

Received March 26, 2021, accepted April 21, 2021, date of publication April 29, 2021, date of current version May 14, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3076448

ProtoPred: Advancing Oncological Research Through Identification of Proto-Oncogene Proteins

SHARAF J. MALEBARY¹, (Member, IEEE), RABIA KHAN²,
AND YASER DAANIAL KHAN², (Member, IEEE)

¹Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Rabigh 21911, Saudi Arabia

²Department of Computer Science, School of Systems and Technology, University of Management and Technology, Lahore 54770, Pakistan

This project was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, under grant No. (DF-291-611-1441). The authors, therefore, gratefully acknowledge DSR technical and financial support.

ABSTRACT Proto-oncogenes are the genes that have the potential to transform normal cells into cancer cells as a result of mutations. They usually contain encoding of proteins whose function is to inhibit cell differentiation, stimulate cell division, and prevent the death of cells. While the prognosis regarding proto-oncogene may occur at varying phases of cancer, the accuracy of the identification method is always questionable. The standard procedure for detecting these genes involves in-vitro experimentations but it proves to be very costly, time taking, and laborious. This problem is addressed by the use of computer-aided approaches established in studies encompassing methods in computational biology and bioinformatics. Early diagnosis of cancer is crucial for the full recovery of the patient. Proto-oncogene proteins are an important biomarker that helps identify the onset of a specific type of cancer. Keeping this in mind, this study proposes an efficient methodology for in-silico identification of proto-oncogenes. The predictor proposed in this study computes position and composition relevant statistical features incorporated into the pseudo-amino-acid composition (PseAAC) based on Chou's 5-step rules. Subsequently, the study finds that the use of a random forest classifier performs the most accurate prediction of proto-oncogene proteins. The method was validated using the 10 folds cross-validation, Jackknife testing, Self-Consistency, and Independent set testing, giving 95.44%, 97.17%, 99.8%, and 96.41% accurate results, respectively. These results imply that the proposed model can play a key role in the early prognosis of cancer and aid scientists in the discovery of mechanisms against cancer.


INDEX TERMS Proto-oncogenes, prediction, PseAAC, 5-steps rule, statistical moments.

I. INTRODUCTION

Every gene is formed by a sequence of nucleotide bases that contain information regarding the growth and working of cells. This essentially materializes when the genetic information is translated into proteins by the cells. Each protein has a specific function in the human body. Proto-oncogenes encode proteins that regulate cell differentiation and growth in humans [1]. The usually encoded proteomic products are DNA-binding proteins, protein kinases involved in signal transduction, growth factors and their receptors, transcription factors, and cell cycle regulators. A mutation in

its genomic sequence can trigger overexpression of proto-oncogenes resulting in proliferation, which renders the cell unresponsive to normal regulatory and growth-inhibitory signals, consequently causing the formation of a tumor. Proto-oncogenes are commonly activated in transformed cells by gene amplification or point mutations [2].

Oncogenomics is the study of genes associated with the onset of cancer. It involves the biological function of many genes. Carcinogenic mutations drive cancer development while mutated forms of proto-oncogenes that cause cancer are called oncogenes. In general, certain proteins have the function of stimulating cell division, preventing cell differentiation, and preventing cell death. These processes are imperative for the protection and normal development of tissues

The associate editor coordinating the review of this manuscript and approving it for publication was Ilaria Boscolo Galazzo .

and organs. However, oncogenes generally increase the translation of these proteins, which leads to increased cell division, reduced cell differentiation, and unmatched cell death. Taken together, these apparent patterns signify cancer development [3]. Oncogenes are currently important molecular targets for designing anti-cancer drugs. Some primary proto-oncogenes negatively regulate cell differentiation. Simultaneously, tumor suppressor genes (TSG) suppress and rectify the activity of oncogenes. Methylation of RNA/DNA affects the activation and expression of genes. N⁶-Methyladenosine (m⁶A) is the most common form of methylation. Cancer occurs when proto-oncogene mutates and becomes an oncogene or when oncogenes are locally hypo-methylated while tumor suppressor genes are hypermethylated [4].

Scientists are working on the personalized treatment of cancer. This involves numerous studies based on tumor suppressor genes, DNA repair genes, oncogenes, proto-oncogenes, and DNA methylation. Numerous computational approaches have been devised as in-silico methodologies for the identification of tumor suppressor genes. Studies have been performed on gene ontology and pathway enrichment analysis of TSG and non-TSG data to devise a methodology for their identification. A weighted graph was set up using protein-protein interaction along which the shortest path method was conceived to identify TSGs while a network diffusion algorithm has been established for this purpose. The role of methylation in the onset of cancer and especially that of RNA m⁶A methylation has been outlined in various studies. Similarly, the identification of proto-oncogene proteins plays a crucial role in devising a personalized treatment. Understanding specific biomarkers that cause cancer can help in the discovery of new drugs and other experimental treatments. The experimental prediction of protein function is a laborious task; thus, data scientists use different computational approaches to design assiduous methods for this purpose (see e.g., [5]–[23]).

Here, we propose a predictor for proto-oncogene identification by integrating Chou's PseAAC with various statistical moments. This work is entirely anticipated by Chou's 5-step rule, followed by various initial studies [9], [10], [22], [24]. Moreover, it encompasses incidence matrices that quantify mutual correlation among all the arbitrary residues for variable-sized primary sequences into a fixed size notation. Chou proposes these 5 steps for any such classification problem in the field of proteomics or genomics: (1) Selection or creation of robust and accurate benchmark dataset for training and testing of predictors. (2) Providing effective formulation to transform the dataset into a meaningful feature set that reflects a clear correlation and uncovers obscure information for accurate prediction. (3) Introduction of powerful classification algorithms. (4) Performing validation tests effectively to evaluate the prediction accuracy of the model. (5) Setting up a web server accessible to general users. Publications that develop robust sequence analysis methodologies have the following substantial advantages: (1) Transparency and understandability in logic development, (2) Complete

clearness in behavior and action, (3) Duplication of reported results is simple for other researchers; (4) It can stimulate newer sequence analysis methodology yielding even better results (5) Very useful when used in scientific research as it immensely reduces the effort of biologists.

II. MATERIAL AND METHODS

Herein, the salient features of Chou's 5-step rule are listed. The overall methodology is highlighted in Figure 1.

