[[1]](#footnote-1)

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*—Background: Protein is a vital component of cells, and it is compulsory for the structure, rule, regulations, and functions of the body’s tissues. Therefore, the identification of a protein function involves a detailed understanding of structure of protein. Stress proteins are the main and essential mediators of several responses to cellular stress and categorized based on their structure or function into two major groups. Only under cellular stress was one group activated, whereas the other operates in both stressed as well as normal cellular function.

Method: 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.

**Data collection:** Data collected from UniProtKB. 7092 reviewed positive protein sequences, and 7500 reviewed Negative protein sequences by searching ‘Stress Response.

Findings: 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.

Interpretation: 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

# INTRODUCTION

P

roteins are large bio-molecules, or macro-molecules, consist of one or more long chains of amino acid residues. Protein is a major nutrient which is essential for muscle bulk. It is a major macronutrient that provide calories or energy. Fifty percent of humans the dry weight is protein. A single protein called hemoglobin that carries oxygen in human body up to 92% of the dry weight of the red blood cell. Proteins are made up of several different amino acids that are bound together. There are twenty distinct building blocks widely found in plants and animals that make protein from these amino acids. A typical protein is composed of 300 or more amino acids, and the specific number and sequence of amino acids make each protein is unique. The amino acid 'letters' can be organized instead of the alphabet in millions of different ways to create 'words' and a whole 'language' protein. The resulting protein, depending on the number and sequence of the amino acids, will fold into a specific form[1]. This form is essential because it defines the proteins function such as muscle building, construction, DNA transcription RNA translation, and structural functions, etc. The structure of the Protein is divided into four levels. The primary level contains residues of amino acid, 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.

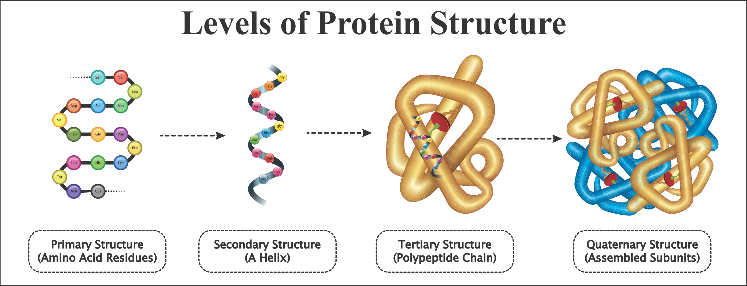


Figure ‎A‑I: Protein Composition

First coined in 1936, the terminology of “Stress” describes the relationship between a force and the resistance to fight or counter that force. Stress is a sensation of physical or emotional discomfort that can make you feel anger, irritated, disturbed, frustrated, depressed, or nervous. The founder of stress theory, “Hans Selye” 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. The heterogeneous group of disorders characterized by the gradual degeneration of the structure and function of the central nervous system (CNS) or peripheral nervous system (PNS) are neurodegenerative diseases. Stress proteins are principal mediators of multiple responses to cellular stress and sub-divided according to their mechanism of action into two categories. One category of these stress proteins is activated only under cellular stressed conditions, whereas others are activated to enhance cell survival in both stressed and normal cellular functions [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 therapy or 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.

# Materials and Methods

This section explains our data or knowledge 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

# 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](https://www.uniprot.org), by using keyword [stress response [KW-0346]](https://www.uniprot.org/uniprot/?query=keyword%3A%22Stress+response+%5BKW-0346%5D%22&sort=score). 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.

