*VFAST Transactions on Software Engineering*
*http://vfast.org/journals/index.php/VTSE@ 2021, ISSN(e): 2309-3978, ISSN(p): 2411-6246*
*Volume 9, Number 2, April-June 2021* *pp: 01-09*

# Prediction of Allergen and Non-Allergen Proteins Sequence via Chou's 5-Step Rule

Sumbal Nawaz, Kalsoom Fatima, Muhammad Adeel Ashraf*

[1]Department of Computer Sciences, University of Management and Technology, Lahore, Pakistan
*Corresponding Author: adeel.ashraf@umt.edu.pk

ABSTRACT
*Some specific kinds of proteins are responsible for the risk of immediate type I allergic reaction. Therefore, the proteins that are made to use in the consumer product should be checked for their allergic reactions before introducing them in the market. The FAO/WHO instructions for the assessment of allergic proteins depend on the linear sequence window identity and short peptide hits misclassify many proteins as allergen proteins. This study introduces the AllerPredictor model that predicts the allergen & non-allergen proteins depending on the sequence of proteins. Data was downloaded from two major databases, FARRP and UniProtKB. The results of this model were validated with the help of self-consistency testing, independence testing, and jackknife testing. The accuracy for self-consistency validation is 99.89%, for the independence testing is 74.23%, and for 10-fold cross-validation, it is 97.17%. To predict the allergen and non-allergen proteins, this AllerPredictor model has a better accuracy than other existing methods.*

KEYWORDS
PseAAC, Statistical moments, Random Forest Method, 5-step rule.

JOURNAL INFO
HISTORY: Received: January 25, 2021 Accepted: March 20, 2021 Published: March 30, 2021

## 1. INTRODUCTION

Allergy is the most frequently occurring chronic disease that is spreading globally. Moreover, this is also the main cause of asthma and of asthma exacerbations that are increasing in many countries' day by day [1].

IgE type hypersensitivity causes allergic hypersensitivity and is caused by allergens. These kinds of allergic reactions often produce in the basophils or the mast cells which release multiple irritant arbitrators such as cytokines, leukotrienes, histamine and chemokines.

Furthermore, these arbitrators can produce severe symptoms such as hives, sneezing, rashes, itching, asthma attacks, difficulty in breathing which can cause death consequently [2].

An antigen is a protein that is made up of lipids, amino acids, and sugar. The antigen is classified as a foreign antigen and an autoimmune antigen. Allergen is the specific antigen that initiates an allergic reaction in the body [3].

On the other hand, antibodies are immunoglobulins which are y shaped. These are produced by the B cells, which exist in the immune system. Antibodies work against antigens. IgA, IgE, IgG, IgM, and IgD are the five types of antibodies [3]. IgE antibodies work against the allergen antigen.

There is a need that all products which come in the market must be checked first for their allergic reactions to avoid the high risk of IgE type 1 severe allergic reaction [4].

Allergens can cause serious type 1 hypersensitivity reactions in individuals and these reactions are very harmful to human health [5]. Allergic proteins in food can cause chronic illness due to the following products mainly soy, wheat, eggs, milk, and different kinds of wheat or peanuts [6].

That's why it is very essential to recognize and remove allergens from all the biotechnological products like vaccines, genetically produced crops, and therapeutics and it is also very necessary to identify the allergens from the sequence of genomes. Though, identifying primary identifier for the allergenicity [7].

The FAO/WHO organization gives guidelines to know the potential of allergenic protein. It tells the rules that the given sequence of amino acids is like the known allergenic protein. According to the FAO/WHO rule 1, a sequence of a protein will be classified as the allergen protein if there are six contiguous amino acids match with each other, and according to the rule 2, at least the sequence of 35 amino acids should be matched at the window of 8 when we compare it with the known allergens. Many pieces of research have been conducted in this area [8]. Allergens have a 3-dimensional structure and most of the protein is stable, small, and well structured. That's why these are perfect for different kinds of structural studies [9]. SORTALLER online allergen classifier was created based on the normalized BLAST E-values and featured peptide dataset [10]. Peptide Locator has been presented for the automated prediction of the peptides within the sequence of a protein and tested in the fivefold cross-validation at the curve of 0.92 [11].