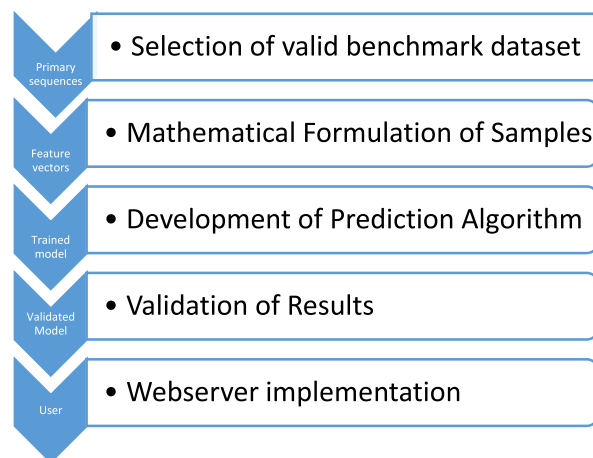


FIGURE 1. Flow of the proposed model.

A. BENCHMARK DATASET

This study is based on Chou's PseAAC which has been extensively used formerly for the prediction of methylation sites, SUMOylation sites, signal peptide cleavage sites, multiple lysine Post Translational modification (PTM) sites, lysine ubiquitination sites, hydroxylysine sites, hydroxyproline, and lysine succinate sites in numerous publications. Uniprot database is a well-known database of proteins containing experimentally proven information regarding numerous proteins collected from text and otherwise. It also contains a huge number of uncategorized protein sequences. Each protein is assigned a unique accession number and has been annotated according to its known attributes using keywords and other characteristics. Protein sequences annotated with the keyword 'proto-oncogene' were obtained from the Uniprot database. Similarly, a converse query was used to generate negative samples. Strings containing spaces, special characters, and characters that are not used for an amino acid (B, J, O, U, X, and Z) were removed from both negative and positive samples. Also, primary structures that were too short or contained ambiguous words like potential, probable, and fragment were excluded. The obtained data still consists of redundancies as many of the sequences maybe be homologous to each other. Such homology is quite problematic especially when it comes to validation and performance evaluation. Usually to alleviate this problem scientists make use of various utilities like CD-HIT that can compute relative similarity among arbitrary sequences. The CD-HIT suite was

used to reduce homology bias within extracted data [25]. Generally, within the scientific community, a 60% cutoff for homology is considered acceptable. Redundant sequences with a similarity of more than 60% were excluded from the dataset by setting the sequence identity cutoff at 0.6.

Let an arbitrary sample within the dataset be expressed as

$$P_{\xi}(\mathbb{Z}) = R_0 R_1 \cdots R_{\xi} \quad (1)$$

where ξ is non-uniform indicating that the length of a sequence may vary. In other words, ξ represents the arbitrary length of the primary sequence, which in this case is variable for each sample.

For the collection of the peptides the following two categories can be defined:

$$P_{\xi}(\text{POG}) \in \begin{cases} P_{\xi}^+(\text{POG}), & \text{if its a protooncogene} \\ P_{\xi}^-(\text{POG}), & \text{otherwise} \end{cases} \quad (2)$$

where true proto-oncogenes and the corresponding false proto-oncogenes are represented as $P_{\xi}^+(\text{POG})$ and $P_{\xi}^-(\text{POG})$, respectively. The symbol \in is from the set theory which depicts ‘‘a member of’’ and suffix ξ signifies that sequence is of arbitrary length. Homology among sequences was reduced to 60% using the CD-HIT suite. A representative was selected from each homologous cluster returned by CD-HIT. Subsequently, the obtained benchmark dataset after preprocessing is denoted as:

$$\mathbb{Z} = \mathbb{Z}^+ \cup \mathbb{Z}^- \quad (3)$$

where \mathbb{Z}^+ and \mathbb{Z}^- contain 252 positive and 630 negative samples respectively. For the ease of readers, $252 + 630 = 882$ samples are listed in Supplementary Information File S1. For such classification problems, the dataset is usually split into a test dataset and a training dataset. The training dataset is used to train the proposed model, whereas the test dataset is used for testing and validation of the model. Furthermore, evidence regarding the effectiveness of the model is accumulated based on various tests such as jackknife and cross-validation tests. Additionally, the benchmark dataset is also used for performing independent set testing by subdividing it into two subsets.

B. SAMPLE FORMULATION

The number of biological sequences reported is on a rapid rise. This scenario poses significant problems for computational biologists. A formulation that can transform these sequences into numeric discrete models and yet keep sequence pattern-related traits intact is in great demand. Most of the machine learning models are not designed to handle raw varying length proteomic data. Hence they need to be transformed into specific size vectors only [26]. To solve this problem, a position and composition-based statistical feature-based model is proposed [27]. Such models are currently being abundantly used in the field of computer-aided proteomics [28], [29], biomedicine, and drug discovery [30].

With the successful development of such models for proteomic analysis, scientists have successfully developed similar models pertaining to DNA / RNA sequences for genome analysis [31]. The peptide sequence samples within the benchmark dataset are transformed based on this concept into a feature vector. Where Ψ_i represents a feature coefficient for an arbitrary sample from the dataset and the transpose operator is denoted by T .

$$V = [\Psi_1 \ \Psi_2 \ \cdots \ \Psi_u \ \cdots \ \Psi_{\Omega}]^T \quad (4)$$

All the features obtained from the primary sequence are furnished into a vector V having a fixed length Ω . The overall set of feature vectors hence obtained is used for training and validation. The computational elements used to compute features in Eq (4) are further discussed in subsections C-F of the current section.

C. STATISTICAL MOMENT'S CALCULATION

Statistical moments were employed for the quantitative analysis of the obtained dataset. Moments are generally used to mine specific properties of data. Few of the moments represent the eccentricity and orientation of data, while some represent properties like its size, skewness, and variance. Numerous moment-defining polynomials exist based on specific distributions [32]–[36]. The central moments, raw moments, and Hahn moments for the proposed predictor were calculated up to order 3. Raw moments exhibit position and scale variant properties. Subsequently, central moments are position invariant and scale variant. Order up to 3 generates sufficient information regarding the nature of data in numeric form as discussed in [37]. Also, the Hahn coefficient was calculated using Hahn polynomials which generates yet another set of moments describing original data [35]. These statistical moments were chosen for their orthogonal properties. Orthogonal moments exhibit varying traits and can be used to reconstruct the original data, hence they encode within themselves crucial innate characteristics that accommodate precise classification. Overall, these moments sufficiently transform information regarding the positioning and composition of residues in the primary structure. The use of location and scale-invariant moments was eluded as scale and location play a primary role in determining a protein function [34]. These moments use a two-dimensional matrix P' of dimension n^*n as the source. It is a sequential transformation of all the amino acid residues of protein covered in P .

$$P' = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1n} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{n1} & \beta_{n2} & \cdots & \beta_{nn} \end{bmatrix} \quad (5)$$

The use of function ω has been reported in another study by Akmal et al., (2017) which transforms P into P' and each of its arbitrary element β_{ij} is an amino acid residue now put into a two-dimensional context. Unique ordinal values of elements of P' were utilized to calculate all the moments up

to degree 3 [38]. The equation (6) mentioned below was used to compute raw moments.

$$M_{ij} = \sum_{p=1}^n \sum_{q=1}^n p^i q^j \beta_{pq} \quad (6)$$

where $i + j$ represents the degree of the moments, β_{pq} is an arbitrary element of matrix P' . Further, raw moments were represented as $M_{00}, M_{01}, M_{02}, M_{03}, M_{10}, M_{11}, M_{12}, M_{20}, M_{21}$ and M_{30} for degree up to 3. Subsequently, central moments are calculated using the following equation:

$$\eta_{ij} = \sum_{p=1}^n \sum_{q=1}^n (p - \bar{x})^i (q - \bar{y})^j \beta_{pq} \quad (7)$$

where $\bar{x} = M_{10}/M_{00}$ and $\bar{y} = M_{01}/M_{00}$ represents the centroid of data.

