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

Abdullah

S2018279003

SUPERVISED BY

DR. YASER DAANIAL KHAN

SCHOOL OF SYSTEMS AND TECHNOLOGY

UNIVERSITY OF MANAGEMENT AND TECHNOLOGY

LAHORE, PAKISTAN

2019

PREDICTION OF STRESS RESPONSE PROTEINS VIA RANDOM FOREST AND NEURAL NETWORK BY INTEGRATION OF STATISTICAL MOMENTS

ABDULLAH

S2018279003

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN SOFTWARE ENGINEERING

SCHOOL OF SYSTEMS AND TECHNOLOGY
UNIVERSITY OF MANAGEMENT AND TECHNOLOGY
LAHORE, PAKISTAN

2019

بسم الله الرحمن الرحيم

UNIVERSITY OF MANAGEMENT AND TECHNOLOGY

ORIGINAL LITERARY WORK DECLARATION

Name of Student: ABDULLAH

Registration No: S2018279003

Name of Degree: MS Computer Science

Title of Dissertation/Thesis: PREDICTION OF STRESS RESPONSE PROTEINS VIA RANDOM FOREST AND NEURAL NETWORK BY INTEGRATION OF STATISTICAL MOMENTS

I, **ABDULLAH,** ID: **S2018279003** Session 2017-2019, hereby certify that this thesis is being submitted in partial fulfillment of the requirement for the MS degree in Software Engineering. This thesis is my original work and the material presented here is not being used for the acquisition of any other degree from any institute.

Student's Signature: ________________ Date: ________________

CERTIFICATE OF APPROVAL

It is certified that the research work presented in this thesis entitled **“Prediction of Stress Response Proteins via Random Forest and Neural Network by Integration of Statistical Moments”** was conducted by **Abdullah, ID: S2018279003** under the supervision of

Dr. YASER DAANIAL KHAN.

No part of this thesis has been submitted anywhere for any other degree.

This thesis is submitted to the Department of Computer Science, University of Management and Technology for the partial fulfillment of the requirement for the degree of Master of Science in Computer Science at the

Department of Software Engineering, University of Management and Technology, Lahore, Pakistan.

Dr. Yaser Daanial Khan ____________________________
Supervisor

Dr. Muzammil Hussain ____________________________
Director of Graduate Studies

ABSTRACT

Protein is a vital component of cells and 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 main and key mediators of several cellular stress responses and are categorized into two major groups based on their structure or function. One group gets activated only under cellular stress whereas the other operates in both stressed as well as normal cellular function. This research has been 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, 99.8 % valid result has been attained by using a random forest classifier, whereas by using an artificial neural network classifier 75.2% precise result was noted and the support vector machine shows 65.2% accuracy results. The authenticity of the random forest classifier was observed as 99.8%, 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%.

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

ACKNOWLEDGEMENTS

In the name of **ALLAH ALMIGHTY**, the Most Merciful and the Most Beneficent who is the entire source of all knowledge and with HIS blessings I became able to complete this research successfully.

I would like to extend my gratitude to the University Of Management and Technology for providing me such an environment and resources which enabled me to complete this research.

I would like to express my special thanks to my Supervisor, **Dr. Yaser Daanial Khan** (Chairperson, Department of Computer Science) who gave me the golden opportunity to do this wonderful thesis on the topic 'Prediction of Stress Response Protein via Random forest and neural network by Integrating of Statistical Moments, which help me in doing a lot of research. I would also like to thank you for their guidance, suggestions, and constructive comments throughout this study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

FV	:	Feature Vector
AA	:	Amino Acid
MCC	:	Mathew’s Correlation Coefficient
ACC	:	Accuracy
SP	:	Specificity
SN	:	Sensitivity
IFM	:	Input Feature Matrix
OM	:	Output Matrix
TP	:	True Positive
TN	:	True Negative
FP	:	False Positive
FN	:	False Negative
AAC	:	Amino Acid Composition
AAF	:	Amino Acid Factors
RF	:	Random Forest
NN	:	Nearest Neighbor
SVM	:	Support Vector Machine
PseAAC	:	Pseudo Amino Acid Composition
FV	:	Frequency Vector
ROC	:	Receiver Operation Characteristic Curve
PRIM	:	Position Relative Incidence Matrix
RPRIM	:	Reverse Position Relative Incidence Matrix
AAPIV	:	Accumulation Absolute Positions Incidence Vector
RAAPIV	:	Reverse Accumulation Absolute Positions Incidence Vector
AUC	:	Area under Curve

CHAPTER 01: INTRODUCTION

1. INTRODUCTION

Proteins are large bio-molecules, or macro-molecules, consist of one or more long chains of amino acid residues. 50% 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 which transports oxygen in the human body. Proteins consist of many different amino acids that are 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 be arranged in millions of different ways to create 'words' and a whole protein 'language' rather than the alphabet is. The resulting protein will fold into a particular shape, based on the number and sequence of the amino acids[1]. This form is very important because it will determine the function of the protein like muscle construction, DNA transcription RNA translation, and structural functions, etc. protein structure divided into 4 levels, Primary Level which contains amino acid residues, Secondary Level which contain alpha helices and beta sheets, Tertiary Level which contains polypeptide chain also known as the native structure of protein and Quaternary Level which contains assembled subunits as shown in Figure 1-A.

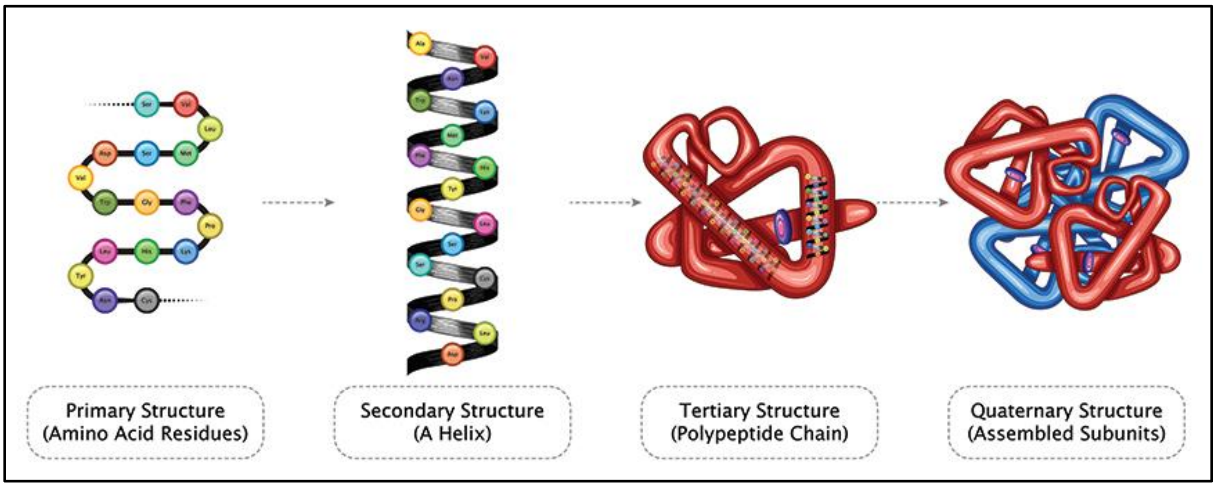


Figure 1-A: Protein Composition

The terminology of “Stress” has been first coined in 1936 to define the interaction between a force and the résistance 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 are primary mediators of several cellular stress responses and are subdivided into two categories based on their mechanism of action. One group of these stress proteins is activated only under cellular stressed conditions whereas others function in both, stressed and normal cellular functions, enhancing cell survival[3]. These proteins were found to be 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 [4] protein degradation, affecting their survival at elevated temperatures. Classical 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 below mentioned steps proposed by Kuo-Chen Chou in 2011[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.

CHAPTER 02: MATERIAL & METHODS

2. MATERIALS AND METHODS:

This Section describes our data and implements Chou's five-step rules. We collect data from UniProtKB. Take data of 7092 reviewed positive protein sequences and 7500 reviewed Negative protein sequences by searching ‘Stress Response [KW-0346]’ in keywords. The ratio of positive and negative proteins sequence is almost equal, as shown in Figure 2-A.

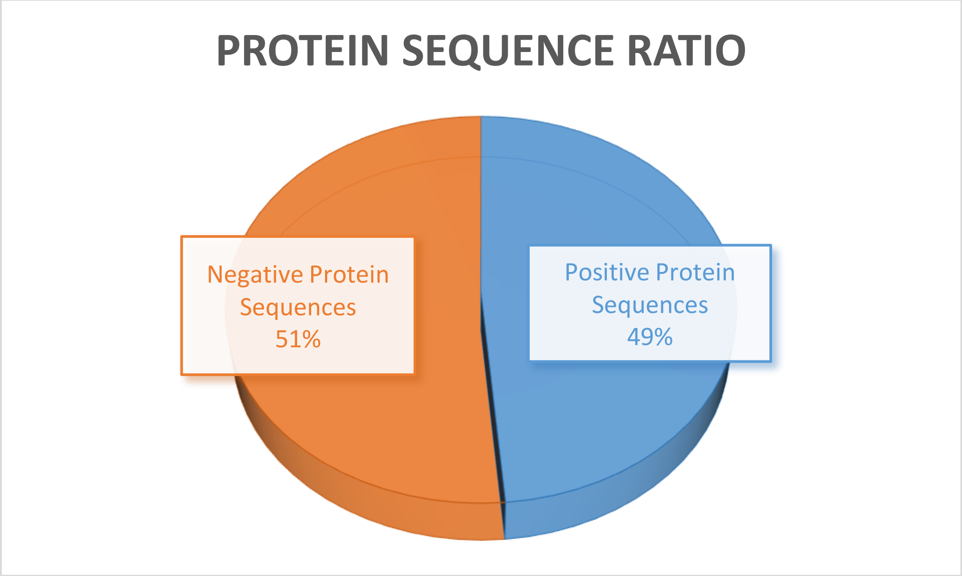


Figure 2-A: Ratio of Positive & Negative Dataset

Implementat chou’s five steps as following. The first step is the creation of the data-set collection, the second step is feature extraction from samples, the third step is to develop and train the predicted model using features selection as shown in Figure 2-B.