To differentiate the allergens and non-allergens, a fingerprint approach is presented [12]. Text classification technique is used to predict allergen. The overall accuracy

rate was 77% [13]. Identifying protein is very helpful for this Multip-SChlo method in which accuracy rate reaches to 55.52% [14]. Then Cross React technique was applied to the distinct set of the seven allergens to identify the cross-reactive allergen [15].

The AllerCatPro method was developed to predict the allergenic protein based on the protein structure. The major databases that were used are: COMPARE, FARRP, WHO/IUIS, Allergome and UniProtKB. The overall accuracy rate of this method was 84% [4].

Though, allergic proteins need to be identified by B and T cells which are used to initiate the evolution of protein particularly IgE. In this study, we introduce the new model to recognize protein structure from its amino acid sequence. This model will predict allergen and non-allergen proteins from the protein sequence. Firstly, we collected all the reliable and available data of allergen protein from various databases. After that, we combined it into a single, unique, and complete form of data set.

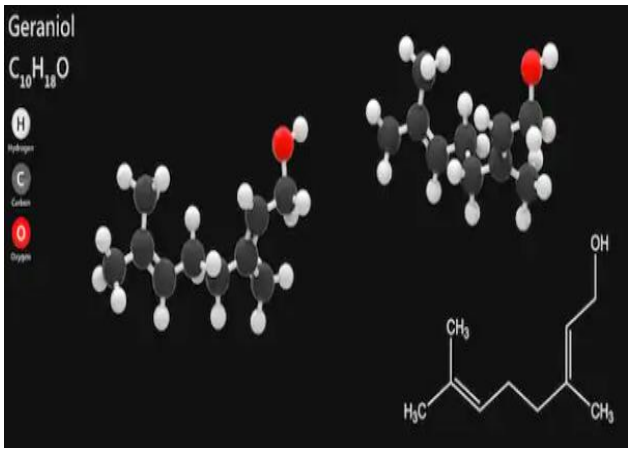


**Figure-1:** Structural Formula of allergen

## 2. MATERIAL AND METHOD.

In this section, we have described the overall process of identification of allergen and non-allergen protein. It includes primary steps such as collection of data, extraction of features, training Random forest model, and then validation of training. All steps are done according to Chou's 5 step rule.

The first step is the collection of the benchmark dataset from the well-known databases UniProt and Allergen online. All the sequences related to the allergen were identified. CD-Hit with the threshold value of 60% was used to remove the redundancy from the dataset.

After this, the second step of Chou's rule consists of feature extraction using the wide range techniques of the feature vector. In the third step, the input and output matrix of feature vectors were used to train the network using random forest.

In the fourth step, accuracy of the predicted model was validated using the test datasets. All these steps are shown in Figure 2.

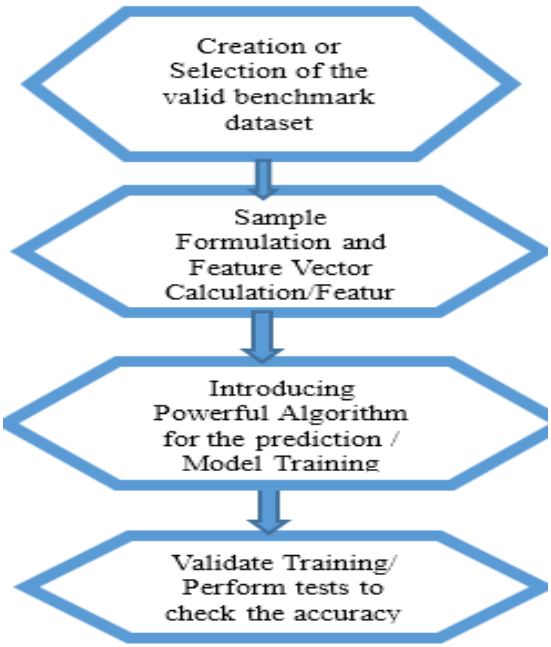


**Figure-2:** First Four steps of the Chou's 5-steps rules.

### 2.1. COLLECTION OF BENCHMARK DATASET.

As stated by the five step rules, the firsts step is the benchmark dataset collection. For the development of the statistical model, the most important thing is to start the standard dataset which will be used for the training and for the testing of the predicted model. The model accuracy will be unpredicted if the collected dataset is full of errors. Here the dataset collected from the database UniProt and Allergen online for the prediction of allergen non allergen protein sequence.