## 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. Nevertheless, its arrangement, example and highlight fundamental to the objective still been thoroughly broken 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].

|  |  |
| --- | --- |
|  | 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 ​​ 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:

|  |  |
| --- | --- |
|  | Eq. 2 |

## Statistical Moment’s Calculation

Statistical moments that are used for the qua1 explain different prospects based on the functions of distribution and polynomials. Centered on the function of distribution and mathematicians clarify various prospect. Some moments, such as “Raw, Central and Hahn moments.” [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 for the variation of size and position based on “Hahn Polynomials”. 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

|  |  |
| --- | --- |
|  | 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:

|  |  |
| --- | --- |
|  | 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:

|  |  |
| --- | --- |
|  | 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.

|  |  |
| --- | --- |
|  | 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.

|  |  |
| --- | --- |
|  | Eq. 7 |



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

## Determination of Positive Relative Incidence Matrix:

The mathematical calculations for the prediction of the protein were performed in the initial process. The pivotal model is the necessary arrangement of proteins and the relative position of the remaining particles. The total number of proteins used to assemble the “PRIM” in the form of the 20x20 matric and the result extracted. The remaining particles of the amino acid are, thus, as follows:

|  |  |
| --- | --- |
|  | 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

## Determination of Reverse Positive Relative Incidence Matrix:

The viability and rightness of AI procedure are unfathomably dependent upon cautious quality, the watchfulness where the most related pieces of data have hauled out. The PRIM network uncovered or isolates information with respect to the overall circumstance of amino acids. Another lattice RPRIM formed so that it works unclear strides from PRIM, notwithstanding, on the switch fundamental game plan. Introduction of RPRIM uncovers furthermore disguised models and diminishes ambiguities among proteins that currently seem to be indistinguishable from polypeptide game plans.

|  |  |
| --- | --- |
|  | Eq. 9 |

## Determination of Frequency Vector:

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

|  |  |
| --- | --- |
|  | 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).

## Accumulative Absolute Position Incidence Vector Generation:

FM is only used for extraction 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:

|  |  |
| --- | --- |
|  | Eq. 11 |

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

|  |  |
| --- | --- |
|  | Eq. 12 |

## 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:

|  |  |
| --- | --- |
|  | Eq. 13 |

# 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

## 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.

## 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 |

## 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.

|  |  |
| --- | --- |
|  | 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. *θx* and θx+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 + θx+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

|  |  |
| --- | --- |
|  | Eq. 15 |

## 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.

# Results

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

## 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.

## 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:

|  |  |
| --- | --- |
|  | Eq. 16 |

Where represents the total number of Stress responses which were predicted indeed as Stress response and represents the total number of Stress responses which were predicted falsely as the non- Stress response. Also, is the total number of non-stress responses that actually predicted as non-proto-oncogenes, and 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