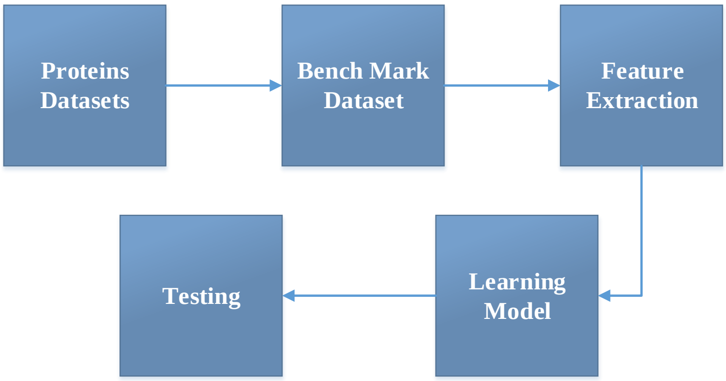


Figure 2-B: Chou's Five Steps

3. BENCHMARK DATASET:

Here we are discussing, the Chou’s 5–step, the protein arrangement of the benchmark was found by using the “chou and Shen”[10] that was exempted in 2006 by UniProtKB. After that more Stress Proteins are added in the dataset that was pulled out from the latest version 2017_03 UniProtKB. The additional extension is used to get different and useful Proto Stress Response Proteins information.

3.1. 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 or highlight fundamental to the objective broke down. This is on grounds that all the current machine-learning calculations can just deal with vector yet not sequence tests, as expounded[11]. To report this issue, the pseudo amino acid structure or “PseAAC” [12]was anticipated. As far back as the idea of “PseAAC” or “Chou's PseAAC” [13, 14] was developed, it has quickly entered into the numerous bio-medicine and medication improvement zones[15, 16] and about every one of the zones of computational proteomics, and also 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[19, 20] Specific components for Protein groups have been saved for vector making and running genetic examinations. These examinations are important as they are used in monitoring various issues in development genomics, such as developmental reviews as described in the paper [10]. According to the demand, a strong web server is launched which is called Pse-In-One[21]. It can be grasped, both protein/peptide and Protein groups can be used to make a perfect Feature vector. As shown by Eq. 1 and the possibilities of PseAAC, additional information can be used for tests in S1 [8].

$$\mathrm{L}_{\xi=7}(\mathrm{I}) = [\Psi_1\Psi_2\ldots\Psi_{\mathrm{u}}\ldots\Psi_{\Omega}]^{\mathrm{D}} \qquad \text{Eq. 1}$$

Where the parts Ψj ($j = 1, 2, \cdots, \Omega$) of the content will be considered as a method of eliminating the properties of the development of Proteins sequence, and T-transfer is considered executive.

Where $A_{20} = D$ and R_x ($x = 1, 2, \cdots, 41$; x20) 20 can be presented before close to 15 amino acids or false code. We approach the math values $[1, 2,3,4,5,6, \ldots \ldots 20]$ local amino acids, depending on the application of their sole value, and used 21 to identify the exposed amino acid x. At this time, we utilize the structure measurable moments way to deal with characterized and its measurement.

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

$$\mathrm{L} = \mathrm{A}_1\mathrm{A}_2\ldots\mathrm{A}_{17}\mathrm{A}_{18}\mathrm{A}_{19}\mathrm{A}_{20}\ldots\mathrm{A}_{40}\mathrm{A}_{41} \qquad \text{Eq. 2}$$

3.2. Statistical Moment's Calculation:

Statistical moments are used for the quantitative collection of data. Different moment's orders are 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. Different prospects are explained by statisticians and mathematicians that base on the functions of distribution and polynomials. Some moments are used such as "Raw, Central and Hahn moments."[22, 23]. The first one is for location and scale variance, it is used for mean calculation, and dataset asymmetry probability distribution. Central moments are location invariant because there is the centric calculation performed but it can be used for mean and variance calculation[24, 25]. Hahn moments are 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. Also, scale-invariant moments are 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" is used by the function as explained by

$$\mathrm{L}' = \begin{pmatrix} \mathrm{c}_{11} & \cdots & \mathrm{c}_{1\mathrm{h}} \\ \vdots & \ddots & \vdots \\ \mathrm{c}_{\mathrm{g}1} & \cdots & \mathrm{c}_{\mathrm{gh}} \end{pmatrix} \qquad \text{Eq. 3}$$

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

$$\mathrm{G}_{\mathrm{xy}} = \sum_{\mathrm{l}=1}^{\mathrm{h}} \sum_{\mathrm{n}=1}^{\mathrm{h}\Sigma} \mathrm{l}^{\mathrm{x}}\mathrm{n}^{\mathrm{y}}\beta_{\mathrm{ln}} \qquad \text{Eq. 4}$$

Where (l+n) is the moments and raw moments were as

G_{00} G_{01}, G_{10}, G_{11}, G_{12}, G_{21}, $\mathrm{G}_{30,}$ and G_{03}.

In the future, the central moments are determined as:

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

The distinction in L into a 2D square cross-segment L offers an advantage that Hahn-moments can be effectively enrolled 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 are protected throughout this time. For figuring the Hahn polynomial of demand n, the below condition is used Eq. 6.

$$N_{h}^{z,t}(j, A) = (A + T - 1)_{a} * \sum_{i=0}^{h} (-1)^{i} \frac{(-h)_{k}(-j)_{i}(2A + z - t - a - 1)_{i}}{(A + t - 1)_{i}(A - 1)_{i}} \frac{1}{I!} \qquad \text{Eq. 6}$$

In the above equation the pochhammer and the Gamma executives are portrayed by [28].

For figuring of symmetrical institutionalized Hahn of the two-dimensional discrete data, following enunciation is used.

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

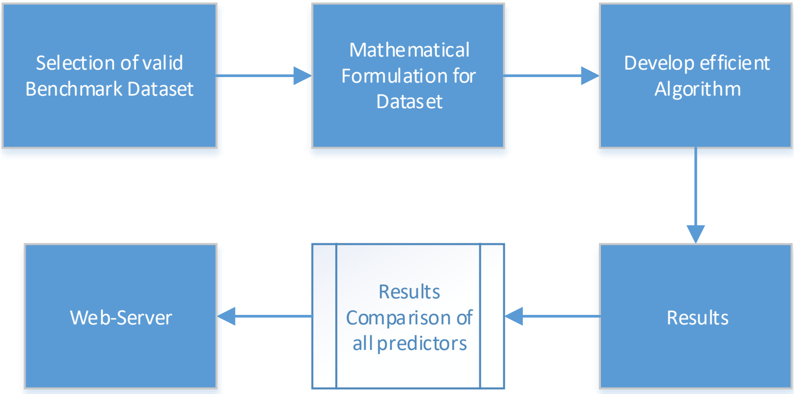

Selection of valid Benchmark Dataset
Mathematical Formulation for Dataset
Develop efficient Algorithm
Results
Results Comparison of all predictors
Web-Server

Figure 3-A: Graphical presentation of Chou's 5-step rule

3.3. Determination of PRIM (Position Relative Incidence Matrix):

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

$$M_{PRIM} = \begin{bmatrix} M_{1\to 1} & M_{1\to 2} \cdots & M_{1\to y} \cdots & M_{1\to 1} \\ M_{2\to 1} & M_{2\to 2} \cdots & M_{2\to y} \cdots & M_{2\to 20} \\ M^{\vdots}_{x\to 1} & M^{\vdots}_{x\to 2} \cdots & M^{\vdots}_{x\to y} \cdots & M^{\vdots}_{x\to 20} \\ M^{\vdots}_{A\to 1} & M^{\vdots}_{A\to 2} \cdots & M^{\vdots}_{A\to y} \cdots & M^{\vdots}_{A\to 20} \end{bmatrix} \qquad \text{Eq. 8}$$

The protein matrix of action demonstrates the remaining ith 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 are 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.

3.4. Determination of RPRIM (Reverse Position 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 been dragged out. The PRIM grid exposes or separates data regarding the general situation of amino acids. Another grid RPRIM is 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 which now look identical to polypeptide arrangements.

$$M_{RPRIM} = \begin{bmatrix} M_{1\to 1} & M_{1\to 2} \cdots & M_{1\to y} \cdots & M_{1\to 20} \\ M_{2\to 1} & M_{2\to 2} \cdots & M_{2\to y} \cdots & M_{2\to 20} \\ M^{\vdots}_{x\to 1} & M^{\vdots}_{x\to 2} \cdots & M^{\vdots}_{x\to y} \cdots & M^{\vdots}_{i\to 20} \\ M^{\vdots}_{A\to 1} & M^{\vdots}_{A\to 2} \cdots & M^{\vdots}_{A\to y} \cdots & M^{\vdots}_{A\to 20} \end{bmatrix} \qquad \text{Eq. 9}$$

3.5. Frequency matrix determination:

Every protein structure contains an instant rate matrix (Q) that generates the structure of datasets that were 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 is considered as:

$$\xi = \{\tau_{1,}\tau_{2,\ldots,}\, \tau_{20}\} \qquad \text{Eq. 10}$$

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

3.6. AAPIV Generation:

The FM is only used for extracting the information of amino acids, it does not determine the relative position of residues. AAPIV (Accumulative Absolute Position Incidence Vector) is used for the length of 20 elements. To 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 is always arranged.

