross validation) tests, independent tests and jackknife tests are most commonly used. Accordingly, we used jackknife and independent tests to evaluate the performance of the classifier.

Accuracy (Acc), Sensitivity (Sn), Specificity (Sp) and Mathew's Correlation Coefficient (MCC) and F-measure metrics are the most commonly used metrics, in binary classification problems, in the proposed prediction model to measure its quality and performance. The eqs.(23) defines these metrics as follows for analyses of the evaluators.

$$\begin{cases} Sn = \dfrac{TP}{TP + FN} \\ Sp = \dfrac{TN}{TN + FP} \\ Accuracy = \dfrac{TP + TN}{TP + TN + FP + FN} \\ MCC = \dfrac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \end{cases} \tag{23}$$

Unfortunately, the conventional formulations used in above metrics lack in intuitiveness and have been difficult to understand as many scientists have faced complex measures in utilizing them especially the MCC. To address this issue, Chou's four intuitive equations were converted by [62], [63] and these conventional equations utilized symbols which were introduced in [64]. The symbols that define these equations are $Q^+, Q^-, Q_-^+$ *and* $Q_+^-$. The description of these symbols is defined in Table.2.

**TABLE 2. Description of symbols used to define these equations.**

| Symbols | Description of Symbols |
|---|---|
| $Q^+$ | The total number of true cancerlectins |
| $Q_-^+$ | The total number of true cancerlectins incorrectly predicted as non-cancerlectins |
| $Q^-$ | The total number of true non-cancerlectins |
| $Q_+^-$ | The total number of non-cancerlectins predicted as cancerlectins |

From the above correspondence in Table.2, we can define eqs.(24):

$$\begin{cases} TP = Q^+ - Q_-^+ \\ TN = Q^- - Q_+^- \\ FP = Q_+^- \\ FN = Q_-^+ \end{cases} \tag{24}$$

By substituting above equation (24) to equation (23) we get,

$$\begin{cases} Sn = 1 - \dfrac{Q^{\pm}}{Q^+} \\ Sp = 1 - \dfrac{Q_+^-}{Q^-} \\ Accuracy = 1 - \dfrac{Q_-^+ + Q_+^-}{Q^+ + Q^-} \\ MCC = \dfrac{1 - \left(\dfrac{Q^{\pm}}{Q^+} + \dfrac{Q_+^-}{Q^-}\right)}{\sqrt{\left(1 + \frac{Q_+^- - Q^{\pm}}{Q^+}\right)\left(1 + \frac{Q^{\pm} - Q_+^-}{Q^-}\right)}} \end{cases} \tag{25}$$

The above eqs.(25) has the same meaning as the eqs.(23) but it is more easy to understand and intuitive. Table.3 defines the detail description of these equations.

The set of metrics used in above Table.3 are not applicable to multi-labeled prediction models rather they are only useful for single labeled-systems. A different set of metrics exists for multi-labeled-systems which have been used by various researchers in [65]–[68]. The comparison of existing models with the proposed model is mentioned in Table.4

## VII. RESULTS AND DISCUSSIONS

It is necessary to compare the proposed novel model with other state-of-the-art models in order to estimate the performance of the proposed prediction model. The proposed model was compared with well-known existing classifiers and Random Forest performed quite better in accuracy and

**TABLE 3.** Description of equations used eqs. (25).

| When, | Then, | Description |
|---|---|---|
| $Q_{-}^{+} = 0$ | Sn = 1 | None of the cancerlectin protein is predicted as a non-cancerlectin protein |
| $Q_{-}^{+} = Q^{+}$ | Sn = 0 | All of the cancerlectin proteins were incorrectly predicted as non-cancerlectin proteins |
| $Q_{+}^{-} = 0$ | Sp = 1 | None of the non-cancerlectin protein is incorrectly predicted as cancerlectin protein |
| $Q_{+}^{-} = Q^{-}$ | Sp = 0 | All of the non-cancerlectin proteins are incorrectly predicted as cancerlectins |
| $Q_{-}^{+} + Q_{+}^{-} = 0$ | ACC = 1, MCC = 1 | None of the cancerlectin proteins and none of the non-cancerlectin proteins were incorrectly predicted |
| $Q_{-}^{+} = Q^{+}$ *and* $Q_{+}^{-} = Q^{-}$ | ACC = 0, MCC = -1 | All of the cancerlectin proteins and all of the non-cancerlectin proteins were incorrectly predicted. |
| $Q_{-}^{+} = \frac{Q^{+}}{2}$ *and* $Q_{+}^{-} = \frac{Q^{-}}{2}$ | ACC = 0.5, MCC = 0 | The overall prediction is not a better than any other random prediction outcome. |