## 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 |

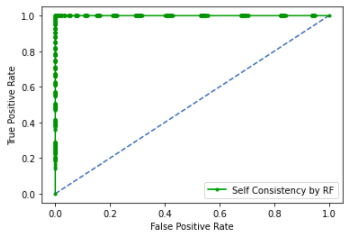


Figure ‎C‑I: Self-Consistency ROC Graph of RF

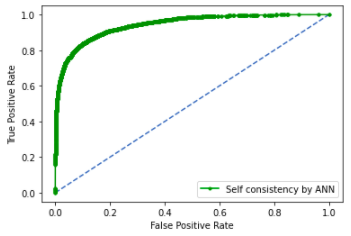


Figure ‎C‑II: Self-Consistency ROC Graph of ANN

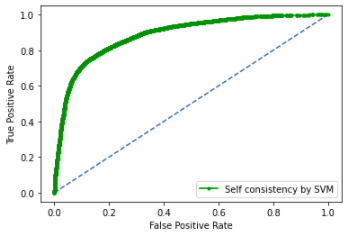


Figure ‎C‑III: Self-Consistency ROC Graph of SVM

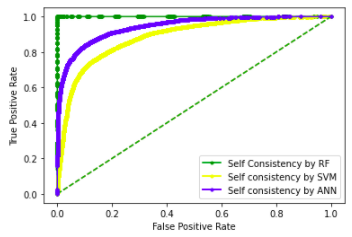


Figure ‎C‑IV: Self-Consistency Comparison ROC

## 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 |

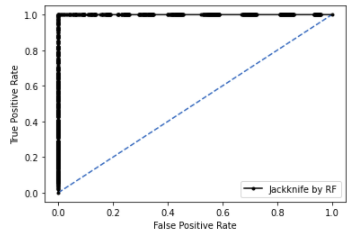


Figure ‎D‑I: Jack-Knife ROC Graph of RF

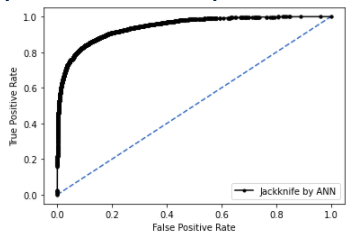


Figure ‎D‑II: Jack-Knife ROC Graph of ANN

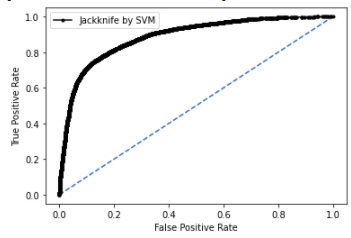


Figure ‎D‑III: Jack-Knife ROC Graph of SVM

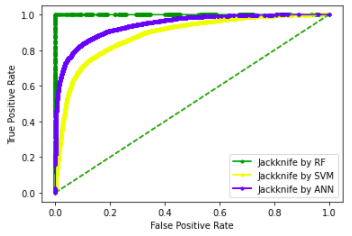


Figure ‎D‑IV: Jack-Knife Comparison ROC

## 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 |

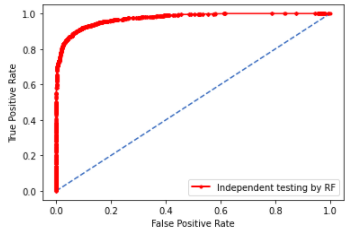


Figure ‎E‑I: Independent Testing ROC Graph of RF

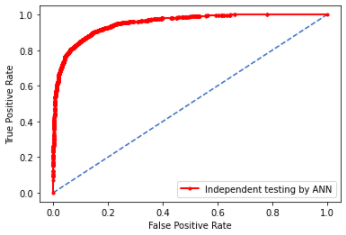


Figure ‎E‑II: Independent Testing ROC Graph of ANN

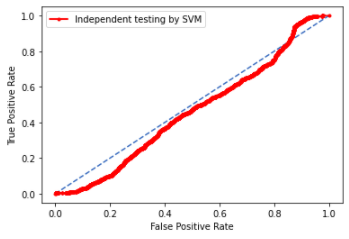


Figure ‎E‑III: Independent Testing ROC Graph of SVM

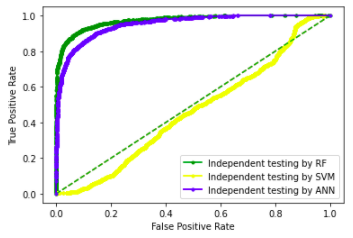


Figure ‎E‑IV: Independent Testing Comparison ROC

## 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 |

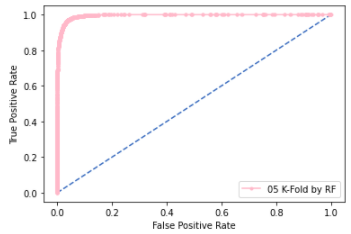


Figure ‎F‑I: 5-Fold ROC Graph of RF

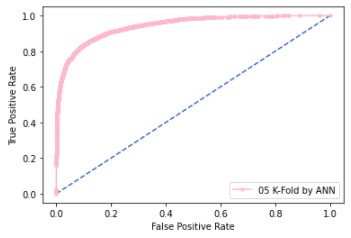


Figure ‎F‑II: 5-Fold ROC Graph of ANN

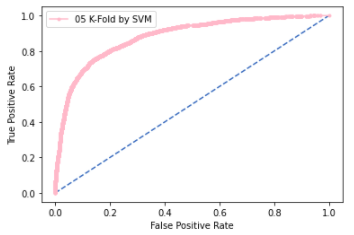


Figure ‎F‑III: 5-Fold ROC Graph of SVM

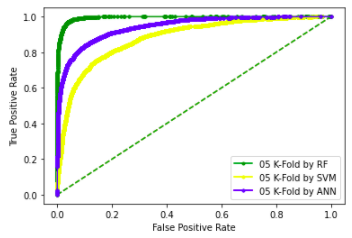