UniProt stands for the Universal protein Resource is a well-known database used for getting the freely available dataset of protein sequences. The Allergen online are another FARRB database which are specifically made for the allergen protein sequence.

For the collection of positive datasets 1534 entries were downloaded from the allergen online and 922 entries were downloaded from the UniProt using the keywords "allergen". Total 2456 positives dataset samples were collected.

From the allergen online, pdf files which had 2128 accession allergen entries were downloaded. Then each accession entry is downloaded individually from UniProt. From 2128 allergen online entries there were some entries which were not found at UniProt, total of 1534 allergen entries were downloaded from UniProt from the allergen online accession entries. Further, the negative dataset was only downloaded from the UniProt.

For the negative dataset the reverse query was used. No operator was used with the allergen keyword. From UniProt

1500 negative data sets are downloaded. Different database sources are shown in Table 1

Table 1. Sources databases for the allergen protein

| Databases | Website Link | Posi data | Neg data |
|---|---|---|---|
| UniProtKB | https://www.UniProt.org/ | 922 | 1500 |
| FARRP | http://www.allergenonline.org/ | 1534 | - |

After downloading positive datasets from two different websites it composed into one file and passed from CD-HIT.CD-HIT used to remove redundancy at threshold 0.6. The result of positive data after CD-HIT was 573.Similarly negative dataset passed from the cd hit. The negative dataset which was left after CD-HIT was 1264. The training and testing dataset were produced for the identification of the statistical model. The predicted model is trained using the training dataset and then it is tested using the testing dataset. The dataset was minimized in this way:

**A= A+ U A-** **(1)**

573 positive dataset union with the 1264 negative dataset samples are included in the dataset of this paper.

## 2.2. SAMPLE FORMULATION.

Nowadays, with the enhancement of biological data and their sequences, the major difficulty is the vector formulation or the distinct models from the sequences of data without the loss of sequence of information and feature for the target analyses. For the vector formulation from the protein sequence there is the main machine learning algorithms present such as the Covariance Discriminate algorithm [18], Support Vector Machine (SVM) algorithm [19], 'Nearest Neighbor (KNN)' algorithm [20], and 'Random Forest (RF)' algorithm [21]. These methods can handle the vector input though there is the chance that the distinct model may lose any pattern of protein sequence.

For the prevention of the loss of pattern sequence which is related to protein sequence PseAAC model was introduced [22]. Then the Chou PseAAC model came into existence [23]. Then various software developed for this model that involves PseAAC general, builder and the propy developed [24], [25], [26]. The idea of PseAAC model was implemented in the PseKNC model for the calculation of various feature vectors related to the sequences of RNA and DNA which proves to be very effective and efficient. In the recent time an amazing efficient and powerful website Pse-in-one and its version 2.0 was developed which can generate needed feature vectors according to the need of users from peptide, RNA, DNA and protein sequences [27].

As stated by the Chou's general PseAAC the formula for the equation for the protein sequence are expressed as

$P_{\S}=7(\mathbb{H}) = [\mathbb{N}1\mathbb{N}2 \cdots \mathbb{N}n \cdots \mathbb{N}\Omega]$ (28).

## 2.3. VECTOR SITE VICINITY.

It is determined from the subgroups of protein sequence. Modifications in the protein sequence are shown here.

$\mathcal{P} = \{\alpha 1 \cdots \alpha q-2, \alpha q-1, \alpha q, \alpha q+1, \alpha q+2, \cdots \alpha n\}$ ( 28)

## 2.4. CALCULATION OF STATISTICAL MOMENT.

These moments are quantities measured which are important to represent data accumulation. The moment is the special quantitative metrics of the different points of shape. Statistical moments are perfect for making a variety of features from the known pattern. Many types of research used these statistical moments to find out the features from any pattern of protein. It describes various functionalities of specific patterns [29]. The purpose of the proposed problem is fulfilled with several moments like central moments, Hahn moments, raw moments along the centroid and origin of data as used in [30], [31]. Apart from the Hahn and central moment, there are many other moments present. Though, in some previous studies it is shown that the distinct orthogonal moments produce the more accurate result than the continuous orthogonal moments for the quantized and distinct data. Orthogonal moments have capability to transform shape with the minimum loss of data [32]. ]. Protein sequence if denoted by the