P was transformed into a square matrix P' as it offers a substantial advantage for enumeration of Hahn moments. Hahn moments are discrete orthogonal moments that require a square matrix of data as input. The orthogonal property of Hahn moments offers several dividends as these moments can help reconstruct data using an inverse function. This essentially implies that it preserves the information pertaining to sequence structure and relative positioning of original data within these moments.

The following formulation is used to compute the Hahn polynomials of order n for a one-dimensional matrix of size N .

$$h_n^{u,v}(r, N) = (N + v - 1)_n (N - 1)_n \times \sum_{k=0}^n (-1)^k \frac{(-n)_k (-r)_k (2N + u + v - n - 1)_k}{(N + v - 1)_k (N - 1)_k} \frac{1}{k!} \quad (8)$$

where n is the order of the moment, N is the size of the data array, u and v are predefined constants. Further, the equation makes use of the Pochhammer symbol which in turn uses the gamma operator as mentioned in [38]. Based on this Hahn coefficient the two dimensional Hahn moments are computed as

$$H_{ij} = \sum_{q=0}^{N-1} \sum_{p=0}^{N-1} \beta_{pq} h_i^{u,v}(q, N) h_j^{u,v}(p, N), \quad (9)$$

where $i + j$ is the order of the moment, u, v are predefined constants and β_{pq} is an arbitrary element of the square matrix P' .

Furthermore, while working with abundant statistical data problems like collinearity or multi-collinearity might be encountered. Multi-collinearity is problematic as it undermines the significance of independent variables and increases overheads. In a feature set, if a coefficient is linearly derived from other coefficients then the feature set is bound to be collinear. One way to ensure that feature set coefficients are not collinear is the correct choice of feature extraction technique. In this study, only those statistical moments are used that are independently formulated. Orthogonal moments quite much ensure that values are not collinear since each coefficient is computed as a bivariate function formulated

such that the sum of the power of variables is the order of the moment. However, since raw, central, and Hahn moments are used, more measures need to be taken to remove collinearity. Principle Component Analysis (PCA) is performed to eliminate collinear coefficients.

D. DETERMINATION OF INCIDENCE MATRICES

A computational model was formed based on the composition and relative positioning of amino acid residues. It played a crucial role in computationally determining the characteristics of proteins. A Position Relative Incident Matrix (PRIM) was formulated as 20×20 matrix to enumerate the relative positioning correlations among amino into a fixed size notation. It is given as

$$S_{PRIM} = \begin{bmatrix} s_{1 \rightarrow 1} & s_{1 \rightarrow 2} \cdots & s_{1 \rightarrow j} \cdots & s_{1 \rightarrow 20} \\ s_{2 \rightarrow 1} & s_{2 \rightarrow 2} \cdots & s_{2 \rightarrow j} \cdots & s_{2 \rightarrow 20} \\ \vdots & \vdots & \vdots & \vdots \\ s_{i \rightarrow 1} & s_{i \rightarrow 2} \cdots & s_{i \rightarrow j} \cdots & s_{i \rightarrow 20} \\ \vdots & \vdots & \vdots & \vdots \\ s_{20 \rightarrow 1} & s_{20 \rightarrow 2} \cdots & s_{20 \rightarrow j} \cdots & s_{20 \rightarrow 20} \end{bmatrix} \quad (10)$$

Each component of this matrix signifies the sum of the positions of the j^{th} residue relative to the first occurrence of the i^{th} residue given as $S_{i \rightarrow j}$. Hence, this matrix encompasses 400 coefficients which is a huge size. Statistical moments provide the prospect to transform this information into a succinct form. Computing raw, central and Hahn moments for PRIM matrix yields 30 coefficients in all for degree up to 3. Similarly, the Reverse Position Relative Incidence Matrix (RPRIM) was formed using a primary sequence of proteins. It was represented as:

$$S_{RPRIM} = \begin{bmatrix} s_{1 \rightarrow 1} & s_{1 \rightarrow 2} \cdots & \cdots & s_{1 \rightarrow 20} \\ s_{2 \rightarrow 1} & s_{2 \rightarrow 2} \cdots & s_{2 \rightarrow j} \cdots & s_{2 \rightarrow 20} \\ \vdots & \vdots & \vdots & \vdots \\ s_{i \rightarrow 1} & s_{i \rightarrow 2} \cdots & s_{i \rightarrow j} \cdots & s_{i \rightarrow 20} \\ \vdots & \vdots & \vdots & \vdots \\ s_{20 \rightarrow 1} & s_{20 \rightarrow 2} \cdots & s_{20 \rightarrow j} \cdots & s_{20 \rightarrow 20} \end{bmatrix} \quad (11)$$

Statistical moments applied for $RPRIM$ reduced the dimensionality of $RPRIM$ and resulted in the formation of a set of 30 elements. The $PRIM$ (Eq. (10)) reveals position relative information of amino acid residues in the polypeptide chain, this information is augmented by the $RPRIM$ (Eq. (11)) matrix which uncovers even further obscured information by performing the same operation on the reverse of primary sequence.

E. FREQUENCY VECTOR DETERMINATION

The frequency vector is simply computed by counting the occurrence of each amino acid residue within the primary sequence. Each element in the frequency vector represents the frequency of occurrence of the corresponding amino acid residue within the given sequence. Hence, the frequency vector yields 20 coefficients.

$$\rho = \{\tau_1, \tau_2, \dots, \tau_{20}\} \quad (12)$$

The frequency of occurrence of the i^{th} residue is represented as τ_i . The frequency matrix provides further compositional information regarding the sequence.

