PREDICTION OF STRESS RESPONSE PROTEINS VIA RANDOM FOREST, SUPPORT VECTOR MACHINE AND NEURAL NETWORK BY INTEGRATION OF STATISTICAL MOMENTS

Abdullah
S2018279003@umt.edu.pk
Department of Computer Science, University of management and Technology

***Abstract*—Protein is a vital component of cells, and it is compulsory for the structure, regulations, and functions of the body's tissues. Therefore, the detection of a protein function requires a thorough understanding of protein structure. Stress proteins are the primary and critical mediators of several cellular stress responses and categorized into two major groups based on their structure or function. One group activated only under cellular stress, whereas the other operates in both stressed as well as normal cellular function. This research has done for the identification of stress in a protein sequence with the help of three different algorithms of the machine learning method, a random forest classifier, an artificial neural network classifier, and a support vector machine. Hence, in our investigation, 99.9 % valid results have attained by using a random forest classifier, whereas by using an artificial neural network classifier, 87.0% precise results noted and the support vector machine shows 80.2% accuracy results. The authenticity of the random forest classifier was observed as 99.9%, sensitivity value was measured as 99.6% while the specificity was calculated as 99.9% all in-inclusive and the MCC value was measured as 0.993%. Live web server is available at http://biopred.org/stressprotiens.**

Keywords: Stress Response Protein, Machine learning, Random Forest, Cross-validation, support vector machine

I. Introduction

Proteins are large bio-molecules, or macro-molecules, consist of one or more long chains of amino acid residues. Fifty percent of the dry weight of humans is protein. Up to 92% of the dry weight of the red blood cell is a single protein called hemoglobin that transports oxygen in the human body. Proteins consist of many different amino acids linked together. There are twenty distinct building blocks of these amino acids commonly found in plants and animals, which make protein. A typical protein is composed of 300 or more amino acids, and each protein is unique in the specific number and sequence of amino acids. The amino acid 'letters' can arranged in millions of different ways to create 'words' and a whole protein 'language' rather than the alphabet. The resulting protein will fold into a particular shape, based on the number and sequence of the amino acids[1]. This form is essential because it will determine the function of the protein like muscle construction, DNA transcription RNA translation, and structural functions, etc. Protein structure is divided into four levels. The primary level contains amino acid residues, Secondary level have alpha helices and beta sheets, Tertiary level contains polypeptide chain also known as the native structure of the protein, and Quaternary level contains assembled subunits as shown in Figure A-I.

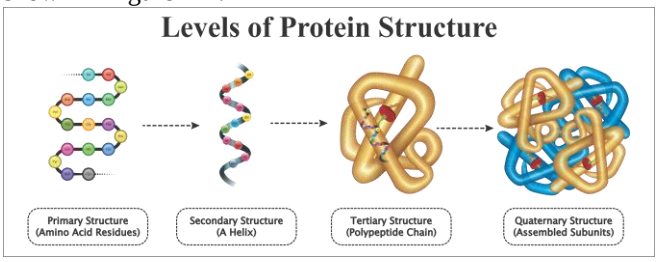

Figure A-I: Protein Composition

The terminology of "Stress" first coined in 1936, define the interaction between a force and the resistance to counter that force. "Hans Selye" the founder of stress theory, described stress as the nonspecific response of any demand upon a body [2]. Stress proteins include HSPs (heart shock proteins), RNPs (RNA chaperone proteins), and proteins primary role in ER (Endoplasmic Reticulum), peptidyl-propyl isomerases, protein disulfide isomerases (PDIs) and the chaperone lectin-binding mechanism. Stress protein is associated with many human diseases, especially cardiac and neurodegenerative diseases, such as heart attack, Alzheimer's disease, Parkinson's disease, and Huntington's disease. Neurodegenerative diseases are a heterogeneous group of disorders characterized by the gradual degeneration of the central nervous system or peripheral nervous system structure and function. Stress proteins are primary mediators of several cellular stress responses and sub-divided into two categories based on their mechanism of action. One group of these stress proteins is activated only under cellular stressed conditions, whereas other activates in both, stressed and normal cellular functions, enhancing cell survival [3]. These proteins found to conserved across many eukaryotic and prokaryotic linkages and demonstrate varied functional activities inside a cell. For instance, mutations in DNA encoding stress proteins of Drosophila hindered with the mitotic division and proteasome-mediated protein degradation [4], affecting their survival at elevated temperatures.

Classic examples of stress proteins include heat shock proteins or molecular chaperones that help to repair cellular damage [5, 6]. Moreover, Chaperons can significantly alter disease progression in the case of chronic injuries, DNA damage, and age-related cellular dysfunction [7]. Their tissue

specificity and selective induction exhibit their potential evolution through micro-environmental changes despite their ubiquity in all organisms. Also, to enhance cell survival, stress response proteins [8] modulate immune responses and function in tissue and organ trauma. Clinical implications of these HSPs account for their structural and functional understanding and their potential roles in treatment.

This study comprises the identification of stress in a protein sequence with the help of a machine learning approach like a random forest approach and neural network approach. By developing, a classifier for identification of stress there will be a comparative analysis of both approaches.

The paper implies five basics steps, as mention below [9]. 1-Benchmark data set, 2-Feature Extraction, 3-Learning models, 4-Testing, 5-Results/Webserver. This five-step rule is widely used for the prediction and analysis of proteins sequence.

II. Materials and Methods

This section describes our data and implements five-step rules. Data collected from UniProtKB. 7092 reviewed positive protein sequences, and 7500 reviewed Negative protein sequences by searching 'Stress Response [KW-0346]' in keywords is taken. The ratio of positive and negative proteins sequence is almost equal, as shown in Figure A-I.

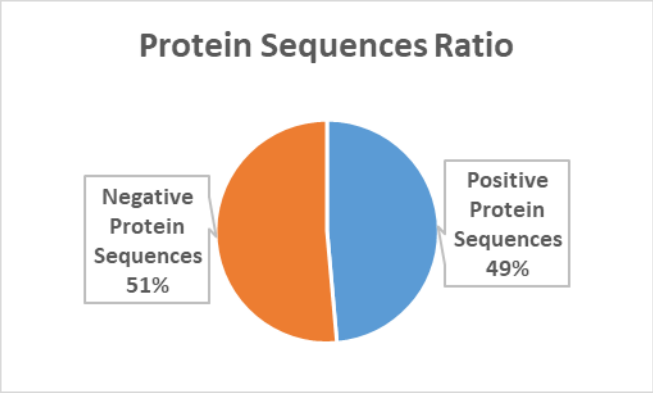

Figure A-I: Ratio of Positive & Negative Dataset

The first step of five-step rule is the creation of the data-set, the second step is benchmarking data-set, the third step is feature extraction from samples, the fourth step is to develop and train the predicted model using features selection, the fifth step is apply testing on all predicted models, as shown in Figure A-II.

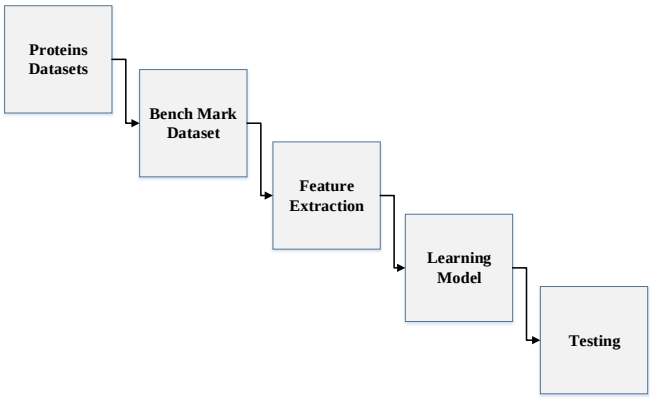

Figure A-II: Proposed Five Steps

III. Benchmark Dataset

The protein arrangement of the benchmark was found by using following paper [10] that was exempted in 2006 by UniProtKB. Data is the most central task. Reviewed data of stress protein sequence of different cell lines (organisms) were retrieved from UniProtKB, by using keyword stress response [KW-0346]. After that more Stress Proteins are added in the dataset that was pulled out from the latest version 2017_03 UniProtKB. The additional extension used to get different and useful proto-stress response proteins information.

A. Sample Formulation:

With the fast development of Biological sequence, the most critical issue in computational science is how to define a natural succession with a discrete model. However, still extensively has its arrangement, example and highlight fundamental to the objective broke down. This is states that all the current machine-learning calculations can just deal with vector not with sequence tests, as expounded [11]. To report this issue, the pseudo amino acid structure [12] was anticipated. As far the idea of these papers [13] - [14] was developed, it has quickly entered into the numerous bio-medicine and medication improvement zones[15] - [16] and every zone of computational proteomics, and an extensive rundown of references referred to in an audit paper [17]. Since it has been generally and progressively utilized, some ground-breaking open-access programming projects[11] - [18] were built to create different methods of extracting features. Enlivened by the achievements of utilizing to manage protein arrangements, four web-servers were proposed by following papers [19] - [20] Specific components for Protein groups have been saved for vector making and running genetic examinations. These examinations are important in monitoring various issues in development genomics, such as developmental reviews as described in the paper [10]. According to the demand, a robust web server is launched called Pse-In-One [21]. Both protein/peptide and Protein groups can use to make a perfect Feature vector. Additional information can use for tests in S1

[8].

$$L_{\xi=7}(I) = [\Psi_1 \Psi_2 \dots \Psi_u \dots \Psi_\Omega]^D \quad \text{Eq. 1}$$

The parts Ψj (j = 1, 2, ⋯, Ω) of the content will be considered as a method of eliminating the properties of the development of Proteins sequence, and T-transfer is considered as an executive.

A20 = D and Rx (x = 1, 2, ⋯, 41; x20) 20 can be presented before close to 15 amino acids or false code. We approach the math values $[1, 2, 3, 4, 5, 6, \dots \dots 20]$ for local amino acids, depending on the application of their sole value, and used 21 to identify the exposed amino acid x. This time, we used the structure measurable moments to deal with the characteristics and its measurement.