P= {α1, α2, α3, ……, αk} (33)

There is an end goal that is two dimensional moments. For the two-dimensional moment's one-dimensional design is converted through the row major scheme. Its length is calculated using the square root of protein length $n = \lceil\sqrt{k}\rceil$ [34]

Here n is a measurement metric of 2-dimensional metric whereas k is protein length. Moreover, to adjust all protein sequence components new matrix P' is formed with the n*n dimensions.

$$\begin{bmatrix} P_{11} & P_{12} & \cdots P_{1k} \\ P_{21} & P_{22} & \cdots P_{2k} \\ \vdots & \vdots & \ddots \vdots \\ P_{k1} & P_{k2} & \cdots P_{kk} \end{bmatrix} \quad (35)$$

A ϖ function used for the conversion of matric P into the

P' $\varpi(au) = \omega ij$ (36).

Where I=u+1 and j=u mod v and P'
is sequential protein sequence. For the calculation of the raw moment following expression used.

$\mathbb{M}ij = \sum\sum \ell im\ \ell m\ \upsilon\ m=1\ \upsilon\ \ell =1$ (37)

Here i+j is moment direction. The following moments are calculated MOO, MO, M10, MO2, M11, M20, M12, M21, M30, and M03 up order three. The central moments are calculated and their centroid points are represented as y bar and z bar

$\bar{y}$ = M10/M00 and $\bar{z}$ = M01/MOO

After these raw moments are calculated using this equation $\eta ij = \sum \sum (\ell - \bar{y}) (m - \bar{z}) j\omega \ell m \, v \, m=1 \, v \, \ell=1$ (38). Similarly, Hahn moments are also calculated up to the order 3.

**2.5.FREQUENCY MATRIX (FM) DETERMINATION.**

There is another matrix called the frequency matrix it is made to cover information about composite of the protein sequence structure. Primary reason for using frequency to extract information about the sequence of proteins. The matrix shown here

$FM = \{r1,2, \cdots, r20]$ (39)

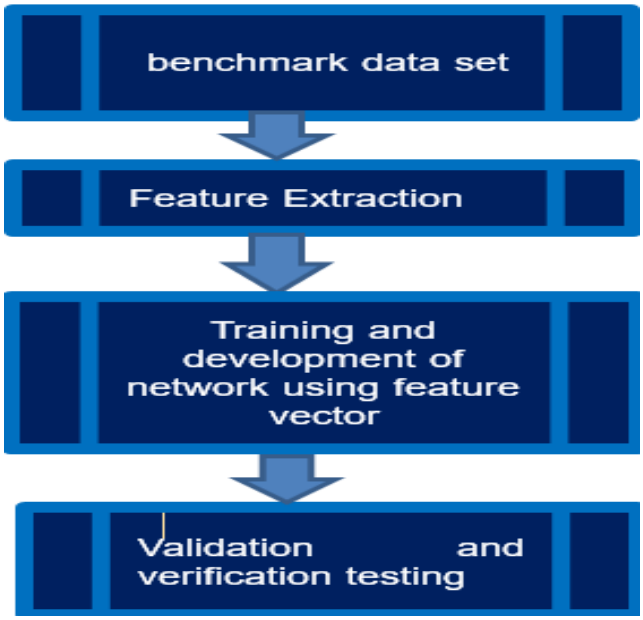


**Figure-3:** Flowchart for proposed methodology.

**2.6. PRIM Calculations.**

The starting point of feature extraction is the calculation of matrix from input protein. Due to this, the length of protein arrangement utilizes PRIM. These matrices are used for finding out those moments that are required for the vector shape.

$$H_{PRIM} = \begin{bmatrix} H_{1\to1} & H_{1\to2} & H_{1\to3} H_{1\to j} & \cdots & H_{1\to20} \\ H_{2\to1} & H_{2\to2} & H_{2\to3} H_{2\to j} & \cdots & H_{2\to20} \\ \vdots & \vdots & \vdots \ \vdots & \cdots & \vdots \\ H_{i\to1} & H_{i\to2} & H_{i\to3} \, H_{i\to j} & \cdots & H_{i\to20} \\ \vdots & \vdots & \vdots \ \vdots & \vdots & \vdots \\ H_{N\to1} & H_{N\to2} & H_{N\to3} H_{N\to j} & \cdots & H_{N\to20} \end{bmatrix} \quad (40)$$