F. ABSOLUTE POSITION INCIDENCE VECTOR

The frequency matrix essentially provides information regarding the composition of amino acid residues. Summarization of positioning of residues is provided by yet another vector namely Accumulative Absolute Position Incidence Vector (AAPIV). It constitutes a single coefficient corresponding to each amino acid residues hence forming a length of 20 elements. Elements of the vector AAPIV contain the sum of the position of occurrence of each natural amino acid in the primary structure given as

$$K = \{\mu_1, \mu_2, \mu_3, \dots, \mu_{20}\} \tag{13}$$

where Equation 14 enables to compute arbitrary i^{th} element of AAPIV

$$\mu_i = \sum_{k=1}^n p_k \tag{14}$$

where p_k is the position of the i^{th} amino acid residue. Subsequently, reverse accumulative absolute position vector (RAAPIV) further evaluated detailed information based upon the absolute position of amino acids in peptide samples. Reversing of primary sequence and then calculating AAPIV yielded RAAPIV, denoted as:

$$\Lambda = \{\gamma_1, \gamma_2, \gamma_3, \dots, \gamma_{20}\} \tag{15}$$

where γ_i represents the sum of all the positions where the i^{th} amino acid residue occurs in the primary structure.

G. FEATURE VECTOR DESCRIPTION

Primary sequences processed through all the above steps are ultimately combined to form an accumulative feature vector. Two-dimensional representation of the primary sequence P' , $PRIM$, and $RPRIM$ matrices are transformed into a succinct form by computing their statistical (raw, central, and Hahn) moments. Consequently, it yields 90 coefficients. Furthermore, the frequency vector (ρ), AAPIV (K) and RAAPIV (Λ) are also pooled into the vector account for 60 more coefficients. Overall, a fixed-sized feature vector with 150 coefficients formulated for each primary structure of arbitrary length as depicted in Fig 2.

Eventually, the combined feature set is subjected to PCA for the removal of collinearity and dimensionality reduction. Varying permutations were probed in several experiments, it was observed reducing the 150-column feature set data into 100 columns provides optimal outcome.

H. CLASSIFICATION ALGORITHM

Random Forest (RF) is a powerful machine learning Classifier that assiduously performs classification or prediction. Random forest operates by constructing a multitude of decision trees during training. The advantage of the Random Forest classifier includes their non-parametric nature,

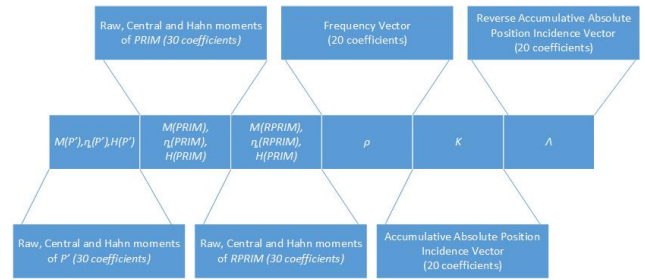


FIGURE 2. Structure of feature vector.

maximum classification accuracy, and capability to adapt to the features that bear importance for maximum accuracy. Assimilated input feature vectors based on primary sequences as described in the previous section are furnished into an input matrix. Similarly, an expected output matrix was also improvised which contained the expected output for the corresponding row in the input matrix. Both the matrices were used to train an RF classifier (Figure 3) constructed using parameters shown in Table 1.

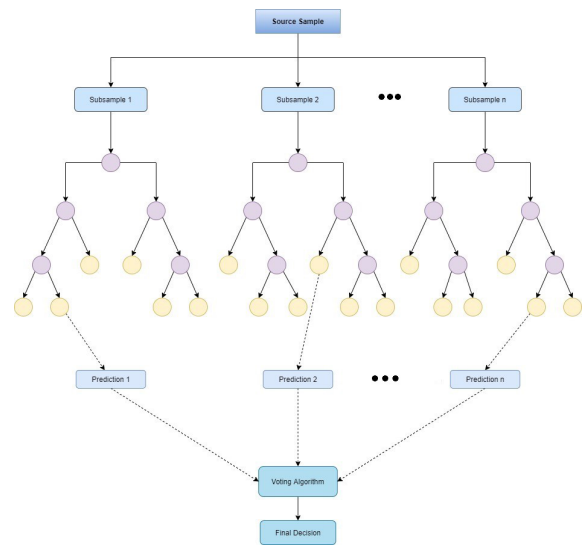


FIGURE 3. Architecture of random forest classifier for the proposed prediction model.

TABLE 1. Probed and optimal set of parameters for RF.

Parameters	Range	Optimized
n_estimators	10 - 100	50
criterion	{“gini”, “entropy”}	gini
max_depth	3 - 20	16
min_samples_split	2 - 10	5
min_samples_leaf	1 - 2	1
max_features	{“auto”, “sqrt”, “log2”}	auto

The details of the most optimal parameters of the RF classifier for the identification of proto-oncogene proteins are shown in Table 1. The number of estimators was set at 50 while the maximum depth was kept at 16. Several classifiers exist and the choice of a robust, efficient, and accurate classifier is a pivotal task. RF classifiers form a decision based on opinion received from a federation of decision trees. The opinions received from each decision tree are combined into a single outcome. The ability of RF classifiers to consult multiple opinions insulates it against over-fitting without compromising on prediction error as compared with other monolithic classifiers like neural networks. Another feature that renders more diversity to RF classifiers is the fact that a decision tree at each node splits based on the best feature rather than the most important one. RF classifier also bears superior performance while dealing with multiclass data as compared to other binary classifiers like Support Vector Machine (SVM) which requires cascading. SVM is binary classifiers that attempt to partition the multi-dimensional feature vector space using a hyper-plane. It makes use of support vectors to determine the most optimal parameters for partitioning the feature vector space. SVM has also been used successfully in machine learning problems. Table 2 provides a range of probed values along with the most optimal values encountered for SVM.

TABLE 2. Probed and optimal set of parameters for SVM.

Parameters	Range	Optimized
Kernel	Linear, RBF(Radial Basis Function)	RBF
C	1 - 1000	10-100
gamma	“scale”, “auto”	“auto”

Artificial neural networks (ANN) are also prevalently used as robust classifiers. A trend in the use of artificial neural networks is seen as multiple researchers have used ANN in many bio-computational decision problems. In this study, ANN is used with a back-propagation algorithm for the prediction of proto-oncogene proteins. ANN has been inspired by the working of the human brain. The human brain consists of neurons that work together to process and receive information and learn skills from experience. ANN algorithm also works similar to the brain as it consists of multiple nodes that are linked with each other. The first layer of nodes in the input layer, the second layer is called the hidden layer while the third layer is the output layer. Data that needs to be modeled is passed onto the input layer while the hidden layer(s) are used for intermittent processing. Subsequently, the output layer shows the resultant outcome. In back-propagation, the values received at the output layer are again used as feedback for the hidden layer for improving its accuracy upon each iteration and also to reduce or minimize the error rate in classification. Subsequently, Table 3 provides the probed set of

parameters for ANN along with the most optimal parameters observed.

TABLE 3. Probed and optimal parameters for ANN.