AAPIV path is represented as:

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

So, an illogical ith element of AAPIV is calculated as:

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

3.7. RAAPIV (Reverse Accumulative Absolute Position Incidence Vector):

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

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

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

CHAPTER 03: PROPOSED METHODOLOGY

4. PROPOSED METHODOLOGY:

In this study, we concentrated on a specific type of protein and try to predict on the base of the protein sequence. For this prediction, we follow these steps, shown in Figure 4-A. We suggested three different algorithms for the prediction of stress response proteins, and after comparing the results of all these algorithms, we find which algorithm is best in accuracy.

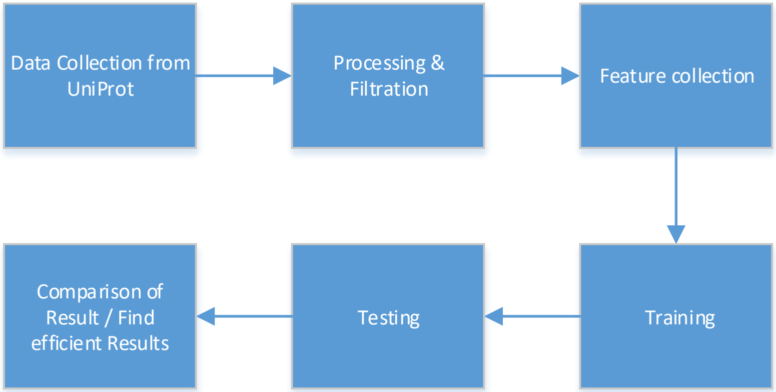

Data Collection from UniProt
Processing & Filtration
Feature collection
Training
Testing
Comparison of Result / Find efficient Results

Figure 4-A: Flowchart of Proposed Methodology

4.1. Prediction Algorithms:

In this study, three different prediction algorithms are proposed. After training and test of all proposed algorithms, we suggest the most efficient prediction algorithm. The proposed algorithms are introduced below.

4.1.1. Random Forest:

Random Forest (RF) is a powerful machine learning classifier that is used for the classification, prediction, and accuracy of the model. It is operated by constructing a multitude of decision trees at training time and outputting the class which is the mode of the classes. The advantages of the Random Forest Classifier are the 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) is designed. FIM was designed by using all the vectors, each line agrees on a single model. A "Frequency output matrix" is a control method, which is followed by a continuous component class ("positive or negative") in the FIM. Mathematical networks were used for both (FIM and EOM) to do the right work.

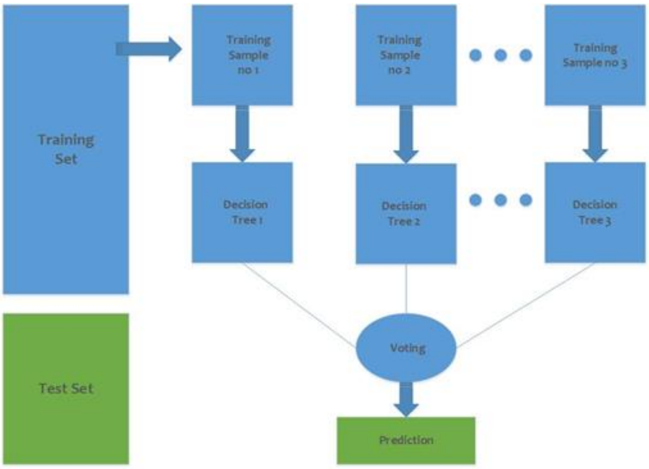


Figure 4-B: Architecture of Random Forest Classifier for the proposed prediction model

A random forest classifier is used with these parameters for the implementation as shown in Table 4.1.1-A.

Table 4.1.1-A: 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

4.1.2. Artificial Neural Network:

It is known as a neuronal construction in which the yield of the preceding neuron is an input of the following neuron. In beginning unit, transformation is made to the consequence of the addition of all previous weighted inputs to the values as shown in Figure 4-C.

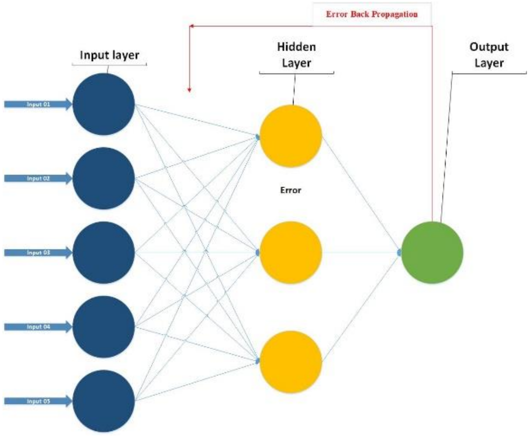


Figure 4-C: Architecture of ANN Classifier for the proposed prediction model

The earlier formed benchmark dataset had both positive samples as well as negative samples. A computed function vector of all the composed samples. For the 2D symbol of the primary protein substructure, PRIM, and RPRIM, all the feature vectors contain 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) is formed. An likely Output matrix built-in a supervised way, which followed to class (negative or positive) of the consistent element in FIM. Both these media (EOM and FIM) were intended to be used 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" was 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*)** is represented by $\boldsymbol{\theta} \in \mathbf{A}^s$. The gradient function is represented as $\nabla\boldsymbol{\theta A}$ **(*θ*)** and **A** shows the learning factor rate. The function depends on the learning rate which 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 are 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 faults 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 as:

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

Where A_g is the learning rate used for the gth period.

4.1.3. 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 normally used in a classification problem[33]. SVMs are based on the idea of finding a hyperactive plane that greatest divides a dataset into two classes.

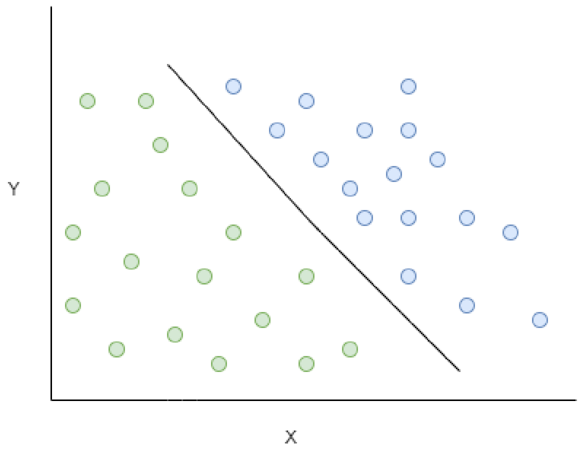


Figure 4-D: Architecture of SVM Classifier for the proposed prediction model

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

CHAPTER 04: RESULTS

5. RESULTS:

This section focuses on the remaining two steps of Chou’s 5-step rule, which are the validation of the proposed model and the webserver development.

5.1. Accuracy Estimation:

To 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 be considered. Thus, we define a set of metrics for the estimation of accuracy.

5.2. Metrics for 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 model) and, (3) Sn (sensitivity of model) [34]. Transformation metrics use the mathematical equations given by [35] are commonly used in the literature to measure the quality of predictive methods. But 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 very important indicator that reflects the stability of the prediction method. Fortunately, four intuitive metric sets were derived based on the Chou symbol introduced to study protein signal peptides[36], and “Xu et al”.[37] described it in four steps,

Eq. 16 Metrics formulation:

$$\begin{gathered}
Sn = 1 - \frac{A_-^+}{A_+^-} \quad 0 \le Sn \le 1 \\
Sp = 1 - \frac{A_+^-}{A^-} \quad 0 \le Sp \le 1 \\
Acc = 1 - \frac{A_-^+ + A_+^-}{A^- + A^-} \quad 0 \le Acc \le 1 \\
MCC = \frac{1-(A_-^+ / A^+ + A_+^- / A^-)}{[\sqrt{[1+(A_+^- - A_-^+)/A^+]*[1+(A_-^+ - A_+^-)/A^-]}}
\end{gathered} \tag*{Eq. 16}$$

Where $\mathbb{N}^+$ represents the total number of Stress responses which were predicted truly as Stress response and $\mathbb{N}_-^+$ represents the total number of Stress responses which were predicted falsely as the non- Stress response. In addition, $\mathbb{N}^-$ is the total number of non-stress responses that were actually predicted as non-proto-oncogenes, and $\mathbb{N}_+^-$ is the total number of non-stress responses that were incorrectly predicted as proto-oncogenes. This equation explains the accuracy, sensitivity, specificity, and stability in terms of MCC, and is reported 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
- 10-Fold Cross-validation testing

5.3. Training Accuracy results through Self-Consistency Testing:

A self-consistency test was performed, and the same benchmark dataset was used for training and testing the proposed predictor. This validation method is used when the actual positive value of the benchmark dataset is known. The self-consistency results of random forest, Support vector machine, and Artificial Neural network are shown in Table 5.3-A. 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 is shown in Figure 5-A, Figure 5-B, and Figure 5-C respectively. In addition comparison, the ROC graph of all these predictors is Figure 5-D.