The standard database length of peptide is 41, Eq. 2 can be denoted:

$$L = A_1 A_2 \dots A_{17} A_{18} A_{19} A_{20} \dots A_{40} A_{41} \quad \text{Eq. 2}$$

B. Statistical Moment's Calculation

Statistical moments used for the quantitative collection of data. The different orders of moments used for the representation of different data properties. Some moments can be used for data size evaluation and others for direction and peculiarity of data indication. Statisticians and mathematicians explain different prospects based on the functions of distribution and polynomials. Some moments such as "Raw, Central and Hahn moments." Are used in following [22, 23]. The raw moment is for location and scale variant, used for mean calculation, and dataset asymmetry probability distribution. Central moments are location invariant because centric calculations are performed there, but it can be used for mean and variance calculation [24, 25]. Hahn moments based on "Hahn Polynomials" used for the variance of scale and location. All these moments provide sensitive information about the sequence order [26]. The explanation for selecting these measurable moments is the arrangement of the sensitive data given by these moments. Besides, scale-invariant moments maintained by a strategic distance. The information is characterized in its specific manner by one of the techniques evaluated by esteems [27]. All these moments are used in the form of a 2D matrix, i.e. p of n*n dimension produced for considerate amino acid residue in proteins p. In L' a "Transformation matrix" used by the function as explained by

$$L' = \begin{pmatrix} c_{11} & \cdots & c_{1h} \\ \vdots & \ddots & \vdots \\ c_{g1} & \cdots & c_{gh} \end{pmatrix} \quad \text{Eq. 3}$$

The framework changes into L' done by capacity ω characterized by "Akmal et al., 2017". Each moment is determined up to degree 3, so components of L′ are used. In conclusion, the raw moments are determined as:

$$G_{xy} = \sum_{l=1}^{h} \sum_{n=1}^{h\Sigma} l^x n^y \beta_{ln} \quad \text{Eq. 4}$$

Where (l+n) is the moments and raw moments were as G00 G01, G10, G11, G12, G21, G30, and G03.

In the future, the central moments are determined as:

$$H_{xy} = \sum_{l=1}^{h} \sum_{n=1}^{h\Sigma} (l - \bar{a})^i (n - w)^y \beta_{ln} \quad \text{Eq. 5}$$

The distinction in L into a 2D square cross-segment L offers an advantage that Hahn-moments can effectively enroll for an even-dimensional information connection. Discrete Hahn moments require square cross-segment as information. Hahn-moments' evenness inferences the alterable property of these moments and further applies it to discrete Hahn-moments through opposite use of them. Repeat the knowledge is feasible, and data concerning movement membership and relative positions protected throughout this time. For figuring the Hahn polynomial of demand n, the below condition is used Eq. 6.

$$N_h^{z,t}(j, A) = (A + T - 1)_a * \sum_{i=0}^{h} (-1)^i \frac{(-h)_k (-j)_i (2A + z - t - a - 1)_i}{(A + t - 1)_i (A - 1)_i} \frac{1}{I!} \quad \text{Eq. 6}$$

In the above equation pochhammer and the Gama executive portrayed by the following paper [28].

For figuring of symmetrical institutionalized Hahn to two-dimensional discrete data, following equation used.

$$E_{xy} = \sum_{n=0}^{H-1} \sum_{l=0}^{H-1} \beta_{xy} e_x^{\widetilde{a,t}}(n, H) e_y^{\widetilde{a,t}}(l, H) \quad g, h = 0, 1, \cdots H\text{-}1 \quad \text{Eq. 7}$$

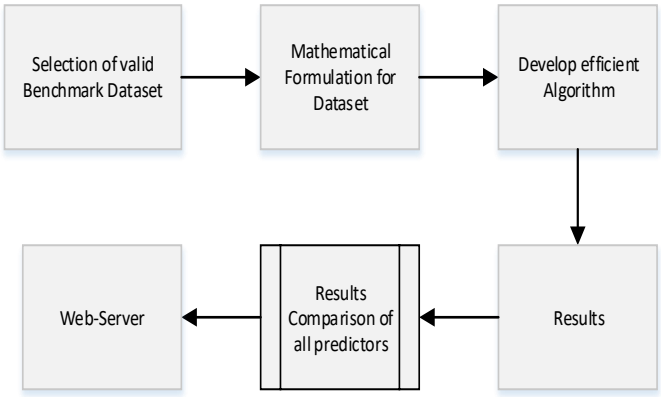

Figure B-I: Graphical presentation of 5-step rule

C. Determination of Positive Relative Incidence Matrix:

In the initial phase, the mathematical calculations made for the protein's prediction. The necessary arrangement of proteins and the relative position of the remaining particles is the pivotal model. The total number of proteins used to assemble the "PRIM" and the result extracted in the form of the 20x20 matric. Therefore, the remaining particles of the amino acid are as follows:

$$M_{PRIM} = \begin{bmatrix} M_{1\to1} & M_{1\to2} \cdots & M_{1\to y} \cdots & M_{1\to1} \\ M_{2\to1} & M_{2\to2} \cdots & M_{2\to y} \cdots & M_{2\to20} \\ M^i_{x\to1} & M^i_{x\to2} \cdots & M^i_{x\to y} \cdots & M^i_{x\to20} \\ M^i_{A\to1} & M^i_{A\to2} \cdots & M^i_{A\to y} \cdots & M^i_{A\to20} \end{bmatrix} \quad \text{Eq. 8}$$

The protein matrix of action demonstrates the remaining i-th area, which is not hindered by Ax-y and y = 1…..20 demonstrates the request to the local amino corrosive. As a result, 400 numbers produced by the use of matrix, if we want to reduce the number, we just compare the statistical moments and PRIM that produce the set number of 24

D. Determination of Reverse Positive Relative Incidence Matrix:

The effectiveness and correctness of machine learning strategy are inconceivably subject to careful quality, the carefulness in which the most related parts of information have dragged out. The PRIM grid exposes or separates data regarding the general situation of amino acids. Another grid RPRIM shaped in such a way that it works indistinguishable steps from PRIM, however, on the switch essential arrangement. Presentation of RPRIM reveals additionally concealed examples and reduces ambiguities among proteins that now look identical to polypeptide arrangements.

$$M_{RPRIM} = \begin{bmatrix} M_{1\to1} & M_{1\to2}\cdots & M_{1\to y}\cdots & M_{1\to20} \\ M_{2\to1} & M_{2\to2}\cdots & M_{2\to y}\cdots & M_{2\to20} \\ M^{i}_{x\to1} & M^{i}_{x\to2}\cdots & M^{i}_{x\to y}\cdots & M^{i}_{i\to20} \\ M^{i}_{A\to1} & M^{i}_{A\to2}\cdots & M^{i}_{A\to y}\cdots & M^{i}_{A\to20} \end{bmatrix} \quad \text{Eq. 9}$$

E. Determination of Frequency Vector:

Every protein structure contains an instant rate matrix (Q) that generates the structure of datasets obtained from amino acid frequency. The order on which amino acid depends is the number of frequencies signified and used for calculating the dispersal of frequency, so it considered as:

$$\xi = \{\tau_1, \tau_2, \cdots, \tau_{20}\} \quad \text{Eq. 10}$$

Where τi characterized by the occurrence of i-th remaining amino acid residue. The compositional data is determined by computing the FM in a Sequence (order).

F. Accumulative Absolute Position Incidence Vector Generation:

The FM only used for extracting the information of amino acids; it does not determine the relative position of residues. AAPIV (Accumulative Absolute Position Incidence Vector) used for the length of 20 elements. Determine the relative position, it uses the methodology of computing the ordinal value and summed up for every amino acid in the group. The primary sequence always arranged.

AAPIV path represented as:

$$K = [23] \quad \text{Eq. 11}$$

So, an illogical i-th element of AAPIV calculated as:

$$\mu_i = \sum_{k=1}^{n} p_k \quad \text{Eq. 12}$$

G. Reverse Accumulative Absolute Position Incidence Vector:

RAAPIV built for retreating the basic model to find out AAPIV. It simplified as a 20-component vector. The reverse operation applied to the primary sequence to generate the AAPIV.

The presence of a craving buildup in the turned around arrangement appeared as:

$$\Lambda = \{\eta_1, \eta_2, \eta_3, \ldots, \eta_{20}\} \quad \text{Eq. 13}$$

IV. PROPOSED METHODOLOGY

This study concentrated on a specific type of protein and try to predict with the help of protein sequence. For this prediction, we follow these steps, shown in Figure G-I. We suggested three different algorithms for the prediction of stress response proteins, after comparing the results of all these algorithms, we found the best algorithm of accuracy.

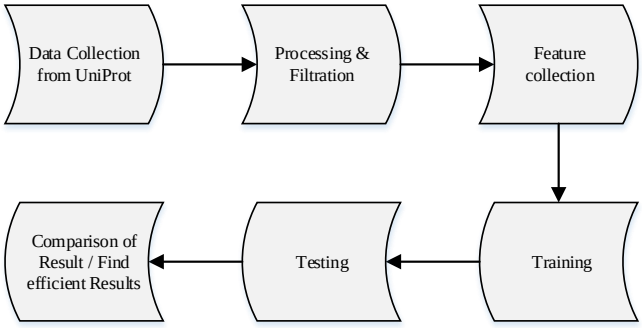

Figure G-I: Flowchart of Proposed Methodology

A. Prediction Algorithms:

This study proposed three different prediction algorithms. After training and testing of all these proposed algorithms, the most efficient prediction algorithm has suggested. The proposed algorithms introduced below.