Parameters	Range	Optimized
max_iter	100 - 1000	437
hidden_layer_sizes	10 - 50	23
activation	{‘identity’, ‘logistic’, ‘tanh’, ‘relu’}	relu
solver	{‘lbfgs’, ‘sgd’, ‘adam’}	adam
learning_rate	0.0001 - 0.1	0.001
momentum	0 - 1	0.7
validation_fraction	0 - 1	0.4

III. RESULTS AND DISCUSSION

This section focuses on analyzing the accuracy of the model using rigorous and standard testing methodologies. Testing for a genomic or proteomic application is quite different from other conventional applications as new data is not readily available. To avoid inconsistencies only non-homologous and naturally occurring experimentally established data is used.

A. ACCURACY ESTIMATION

Objective evaluation of a prediction model requires estimation of its accuracy described within well-defined parameters. The choice of the testing and validation methodology described in form of well-understood accuracy metrics aimed at quantifying the performance of the method is a crucial task. A regular and well-known set of metrics for the estimation of accuracy is defined here. The most commonly used tests are self-consistency, Cross-validation, and Jackknife testing. For each test, various metrics are calibrated to evaluate accuracy.

1) METRICS FOR ACCURACY ESTIMATION

Based on different perspectives, four different metrics are discussed which evaluate the accuracy of the prediction model: (1) Acc for overall accuracy (2) S_p for model specificity (3) S_n for Model sensitivity, and (4) MCC for model stability. These metrics have been prevalently used in literature to gauge the predictive quality of a proposed model. In particular, Matthew’s correlation coefficient (MCC) is a crucial indicator that echoes the stability of the model. The balance between the capability of the model to recognize positive samples and the inability to recognize positive or negative samples is given by the F_{score} . Consequently, these intuitive metrics were derived based on various prediction and

classification studies, as given in Eq.16 below.

$$\left\{ \begin{array}{l}
 S_n = 1 - \frac{N_-^+}{N^+} \\
 0 \leq S_n \leq 1 \\
 S_p = 1 - \frac{N_+^-}{N^-} \\
 0 \leq S_p \leq 1 \\
 Acc = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} \\
 0 \leq Acc \leq 1 \\
 MCC = \frac{1 - \left(\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-} \right)}{\sqrt{\left(1 + \frac{N_+^- - N_-^+}{N^+} \right) \left(1 + \frac{N_-^+ - N_+^-}{N^-} \right)}} \\
 \text{where } -1 \leq MCC \leq 1 \\
 F_{score} = \frac{N^+}{N^+ + \frac{1}{2}(N_-^+ + N_+^-)}
 \end{array} \right. \quad (16)$$

where N^+ signifies the number of proto-oncogenes truly predicted as proto-oncogenes and N_-^+ denotes the number of proto-oncogene proteins predicted inaccurately as the non-proto-oncogene proteins. Moreover, N^- represents the number of non-proto-oncogenes predicted accurately while N_+^- denotes the number of non-proto-oncogenes predicted inaccurately. Equation (16) comprehensively shows how sensitivity, specificity, overall accuracy, stability in terms of MCC and F_{score} are computed [39], and is reported in various studies (see, e.g., [40]–[42]). However, it is for binary class data, and for multi-class, other metrics are proposed.

B. VALIDATION TESTS

Typically, experimentally validated datasets are used for model training and testing. However, readily producing random test cases is not possible since proteomic data is innate to nature. Naturally occurring proteomic sequences are furnished into the study that is supported by some experimental verification. Methodologies still exist that could rigorously test the effectiveness of the model even though the test cases are limited. These tests are designed to evaluate the accuracy and reliability of the model based on its ability to identify unknown data. Generally, for such a bioinformatics problem self-consistency test, k folds cross-validation test, jackknife test, and independent test are used [43]. The jackknife test is very exhaustive and will always give the same results for a particular dataset. A cross-validation test is a convenient choice for estimating that an established model is functioning satisfactorily if new test cases are not readily available to validate the model’s accuracy.

A self-consistency test was performed using the previously derived benchmark dataset for training as well as testing of the proposed predictor using. This is the most trivial sort of test that simply tests the model on the same data that was

used for its training. This test works as a benchmark to gauge the ability of a classifier to identify hidden patterns within a dataset. The self-consistency test results are shown in Table 4. It depicts all the metrics obtained from the self-consistency test. The test yields an MCC equivalent of 0.99 for RF, 0.76 for ANN, and 0.67 for SVM, which clearly shows that the RF model is capable of readily, and quite accurately identify the specific unique patterns within the primary structure of proto-oncogene proteins. Furthermore, the Receiver operating characteristics (ROC) graph obtained from the test for all three classifiers is also depicted in Figure 4. The graph shows a huge area under the RF curve, which signifies the high accuracy of the RF model.

TABLE 4. Self-consistency result via random forest classifier.

Model	TP	FN	FP	TN	Acc (%)	S_p	S_n	MCC	F_{score}
RF	252	1	0	630	99.8	1	0.99	0.99	0.998
NN	213	39	45	585	90.4	0.93	0.85	0.76	0.835
SVM	196	56	63	567	86.5	0.90	0.78	0.67	0.767

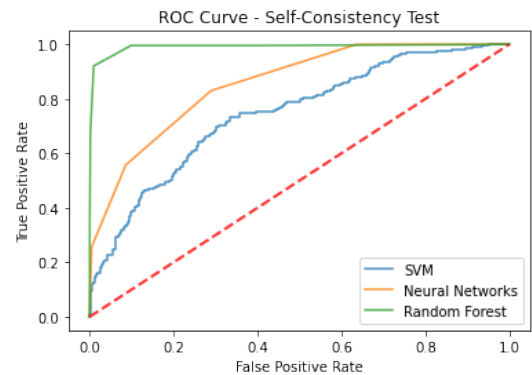


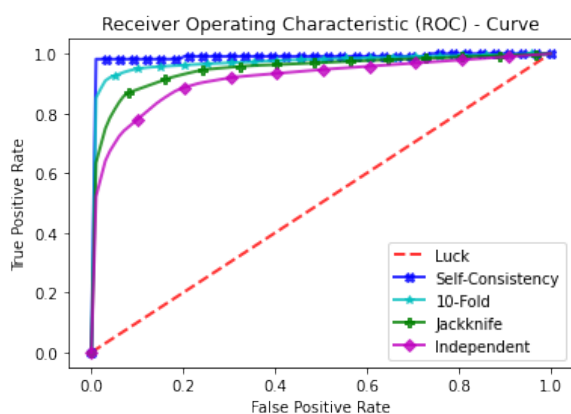
FIGURE 4. ROC curve for self-consistency tests for each classifier.

The self-consistency test is clear proof of the appropriateness of the RF classifier for the problem. Hence, further validations are performed using only the RF classifier.