Table 5.3-A: Self-Consistency testing results of all predictors

Predictor	Accuracy (%)	Sp (%)	Sn (%)	MCC
Random Forest	99.8	99.9	99.7	0.99
ANN	75.2	74.8	75.7	0.50
SVM	65.2	68.7	61.2	0.29

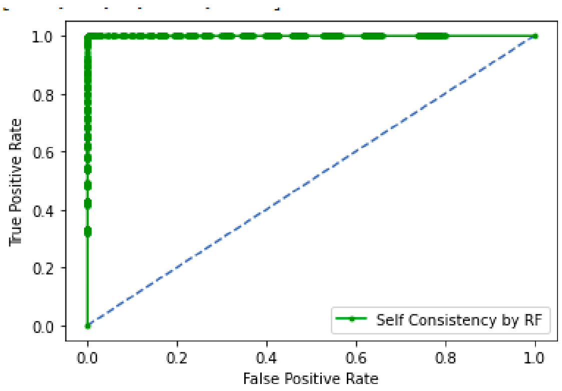


Figure 5-A: Self-Consistency ROC Graph of RF

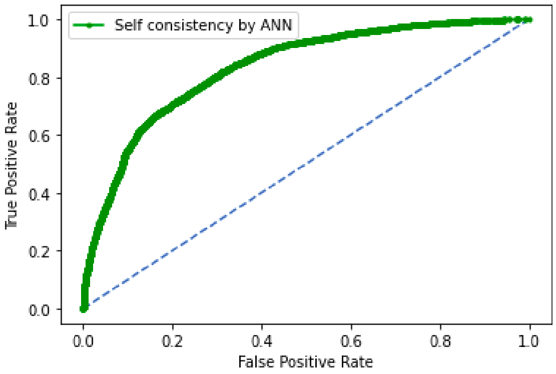


Figure 5-B: Self-Consistency ROC Graph of ANN

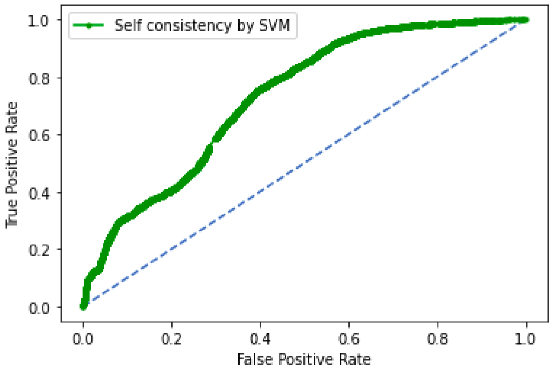


Figure 5-C: Self-Consistency ROC Graph of SVM

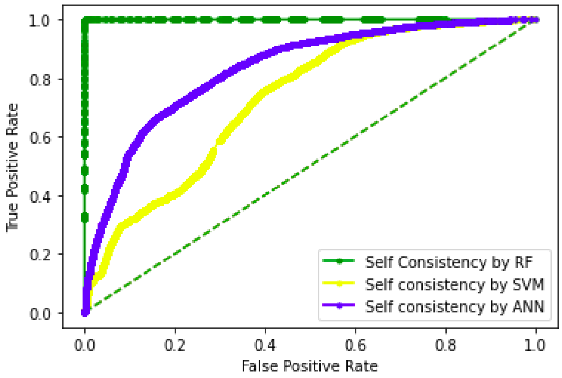


Figure 5-D: Self-Consistency Comparison ROC

5.4. Training accuracy results through Jack-Knife testing:

In Jack-knife, for the training and testing of the validation, all the 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 a folding-knife intentional problem, due to subsampling and independent testing, can be completely avoided. A jack-knife test was performed on RF, ANN, and SVM, the results are shown in Table 5.4-A. It displays the overall accuracy, specificity, sensitivity, and stability of the predictive model while the ROC curve is shown in Figure 5-E, Figure 5-F, and Figure 5-G, respectively. The comparison ROC graph is shown in Figure 5-H.

Table 5.4-A: Jack-Knife testing results of all predictors

Predictor	Accuracy (%)	Sp (%)	Sn (%)	MCC
Random Forest	99.8	99.9	99.6	0.99
ANN	75.2	74.3	76.3	0.5
SVM	65.3	68.8	61.2	0.3

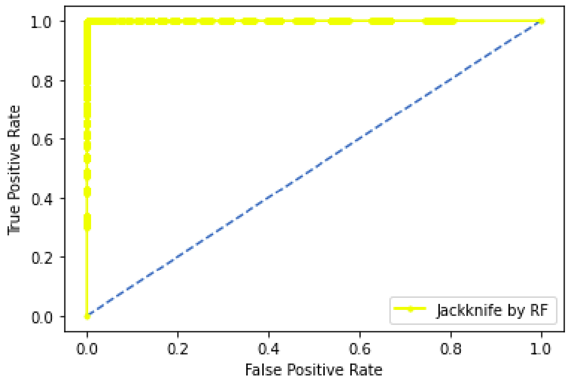


Figure 5-E: Jack-Knife ROC Graph of RF

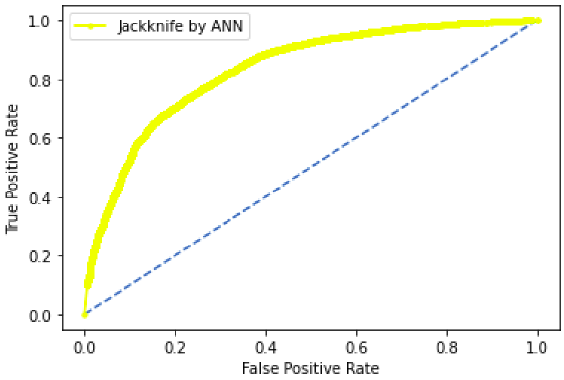


Figure 5-F: Jack-Knife ROC Graph of ANN

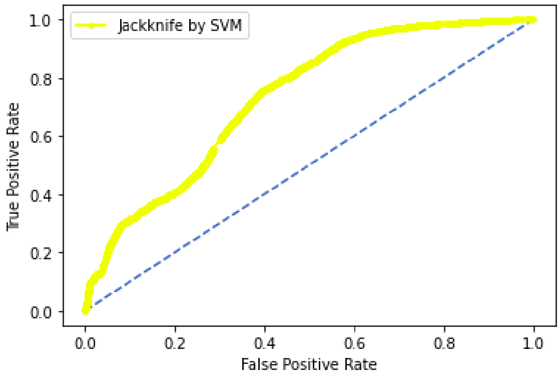


Figure 5-G: Jack-Knife ROC Graph of SVM

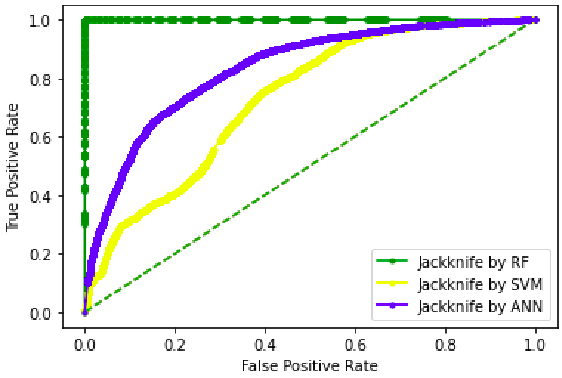


Figure 5-H: Jack-Knife Comparison ROC

5.5. Training Accuracy through Independent Dataset Testing:

Independent dataset testing was performed by conducting 70-30 splits on the original dataset. RF classifier was trained by a 70% dataset and was tested by using the remaining 30% dataset. Table 5.5-A displays the overall accuracy, specificity, sensitivity, and stability of the random forest, artificial neural network, and SVM, Figure 5-I, Figure 5-J, and Figure 5-K shows the ROC graph of testing respectively. While the comparison ROC graph of all testing is shown in Figure 5-L.

Table 5.5-A: Independent Dataset Testing results of all

Predictor	Accuracy (%)	Sp (%)	Sn (%)	MCC
Random Forest	84.5	89.5	78.7	0.68
ANN	73.5	71.6	75.6	0.47
SVM	53.57	78.4	7.6	0.19