efficiency in predicting cancerlectins from non-cancerlectins. The performance of all the compared classifiers is listed in Table 6. Furthermore, some models based on computational methodology have been developed in recent past using the same benchmark dataset of cancerlectins [27]. In current comparison, [28] was the pioneering work in development of a prediction model for cancer-lectins based on amino-acid compositions, dipeptide-compositions and evolutionary information. They utilized SVM with PROSITE domain information with integration of PSSM (position-specific-scoring-matrix). The maximum accuracy of 69.09% was achieved during their study. Similarly, [29] utilized g-gap dipeptides with obtaining an accuracy of 75.19%.

**TABLE 4.** Comparison of the proposed model with state-of-the-art models based on Jackknife tests.

| Classifiers | Sn (%) | Sp (%) | ACC (%) | MCC |
|---|---|---|---|---|
| CancerPred (AAC) [28] | 68.0 | 64.2 | 65.8 | 0.32 |
| CancerPred (DC) [28] | 67.3 | 62.8 | 64.8 | 0.30 |
| CancerPred (Split 2-part) [28] | 66.3 | 64.2 | 65.1 | 0.31 |
| CancerPred (Split 4-part) [28] | 65.1 | 66.9 | 66.1 | 0.32 |
| CancerPred (PSSM) [28] | 67.9 | 68.6 | 68.3 | 0.36 |
| CancerPred (PSSM-14-PROSITES) [28] | 68.0 | 69.9 | 69.1 | 0.38 |
| CaLecPred (G-gap dipeptide) [29] | 69.1 | 80.1 | 75.2 | - |
| PSSM-CTD-PseAAC [71] | 77.9 | 71.7 | 74.8 | 0.497 |
| Tripeptide [71] | 75.28 | 80.53 | 77.48 | - |
| Proposed Model | 91.57 | 85.84 | 88.36 | - |

Recently, [69] enhanced the accuracy to 77.48% and used tri-peptide compositions based on optimal feature subset selection methodology. However, using these features was still not useful to obtain an efficient and high accuracy. In contrast to the above models, our proposed model has achieved 88.36% accuracy using statistical moments based features and random forest based classification. The performance of the current prediction model using jackknife tests on benchmark dataset is listed in Table 4. In addition to jackknife tests, an independent test was also performed using the independent dataset. The independent dataset consists of 40 cancerlectins and 40 non-cancerlectins. The comparison of proposed model and CaLecPred [29] webserver using independent dataset is listed in Table 5 and Table 7. Due to unavailability of the webservers, other methods such as [28], [70] could not be utilized for independent tests. Furthermore, 10-fold cross-validation test was also conducted using random forest classifier on benchmark dataset and obtained the accuracy of 90.39% listed in Table 6. The ROCs of 5-fold and 10-fold cross-validation tests are shown in (Figure.5) and (Figure.6) respectively. Hahn moments based feature sets utilized in the currently proposed model are easier for the random forest

**TABLE 5.** Comparison of proposed model with state-of-the-art models based on Independent test.

| Classifiers | Sn(%) | Sp(%) | ACC(%) | F-Measure | MCC |
|---|---|---|---|---|---|
| CaLecPred (G-gap dipeptide) [29] | 59.46 | 58.14 | 58.75 | 0.5714 | 0.1755 |
| Proposed Model | 61.54 | 71.43 | 65.0 | 0.6957 | 0.3145 |