**2.7. RPRIM CALCULATIONS.**

RPRIM matrix calculations are exactly like PRIM. HPRIM matrix is calculated such as the input of a sample of reverse protein.

$$H_{PRIM} = \begin{bmatrix} H_{1\to1} & H_{1\to2} & H_{1\to3} H_{1\to j} & \cdots & H_{1\to20} \\ H_{2\to1} & H_{2\to2} & H_{2\to3} H_{2\to j} & \cdots & H_{2\to20} \\ \vdots & \vdots & \vdots \ \vdots & \cdots & \vdots \\ H_{i\to1} & H_{i\to2} & H_{i\to3} \, H_{i\to j} & \cdots & H_{i\to20} \\ \vdots & \vdots & \vdots \ \vdots & \vdots & \vdots \\ H_{N\to1} & H_{N\to2} & H_{N\to3} H_{N\to j} & \cdots & H_{N\to20} \end{bmatrix} \quad (41)$$

**2.8. FINDING THE AAPIV.**

Frequency matrix determined for the extraction of compositional informational. It is formed when the length of 20 elements of primary sequences of protein are combined at their own location.

$$AAPIV = [u_1, u_2, u_3, \ldots, u_{20}] \quad (42)$$

**2.9. FINDING THE RAAPIV.**

To take out the obscure and deep information about relative position of individual amino acid residue this technique is used.

RAAPIV= [u1, u2, u3, …, u20] (43)

**2.10. PREDICTION MODEL.**

Here, a random forest or the random decision tree model used to find the results such as used in the [44]. Datasets are constructed which contain both positive sample and negative sample. Then feature vectors constructed through that dataset which is used for the identification of allergy & non-allergen protein sequence. The two feature vectors are merged to form input matrix that consist of input vectors as both negative and positive samples in addition to the output matrix.

It consists of a lot of individual decision trees which operate combines. In it each tree of the random forest is considered as the class prediction. The class who gets more votes becomes the prediction of the model. Datasets were provided separately, positive and negatively. It gives the result in false positive, true negative, true positive and false negative.

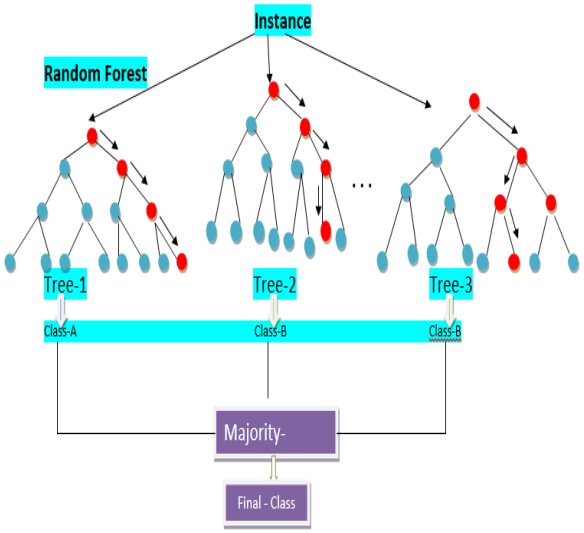


**Figure-4:** Working of Random Forest Simplified

## 2.11. DECISION TREE LEARNING.

The famous algorithm used for many machines learning work. Particularly, those trees which grow more they get more irregular patterns. They have high variance and low bias. They are more accurate. Random forests are the technique in which multiple decision trees are averaged, which are trained on various chunks of the same training set with the purpose of minimizing the variance. Leo Briman was the first man who noticed that link between the kernel method and random forest. He noticed that random forests which grow faster are equivalent to the kernel true argin. It is based on uniform random forest and centered random forest.

## 3. RESULT AND DISCUSSION.

The proposed model is used for the prediction of allergen & non-allergen from the protein sequences. The prediction of allergenicity depends on the variant techniques of feature extraction. In this section details of the validation tests and results are desired.

## 3.1. ESTIMATED ACCURACY.

The purpose of this predictor is very essential that helps to find out the model success rate [47]. Though, for such analysis, it needs to focus on the two essential factors which are (1) choosing the accuracy metrics (2) selection of the testing method which are used for model validation. Here we will produce metrics for the analysis and then we will do the validation of these methods.