Figure ‎F‑IV: 5-Fold Cross-Validation Comparison ROC

## Validation through 10-Fold Cross-Validation:

The tentatively approved datasets are utilized for model forecast and testing. Be that as it may, no tentatively approved dataset contrasts the model and genuine information. In the event that the information is accessible, it may not be sufficient to test the exactness of the prescient model. These tests need to perform to assess the four measurements in Eq. 16 to guarantee adequate exactness and dependability. All in all, you can test through k-crease (sub-check), folding blade, and autonomous test. The folding blade model test is thorough and may give various outcomes for a specific benchmark record. Cross-approval is the most ideal choice to affirm the best possible working of the created model when there is no unmistakable record to approve the model's forecast.

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 |

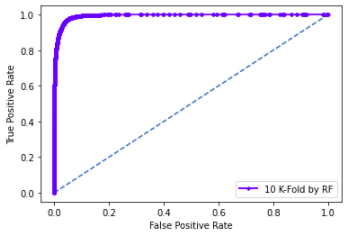


Figure ‎G‑I: 10-Fold ROC Graph of RF

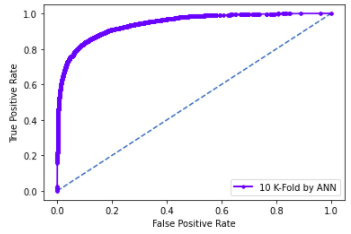


Figure ‎G‑II: 10-Fold ROC Graph of ANN

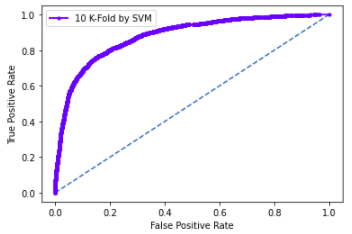


Figure ‎G‑III: 10-Fold ROC Graph of SVM

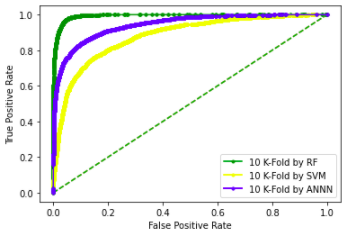


Figure ‎G‑IV: 10-Fold Comparison ROC

# Comparison

In this part of the examination, we look at the testing consequences of each anticipated calculation. For each calculation, we perform five distinct tests, a Self-Consistency test, Jack-Knife testing, Independent testing, 10-Fold Cross-Validation testing, and 5-Fold Cross-Validation testing. Beneath we think about the consequences of all testing procedures for each indicator.