Cross-validation requires that the benchmark dataset is split into k disjoint partitions. Each fold or partition is randomly selected as a mutually disjoint data partition for validation while the rest of the data was used to train the model. Consequently, all partitions of the dataset are exhaustively used for both training and testing. The overall result is reported as the mean of all the results received from each fold. The method is applied uniformly to negative and positive data samples. Random partitions were formed using $k=10$. Cross-validation is potentially a better verification method since it exhaustively tests all the data. Table 5 depicts the outcomes obtained with 10 folds cross-validation while Figure 5 shows the ROC curve for 10 folds. The overall MCC score for 10 folds cross-validation is 0.91 which signifies a high rate of

TABLE 5. Cross-validation result via random forest classifier.

Fold #	TP	F N	F P	TN	Acc (%)	S_p	S_n	MCC	F_{score}
1	20	5	0	63	92.4	0.8	1.0	0.86	0.88
2	24	2	0	63	96.1	0.93	1.0	0.94	0.96
3	20	5	0	63	93.6	0.8	1.0	0.86	0.88
4	23	2	0	63	96.1	0.92	1.0	0.94	0.95
5	20	6	0	63	95.3	0.77	1.0	0.83	0.87
6	24	1	0	63	97.4	0.96	1.0	0.97	0.98
7	20	5	0	63	95.4	0.8	1.0	0.86	0.88
8	22	3	0	63	95.7	0.88	1.0	0.92	0.94
9	25	1	0	63	97.6	0.95	1.0	0.94	0.98
10	24	0	0	63	94.8	0.96	1.0	0.97	1.0
Final 10 fold -CV Score					95.4	0.87	1.0	0.91	0.92
Standard Deviation (σ^2)									7.39

**FIGURE 5.** ROC curve of validation tests.

accuracy. Subsequently, the standard deviation (σ^2) obtained from the set of experiments carried is also depicted as 7.39

Jackknife testing is the most rigorous and consistent testing technique. In each iteration, it leaves out a sample and trains the model on the rest of the samples. The trained model was later tested using the left-out sample. In this way, this exhaustive technique evaluates the behavior of the classifier for each sample. Jackknife always returns a unique output for a specific dataset. Jackknife test completely evades intentional problems due to inconsistencies in subsampling and independence. The results of the jackknife test for the proposed predictor are shown in Table 6 while the ROC curve is shown in Figure 4. The model exhibits an MCC of 0.931 showing that the model performs quite accurately for unknown data. Continually, the standard deviation, for all the results received from each iteration of jackknife testing is also listed and measured as 24.4.

Subsequently, independent set testing was performed by splitting the dataset into 70-30% partitions. RF classifier was trained on 70% partition, which was later tested using the remaining 30% samples. The results are shown in Table 7 while ROC is shown in Figure 6. It depicts the number of samples identified as TP, FN, TN, FP along with the accuracy metric. Independent data set is a simple but

TABLE 6. Jackknife test result via random forest classifier.

TP	FN	FP	TN	Acc (%)	S_p	S_n	MCC	F_{score}	σ^2
229	24	1	629	97.17	0.99	0.91	0.931	0.94	24.4

TABLE 7. Independent dataset testing.

Training Dataset		Testing Dataset	
TN	168	TN	58
FP	13	FP	8
FN	0	FN	0
TP	432	TP	198
Acc (%)	94.45 (%)	Acc (%)	96.41
F_{score}	0.986	F_{score}	0.98

adequate benchmark to establish the accuracy of the model on bulk unknown data. Results show that the predictor performs satisfactorily on independent set testing as well.

Because of rigorous scrutiny of the proposed model based on these validation tests, it is concluded that the proposed predictor is accurate and an efficient way of identifying proto-oncogene peptides based on their primary structures. An overview of all the results is shown in Figure 4 in form of ROC graphs. The area under the curve for these tests is considerably high which signifies the high accuracy of the proposed predictor.

Furthermore, the proposed model is also evaluated in comparison with other baseline and state of art feature extraction techniques. PseAAC is one of the most popular and predominantly used feature extraction techniques for proteomic identification in unison with a multitude of classifiers. Based on the correlation of amino acid residue positions and their composition this model yields a set of coefficients [7] which are used as a feature set. Scientists have used a multitude of classifiers in combination with PseAAC. Subsequently, another method namely Position Specific Scoring Matrix (PSSM) [44] is used to extract features based on basic sequence-related statistics. Here, PSSM is being used as a baseline method as it merely bears coefficients dependent on the positioning of amino acid residues. PSSM is one of the most basic and simplistic feature extraction methodologies. Comparison of the proposed methodology is performed with PseAAC and PSSM based feature extraction techniques while RF is employed as the classifier. The results of independent set testing for all these techniques are illustrated in the ROC curve below.

The ROC in Figure 6 depicts that the proposed feature extraction technique works better than classical PseAAC and PSSM which means it is more capable of sieving out the most momentous features required for the identification of proto-oncogenes proteins.

IV. WEBSERVER

Another aspect of such computational studies is the development of a user-friendly public web-server for biologists as shown in recent publications by various researchers. As stated in a previous study [45], a publicly available web-server is the imminent course of action for reporting recent

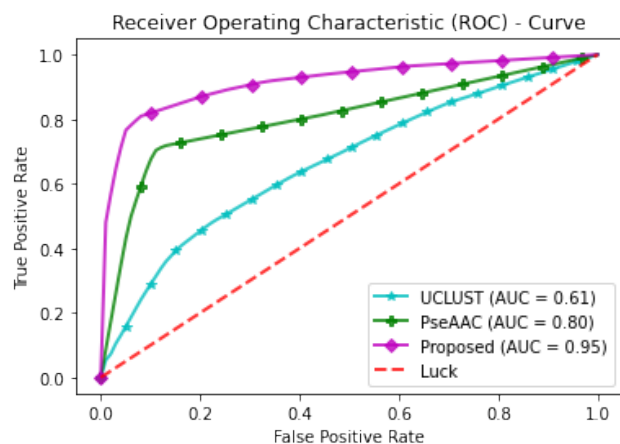


FIGURE 6. Comparative analysis with state of art and baseline models.

significant computational insights and analyses. They have greatly helped to advance the usefulness of computational methods in medicine, leading drug discovery to an unprecedented revolution [46]. Thus, the proposed work also endeavors to make available a webserver providing a web-based implementation of the methodology, in the future. However, all the source code and other materials have been made available at <https://github.com/csbioinfopk/protoncogene>.