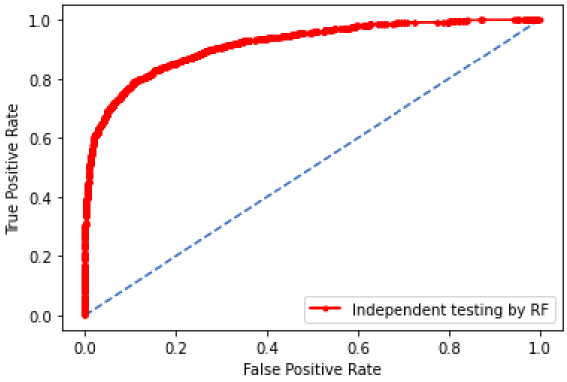


Figure 5-I: Independent Testing ROC Graph of RF

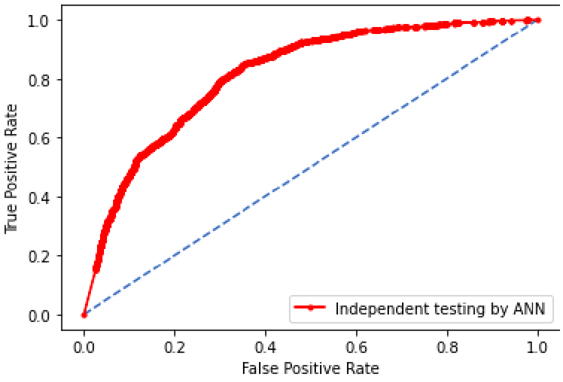


Figure 5-J: Independent Testing ROC Graph of ANN

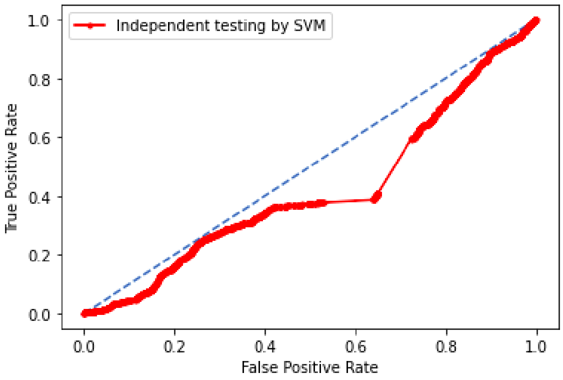


Figure 5-K: Independent Testing ROC Graph of SVM

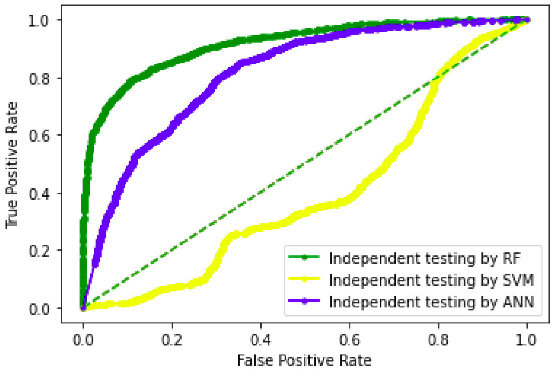


Figure 5-L: Independent Testing Comparison ROC

5.6. Validation through 10-Fold Cross-Validation:

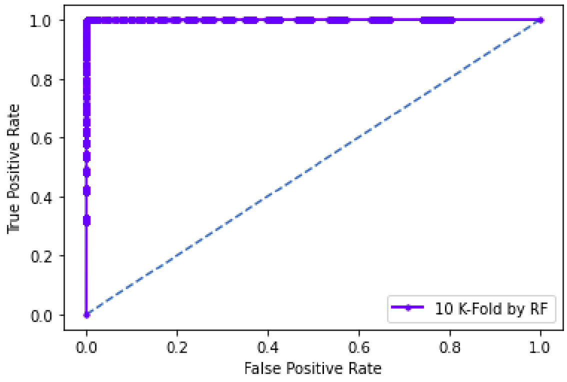
The experimentally validated datasets are used for model prediction and testing. However, no experimentally validated dataset compares the model with actual data. If the data is available, it may not be enough to test the accuracy of the predictive model. Which tests need to be performed to evaluate 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 is 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 parts of the dataset are used for both testing and training. The result obtained is the average of all accuracies. Similar methods were 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 are used to select partition data or random data for testing.

Table 5.6-A shows the results obtained with 10-fold cross-validation via the Random Forest, ANN, and SVM, while Figure 5-M, Figure 5-N, and Figure 5-O shows the ROC curve of all predictors respectively. Figure 5-P shows the comparison ROC curve of all.

Table 5.6-A: 10-Fold Cross-Validation Result of all

Predictor	Accuracy (%)	Sp (%)	Sn (%)	MCC
Random Forest	99.8	99.9	99.6	0.99
ANN	61.8	72.4	49.5	0.22
SVM	65.3	68.8	61.2	0.3



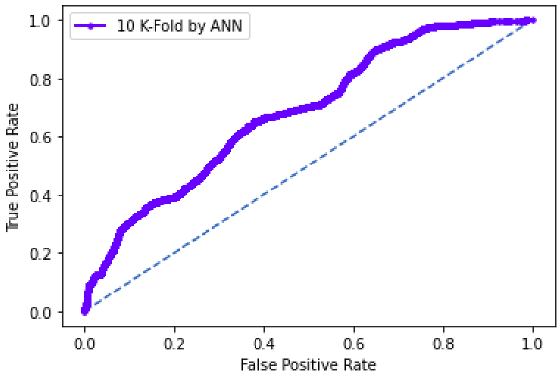
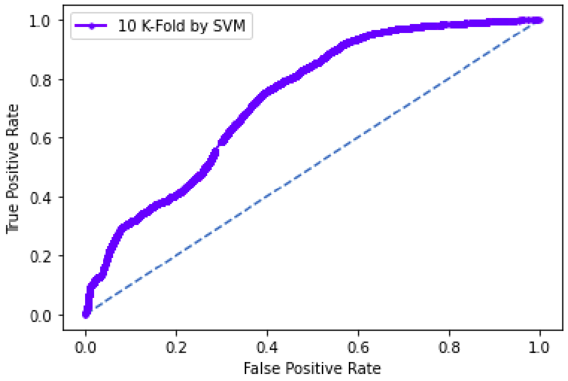


Figure 5-M: 10-Fold ROC Graph of RF

Figure 5-N: 10-Fold ROC Graph of ANN



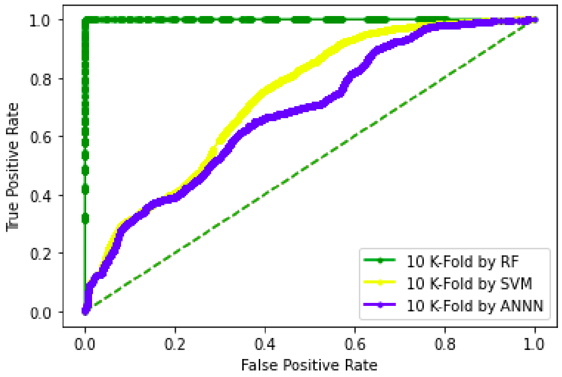


Figure 5-O: 10-Fold ROC Graph of SVM

Figure 5-P: 10-Fold Comparison ROC

5.7. Validation through 5-Fold Cross-Validation:

In cross-validation, the dataset is divided into k sets, and k is chosen at the start and then it is kept constant. Usually, k is kept 5 or 10 but, in this problem, k was set to 5. The model is tested k times, and, in each iteration, 4 sets are used as a training set, and the one set (k set) is treated as a testing set.

Each set can be used as a testing set and the problems like under fitting and overfitting can easily be removed. After performing k iterations, the model accuracy is 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 obtained by ANN, and 65% of overall accuracy is obtained by SVM as shown in Table 5.7-A. While the ROC graphs are shown in Figure 5-Q, Figure 5-R, and Figure 5-S respectively and the comparison ROC graph is shown in Figure 5-T.

Table 5.7-A: 5-Fold Cross Validation Result of all

Predictor	Accuracy (%)	Sp (%)	Sn (%)	MCC
Random Forest	99.8	99.9	99.7	0.99
ANN	61.8	72.4	49.5	0.22
SVM	65.3	68.8	61.2	0.3

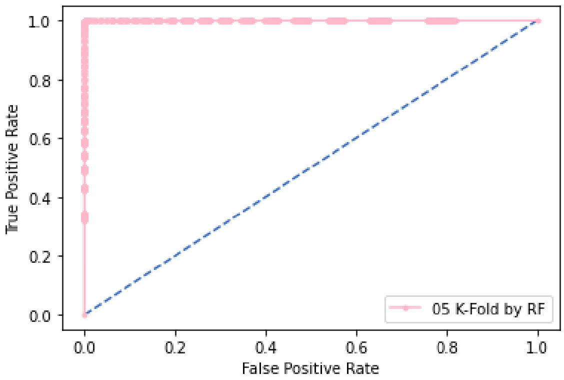


Figure 5-Q: 5-Fold ROC Graph of RF

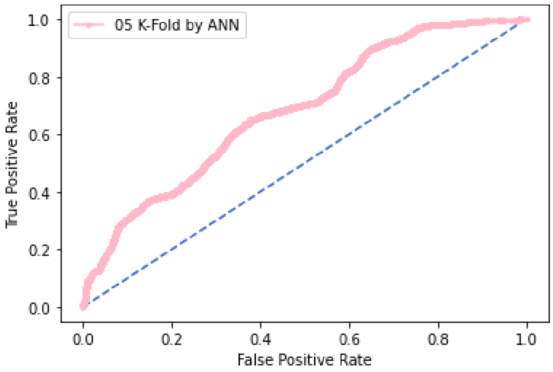


Figure 5-R: 5-Fold ROC Graph of ANN

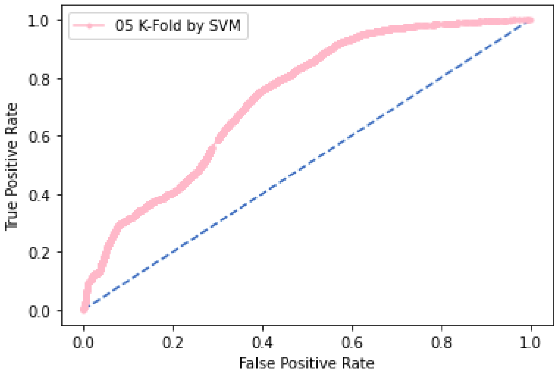


Figure 5-S: 5-Fold ROC Graph of SVM

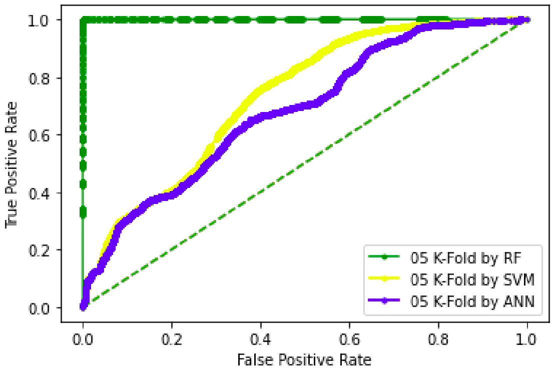


Figure 5-T: 5-Fold Cross-Validation Comparison ROC

CHAPTER 05: COMPARISON

6. COMPARISON:

In this section of the study, we compare the testing result of every predicted algorithm. For every algorithm, we perform five different tests name as Self-Consistency test, Jack-Knife test, 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.