## 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 |

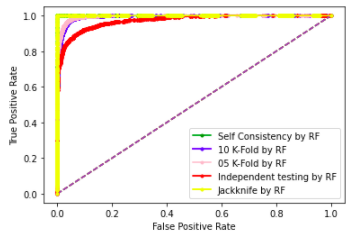


Figure ‎A‑I: Comparison Testing ROC of RF

## 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 |

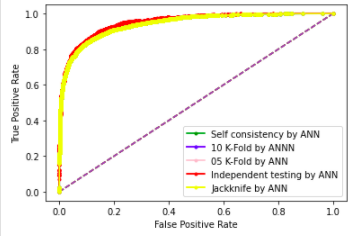


Figure ‎B‑I: Comparison Testing ROC of ANN

## 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 |

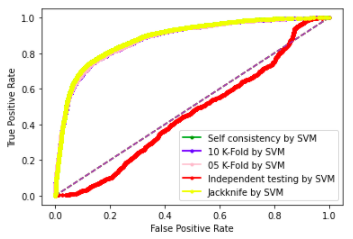


Figure ‎C‑I: Comparison Testing ROC of SVM

## 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.

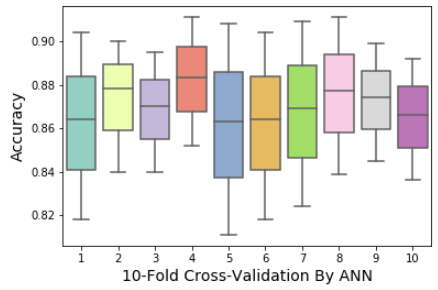


Figure ‎D‑I: ANN 10-Fold Cross-Validation

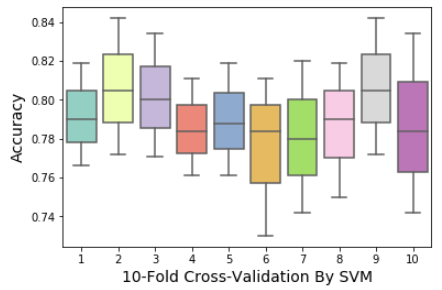


Figure ‎D‑II: SVM 10-Fold Cross-Validation

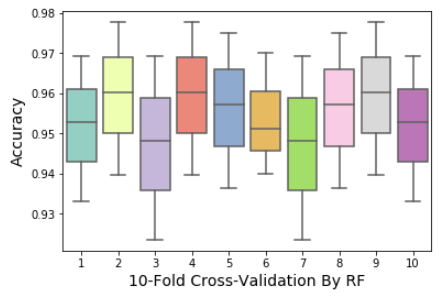


Figure ‎D‑III: RF 10-Fold Cross-Validation

## 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.

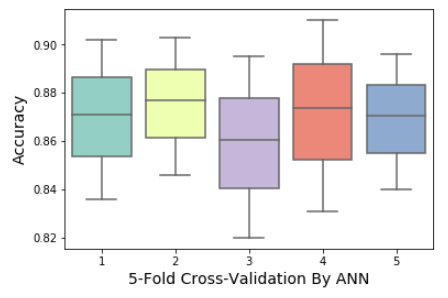


Figure ‎E‑I: ANN 5-Fold Cross-Validation

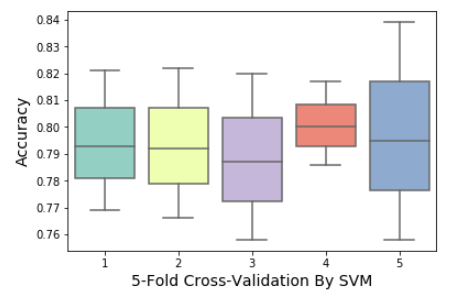


Figure ‎E‑II: SVM 5-Fold Cross-Validation

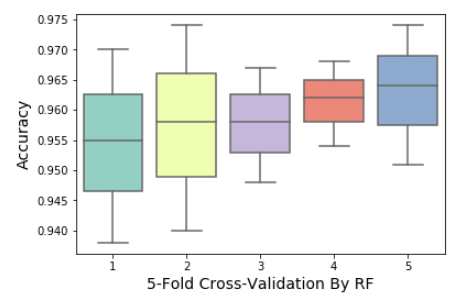