V. CONCLUSION AND DISCUSSION

Mutations in proto-oncogenes are one of the major causes of cancer due to exposure to a mutagen. Proto-oncogenes translate to form proto-oncogene proteins. These proteins work as a biomarker for such susceptibility to cancer. The proposed work presents a robust in-silico technique for the identification of such proteins. Scientists are working to find intelligent and personalized ways to predict the onset of cancer. The identification of proto-oncogene proteins works as a component of such prognosis. The proposed technique adapts all the state of art recommendations to form a computationally intelligent predictor. Robust and non-homologous data is collected supported experimental evidence only from the well-known Uniprot database. Features like PRIM, RPRIM, AAIV, FM, and statistical moments of a two-dimensional representation of the primary structure of proteins are gathered to form feature vectors. Random forest classifier is used for the training of data because it exhibits less susceptibility to overfitting. The yielded results are tested using rigorous tests like self-consistency, cross-validation, independent set, and jackknife testing. All these tests except the self-consistency test partition the actual data into different partitions using diverse methodologies to evaluate the performance of the predictor. The results of these tests show that the model performs well for unknown data. The validation tests of 10 folds cross-validation, jackknife testing, and independent set testing yielding an accuracy of 95.44%, 99.81%, and 96.41 respectively. The remarkable yield of the model in terms of accuracy can be attributed to the combined

capability of the feature extraction technique and random forest classifier. The feature extraction methodology enriched into the model exhibits the capability of extracting obscure patterns comprised of the sequence and composition of the primary structures. Thus, the proposed predictors help predict proto-oncogenes efficiently and accurately and provide baseline data for the discovery of new drugs and biomarkers against cancer. Furthermore, these results also suggest that the proposed predictor is a potent computational tool for rapid and cost-effective identification of proto-oncogene proteins. Proto-oncogene proteins are an important biomarker that can prove useful in detecting the early onset of cancer. The proposed predictor is also a potential tool for researchers to devise diagnostic tests for cancer-related disorders.

ACKNOWLEDGMENT

The authors gratefully acknowledge Deanship of Scientific Research (DSR) technical support.

SUPPLEMENTARY MATERIALS

Supporting Information S1: Benchmark dataset comprising 252 positive and 630 negative samples.

REFERENCES

- [1] D. E. Williams, J. Eisenman, A. Baird, C. Rauch, K. Van Ness, C. J. March, L. S. Park, U. Martin, D. Y. Mochizuki, H. S. Boswell, G. S. Burgess, D. Cosman, and S. D. Lyman, "Identification of a ligand for the C-kit proto-oncogene," *Cell*, vol. 63, no. 1, pp. 167–174, Oct. 1990.
- [2] L. M. Mulligan, J. B. J. Kwok, C. S. Healey, M. J. Elsdon, C. Eng, E. Gardner, D. R. Love, S. E. Mole, J. K. Moore, L. Papi, M. A. Ponder, H. Telenius, A. Tunnacliffe, and B. A. J. Ponder, "Germ-line mutations of the RET proto-oncogene in multiple endocrine neoplasia type 2A," *Nature*, vol. 363, no. 6428, pp. 458–460, Jun. 1993.
- [3] C. A. Finlay, P. W. Hinds, and A. J. Levine, "The p53 proto-oncogene can act as a suppressor of transformation," *Cell*, vol. 57, no. 7, pp. 1083–1093, Jun. 1989.
- [4] P. Edery, S. Lyonnet, L. M. Mulligan, A. Pelet, E. Dow, L. Abel, S. Holder, C. Nihoul-Fékété, B. A. Ponder, and A. Munnich, "Mutations of the RET proto-oncogene in Hirschsprung's disease," *Nature*, vol. 367, no. 6461, p. 378, 1994.
- [5] S. Akbar and M. Hayat, "iMethyl-STNC: 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.
- [6] W. Chen, P. Feng, H. Yang, H. Ding, H. Lin, and K.-C. Chou, "IRNA-3typeA: Identifying three types of modification at RNA's adenosine sites," *Mol. Therapy Nucleic Acids*, vol. 11, pp. 468–474, Jun. 2018.
- [7] J. Jia, Z. Liu, X. Xiao, B. Liu, and K.-C. Chou, "ISuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset," *Anal. Biochem.*, vol. 497, pp. 48–56, Mar. 2016.
- [8] Z. Ju and S.-Y. Wang, "Prediction of citrullination sites by incorporating k-spaced amino acid pairs into Chou's general pseudo amino acid composition," *Gene*, vol. 664, pp. 78–83, Jul. 2018.
- [9] Y. D. Khan, N. Rasool, W. Hussain, S. A. Khan, and K.-C. Chou, "iPhosT-PseAAC: Identify phosphothreonine sites by incorporating sequence statistical moments into PseAAC," *Anal. Biochem.*, vol. 550, pp. 109–116, Jun. 2018.
- [10] 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, 2018.
- [11] L.-M. Liu, Y. Xu, and K.-C. Chou, "iPGK-PseAAC: Identify lysine phosphoglycylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC," *Medicinal Chem.*, vol. 13, no. 6, pp. 552–559, Aug. 2017.