B. Random Forest:

Random Forest (RF) is a powerful machine learning classifier used for the classification, prediction, and accuracy of the model. It operated by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes. The advantages of the Random Forest Classifier include non-parametric nature, maximum classification accuracy, and capability to determine the variable, which is important in predicting the maximum accuracy [29]. Every feature vector consists of “raw”, “central” and “Hahn-moments”. For a 2D depiction of protein arrangement, "PRIM" and "RPRIM" are used. The data includes the "Frequency Matrix (FM)", "AAPIV", "RAAPIV" and "SVV" method on the vector. Finally, a vector (153 + 2r) designed. FIM designed by using all the vectors, each line agrees on a single model. A "Frequency output matrix" is a control method, which followed by a continuous component class ("positive or negative") in the FIM. Mathematical networks used for both (FIM and EOM) to do the good work.

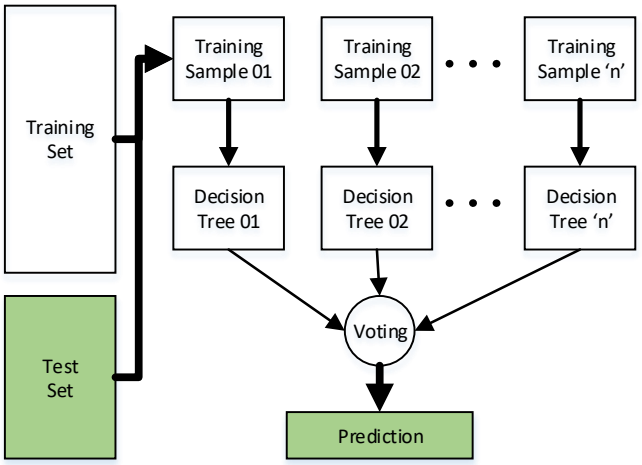

Figure B-I: Architecture of Random Forest Classifier for the proposed prediction model

A random forest classifier is used by these parameters for the implementation as shown in Table B-I.

Table B-I: Parameter for Random Forest Classifier

Parameter	Values	Parameter	Values
n_estimators	50	n_jobs	1
Criterion	Gini	random_state	0
max_depth	16	Verbose	0
warm_start	TRUE	class_weight	Balance

C. Artificial Neural Network:

It is known as a neuronal construction in which the yield of the primary neuron is an input of the following neuron. In the beginning unit, transformation is made to the consequence of the addition of all previous weighted inputs to the values as shown in Figure C-I.

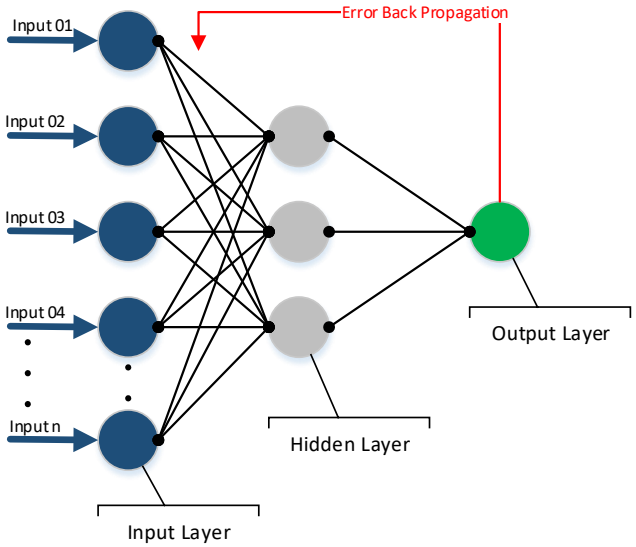

Figure C-I: Architecture of ANN Classifier for the proposed prediction model

The earlier formed benchmark dataset had both positive samples as well as negative samples. For the 2D symbol of the primary protein substructure, PRIM, and RPRIM, all the feature vectors contain a core, new, and Hahn-moments. Similarly, the structure and location data were applied to the function vector in the form of the FM (Frequency Matrix), AAPIV and RAAPIV. Resultantly we get feature vector covering 153+2r elements[30]. When the entire feature vectors combine in such a way that every row agreed to a single sample, as a result, a Feature Input Matrix (FIM) formed. A likely Output matrix built-in a supervised way, which followed the class (negative or positive) of the consistent element in FIM. Both these media (EOM and FIM) intended to use for the training of an artificial neural network. The FIM has adjusted to the neural network data, while the EOM is used to measure information errors using a back-propagation methodology.

The "gradient descent approach" used in the technique, which reduces the role in the contradictory incline track of the purpose and calculates the ratio of variance in the following results.

$$\theta = \theta - a\nabla\theta M\ (\theta) \qquad \text{Eq. 14}$$

Here, **M (θ)** represented by **θ ∈ A**. The gradient function described as **∇θA (θ)** and **A** show the learning factor rate. The function depends on the learning rate maintains the performance, and the learning rate must be a number which does not affect by any minor change that may occur several times. The learning rate differential is accepted by the 'adaptive learning algorithm' [31]. It is dependent on the procedure of the function. The faults of successive repetitions linked as if any error occurs in the succeeding solitary then the limitations aimed at repetition are useless, and the learning rate weakens the performance. By two consecutively calculated limitations, i.e. θ_x and θ_{x+1}, the masses were recalculated, and the production and following faults were designed for succeeding it [32]. If the flaws and learning rate are "Inversely Proportional" then the value of + θ_{x+1} is calculated later by the removal of masses and vice versa.

So it is shown in Eq. 15. Where Ag is the learning rate used for the g-th period

$$\theta g + 1 = \theta g - Ag\ \nabla M\ (\theta g) \qquad \text{Eq. 15}$$

D. Support Vector Machine:

SVM is a supervised learning model with the associated learning algorithm that analyzes data used for regression analysis and classification. SVM is generally used in a classification problem [33]. SVMs based on the idea of finding a hyperactive plane that divides a dataset into two classes.

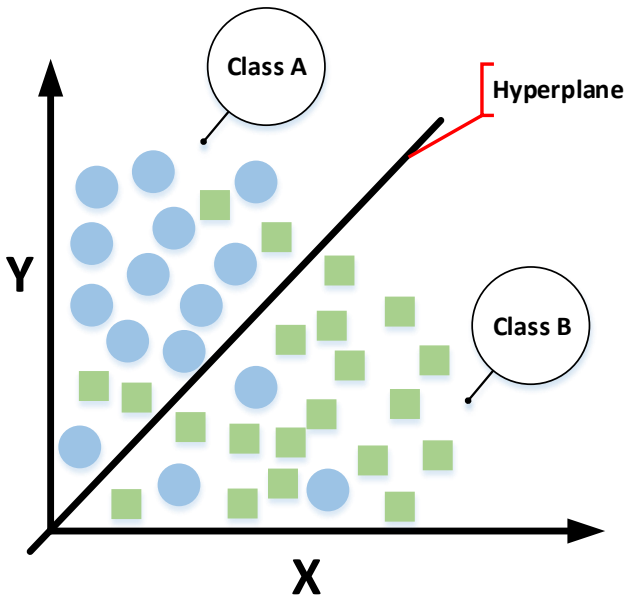

Figure D-I: Architecture of SVM Classifier for the proposed prediction model

Support vectors are the data points nearest to the hyperactive plane, the attributes of a data set, if removed, would alter the position of the hyperactive dividing plane. Because of this, they have considered the critical elements of a data set.

V. RESULTS

This section focuses on validation of the proposed model and shows the results of every proposed model, along with testing.

A. Accuracy Estimation:

Objectively evaluate a predictor, it is necessary to estimate the accuracy measures for that model. The selection of the testing method and the accuracy metrics to score that method is a crucial task and is necessary to consider. Thus, we define a set of metrics for the estimation of accuracy.

B. Metrics of Accuracy Estimation:

In general, The following metrics are used to determine the accuracy of the predictive model from four different perspectives: (1) MCC for model stability (2) Sp (specificity of the model) and, (3) Sn (sensitivity of model) [34]. Transformation metrics use the mathematical equations given by the following paper [35] are commonly used in the literature to measure the quality of predictive methods. However, they are no longer in use because most biologists are intuitive, and these methods are difficult to understand. In particular, MCC (Matthews's correlation coefficient) is a significant indicator that reflects the stability of the prediction method. Fortunately, four intuitive metric sets were derived based on the symbol introduced to study protein signal peptides[36]. [37] Described it in four steps, Eq. 16 Metrics formulation:

$$Sn = 1 - \frac{A_-^+}{A_+^-} \quad 0 \le Sn \le 1$$

$$Sp = 1 - \frac{A_+^-}{A^-} \quad 0 \le Sp \le 1$$

$$Acc = 1 - \frac{A_-^+ + A_+^-}{A^- + A^-} \quad 0 \le Acc \le 1 \qquad \text{Eq. 16}$$

$$MCC = \frac{1 - (A_-^+ / A^+ + A_+^- / A^-)}{[\sqrt{[1 + (A_+^- - A^+)/A^+] * [1 + (A_-^+ - A^-)/A^-]}}$$

Where N^+ represents the total number of Stress responses which were predicted indeed as Stress response and N_-^+ represents the total number of Stress responses which were predicted falsely as the non- Stress response. Also, N^- is the total number of non-stress responses that actually predicted as non-proto-oncogenes, and N_+^- is the total number of non-stress responses that have incorrectly predicted as proto-oncogenes. This equation explains the accuracy, sensitivity, specificity, and stability in terms of MCC, and report in various studies. However, it is used for binary class data, and other metrics are proposed for multi-class data[38, 39].

Types of testing via Random Forest Classifier, Artificial Neural Network, and SVM

- Self-consistency testing
- Jack-knife testing
- Independent Dataset testing
- 05-Fold Cross-Validation
- 10-Fold Cross-validation testing

C. Training Accuracy results through Self-Consistency Testing:

A self-consistency test performed and the same benchmark dataset used for training and testing the proposed predictor. This validation method used when the actual positive value of the benchmark dataset has known. The self-consistency results of random forest, Support vector machine and Artificial Neural network are shown in Table C-I. This shows the actual and predicted classification values for the proposed predictor. It displays the overall accuracy, specificity, sensitivity, and stability of the predictive model. While the ROC curve of these predictors shown in Figure C-I, Figure C-II, and Figure C-III, respectively. Besides, ROC comparison graph of all these predictors is shown in Figure C-IV.