Figure ‎E‑III: RF 5-Fold Cross-Validation

# 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](http://biopred.org/stressprotiens).

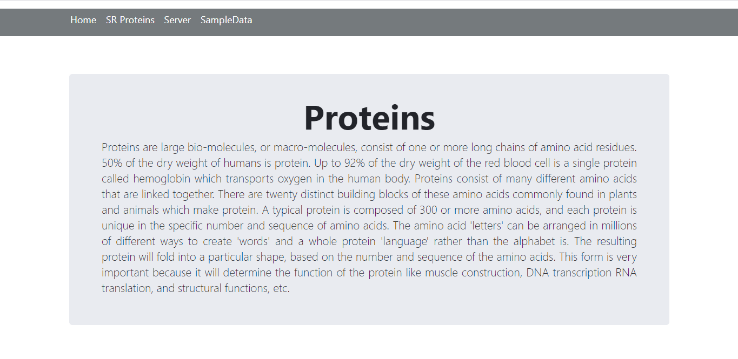


Figure ‎E‑I: Home Page of Web-Server



Figure ‎E‑II: Introduction Page of Web-Server

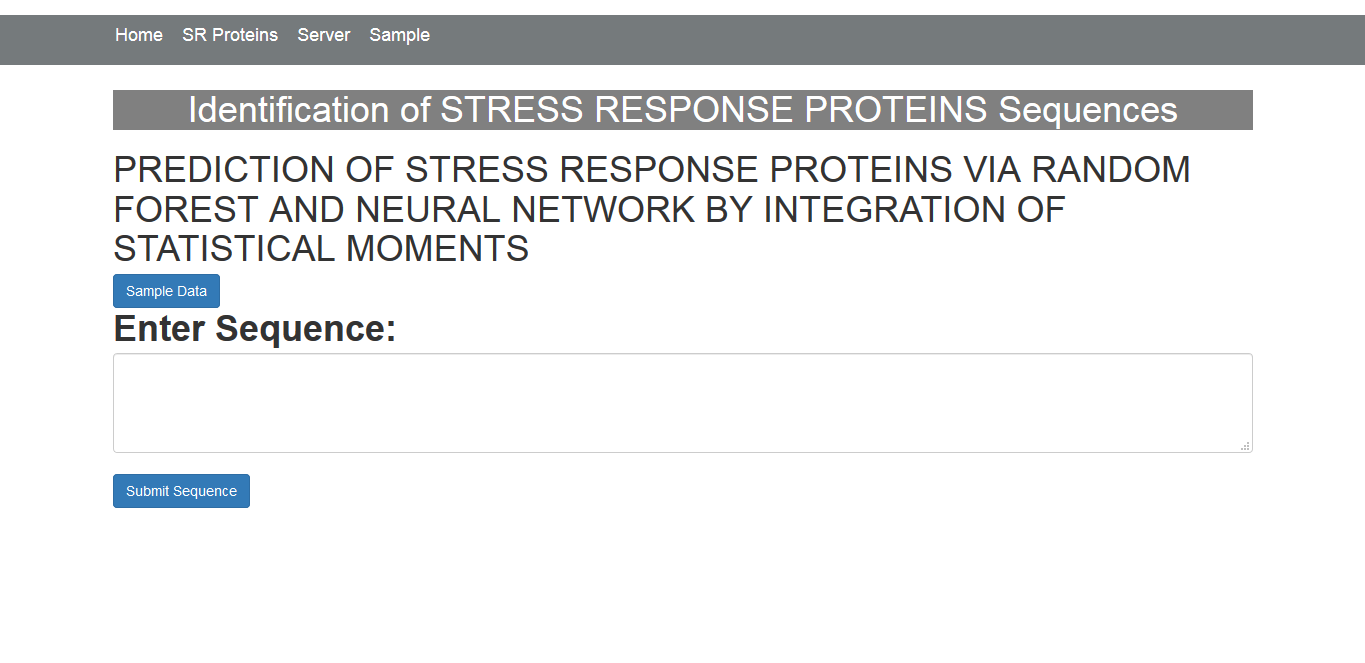


Figure ‎E‑III: Prediction Server Page

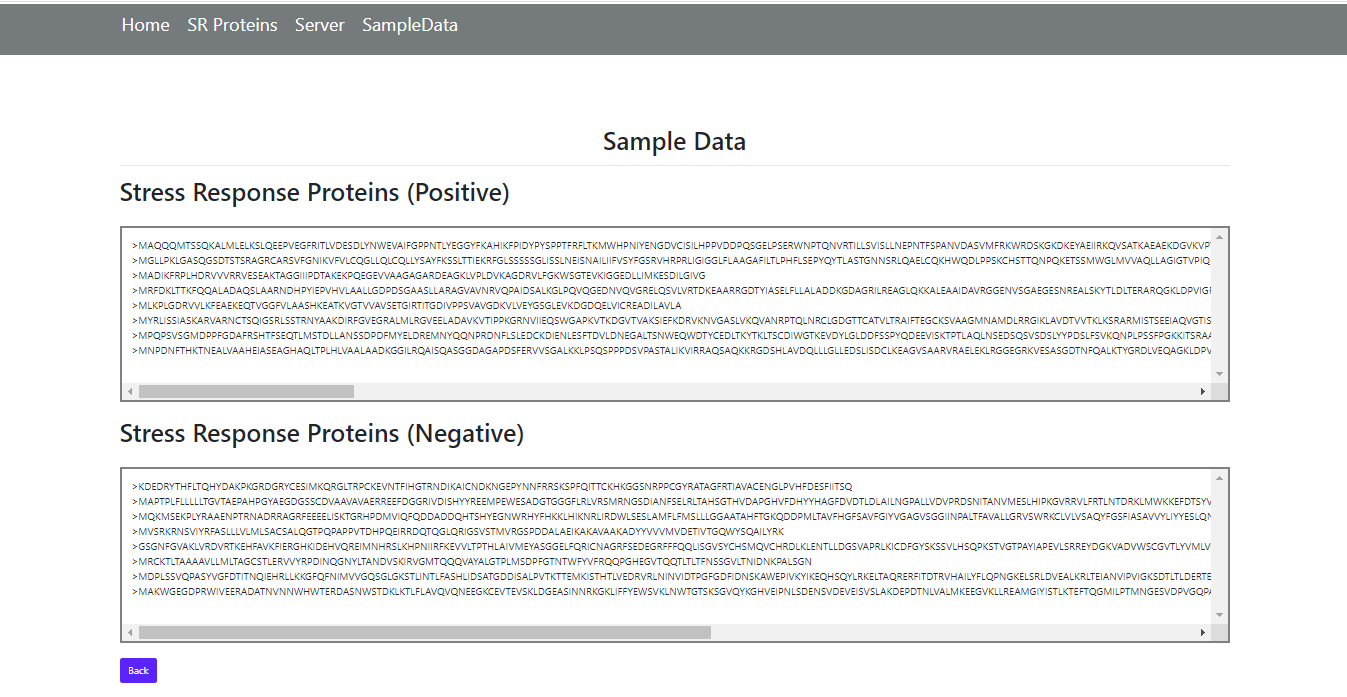


Figure ‎E‑IV: Sample Data Page

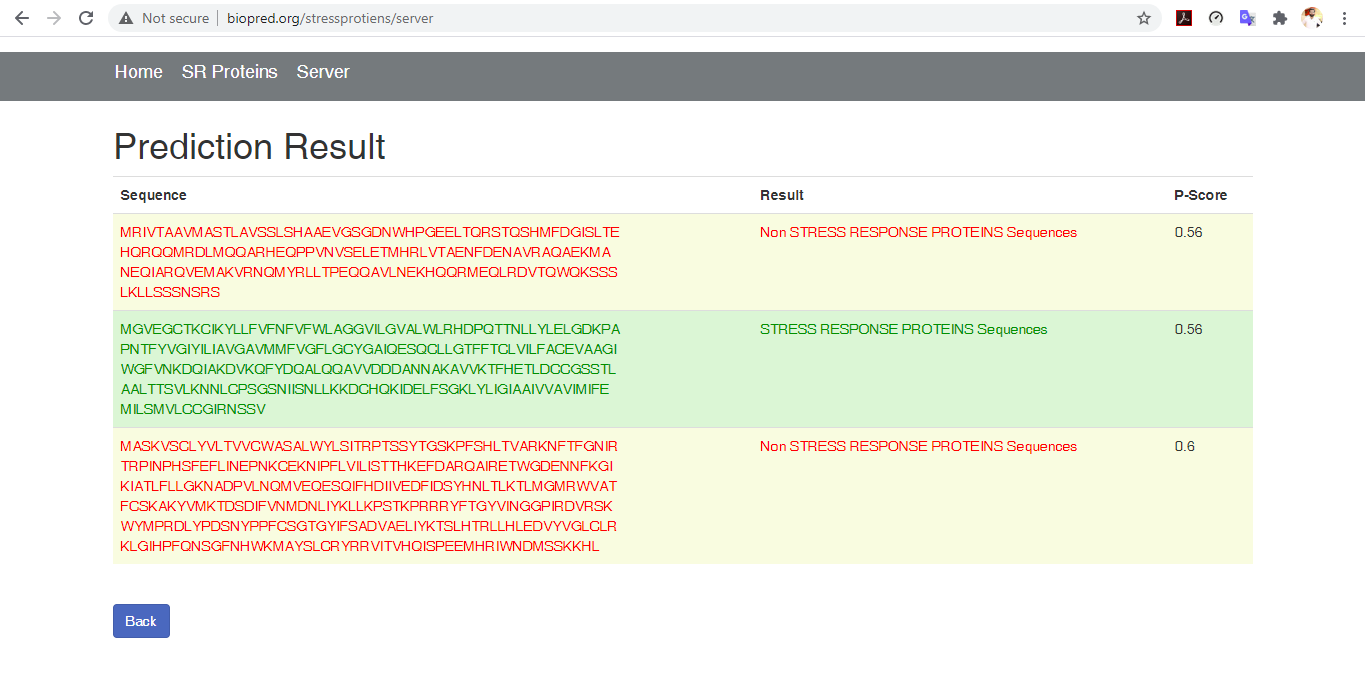


Figure ‎E‑V: Results Page of Web-Server

# Discussion

The feature vector used to describe the features of Amino acid and run algorithms on these features vectors can be found in section 3, a current study based on three classifiers named as Random Forest, Support Vector Machine and Artificial Neural Network. The highest accuracy obtained between the three classifiers by random forest as shown in Table ‎A‑I. The stress response proteins benchmark dataset was very stable, which is why dataset was clean from redundancy and then