**TABLE 6.** Comparison of classifiers for predicting Cancerlectins using 10-fold cross validations.

| Classifier | Sn (%) | Sp (%) | ACC (%) | MCC |
|---|---|---|---|---|
| Naïve Bayes | 77.48 | 33.20 | 53.97 | 0.1182 |
| PNN | 63.14 | 54.86 | 58.74 | 0.1839 |
| Ensemble (AdaBoost) | 66.64 | 77.22 | 72.23 | 0.4430 |
| SVM | 74.78 | 72.80 | 73.73 | 0.4767 |
| KNN | 68.80 | 79.62 | 74.55 | 0.4908 |
| Random Forest | 84.96 | 95.20 | 90.39 | 0.8101 |

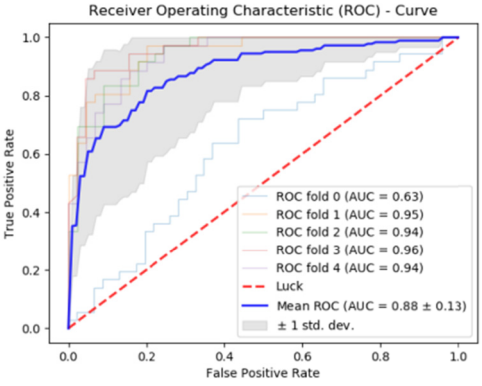


**FIGURE 5.** ROC Curve for 5-Fold cross-validation using random forest classifier.

**TABLE 7.** List and comparison of 15 random Cancerlectins tested on CaLecPred [29] and CanLect webservers.

| UniProtKB ID | CaLecPred [29] Prediction | CanLect Prediction |
|---|---|---|
| P49257 | Incorrect | Correct |
| Q86SR1 | correct | Incorrect |
| P11226 | Incorrect | Incorrect |
| P05162 | Incorrect | Correct |
| P0CG48 | correct | Correct |
| P82683 | Incorrect | Correct |
| P0CG47 | correct | Correct |
| P62987 | correct | Incorrect |
| Q9UHV8 | Incorrect | Correct |
| Q13404 | correct | Correct |
| Q5NKN4 | Incorrect | Correct |
| A4KWA1 | Incorrect | Correct |
| P09382 | correct | Correct |
| Q9Y286 | correct | Incorrect |
| Q7Z7M9 | correct | Correct |

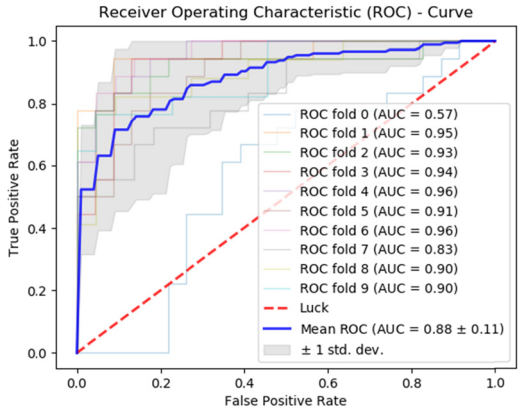


**FIGURE 6.** ROC curve for 10-Fold cross-validation using random forest classifier.

based classifier to classify the feature vectors in acute time. Using the computational cost of training and testing, previous models were not able to produce better results during the classification process. However, the proposed model results are highly efficient as compared to previous models.

## VIII. WEBSERVER

As observed in past studies by many researchers [72]–[76], the development of a web-server is highly significant and useful for building more useful prediction methodologies. Thus like most of the research studies by many scientists in past [49], [51], [54], efforts for a user friendly webserver have been made to provide ease to biologists and scientists in drug discovery. The webserver for cancerlectin predictions is freely available at https://www.biopred.org/canlect which is developed using a web development framework for python known as Flask (version 1.1.1). The step-wise instructions to interact with the webserver are provided below.