Table C-I: Self-Consistency testing results of all predictors

	TN	FP	FN	TP	Acc (%)	Sp (%)	Sn (%)	MCC
RF	6140	0	2	7161	99.9	99.9	99.9	0.99
ANN	5125	1015	708	6455	87.0	90.1	83.5	0.73
SVM	5050	1090	1534	5629	80.2	78.6	82.2	0.61

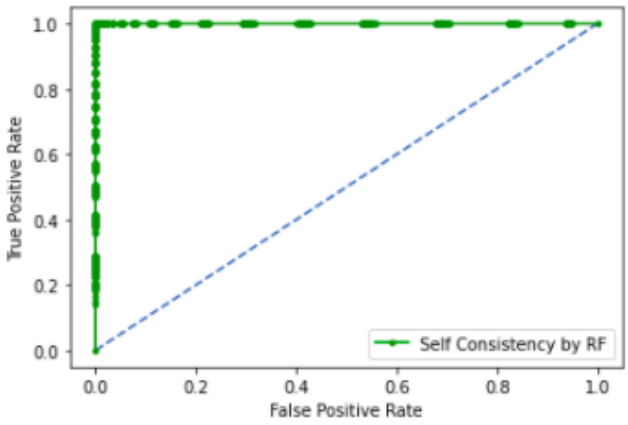

Figure C-I: Self-Consistency ROC Graph of RF

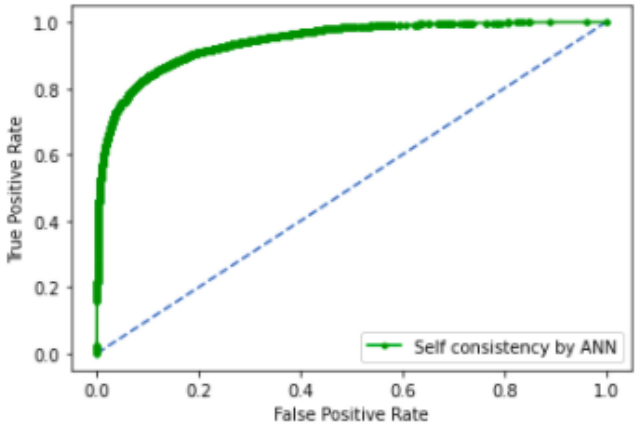

Figure C-II: Self-Consistency ROC Graph of ANN

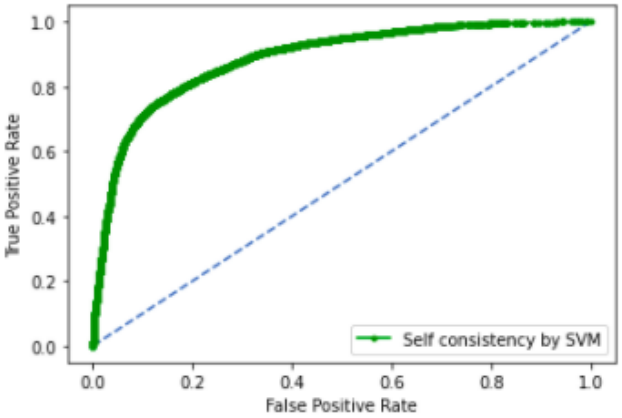

Figure C-III: Self-Consistency ROC Graph of SVM

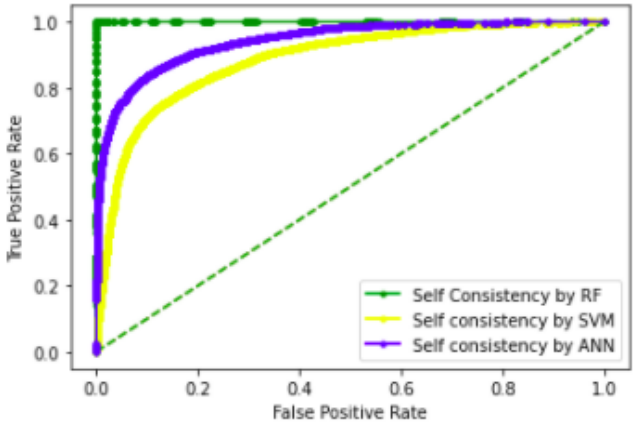

Figure C-IV: Self-Consistency Comparison ROC

D. Training accuracy results through Jack-Knife testing:

In Jack-knife, for training and testing of validation, all records are open even if the data instance is on or off completely. Jack-knife cross-validation always returns a unique output for a record. By using, jack-knife the intentional problems like subsampling and independent testing can completely be avoided. A jack-knife test was performed on RF, ANN, and SVM, the results are shown in Table D-I. It displays the overall accuracy, specificity, sensitivity, and stability of the predictive model, and the ROC curve is shown in Figure D-I, Figure D-II, and Figure D-III, respectively. The ROC comparison graph is shown in Figure D-IV.

Table D-I: Jack-Knife testing results of all predictors

	TN	FP	FN	TP	Acc (%)	Sp (%)	Sn (%)	MCC
RF	6139	1	2	7161	99.9	99.9	99.9	0.99
ANN	5125	1015	708	6455	87.0	90.1	83.5	0.73
SVM	5050	1090	1534	5629	80.3	78.6	82.2	0.61

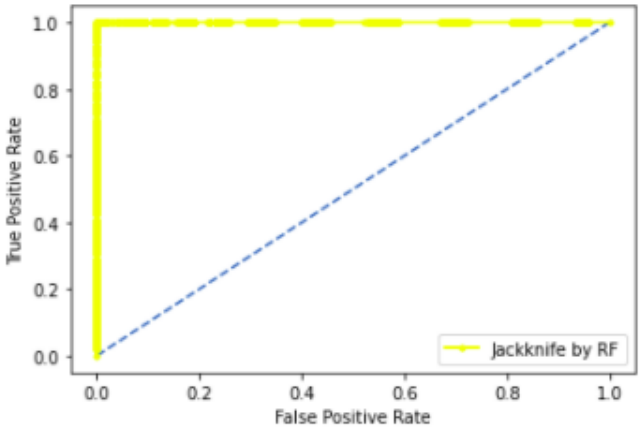

Figure D-I: Jack-Knife ROC Graph of RF

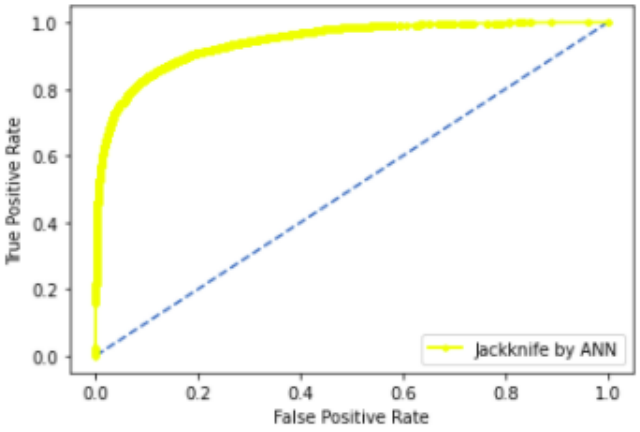

Figure D-II: Jack-Knife ROC Graph of ANN

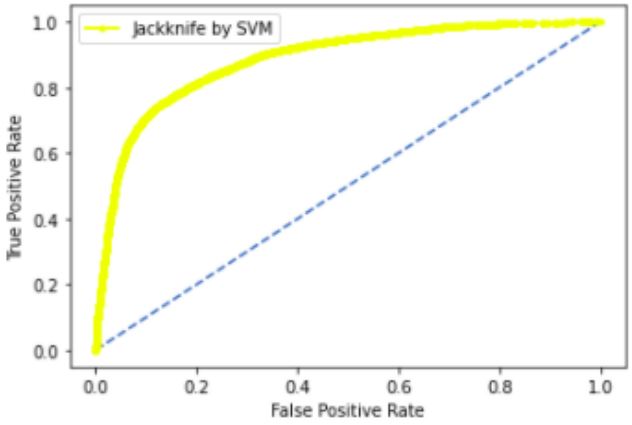

Figure D-III: Jack-Knife ROC Graph of SVM

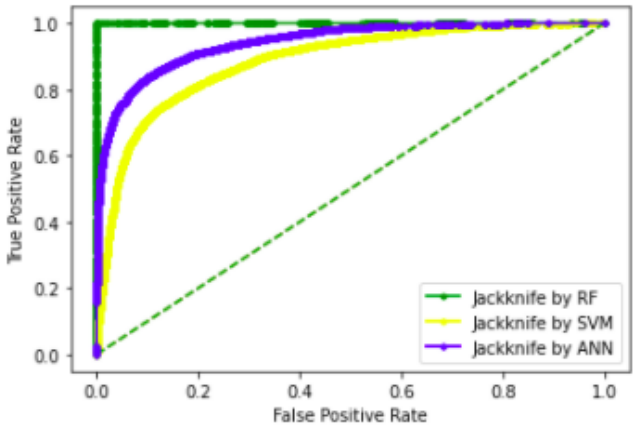

Figure D-IV: Jack-Knife Comparison ROC

E. Training Accuracy through Independent Dataset Testing:

Independent dataset testing performed by conducting 70-30 splits on the original dataset. RF classifier trained by a 70% dataset and tested by using the remaining 30% dataset. Table E-I displays the overall accuracy, specificity, sensitivity, and stability of the random forest, artificial neural network, and SVM, Figure E-I, Figure E-II, and Figure E-III shows the ROC graph of testing respectively. While the ROC comparison graph of all testing is shown in Figure E-IV.