6.1. Comparison of testing on RF:

It is observed that the Random forest algorithm gives the most accurate and efficient results in the prediction of stress response proteins. All testing results are shown in Table 6.1-A, and Figure 6-A shows the ROC graph of these testing results.

Table 6.1-A: All Testing Result Comparison of Random Forest

Random Forest				
Testing Method	**Accuracy (%)**	**Sp (%)**	**Sn (%)**	**MCC**
Self-Consistency Testing	99.8	99.9	99.7	0.99
Jack-Knife Testing	99.7	99.9	99.6	0.99
Independent Testing	84.5	89.5	78.7	0.68
10-Fold Testing	99.8	99.9	99.6	0.99
5-Fold Testing	99.8	99.9	99.7	0.99

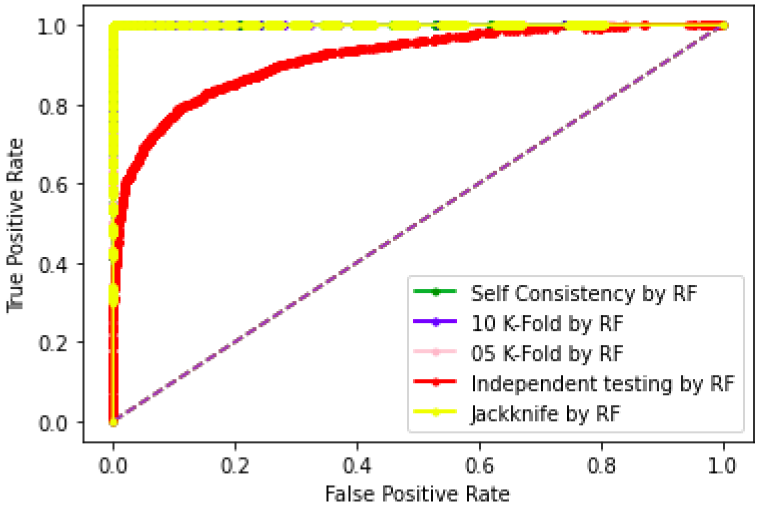


Figure 6-A: Comparison Testing ROC of RF

6.2. Comparison of testing on ANN:

Artificial Neural Network algorithm gives the 2nd best accurate and efficient results in the prediction of stress response proteins. All testing results of ANN are shown in Table 6.2-A, and Figure 6-B shows the ROC graph of these testing results.

Table 6.2-A: All Testing Result Comparison of ANN

Artificial Neural Network				
Testing Method	**Accuracy (%)**	**Sp (%)**	**Sn (%)**	**MCC**
Self-Consistency Testing	75.2	74.8	75.7	0.50
Jack-Knife Testing	75.2	74.3	76.3	0.5
Independent Testing	73.5	71.6	75.6	0.47
10-Fold Testing	61.8	72.4	49.5	0.22
5-Fold Testing	61.8	72.4	49.5	0.22

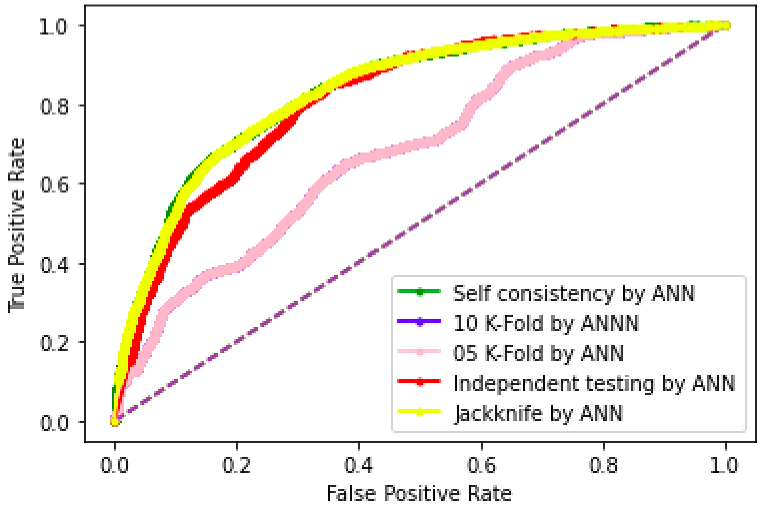


Figure 6-B: Comparison Testing ROC of ANN

6.3. 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 6.3-A, and Figure 6-C shows the ROC graph of these testing results.

Table 6.3-A: All Testing Result Comparison of SVM

Support Vector Machine				
Testing Method	**Accuracy (%)**	**Sp (%)**	**Sn (%)**	**MCC**
Self-Consistency Testing	65.2	68.7	61.2	0.29
Jack-Knife Testing	65.3	68.8	61.2	0.3
Independent Testing	53.57	78.4	7.6	0.19
10-Fold Testing	65.3	68.8	61.2	0.3
5-Fold Testing	65.3	68.8	61.2	0.3

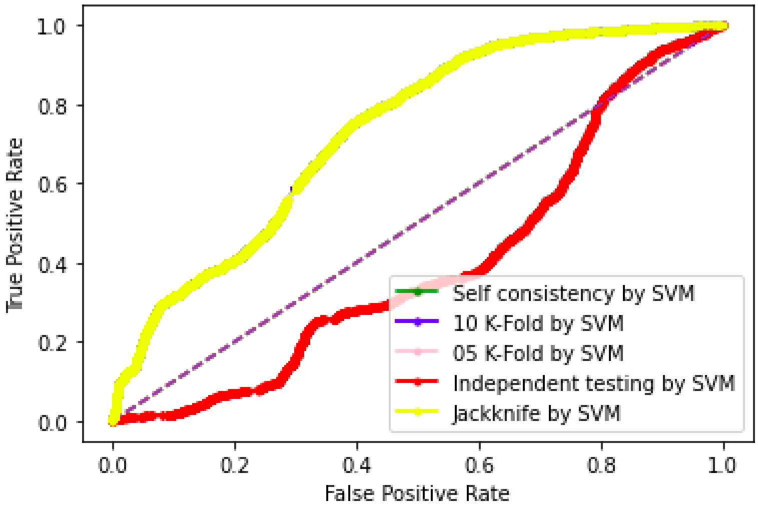


Figure 6-C: Comparison Testing ROC of SVM

6.4. BoxPlot Comparison 10-Fold Cross Validation:

In this section, we discuss boxplot comparison of 10-Fold Cross Validation of RF, ANN, and SVM. Figure 6-F shows boxplot graph of 10-Fold Cross validation by random forest, Figure 6-D show boxplot of 10-Fold cross validation by artificial neural network, and Figure 6-E shows boxplot of 10-Fold cross validation by support vector machine.

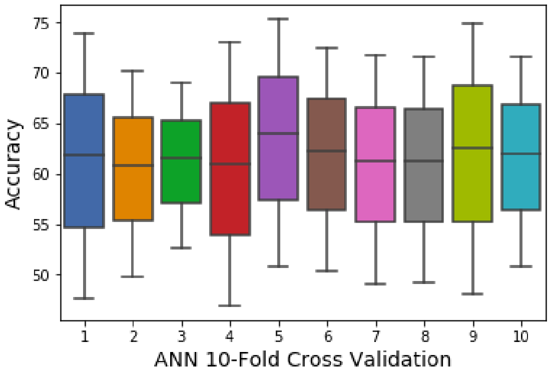


Figure 6-D: ANN 10-Fold Cross Validation

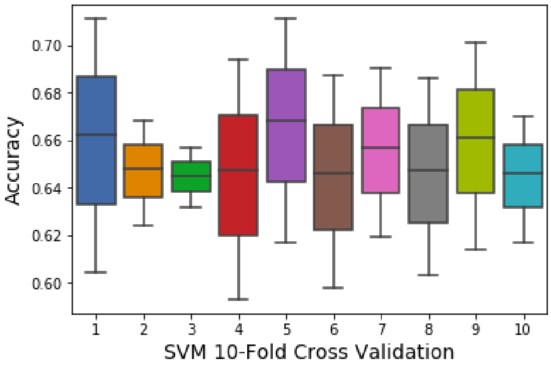


Figure 6-E: SVM 10-Fold Cross Validation

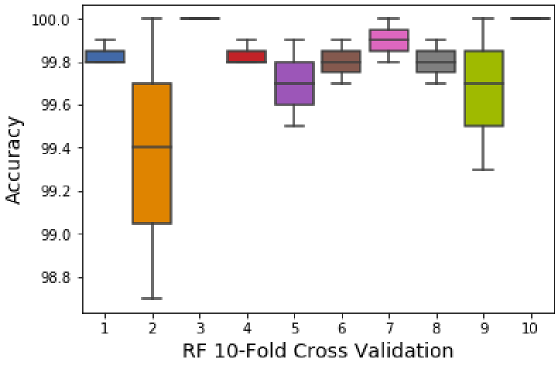


Figure 6-F: RF 10-Fold Cross Validation

6.5. BoxPlot Comparison 5-Fold Cross Validation:

This section discuss about boxplot comparison of 5-Fold cross validation of RF, ANN, and SVM. Figure 6-I shows boxplot graph of 5-Fold Cross validation by random forest, Figure 6-G show boxplot of 5-Fold cross validation by artificial neural network, and Figure 6-H shows boxplot of 5-Fold cross validation by support vector machine.