### A. STEP-1

Open your web browser and navigate to www.biopred.org/canlect. The first page (see Figure.7) of the webserver is **Home**, page from where you can proceed to **ReadMe, Server**, **Data** and **Citations** pages through provided navigation links. The **Server** tab navigates to the portal for

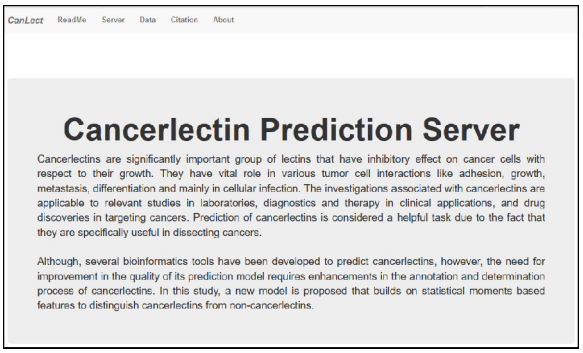


**FIGURE 7. The GUI of www.biopred.org/canlect webserver for cancerlectin predictions.**

prediction process. **Data** tab facilitates the user with links to download the benchmark datasets used during the training and testing process. Finally, the **Citations** tab redirects the user to the page where information about the relevant paper and its citation is provided. In order to perform the prediction, kindly click on the **Server** tab.

### B. STEP-2

On the **Server** page, the user is provided with an empty text-area where the user can input the cancerlectin or non-cancerlectin sequence for prediction. The sequence input to the webserver is required to be in FASTA format. The **Submit** button will proceed to the prediction process for the input sequence. The results of the prediction process will appear on the **Results** page. The time of the prediction process totally depends upon the length of the input sequence.

### C. STEP-3

On the **Data** page, the user is provided with links to download the benchmark datasets for future experimentations.

### D. STEP-4

On the **ReadMe** page, the user is provided with relevant information about the current model which includes the details about the operation algorithm used for predictions.

## IX. CONCLUSION

An efficient and reliable model has been developed in the current study for predicting cancerlectins using statistical moments and random forest classifier. Several classification techniques were proposed to predict cancerlectins, but currently proposed model proved better in accuracy than the existing models. Among these models, our proposed model has achieved highest accuracies of 88.36% using jackknife tests for the benchmark dataset of cancerlectins and 65.0% for independent dataset of cancerlectins in classifications which are respectively so far the best classifications for prediction of cancerlectin proteins. Table.7 shows the comparison of 15 randomly selected cancerlectins prediction results between the CaLectPred and proposed CanLect-Pred webservers.

## X. COMPLIANCE WITH ETHICAL STANDARDS

### A. CONFLICT OF INTEREST

Authors have declared that they have no conflict of interest.

### B. ETHICAL APPROVAL

This article does not contain any studies involved with human participants or animals performed by any of the authors.

## REFERENCES

[1] R. Macholz, "The Lectins. Properties, functions, and applications in biology and medicine. Herausgegeben von I. E. Liener, N. Sharon und I. J. Goldstein. 600 Seiten, zahlr. Abb. und Tab. Academic Press, Inc., Orlando u. A. 1986. Preis: 54,–£; 72,50 $," *Nahrung*, vol. 32, no. 2, pp. 212–213, 1988.

[2] H. Lis and N. Sharon, "Lectins: Carbohydrate-specific proteins that mediate cellular recognition," *Chem. Rev.*, vol. 98, no. 2, pp. 637–674, Apr. 1998.

[3] M. Vijayan and N. Chandra, "Lectins," *Current Opinion Struct. Biol.*, vol. 9, no. 6, pp. 707–714, 1999.

[4] S. Hu and D. T. Wong, "Lectin microarray," *Proteomics-Clin. Appl.*, vol. 3, no. 2, pp. 148–154, 2009.

[5] S.-Y. Jiang, Z. Ma, and S. Ramachandran, "Evolutionary history and stress regulation of the lectin superfamily in higher plants," *BMC Evol. Biol.*, vol. 10, no. 1, p. 79, 2010.

[6] G. R. Vasta and H. Ahmed, Eds., *Animal Lectins: A Functional View*. Boca Raton, FL, USA: CRC Press, 2008.

[7] N. Sharon, "Lectins: Carbohydrate-specific reagents and biological recognition molecules," *J. Biol. Chem.*, vol. 282, no. 5, pp. 2753–2764, Feb. 2007.