Table E-I: Independent Dataset Testing results of all

	TN	FP	FN	TP	Acc (%)	Sp (%)	Sn (%)	MCC
RF	1577	276	81	2057	91.1	96.2	85.1	0.82
ANN	1601	269	222	1899	87.7	89.5	85.6	0.75
SVM	1081	789	1326	795	47.0	37.5	57.8	0.048

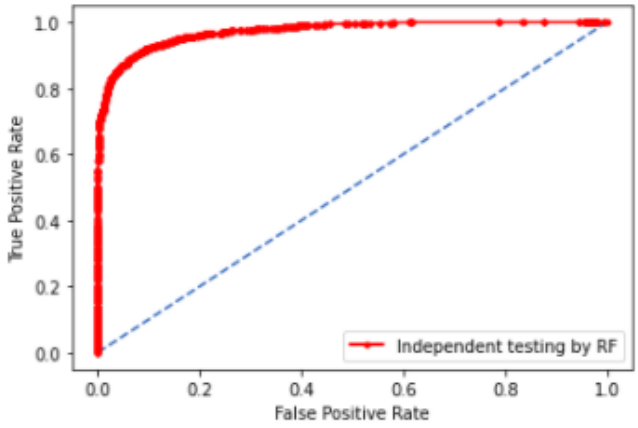

Figure E-I: Independent Testing ROC Graph of RF

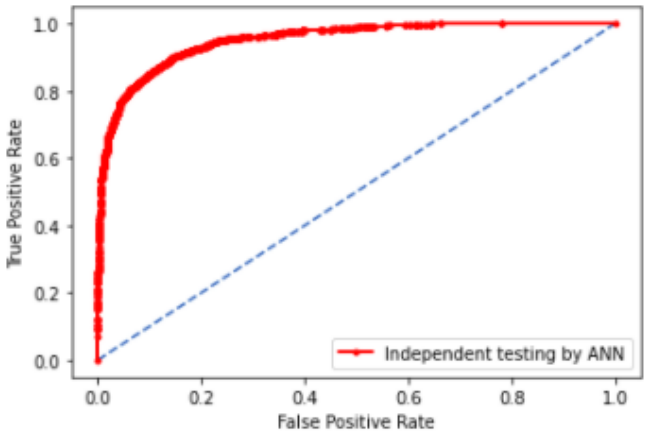

Figure E-II: Independent Testing ROC Graph of ANN

Figure E-III: Independent Testing ROC Graph of SVM

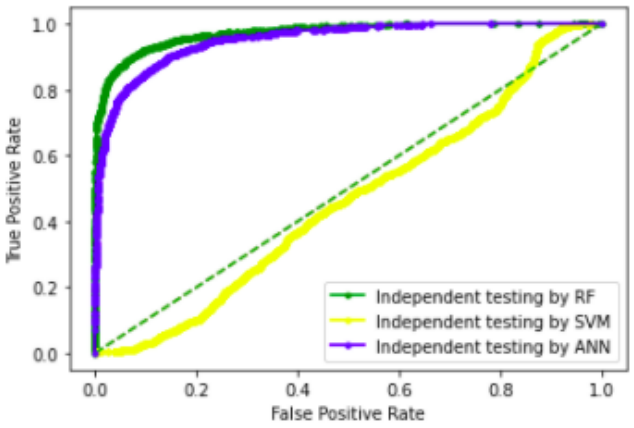

Figure E-IV: Independent Testing Comparison ROC

F. Validation through 5-Fold Cross-Validation

In cross-validation, the dataset divided into k sets, and k has chosen at the start, and then it kept constant. Usually, k kept 5 or 10 but, in this problem, k was set to five. The model is tested k times, and, in each iteration, four sets used as a training set, and the one set (k set) treated as a testing set.

Each set can be used as a testing set, and problems like underfitting and overfitting can easily remove. After performing k iterations, the model accuracy computed by taking averaging of each iteration. The average accuracy will be a result of cross-validation.

In the case of 5-Fold Cross-Validation, 99.8% of overall accuracy is obtained by the random-forest classifier, 61% of overall accuracy is brought by ANN, and 65% of overall accuracy is obtained by SVM as shown in Table F-I. While the ROC graphs are shown in Figure F-I, Figure F-II, and Figure F-III respectively and the ROC comparison graph is shown in Figure F-IV.

Table F-I: 5-Fold Cross Validation Result of all

	TN	FP	FN	TP	Acc (%)	Sp (%)	Sn (%)	MCC
RF	5810	330	211	6952	95.9	97.1	94.6	0.91
ANN	5125	1015	708	6455	87.0	90.1	83.5	0.73
SVM	5058	1082	1665	5498	79.4	76.8	82.4	0.58

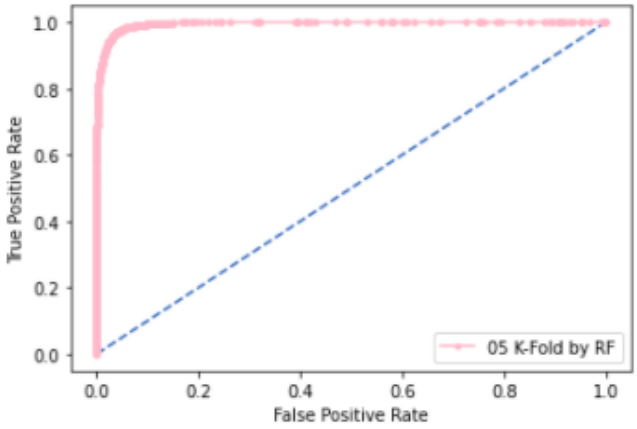

Figure F-I: 5-Fold ROC Graph of RF

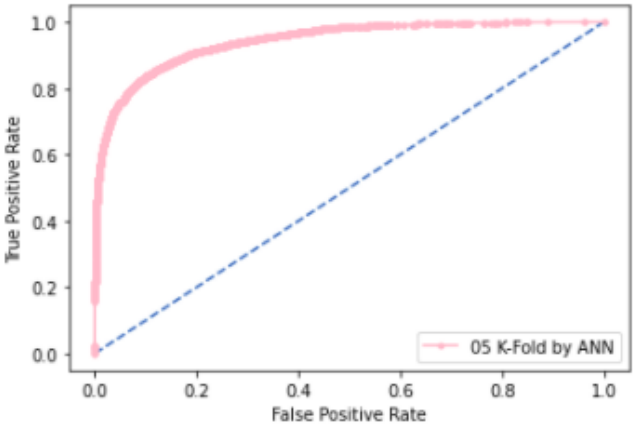

Figure F-II: 5-Fold ROC Graph of ANN

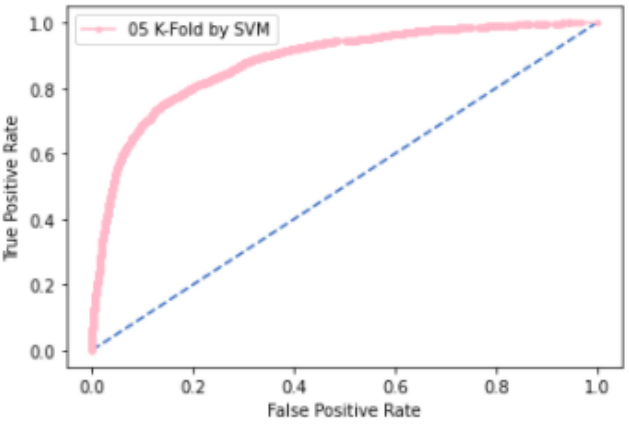

Figure F-III: 5-Fold ROC Graph of SVM

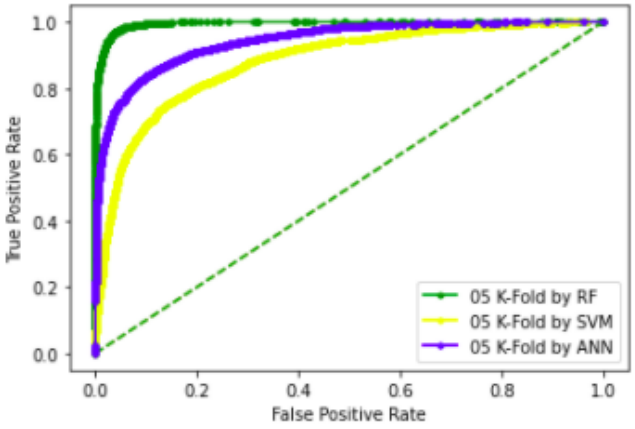

Figure F-IV: 5-Fold Cross-Validation Comparison ROC

G. Validation through 10-Fold Cross-Validation:

The experimentally validated datasets are used for model prediction and testing. However, no experimentally validated dataset compares the model with actual data. If the data is available, it may not be enough to test the accuracy of the predictive model. These tests need to perform to evaluate the four metrics in Eq. 16 to ensure sufficient accuracy and reliability. In general, you can test through k-fold (sub-scan), jack-knife, and independent test. The jack-knife model test is very exhaustive and may give different results for a particular benchmark record. Cross-validation is the best option to confirm the proper functioning of the developed model when there is no clear record to validate the model's prediction.

In cross-validation, the benchmark dataset divided into k single folds. 'K' is the number of parts and it is currently divided into k = 10. Each validator randomly selects a different data slice to validate the rest of the data, so all aspects of the dataset used for both testing and training. The result obtained is the average of all accuracies. Similar methods applied to negative and positive data samples. A random selection was made, forming a k = 10 subset. Cross-validation works better than other verification methods. These methods used to select partition data or random data for testing.

Table G-I shows the results obtained with 10-fold cross-validation via the Random Forest, ANN, and SVM, while Figure G-I, Figure G-II, and Figure G-III show the ROC curve of all predictors respectively. Figure G-IV shows the ROC comparison curve of all.