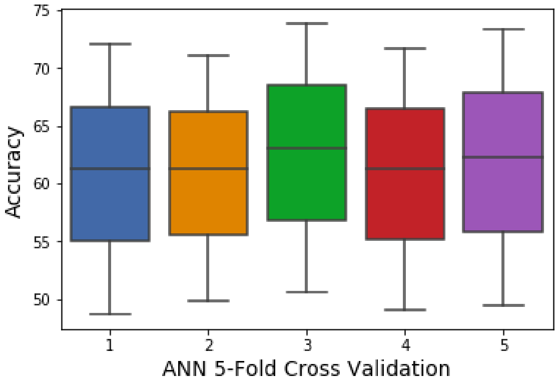


Figure 6-G: ANN 5-Fold Cross Validation

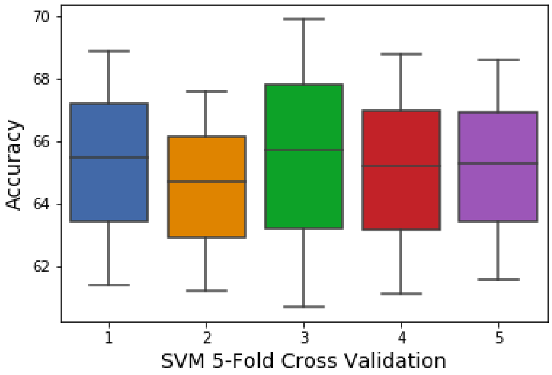


Figure 6-H: SVM 5-Fold Cross Validation

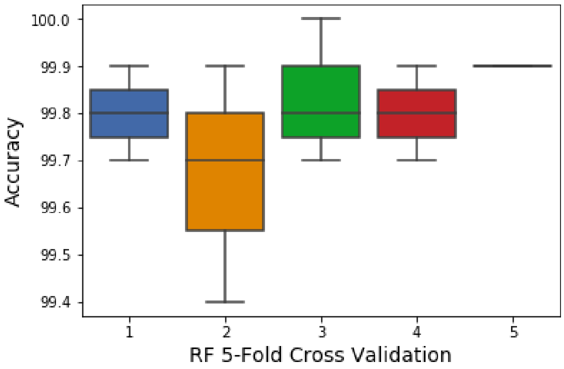


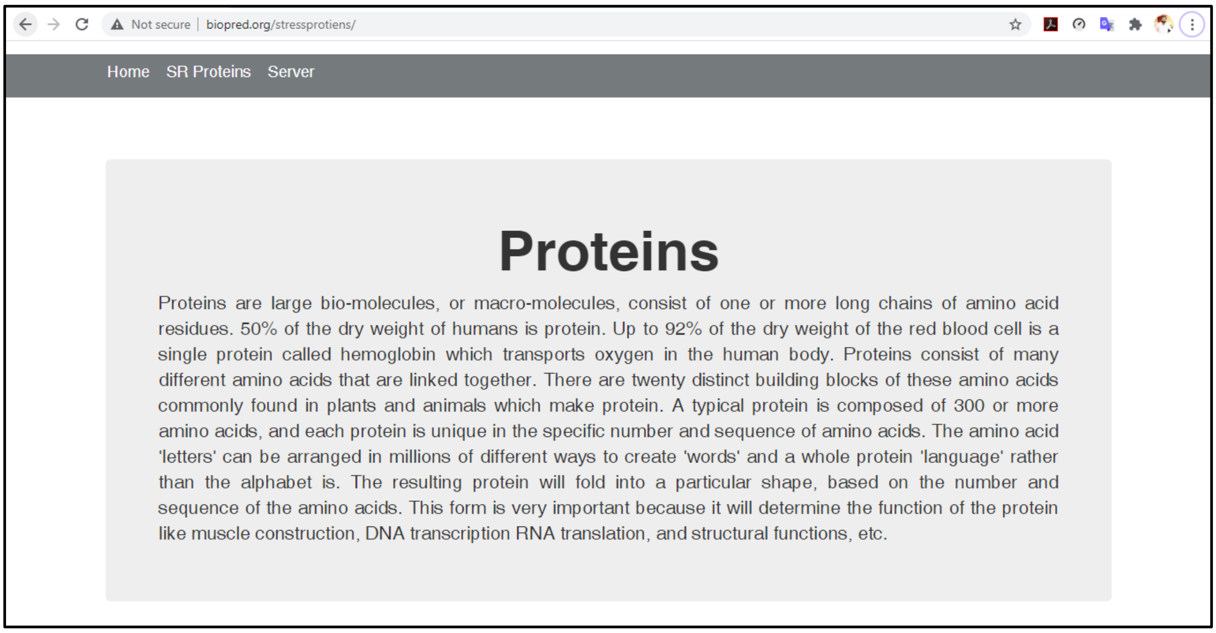
Figure 6-I: RF 5-Fold Cross Validation

CHAPTER 06: WEB-SERVER

7. WEB-SERVER:

The final phase of the Chou's 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 is 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 are shown below. Figure 7-A shows the Home page, Figure 7-B shows the introduction page, Figure 7-C shows the Prediction Server page, and Figure 7-D shows the results page. Live webserver is available at http://biopred.org/stressprotiens.

7.1. Home Page:

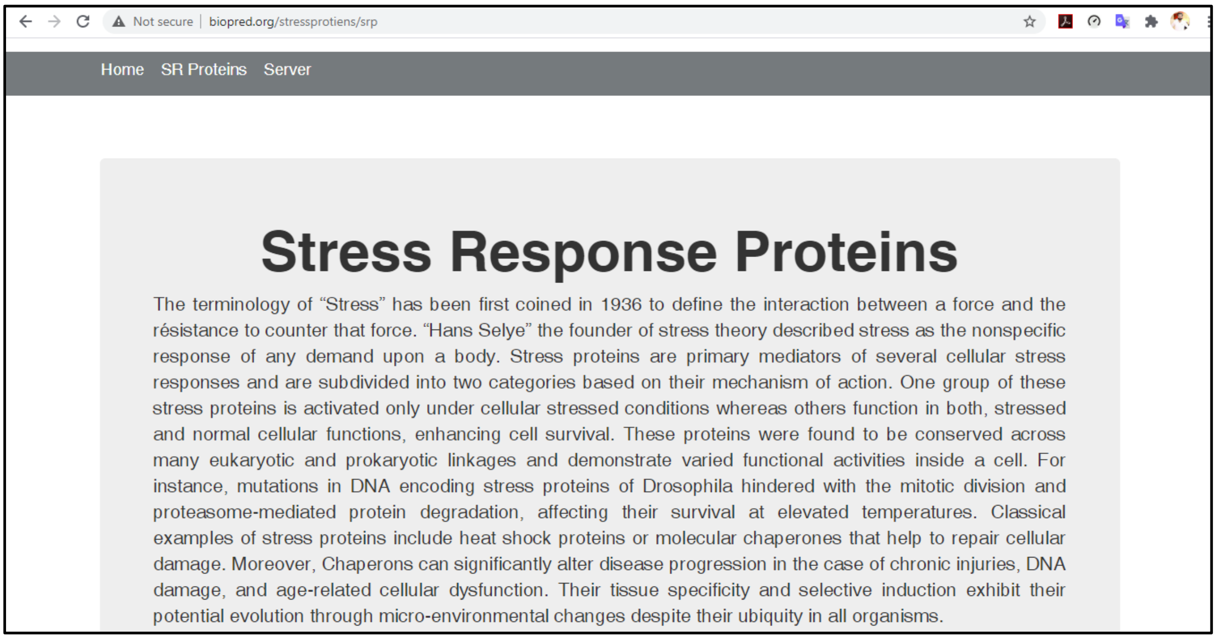


Proteins

Proteins are large bio-molecules, or macro-molecules, consist of one or more long chains of amino acid residues. 50% 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 which transports oxygen in the human body. Proteins consist of many different amino acids that are 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 be arranged in millions of different ways to create 'words' and a whole protein 'language' rather than the alphabet is. The resulting protein will fold into a particular shape, based on the number and sequence of the amino acids. This form is very important because it will determine the function of the protein like muscle construction, DNA transcription RNA translation, and structural functions, etc.