[8] D. Hu, H. Tateno, and J. Hirabayashi, "Lectin engineering, a molecular evolutionary approach to expanding the lectin utilities," *Molecules*, vol. 20, no. 5, pp. 7637–7656, Apr. 2015.

[9] G. R. Vasta, H. Ahmed, and E. W. Odom, "Structural and functional diversity of lectin repertoires in invertebrates, protochordates and ectothermic vertebrates," *Current Opinion Struct. Biol.*, vol. 14, no. 5, pp. 617–630, Oct. 2004.

[10] N. Sharon and H. Lis, "Lectins as cell recognition molecules," *Science*, vol. 246, no. 4927, pp. 227–234, Oct. 1989.

[11] F.-T. Liu and G. A. Rabinovich, "Galectins as modulators of tumour progression," *Nature Rev. Cancer*, vol. 5, no. 1, pp. 29–41, Jan. 2005.

[12] K. L. Abbott and J. M. Pierce, "Lectin–based glycoproteomic techniques for the enrichment and identification of potential biomarkers," *Methods Enzymol. Glycobiol.*, vol. 480, pp. 461–476, 2010.

[13] G. R. Vasta, "Roles of galectins in infection," *Nature Rev. Microbiol.*, vol. 7, no. 6, pp. 424–438, Jun. 2009.

[14] S. Jamal, V. Lavanya, A. Mohamed Adil, and N. Ahmed, "Lectins—The promising cancer therapeutics," *Oncobiol. Targets*, vol. 1, no. 1, p. 12, 2014.

[15] R. Lotan and A. Raz, "Lectins in cancer cells," *Ann. New York Acad. Sci.*, vol. 551, no. 1, pp. 385–398, 1988.

[16] E. G. De Mejía and V. I. Prisecaru, "Lectins as bioactive plant proteins: A potential in cancer treatment," *Crit. Rev. Food Sci. Nutrition*, vol. 45, no. 6, pp. 425–445, Sep. 2005.

[17] H. Ghazarian, B. Idoni, and S. B. Oppenheimer, "A glycobiology review: Carbohydrates, lectins and implications in cancer therapeutics," *Acta Histochemica*, vol. 113, no. 3, pp. 236–247, May 2011.

[18] Y. K. Song, T. R. Billiar, and Y. J. Lee, "Role of galectin-3 in breast cancer metastasis: Involvement of nitric oxide," *Amer. J. Pathol.*, vol. 160, no. 3, pp. 1069–1075, Mar. 2002.

[19] M. D. Swanson, H. C. Winter, I. J. Goldstein, and D. M. Markovitz, "A lectin isolated from bananas is a potent inhibitor of HIV replication," *J. Biol. Chem.*, vol. 285, no. 12, pp. 8646–8655, Mar. 2010.

[20] C. Miller, S. Wilgenbusch, M. Michaels, D. S. Chi, G. Youngberg, and G. Krishnaswamy, "Molecular defects in the mannose binding lectin pathway in dermatological disease: Case report and literature review," *Clin. Mol. Allergy*, vol. 8, no. 1, p. 6, 2010.

[21] S. H. Choi, S. Y. Lyu, and W. B. Park, "Mistletoe lectin induces apoptosis and telomerase inhibition in human A253 cancer cells through dephosphorylation of akt," *Arch. Pharm. Res.*, vol. 27, no. 1, pp. 68–76, Jan. 2004.