Table G-I: 10-Fold Cross-Validation Result of all

	TN	FP	FN	TP	Acc (%)	Sp (%)	Sn (%)	MCC
RF	5804	336	221	6942	95.8	96.9	94.5	0.91
ANN	5125	1015	708	6455	87.0	90.1	83.5	0.73
SVM	5054	1086	1675	5488	79.2	76.6	82.3	0.58

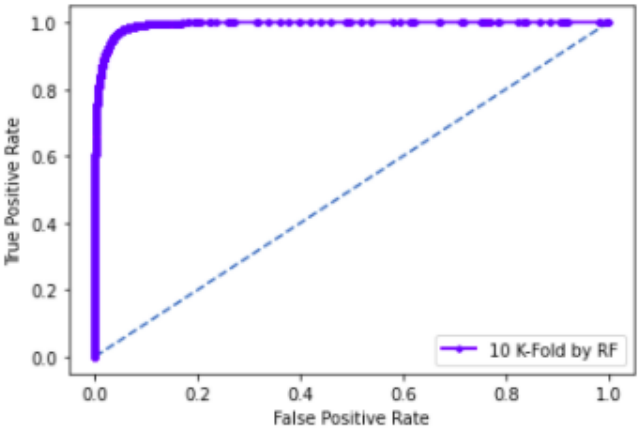

Figure G-I: 10-Fold ROC Graph of RF

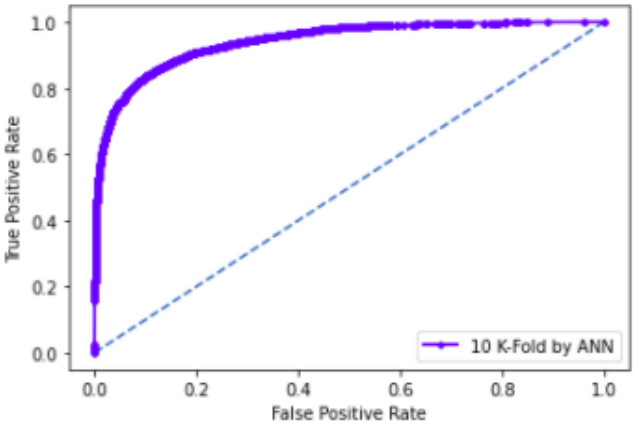

Figure G-II: 10-Fold ROC Graph of ANN

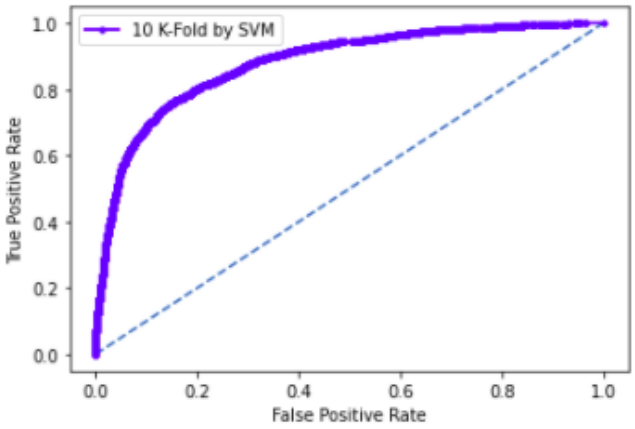

Figure G-III: 10-Fold ROC Graph of SVM

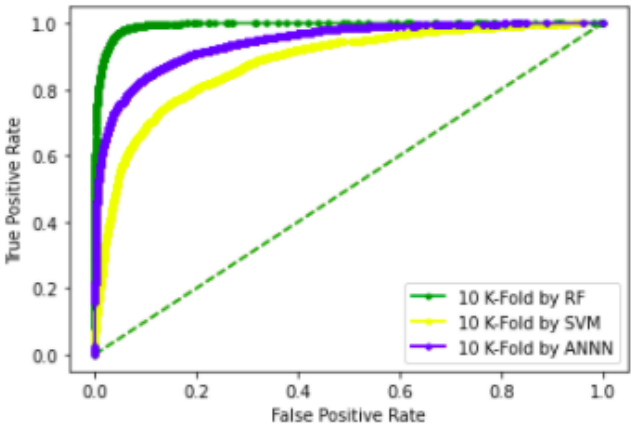

Figure G-IV: 10-Fold Comparison ROC

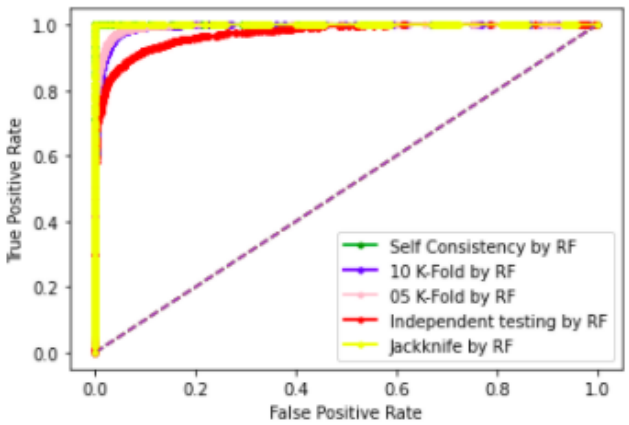

Figure A-I: Comparison Testing ROC of RF

VI. COMPARISON

In this section of the study, we compare the testing results of every predicted algorithm. For every algorithm, we perform five different tests, a Self-Consistency test, Jack-Knife testing, Independent testing, 10-Fold Cross-Validation testing, and 5-Fold Cross-Validation testing. Below we compare the results of all testing techniques for every predictor.

A. Comparison of testing on RF:

It observed that the Random forest algorithm gave the most accurate and efficient results in the prediction of stress response proteins. All testing results are shown in Table A-I, and Figure A-I shows the ROC graph of these testing results

Table A-I: All Testing Result Comparison of Random Forest

Random Forest				
Testing Method	**Accuracy (%)**	**Sp (%)**	**Sn (%)**	**MCC**
Self-Consistency	99.9	99.9	99.9	0.99
Jack-Knife	99.9	99.9	99.9	0.99
Independent	91.1	96.2	85.1	0.82
10-Fold Testing	95.8	96.9	94.5	0.91
5-Fold Testing	95.9	97.1	94.6	0.91

B. Comparison of testing on ANN:

Artificial Neural Network algorithm gives the second-best accurate and efficient results in the prediction of stress response proteins. All testing results of ANN are shown in Table B-I, and Figure B-I shows the ROC graph of these testing results.

Table B-I: All Testing Result Comparison of ANN

Artificial Neural Network				
Testing Method	**Accuracy (%)**	**Sp (%)**	**Sn (%)**	**MCC**
Self-Consistency	87.0	90.1	83.5	0.73
Jack-Knife	87.0	90.1	83.5	0.73
Independent	87.7	89.5	85.6	0.75
5-Fold Testing	87.0	90.1	83.5	0.73
10-Fold Testing	87.0	90.1	83.5	0.73

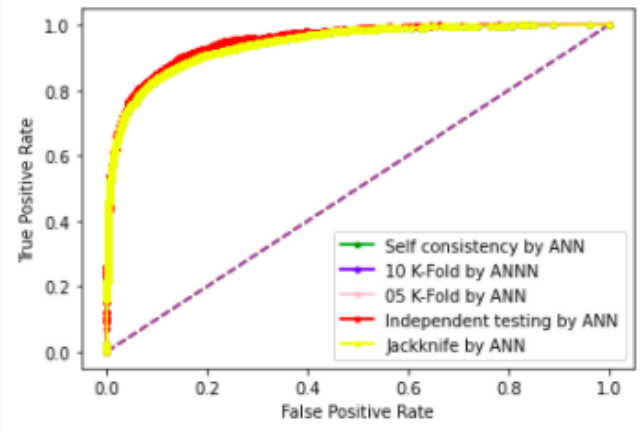

Figure B-I: Comparison Testing ROC of ANN

C. Comparison of testing on SVM:

Support Vector Machine algorithm gives the lowest accurate results in the prediction of stress response proteins. All testing results of SVM are shown in Table C-I, and Figure C-I shows the ROC graph of these testing results.

Table C-I: All Testing Result Comparison of SVM

Support Vector Machine				
Testing Method	**Accuracy (%)**	**Sp (%)**	**Sn (%)**	**MCC**
Self-Consistency	80.2	78.6	82.2	0.61
Jack-Knife	80.3	78.6	82.2	0.61
Independent	47.0	37.5	57.8	0.048
5-Fold	79.4	76.8	82.4	0.58
10-Fold	79.2	76.6	82.3	0.58

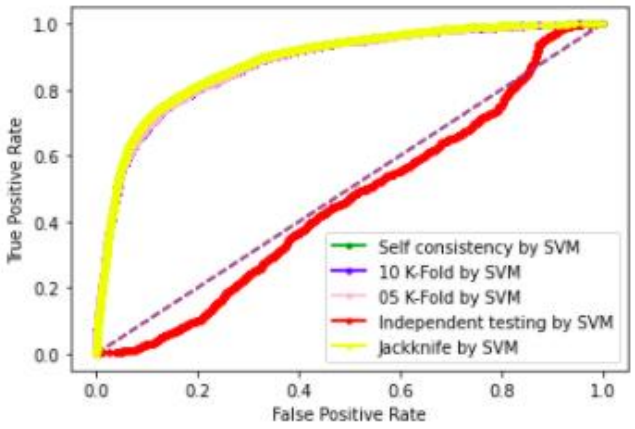

Figure C-I: Comparison Testing ROC of SVM

D. BoxPlot Comparison 10-Fold Cross-Validation:

In this section, we discuss the boxplot comparison of 10-Fold Cross-Validation of RF, ANN, and SVM. Figure D-I show boxplot of 10-Fold cross-validation by an artificial neural network. Figure D-II shows a boxplot of 10-Fold cross-validation by support vector machine. And Figure D-III shows the boxplot graph of 10-Fold Cross-validation by random forest.