applied classifier on that data collection. By utilizing Self-consistency and pocketknife testing, the planned model has accomplished 99.9 % genuine outcome while 95.8% exact outcome has been noted by utilizing k-overlay cross-approval. The validness of the arbitrary woods classifier was discovered to be 99.9%, the affectability esteem was estimated as 99.9% while the explicitness was determined as 99.8% and the MCC esteem was estimated as 0.9873%. Without anyone else consistency test the ANN got 87.0%, and SVM has 80.2% of accuracy without anyone else consistency testing. As opposed to ANN and SVM, Random woodland gave the best outcomes. Stress reaction Predictor is required to turn into a convenient high throughput apparatus for considering or if nothing else exploring Stress Proteins. It is the absolute first marker to foresee Protein Stress reaction.

# Conclusion

In this examination, the Stress Response Protein Sequence named as a counterfeit neural organization, uphold vector machine, and irregular woods classifier have been proposed by three Classifiers to anticipate. It dependent on five-venture rule. The irregular woods classifier delivered the most elevated outcome. 99.9% accuracy was shown by the discoveries of the consequences of the irregular woods model. This shows that for the assessment of Stress Protein Sequences, the Random Forest Classifier shows better results or results. Pocketknife testing and 10-overlap cross-approval affirmed the marker, producing 99.9%, and 95.8%, precise outcomes, individually. Simultaneously, the proposed indicator serves to adequately or dependably anticipate pressure reaction proteins and give standard proof to the disclosure or advancement of new medications and biomarkers against clinical issues of this sort. While our other forecast calculation as talks about in the correlation area, shows low accuracy execution. All things considered, the hugest indicator among anticipated classifiers is an irregular woodland classifier.

1. [↑](#footnote-ref-1)