Figure 7-A: Home Page of Web-Server

7.2. Introduction Page:

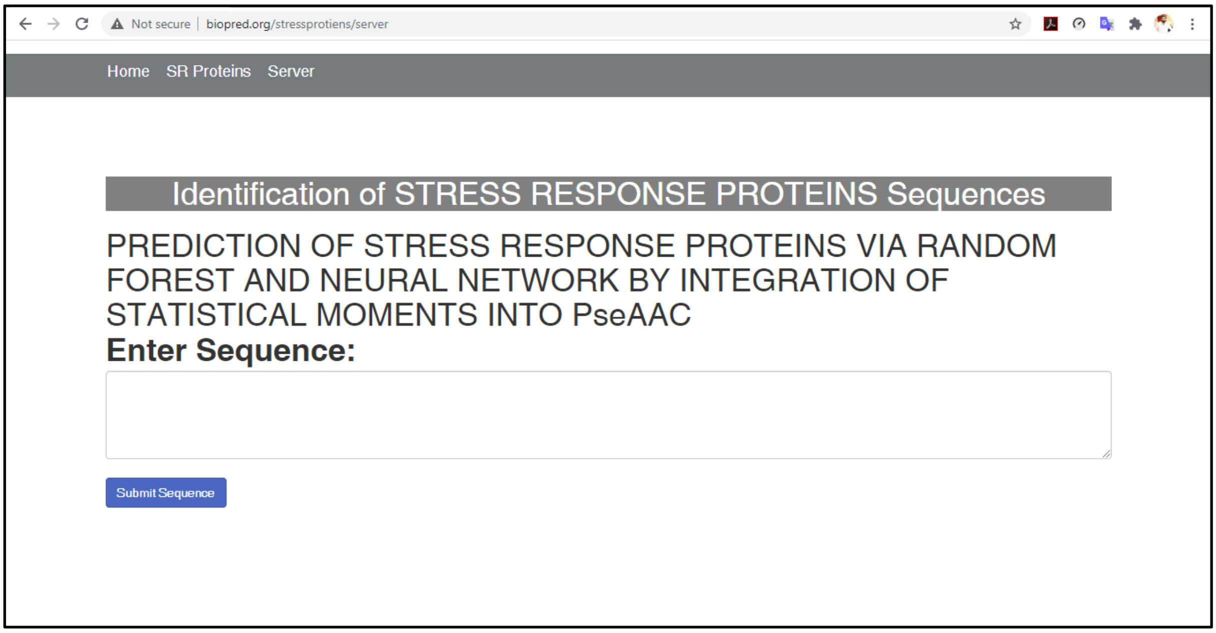


Stress Response Proteins

The terminology of “Stress” has been first coined in 1936 to define the interaction between a force and the résistance to counter that force. “Hans Selye” the founder of stress theory described stress as the nonspecific response of any demand upon a body. Stress proteins are primary mediators of several cellular stress responses and are subdivided into two categories based on their mechanism of action. One group of these stress proteins is activated only under cellular stressed conditions whereas others function in both, stressed and normal cellular functions, enhancing cell survival. These proteins were found to be 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, affecting their survival at elevated temperatures. Classical examples of stress proteins include heat shock proteins or molecular chaperones that help to repair cellular damage. Moreover, Chaperons can significantly alter disease progression in the case of chronic injuries, DNA damage, and age-related cellular dysfunction. Their tissue specificity and selective induction exhibit their potential evolution through micro-environmental changes despite their ubiquity in all organisms.

Figure 7-B: Introduction Page of Web-Server

7.3. Prediction Server Page:



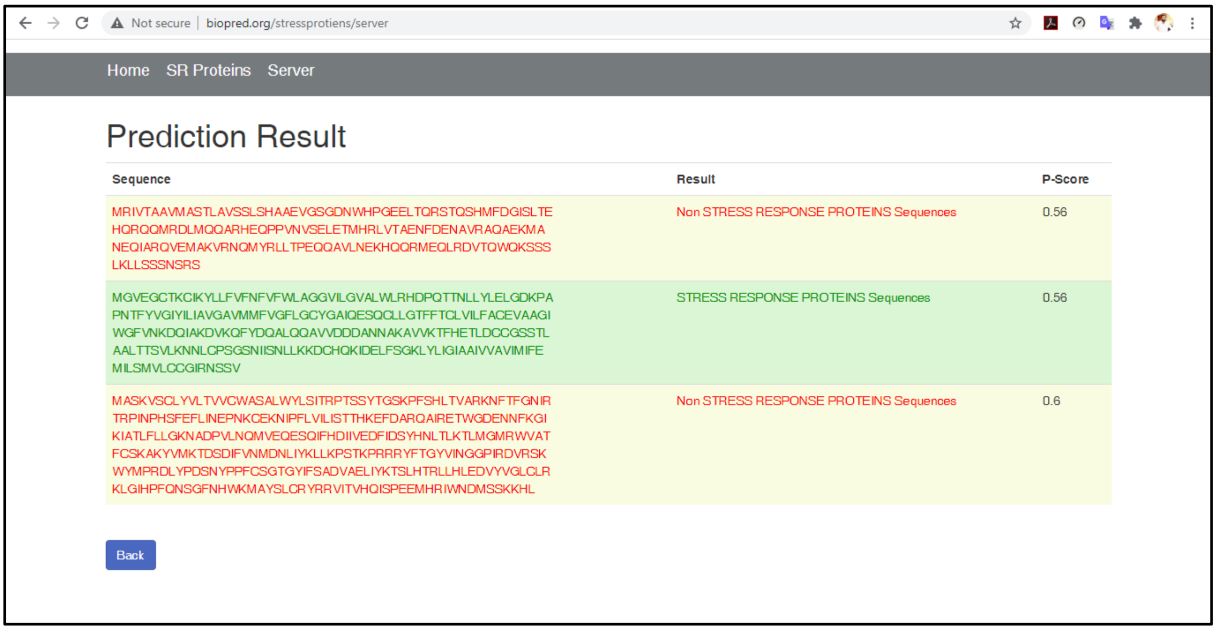
Identification of STRESS RESPONSE PROTEINS Sequences

PREDICTION OF STRESS RESPONSE PROTEINS VIA RANDOM FOREST AND NEURAL NETWORK BY INTEGRATION OF STATISTICAL MOMENTS INTO PseAAC

Enter Sequence:

Figure 7-C: Prediction Server Page

7.4. Results Page:



Prediction Result

Sequence	Result	P-Score
MRIVTAAVMASTLAVSSLSHAAEVGSGDNWHPGEELTQRSTQSHMFDGISLTEHQRQQMRDLMQQARHEQPPVNVSELETMHRLVTAENFDENAVRAQAEKMANEQIARQVEMAKVRNQMYRLLTPEQQAVLNEKHQQRMEQLRDVTQWQKSSSLKLLSSSNSRS	Non STRESS RESPONSE PROTEINS Sequences	0.56
MGVEGCTKCIKYLLFVFNFVFWLAGGVILGVALWLRHDPQTTNLLYLELGDKPAPNTFYVGIYILIAVGAVMMFVGFLGCYGAIQESQCLLGTFFTCLVILFACEVAAGIWGFVNKDQIAKDVKQFYDQALQQAVVDDDANNAKAVVKTFHETLDCCGSSTLAALTTSVLKNNLCPSGSNIISNLLKKDCHQKIDELFSGKLYLIGIAAIVVAVIMIFEMILSMVLCCGIRNSSV	STRESS RESPONSE PROTEINS Sequences	0.56
MASKVSCLYVLTVVCWASALWYLSITRPTSSYTGSKPFSHLTVARKNFTFGNIRTRPINPHSFEFLINEPNKCEKNIPFLVILISTTHKEFDARQAIRETWGDENNFKGIKIATLFLLGKNADPVLNQMVEQESQIFHDIIVEDFIDSYHNLTLKTLMGMRWVATFCSKAKYVMKTDSDIFVNMDNLIYKLLKPSTKPRRRYFTGYVINGGPIRDVRSKWYMPRDLYPDSNYPPFCSGTGYIFSADVAELIYKTSLHTRLLHLEDVYVGLCLRKLGIHPFQNSGFNHWKMAYSLCRYRRVITVHQISPEEMHRIWNDMSSKKHL	Non STRESS RESPONSE PROTEINS Sequences	0.6

Figure 7-D: Results Page of Web-Server

CHAPTER 07: DISCUSSION & CONCLUSION

8. DISCUSSION:

It can be in section 3 that the feature vector is 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. A highest accuracy achieved by random forest among three classifier as shown in Table 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 99.8 % valid result has been attained by using K-Fold Cross Validation, whereas by using Jack-knife testing 99.8% precise result was noted. The authenticity of the random forest classifier was observed as 99.8%, sensitivity value was measured as 99.6% while the specificity was calculated as 99.5% all in-inclusive and the MCC value was measured as 0.9873%. The ANN has achieved 61.8% by k fold validation and SVM has 53.5% of Accuracy by K-fold validation. Random forest gave best results as compare to ANN and SVM. It is anticipated that Stress response Predictor will become a very useful high throughput tool for studying Stress Proteins or, at the very least. It is very first predictor to predict Stress response Protein.

9. CONCLUSION:

In this study, three Classifiers were proposed to predict the Stress Response Protein Sequence named as an artificial neural network, support vector machine, and random forest classifier. It is based on Chou’s five-step rule. The highest result was achieved by a random forest classifier. The results of the random forest model showed 99.8% accuracy. This proves that the Random Forest Classifier shows better outcomes for the prediction of Stress Protein Sequences. The predictor was verified by 10-fold cross-validation and jack-knife tests, yielding accurate results of 99.7%, and 99.8%, respectively. While 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 effective predictor among predicted classifiers.

CHAPTER 08: REFERENCES

10.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.

CHAPTER 09: APPENDIX

11. APPENDIX

1.	ECK125137717	gcgagagcaacattgctgtagattgatatttaatatattagcgtaactgttatgctgttaTctatattatgtgatctaaat
2.	ECK125137718	gggcatggaaagccgggcgagagcaacattgctgtagattgatatttaatatattagcgtAactgttatgctgttatctat
3.	ECK125137719	gggaaacagactcatgttgaccttggttgtaaagagagagcaggcgttattattttcagcAtctgtcgccgcagagaaggg
4.	ECK125136673	tgcttgccagacaggggcgttattaccagttcaagcagggtttgtaagctattattgaacGatccgacttgcgtggagttt
5.	ECK120010467	gcccgggaaaaatatgctcgcgggcttgctatctcgctgacggacaggcaaattgatgacCagcttttaaaccgactccgt