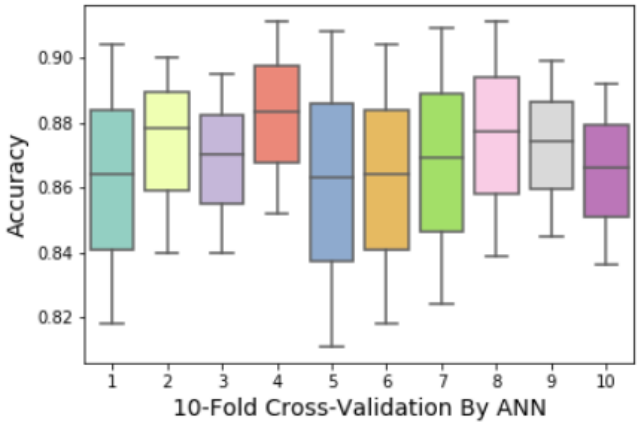

Figure D-I: ANN 10-Fold Cross-Validation

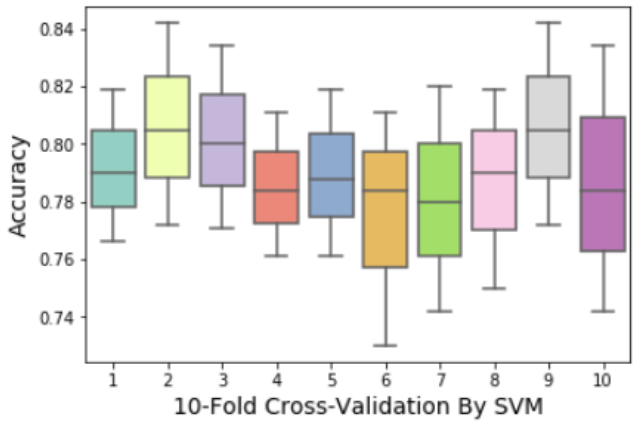

Figure D-II: SVM 10-Fold Cross-Validation

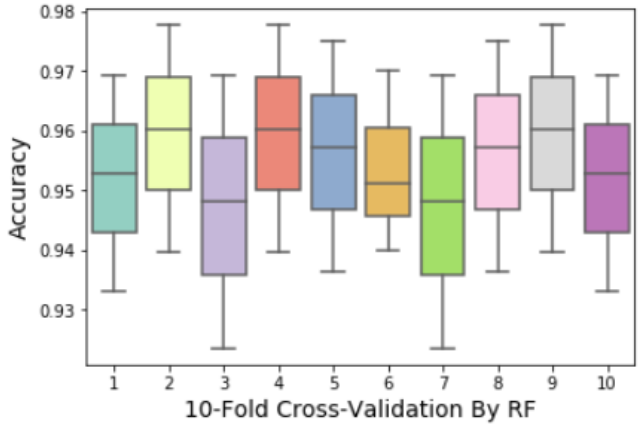

Figure D-III: RF 10-Fold Cross-Validation

E. BoxPlot Comparison 5-Fold Cross-Validation:

This section discusses the boxplot comparison of 5-Fold cross-validation of RF, ANN, and SVM. Figure E-I show boxplot of 5-Fold cross-validation by an artificial neural network. Figure E-II shows a boxplot of 5-Fold cross-validation by support vector machine. And Figure E-III shows the boxplot graph of 5-Fold Cross-validation by random forest.

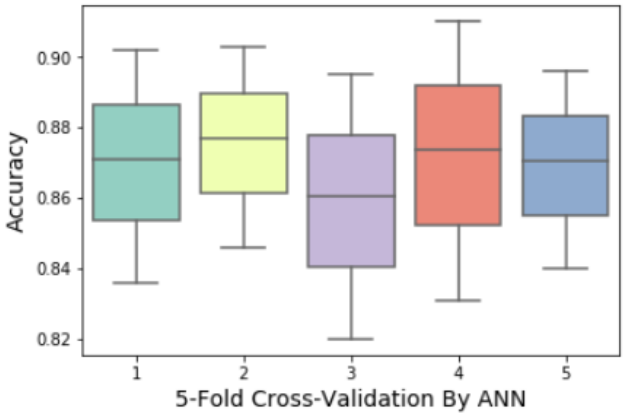

Figure E-I: ANN 5-Fold Cross-Validation

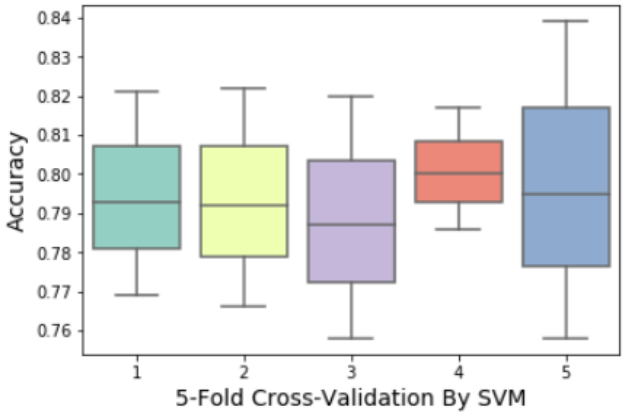

Figure E-II: SVM 5-Fold Cross-Validation

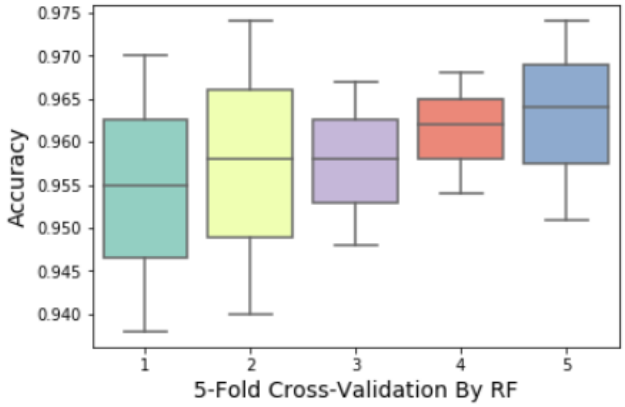

Figure E-III: RF 5-Fold Cross-Validation

VII. WEB-SERVER

The final phase of the five-step rule is the improvement of the web-server. It is showed in a progression of ongoing productions [40, 41], the webserver is easy to use. The openly available web-servers show the practical development that must be accurate and useful in the future for prediction. For the development of a web server, the Flask 1.0.2 used for the Stress response prediction, and the neural networks sklearn 0.0.0, wtform 2.2.1, NumPy 1.16.3, Tensorflow 2.0.0, and Keras 2.2.4 libraries used for the backend and interface. The screens of the webserver shown below. Figure E-I shows the Home page, Figure E-II shows the introduction page, Figure E-III shows the Prediction Server page, Figure E-IV shows the sample sequence data page, and Figure E-V shows the results page. A live webserver is available at Biopred.

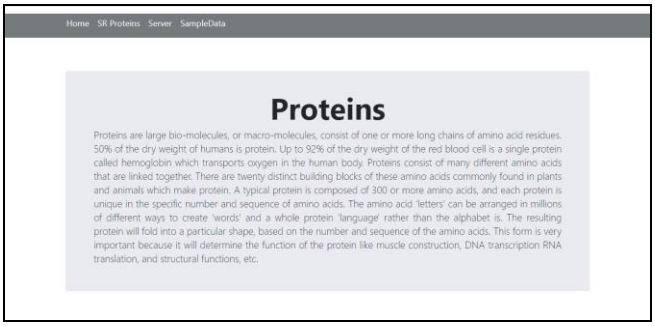

Figure E-I: Home Page of Web-Server

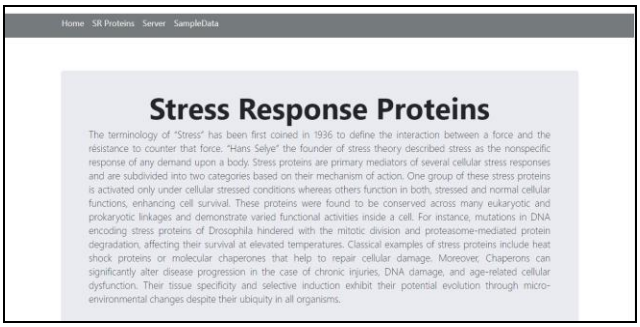

Figure E-II: Introduction Page of Web-Server

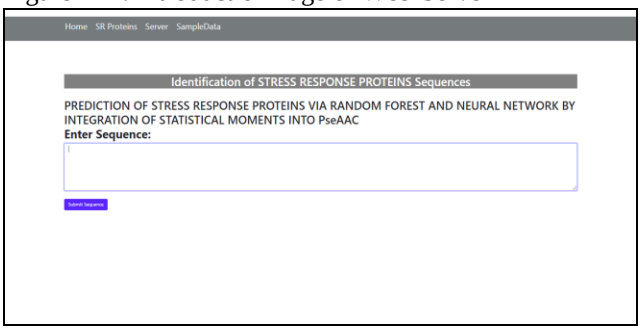

Figure E-III: Prediction Server Page

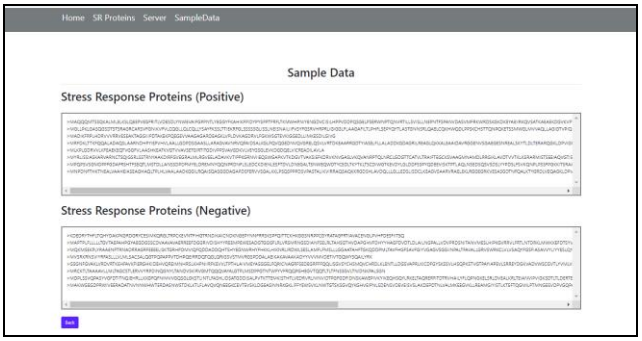

Figure E-IV: Sample Data Page

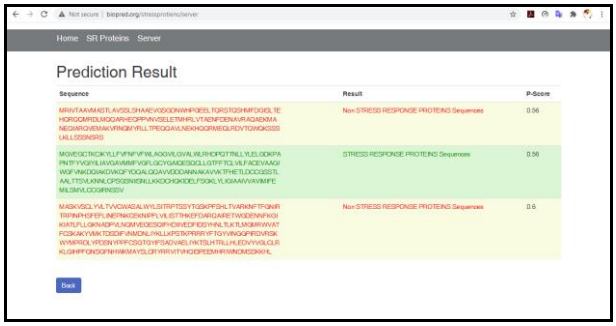

Figure E-V: Results Page of Web-Server

VIII. DISCUSSION

It can be in section 3 that the feature vector used to define the Amino acid features and run algorithms on these features vectors, a current study based on three classifiers named as Random Forest, Support Vector Machine and Artificial Neural Network. The highest accuracy achieved by random forest among three classifiers as shown in Table A-I. The benchmark dataset of stress response proteins was very robous, that is why dataset was clean from redundancy and then applied classifier on that data set. The purposed model has achieved 99.9 % valid result attained by using Self-consistency and jack-knife testing, whereas by using k-fold cross-validation 95.8% precise result has noted. The authenticity of the random forest classifier was observed as 99.9%, sensitivity value was measured as 99.9% while the specificity was calculated as 99.8% all in-inclusive and the MCC value was measured as 0.9873%. The ANN has achieved 87.0% by self-consistency test, and SVM has 80.2% of accuracy by self-consistency testing. Random forest gave the best results as compared to ANN and SVM. It anticipated that Stress response Predictor becomes a handy high throughput tool for studying Stress Proteins or, at the very least. It is a very first predictor to predict Stress response Protein.