[22] A. Gomez-Brouchet, F. Mourcin, P.-A. Gourraud, C. Bouvier, G. De Pinieux, S. Le Guelec, P. Brousset, M.-B. Delisle, and C. Schiff, "Galectin–1 is a powerful marker to distinguish chondroblastic osteosarcoma and conventional chondrosarcoma," *Hum. Pathol.*, vol. 41, no. 9, pp. 1220–1230, Sep. 2010.
[23] S. Nakahara, N. Oka, and A. Raz, "On the role of galectin-3 in cancer apoptosis," *Apoptosis*, vol. 10, no. 2, pp. 267–275, Mar. 2005.
[24] G. Canesin, P. Gonzalez-Peramato, J. Palou, M. Urrutia, C. Cordón-Cardo, and M. Sánchez-Carbayo, "Galectin–3 expression is associated with bladder cancer progression and clinical outcome," *Tumor Biol.*, vol. 31, no. 4, pp. 277–285, Aug. 2010.
[25] Y. Hu, Y. Lu, S. Wang, M. Zhang, X. Qu, and B. Niu, "Application of machine learning approaches for the design and study of anticancer drugs," *Current Drug Targets*, vol. 20, no. 5, pp. 488–500, Mar. 2019.
[26] M. Jordinson, J. Calam, and M. Pignatelli, "Lectins: From basic science to clinical application in cancer prevention," *Expert Opinion Invest. Drugs*, vol. 7, no. 9, pp. 1389–1403, Sep. 1998.
[27] D. Damodaran, J. Jeyakani, A. Chauhan, N. Kumar, N. R. Chandra, and A. Surolia, "CancerLectinDB: A database of lectins relevant to cancer," *Glycoconj J.*, vol. 25, no. 3, pp. 191–198, Apr. 2008.
[28] R. Kumar, B. Panwar, J. S. Chauhan, and G. P. Raghava, "Analysis and prediction of cancerlectins using evolutionary and domain information," *BMC Res. Notes*, vol. 4, no. 1, p. 237, 2011.
[29] H. Lin, W. X. Liu, J. He, X. H. Liu, H. Ding, and W. Chen, "Predicting cancerlectins by the optimal g-gap dipeptides," *Sci. Rep.*, vol. 5, no. 1, 2015, Art. no. 16964.
[30] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD–HIT: Accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, Dec. 2012.
[31] K.-C. Chou, "Impacts of bioinformatics to medicinal chemistry," *Med. Chem.*, vol. 11, no. 3, pp. 218–234, Mar. 2015.
[32] K.-C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins Struct. Funct. Genet.*, vol. 44, no. 1, p. 60, Jul. 2001.
[33] B. Liu, H. Wu, and K.-C. Chou, "Pse-in-One 2.0: An improved package of Web servers for generating various modes of pseudo components of DNA, RNA, and Protein Sequences," *Nature Sci.*, vol. 09, no. 04, pp. 67–91, 2017.
[34] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K.-C. Chou, "Pse-in-One: A Web server for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Nucleic Acids Res*, vol. 43, no. W1, pp. W65–W71, Jul. 2015.
[35] R. C. Papademetriou, "Reconstructing with moments," in *Proc. Int. Conf. Pattern Recognit.*, vol. 3, 1992, pp. 476–480.
[36] A. H. Butt, S. A. Khan, H. Jamil, N. Rasool, and Y. D. Khan, "A prediction model for membrane proteins using moments based features," *BioMed. Res. Int.*, vol. 2016, pp. 1–7, 2016.
[37] A. H. Butt, N. Rasool, and Y. D. Khan, "A treatise to computational approaches towards prediction of membrane protein and its subtypes," *J. Membrane Biol.*, vol. 250, no. 1, pp. 55–76, Feb. 2017.
[38] A. H. Butt, N. Rasool, and Y. D. Khan, "Predicting membrane proteins and their types by extracting various sequence features into Chou's general PseAAC," *Mol. Biol. Rep.*, vol. 45, no. 6, pp. 2295–2306, Dec. 2018.
[39] A. H. Butt, N. Rasool, and Y. D. Khan, "Prediction of antioxidant proteins by incorporating statistical moments based features into Chou's PseAAC," *J. Theor. Biol.*, vol. 473, pp. 1–8, Jul. 2019.
[40] A. H. Butt and Y. D. Khan, "Prediction of S-sulfenylation sites using statistical moments based features via CHOU'S 5-step rule," *Int. J. Peptide Res. Ther.*, 2019.
[41] Y. D. Khan, S. A. Khan, F. Ahmad, and S. Islam, "Iris recognition using image moments and k-means algorithm," *Sci. World J.*, vol. 2014, pp. 1–9, 2014.
[42] H. Zhu, H. Shu, J. Zhou, L. Luo, and J.-L. Coatrieux, "Image analysis by discrete orthogonal dual Hahn moments," *Pattern Recognit. Lett.*, vol. 28, no. 13, pp. 1688–1704, 2007.
[43] H. Zhu, H. Shu, J. Zhou, L. Luo, and J. Coatrieux, "Image analysis by discrete orthogonal dual Hahn moments," *Pattern Recognit. Lett.*, vol. 28, no. 13, pp. 1688–1704, Oct. 2007.
[44] P.-T. Yap, R. Paramesran, and S.-H. Ong, "Image analysis using Hahn moments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 11, pp. 2057–2062, Nov. 2007.