## 3.2. METRIC FORMULATION FOR THE OBJECTIVE EVALUATION.

It is necessary to think about that different metrics which is used for the evaluation. The best metrics use to find out the model accuracy are the (1) Accuracy metric which use for the evaluations of model accuracy, (2) Sensitivity which is use for the evaluation of model capability to predict positive samples, (3) Specificity which is used for the evaluation of model capability to predict negative samples, Matthews Correlation Coefficient is used for measurement of stability of prediction model.

$$Sn=1-N\pm N+ \quad (44)$$

$$SP=1-N\mp N- \quad (45)$$

$$ACC=1-N+NN++N- \quad (46)$$

$$MCC=1-NN++NN-1+N+--NN+1+N-Nt-N \quad (47)$$

Here N- exhibits a total amount of the non-allergen which is perfectly predicted as the non-allergen by this model. N+ shows the total amount of positives which are correctly predicted as positive. N+- shows total number of non-allergens which incorrectly predicted as allergen. N-+ shows the total amount of allergen which is incorrectly identified as non-allergen [35].

## 3.3. SELF-CONSISTENCY TEST.

This test is mainly used to get the confusion matrix. This is an essential test which is used to calculate the effectiveness of predictive models in which complete training datasets are used for model testing [49-72]. This model got the 99.89% ACC, 0.9975 MCC, 99.65% Sn, 100% Sp. The result shows in Table 2 and Figure 5.

**Table 2.** Self-Consistency Testing Results

| Predictor Accuracy Metric | | | | |
|---|---|---|---|---|
| Aller Predictor | ACC (%) | MCC | Sp (%) | Sn (%) |
| | 99.89 | 0.9975 | 100 | 99.65 |


**Figure-5:** Self-Consistency Testing Results

## 3.4. INDEPENDENT DATASET TEST.

This test is basically dividing the overall sample into two partitions. The dataset which is used for training is not

used for the independent testing [49-72]. In this test this model got 74.23% ACC, 0.4021 MCC, 60.36% Sn, and 80.37% Sp. The result shows in Table 3.

**Table 3. Independent Testing Results**

| Predictor Accuracy Metrics | | | | |
|---|---|---|---|---|
| Aller Predictor | ACC (%) | MCC | Sp (%) | Sn (%) |
| | 74.23 | 0.4021 | 80.37 | 60.36 |

Chart Area

### 3.5. MODEL VALIDATION.

### 3.5.1.CROSS VALIDATION MODEL TESTING.

This is the procedure which is used to evaluate the predictive model at dividing actual samples into the training set that is used for model training and into the testing set which is used to analyses the model. It is used in the situation when we want the perfect accuracy to estimate of our predictive model [49-72] .

In an expectation issue, the model is normally given the dataset of information in which preparing is performed (preparing dataset), and a dataset of obscure information against which the model is tried (trying dataset). Cross-validation is a procedure to build up a likelihood which recommended strategy is smooth while a perceptible approval test set isn't available. In k-crease cross-approval, the first example is arbitrarily parted into $k$ equivalent size small samples.

Along these lines, let $Z$ considered as the number of inhabitants in tests which contains similarly negative and positive samples.

$$Z = \{z1, z2, z3 \cdots, zn\} \text{ [16]}$$

Where $Zi$ is any arbitrary positive and negative groupings. Informational index is part into $k$ proportionate size subgroups

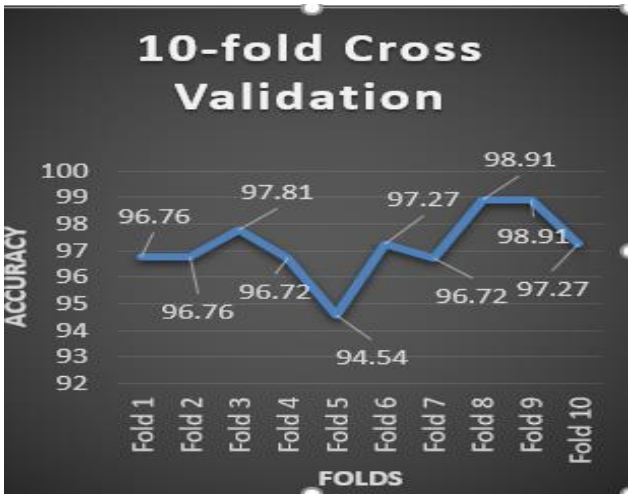


**Figure-6:** 10-fold Cross Validation results (Average of 10-folds)

Additionally, the subsets are selected discretionarily with the end goal that the past measurements are equal. For example