IX. CONCLUSION

In this study, three Classifiers has proposed to predict the Stress Response Protein Sequence named as an artificial neural network, support vector machine, and random forest classifier. It based on five-step rule. The highest result achieved by random forest classifier. The results of the random forest model showed 99.9% accuracy. This proves that the Random Forest Classifier shows better outcomes for the prediction of Stress Protein Sequences. The predictor verified by jack-knife testing and 10-fold cross-validation, yielding accurate results of 99.9%, and 95.8%, respectively. At the same time, the proposed predictor helps to predict stress response proteins efficiently or accurately and provide baseline data for the discovery of new drugs and biomarkers against medical issues of this category. While our other prediction algorithm shows low accuracy results as discusses in the comparison section. In that case, a random forest classifier is the most significant predictor among predicted classifiers.

REFERENCES

[1] A. M. Lesk, *Introduction to protein architecture: the structural biology of proteins*. Oxford University Press Oxford, 2001.

[2] S. Y. Tan and A. J. S. m. j. Yip, "Hans Selye (1907–1982): Founder of the stress theory," vol. 59, no. 4, p. 170, 2018.

[3] T. J. Little, L. Nelson, and T. J. P. O. Hupp, "Adaptive evolution of a stress response protein," vol. 2, no. 10, p. e1003, 2007.

[4] C. N. Rokde and M. Kshirsagar, "Bioinformatics: protein structure prediction," in *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, 2013, pp. 1-5: IEEE.

[5] K.-C. Chou, C.-T. J. C. r. i. b. Zhang, and m. biology, "Prediction of protein structural classes," vol. 30, no. 4, pp. 275-349, 1995.

[6] J. Cheng, A. N. Tegge, and P. J. I. r. i. b. e. Baldi, "Machine learning methods for protein structure prediction," vol. 1, pp. 41-49, 2008.

[7] W. J. J. P. r. Welch, "Mammalian stress response: cell physiology, structure/function of stress proteins, and implications for medicine and disease," vol. 72, no. 4, pp. 1063-1081, 1992.

[8] M. R. Hemm, B. J. Paul, J. Miranda-Ríos, A. Zhang, N. Soltanzad, and G. J. J. o. b. Storz, "Small stress response proteins in Escherichia coli: proteins missed by classical proteomic studies," vol. 192, no. 1, pp. 46-58, 2010.

[9] K.-C. J. J. o. t. b. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," vol. 273, no. 1, pp. 236-247, 2011.

[10] K.-C. Chou, H.-B. J. B. Shen, and b. r. communications, "MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM," vol. 360, no. 2, pp. 339-345, 2007.

[11] K.-C. J. M. c. Chou, "Impacts of bioinformatics to medicinal chemistry," vol. 11, no. 3, pp. 218-234, 2015.

[12] K. C. J. P. S. Chou, Function, and Bioinformatics, "Prediction of protein cellular attributes using pseudo-amino acid composition," vol. 43, no. 3, pp. 246-255, 2001.

[13] D.-S. Cao, Q.-S. Xu, and Y.-Z. J. B. Liang, "propy: a tool to generate various modes of Chou's PseAAC," vol. 29, no. 7, pp. 960-962, 2013.

[14] S.-X. Lin and J. Lapointe, "Theoretical and experimental biology in one—A symposium in honour of Professor Kuo-Chen Chou's 50th anniversary and Professor Richard Giegé's 40th anniversary of their scientific careers," 2013.

[15] W.-Z. Zhong and S.-F. Zhou, "Molecular science for drug development and biomedicine," ed: Multidisciplinary Digital Publishing Institute, 2014.

[16] G.-P. J. C. T. i. M. C. Zhou, "Impact of biological science to medicinal chemistry," vol. 17, no. 21, pp. 2335-2336, 2017.

[17] K.-C. J. C. P. Chou, "Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology," vol. 6, no. 4, pp. 262-274, 2009.

[18] P. Du, S. Gu, and Y. J. I. j. o. m. s. Jiao, "PseAAC-General: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets," vol. 15, no. 3, pp. 3495-3506, 2014.

[19] W. Chen, T.-Y. Lei, D.-C. Jin, H. Lin, and K.-C. J. A. b. Chou, "PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition," vol. 456, pp. 53-60, 2014.

[20] B. Liu, F. Liu, L. Fang, X. Wang, K.-C. J. M. G. Chou, and Genomics, "repRNA: a web server for generating various feature vectors of RNA sequences," vol. 291, no. 1, pp. 473-481, 2016.

[21] W. Chen, P. Feng, H. Ding, H. Lin, and K.-C. J. G. Chou, "Using deformation energy to analyze nucleosome positioning in genomes," vol. 107, no. 2-3, pp. 69-75, 2016.

[22] Y. D. Khan, F. Ahmad, and M. W. J. W. A. S. J. Anwar, "A neuro-cognitive approach for iris recognition using back propagation," vol. 16, no. 5, pp. 678-685, 2012.

[23] Y. D. Khan, F. Ahmed, S. A. J. N. C. Khan, and Applications, "Situation recognition using image moments and recurrent neural networks," vol. 24, no. 7-8, pp. 1519-1529, 2014.

[24] A. H. Butt, S. A. Khan, H. Jamil, N. Rasool, and Y. D. J. B. r. i. Khan, "A prediction model for membrane proteins using moments based features," vol. 2016, 2016.

[25] A. H. Butt, N. Rasool, and Y. D. J. T. J. o. m. b. Khan, "A treatise to computational approaches towards prediction of membrane protein and its subtypes," vol. 250, no. 1, pp. 55-76, 2017.

[26] Y. D. Khan, S. A. Khan, F. Ahmad, and S. J. T. S. W. J. Islam, "Iris recognition using image moments and k-means algorithm," vol. 2014, 2014.

[27] Y. D. Khan *et al.*, "An efficient algorithm for recognition of human actions," vol. 2014, 2014.

[28] M. A. Akmal, N. Rasool, and Y. D. J. P. o. Khan, "Prediction of N-linked glycosylation sites using position relative features and statistical moments," vol. 12, no. 8, p. e0181966, 2017.

[29] W.-R. Qiu, S.-Y. Jiang, Z.-C. Xu, X. Xiao, and K.-C. J. O. Chou, "iRNAm5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition," vol. 8, no. 25, p. 41178, 2017.

[30] B. Liu, F. Liu, L. Fang, X. Wang, and K.-C. J. B. Chou, "repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects," vol. 31, no. 8, pp. 1307-1309, 2015.

[31] X. Cheng, S.-G. Zhao, X. Xiao, and K.-C. J. B. Chou, "iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals," vol. 33, no. 3, pp. 341-346, 2017.

[32] P. Feng, H. Ding, H. Yang, W. Chen, H. Lin, and K.-C. J. M. T.-N. A. Chou, "iRNA-PseColl: identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC," vol. 7, pp. 155-163, 2017.

[33] E. S. Sankari and D. J. J. o. t. b. Manimegalai, "Predicting membrane protein types by incorporating a novel feature set into Chou's general PseAAC," vol. 455, pp. 319-328, 2018.

[34] J. Chen, H. Liu, J. Yang, and K.-C. J. A. a. Chou, "Prediction of linear B-cell epitopes using amino acid pair antigenicity scale," vol. 33, no. 3, pp. 423-428, 2007.

[35] B. Liu, F. Yang, and K.-C. J. M. T.-N. A. Chou, "2L-piRNA: a two-layer ensemble classifier for identifying piwi-interacting RNAs and their function," vol. 7, pp. 267-277, 2017.

[36] K.-C. J. P. E. Chou, "Using subsite coupling to predict signal peptides," vol. 14, no. 2, pp. 75-79, 2001.

[37] Y. Xu, J. Ding, L.-Y. Wu, and K.-C. J. P. o. Chou, "iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition," vol. 8, no. 2, p. e55844, 2013.

[38] J. Jia, Z. Liu, X. Xiao, B. Liu, and K.-C. J. J. o. t. b. Chou, "iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC," vol. 377, pp. 47-56, 2015.

[39] W. R. Qiu, B. Q. Sun, X. Xiao, D. Xu, and K. C. J. M. I. Chou, "iPhos-PseEvo: identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory," vol. 36, no. 5-6, p. 1600010, 2017.

[40] W.-R. Qiu, B.-Q. Sun, X. Xiao, Z.-C. Xu, J.-H. Jia, and K.-C. J. G. Chou, "iKcr-PseEns: Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier," vol. 110, no. 5, pp. 239-246, 2018.

[41] X. Cheng, X. Xiao, and K.-C. J. G. Chou, "pLoc-mEuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC," vol. 110, no. 1, pp. 50-58, 2018.