[45] H.-A. Goh, C.-W. Chong, R. Besar, F. S. Abas, and K.-S. Sim, "Translation and scale invariants of HAHN moments," *Int. J. Image Graph.*, vol. 09, no. 02, pp. 271–285, Apr. 2009.
[46] Y. D. Khan, N. Rasool, W. Hussain, S. A. Khan, and K.-C. Chou, "IPhosY-PseAAC: Identify phosphotyrosine sites by incorporating sequence statistical moments into PseAAC," *Mol. Biol. Rep.*, vol. 45, no. 6, pp. 2501–2509, Dec. 2018.
[47] C.-H. Yang, Y.-D. Lin, and L.-Y. Chuang, "TRNAfeature: An algorithm for tRNA features to identify tRNA genes in DNA sequences," *J. Theor. Biol.*, vol. 404, pp. 251–261, Sep. 2016.
[48] 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.
[49] Y. D. Khan, M. Jamil, W. Hussain, N. Rasool, S. A. Khan, and K.-C. Chou, "PSSbond-PseAAC: Prediction of disulfide bonding sites by integration of PseAAC and statistical moments," *J. Theor. Biol.*, vol. 463, pp. 47–55, Feb. 2019.
[50] Y. D. Khan, A. Batool, N. Rasool, S. A. Khan, and K.-C. Chou, "Prediction of nitrosocysteine sites using position and composition variant features," *Lett. Org. Chem.*, vol. 16, no. 4, pp. 283–293, Mar. 2019.
[51] W. Hussain, Y. D. Khan, N. Rasool, S. A. Khan, and K.-C. Chou, "SPrenylC–PseAAC: A sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-prenylation sites in proteins," *J. Theor. Biol.*, vol. 468, pp. 1–11, May 2019.
[52] T. H. Reiss, "Features invariant to linear transformations in 2D and 3D," in *Proc. Int. Conf. Pattern Recognit.*, vol. 3, 1992, pp. 493–496.
[53] M. Pawlak and X. Liao, "On image analysis by orthogonal moments," in *Proc. Int. Conf. Pattern Recognit.*, vol. 3, 1992, pp. 549–552.
[54] M. Awais, W. Hussain, Y. D. Khan, N. Rasool, S. A. Khan, and K.-C. Chou, "IPhosH-PseAAC: Identify phosphohistidine sites in proteins by blending statistical moments and position relative features according to the Chou's 5-step rule and general pseudo amino acid composition," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published.
[55] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.
[56] R. E. Schapire, "Theoretical views of boosting and applications," in *Proc. Int. Conf. Algorithmic Learn. Theory*. Berlin, Germany: Springer, 1999, pp. 13–25.
[57] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.
[58] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[59] A. Tyryshkina, N. Coraor, and A. Nekrutenko, "Predicting runtimes of bioinformatics tools based on historical data: Five years of Galaxy usage," *Bioinformatics*, vol. 35, no. 18, pp. 3453–3460, Sep. 2019.
[60] F. Pedregosa, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
[61] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?" in *Proc. Int. Workshop Mach. Learn. Data Mining Pattern Recognit.* Berlin, Germany: Springer, 2012, pp. 154–168.
[62] Y. Xu, X.-J. Shao, L.-Y. Wu, N.-Y. Deng, and K.-C. Chou, "ISNO-AAPair: Incorporating amino acid pairwise coupling into PseAAC for predicting cysteineS-nitrosylation sites in proteins," *PeerJ*, vol. 1, p. e171, Oct. 2013.
[63] P.-M. Feng, H. Ding, W. Chen, and H. Lin, "Naïve Bayes classifier with feature selection to identify phage virion proteins," *Comput. Math. Methods Med.*, vol. 2013, pp. 1–6, 2013.
[64] K.-C. Chou, "Prediction of signal peptides using scaled window," *Peptides*, vol. 22, no. 12, pp. 1973–1979, Dec. 2001.
[65] X. Xiao, P. Wang, W.-Z. Lin, J.-H. Jia, and K.-C. Chou, "IAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types," *Anal. Biochem.*, vol. 436, no. 2, pp. 168–177, May 2013.
[66] X. Xiao, Z.-C. Wu, and K.-C. Chou, "ILoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites," *J. Theor. Biol.*, vol. 284, no. 1, pp. 42–51, Sep. 2011.
[67] W.-Z. Lin, J.-A. Fang, X. Xiao, and K.-C. Chou, "ILoc-animal: A multi-label learning classifier for predicting subcellular localization of animal proteins," *Mol. BioSyst.*, vol. 9, no. 4, p. 634, 2013.
[68] K.-C. Chou, Z.-C. Wu, and X. Xiao, "ILoc-hum: Using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites," *Mol. BioSyst.*, vol. 8, no. 2, pp. 629–641, Dec. 2011.
[69] H. Y. Lai, X. X. Chen, W. Chen, H. Tang, and H. Lin, "Sequence-based predictive modeling to identify cancerlectins," *Oncotarget*, vol. 8, no. 17, pp. 28169–28175, 2017.