- [12] W.-R. Qiu, S.-Y. Jiang, B.-Q. Sun, X. Xiao, X. Cheng, and K.-C. Chou, "iRNA-2methyl: Identify RNA 2'-O-methylation sites by incorporating sequence-coupled effects into general PseKNC and ensemble classifier," *Medicinal Chem.*, vol. 13, no. 8, pp. 734–743, Nov. 2017.
- [13] M. F. Sabooh, N. Iqbal, M. Khan, M. Khan, and H. F. Maqbool, "Identifying 5-methylcytosine sites in RNA sequence using composite encoding feature into Chou's PseKNC," *J. Theor. Biol.*, vol. 452, pp. 1–9, Sep. 2018.
- [14] 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.*, vol. 18, no. 2, pp. 596–610, Mar./Apr. 2021.
- [15] 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.
- [16] A. Ehsan, M. K. Mahmood, Y. D. Khan, O. M. Barukab, S. A. Khan, and K.-C. Chou, "iHyd-PseAAC (EPSV): Identifying hydroxylation sites in proteins by extracting enhanced position and sequence variant feature via Chou's 5-step rule and general pseudo amino acid composition," *Current Genomics*, vol. 20, no. 2, pp. 124–133, May 2019.
- [17] A. W. Ghauri, Y. D. Khan, N. Rasool, S. A. Khan, and K.-C. Chou, "PNitro-Tyr-PseAAC: Predict nitrotyrosine sites in proteins by incorporating five features into Chou's general PseAAC," *Current Pharmaceutical Des.*, vol. 24, no. 34, pp. 4034–4043, Jan. 2019.
- [18] W. Hussain, Y. D. Khan, N. Rasool, S. A. Khan, and K.-C. Chou, "SPalmitoylC-PseAAC: A sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-palmitoylation sites in proteins," *Anal. Biochem.*, vol. 568, pp. 14–23, Mar. 2019.
- [19] S. A. Khan, Y. D. Khan, S. Ahmad, and K. H. Allehaibi, "N-MyristoylG-PseAAC: Sequence-based prediction of N-Myristoyl glycine sites in proteins by integration of PseAAC and statistical moments," *Lett. Organic Chem.*, vol. 16, no. 3, pp. 226–234, Feb. 2019.
- [20] Y. D. Khan, A. Batoool, N. Rasool, S. A. Khan, and K.-C. Chou, "Prediction of nitrosocysteine sites using position and composition variant features," *Lett. Organic Chem.*, vol. 16, no. 4, pp. 283–293, Mar. 2019.
- [21] 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.
- [22] Y. D. Khan, N. Amin, W. Hussain, N. Rasool, S. A. Khan, and K.-C. Chou, "IProtease-PseAAC(2L): A two-layer predictor for identifying proteases and their types using Chou's 5-step-rule and general PseAAC," *Anal. Biochem.*, vol. 588, Jan. 2020, Art. no. 113477.
- [23] O. Barukab, Y. D. Khan, S. A. Khan, and K.-C. Chou, "ISulfoTyr-PseAAC: Identify tyrosine sulfation sites by incorporating statistical moments via Chou's 5-steps rule and pseudo components," *Current Genomics*, vol. 20, no. 4, pp. 306–320, Oct. 2019.
- [24] M. A. Akmal, W. Hussain, N. Rasool, Y. D. Khan, S. A. Khan, and K.-C. Chou, "Using Chou's 5-steps rule to predict O-linked serine glycosylation sites by blending position relative features and statistical moment," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Jan. 21, 2020, doi: 10.1109/TCBB.2020.2968441.
- [25] 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.
- [26] K.-C. Chou, "Impacts of bioinformatics to medicinal chemistry," *Medicinal Chem.*, vol. 11, no. 3, pp. 218–234, Mar. 2015.
- [27] C.-T. Zhang and K.-C. Chou, "An optimization approach to predicting protein structural class from amino acid composition," *Protein Sci.*, vol. 1, no. 3, pp. 401–408, Mar. 1992.
- [28] F. Ali and M. Hayat, "Classification of membrane protein types using voting feature interval in combination with Chou's pseudo amino acid composition," *J. Theor. Biol.*, vol. 384, pp. 78–83, Nov. 2015.
- [29] M. Kabir and M. Hayat, "IRSpot-GAEnsC: Identifying recombination spots via ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples," *Mol. Genet. Genomics*, vol. 291, no. 1, pp. 285–296, Feb. 2016.
- [30] W.-Z. Zhong and S.-F. Zhou, "Molecular science for drug development and biomedicine," *Int. J. Mol. Sci.*, vol. 15, no. 11, pp. 20072–20078, 2014, doi: 10.3390/ijms151120072.
- [31] W. Chen, H. Lin, and K.-C. Chou, "Pseudo nucleotide composition or PseKNC: An effective formulation for analyzing genomic sequences," *Mol. BioSyst.*, vol. 11, no. 10, pp. 2620–2634, 2015.
- [32] Y. D. Khan, F. Ahmad, and M. W. Anwar, "A neuro-cognitive approach for iris recognition using back propagation," *World Appl. Sci. J.*, vol. 16, no. 5, pp. 678–685, 2012.
- [33] Y. D. Khan, F. Ahmed, and S. A. Khan, "Situation recognition using image moments and recurrent neural networks," *Neural Comput. Appl.*, vol. 24, nos. 7–8, pp. 1519–1529, Jun. 2014.
- [34] Y. D. Khan, N. S. Khan, S. Farooq, A. Abid, S. A. Khan, F. Ahmad, and M. K. Mahmood, "An efficient algorithm for recognition of human actions," *Sci. World J.*, vol. 2014, pp. 1–11, Jan. 2014.
- [35] 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, Apr. 2014.
- [36] S. Mahmood, Y. D. Khan, and M. K. Mahmood, "A treatise to vision enhancement and color fusion techniques in night vision devices," *Multimedia Tools Appl.*, vol. 77, no. 2, pp. 2689–2737, 2018.
- [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] 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, 2017, Art. no. e0181966.
- [39] W. R. Qiu, B. Q. Sun, X. Xiao, D. Xu, and K. C. Chou, "iPhos-PseEvo: Identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory," *Mol. Informat.*, vol. 36, nos. 5–6, 2017, Art. no. 1600010.
- [40] B. Liu, F. Yang, D.-S. Huang, and K.-C. Chou, "iPromoter-2L: A two-layer predictor for identifying promoters and their types by multi-window-based PseKNC," *Bioinformatics*, vol. 34, no. 1, pp. 33–40, Jan. 2018.
- [41] A. Ehsan, K. Mahmood, Y. D. Khan, S. A. Khan, and K.-C. Chou, "A novel modeling in mathematical biology for classification of signal peptides," *Sci. Rep.*, vol. 8, no. 1, p. 1039, Dec. 2018.
- [42] P. Feng, H. Yang, H. Ding, H. Lin, W. Chen, and K.-C. Chou, "iDNA6 mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC," *Genomics*, vol. 111, no. 1, pp. 96–102, Jan. 2019, doi: 10.1016/j.ygeno.2018.01.005.
- [43] K.-C. Chou and C.-T. Zhang, "Prediction of protein structural classes," *Crit. Rev. Biochem. Mol. Biol.*, vol. 30, no. 4, pp. 275–349, 1995.
- [44] M. Delorenzi and T. Speed, "An HMM model for coiled-coil domains and a comparison with PSSM-based predictions," *Bioinformatics*, vol. 18, no. 4, pp. 617–625, Apr. 2002.
- [45] K. C. Chou and H. B. Shen, "Recent advances in developing Web-servers for predicting protein attributes," *Natural Sci.*, vol. 1, no. 2, pp. 63–92, 2009.
- [46] K.-C. Chou, "An unprecedented revolution in medicinal chemistry driven by the progress of biological science," *Current Topics Medicinal Chem.*, vol. 17, no. 21, pp. 2337–2358, Jul. 2017.



SHARAF J. MALEBARY (Member, IEEE) received the Ph.D. degree in computer science and engineering from the University of South Carolina, USA. He is currently an Assistant Professor with King Abdulaziz University, Rabigh. He wrote several articles in information technology. His research interests include autonomous systems, wireless communications, artificial intelligence, and machine learning.

RABIA KHAN is currently pursuing the M.S. degree with the Department of Computer Science, University of Management and Technology (UMT). Her major research interests include pattern recognition, image processing, bioinformatics, and computer vision.



YASER DAANIAL KHAN (Member, IEEE) is currently working as a Full Professor with the University of Management and Technology (UMT), where he is leading the Research Group on Pattern Recognition and Bioinformatics. He has published articles in well-reputed journals and conferences. His major research interests include pattern recognition, image processing, computer vision, and bioinformatics.