[70] J. Zhang, Y. Ju, H. Lu, P. Xuan, and Q. Zou, "Accurate identification of cancerlectins through hybrid machine learning technology," *Int. J. Geno.*, vol. 2016, pp. 1–11, 2016.
[71] R. Yang, C. Zhang, L. Zhang, and R. Gao, "A two-step feature selection method to predict cancerlectins by multiview features and synthetic minority oversampling technique," *BioMed Res. Int.*, vol. 2018, pp. 1–10, 2018.
[72] X. Cheng, X. Xiao, and K.-C. Chou, "PLoc_bal-mGneg: Predict subcellular localization of Gram-negative bacterial proteins by quasi-balancing training dataset and general PseAAC," *J. Theor. Biol.*, vol. 458, pp. 92–102, Dec. 2018.
[73] K.-C. Chou, "Proposing pseudo amino acid components is an important milestone for proteome and genome analyses," *Int. J. Peptide Res. Ther.*, 2019.
[74] B. Liu, H. Wu, D. Zhang, X. Wang, and K. C. Chou, "PSE-analysis: A Python package for DNA/RNA and protein/ peptide sequence analysis based on pseudo components and kernel methods," *Oncotarget*, vol. 8, no. 8, pp. 13338–13343, 2017.
[75] Z. Liu, X. Xiao, D.-J. Yu, J. Jia, W.-R. Qiu, and K.-C. Chou, "pRNAm-PC: Predicting N6-methyladenosine sites in RNA sequences via physical–chemical properties," *Anal. Biochem.*, vol. 497, pp. 60–67, Mar. 2016.
[76] P. Feng, H. Yang, H. Ding, H. Lin, W. Chen, and K.-C. Chou, "IDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC," *Genomics*, vol. 111, no. 1, pp. 96–102, Jan. 2019.

**AHMAD HASSAN BUTT** received the B.S. degree in computer science from the University of Central Punjab, Lahore, Pakistan, and the M.S. degree in computer science from the University of Management and Technology, Lahore. He is currently a member of the Research Group on pattern recognition and bioinformatics with the University of Management and Technology. His major areas of interests include machine learning, pattern recognition, image processing, and bioinformatics.

**YASER DAANIAL KHAN** is currently working as a Professor with the University of Management and Technology (UMT), Lahore, Pakistan. He is leading the Research Group on pattern recognition and bioinformatics with the University of Management and Technology. He has published articles in well reputed journals and conferences. His major areas of interests include pattern recognition, image processing, computer vision, and bioinformatics.