$$|Zi| \simeq |Zj| \text{ [45]}$$

The impacts of the cross-validation demonstrate that the proposed model is adequately performed well than different indicators. Utilizing cross-validation, benchmark dataset is disseminated into all out-k number of exceptional folds, where k is the number where benchmark dataset is partitioned, for the present, k=10. In each step of testing, an alternate subset of information is chosen arbitrarily for approval over the remainder of the information, by this, each piece of the dataset is utilized for preparing and testing both. Toward the finish of the last pass of cross-validation, the cumulated exactness for k=10 is determined by including the precision of every step and then dividing it by 10 and the accuracy which finally comes to 97.17%. The results are shown in Table 4.

**Table 4.** Average of 10-fold cross validation results

| Accuracy Metrics | | | |
|---|---|---|---|
| ACC (%) | MCC | Sp (%) | Sn (%) |
| 97.17 | 0.934 | 99.5 | 91.96 |

Similarly, the result of 10-fold cross validation sensitivity, specificity, accuracy and MCC are shown in Figure 7. ACC is 97.17%, MCC is 93.399, Sp is 99.5% and Sn is 91.96%.

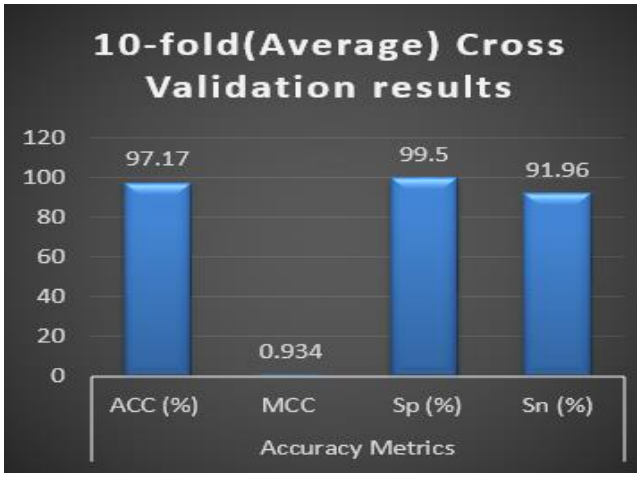


**Figure-7:** 10-fold cross validation for benchmark dataset.

It shows that accuracy of the Allerpredictor method is greater than the accuracy of all previous methods which were used for the prediction of allergen & non-allergen protein. Accuracy of 10-fold cross validation separately shown in Figure 7.

### 3.6. COMPARATIVE ANALYSIS.

In this section, a vast analysis on the different methods used for the prediction of allergen and non-allergen protein. A method named AllerCatPro got the 84 percent accuracy, 0.81 MCC, 67 percent Specificity and 100 percent Sensitivity [4]. Another method EVALLER got 92.8 percent accuracy, 0.863 MCC, 99.0 percent Specificity and 86.6 percent Sensitivity [46]. Similarly, another method AllerHunterd got the 90.7 percent accuracy, 0.826 MCC, 99.2 percent Specificity and 82.2 percent Sensitivity [41]. While the method I proposed in this study is Allerpredictor with 99.89 percent accuracy, 0.9975 MCC, 100 percent Specificity and 99.65 percent Sensitivity. It got the largest values for all metrics Specificity, Sensitivity, Accuracy, and for MCC. This shows that Allerpredictor is better than other methods.

**Table 5.** Comparison of different methods

| Methods | ACC% | MCC | Sp% | Sn% |
|---|---|---|---|---|
| Aller prdictor | 99.89 | 0.9975 | 100 | 99.65 |
| EVALLER [46] | 92.8 | 0.863 | 99.0 | 86.6 |
| AllerHunterd [41] | 90.7 | 0.826 | 99.2 | 82.2 |
| AllerCatPro [4] | 84 | 0.81 | 67 | 100 |

### 3.6 CONCLUSION.

Hence in this model random forest methods are used for the prediction of allergenicity. The AllerPredictor model produced which predicts the allergen & non-allergen protein depends on the sequence of proteins. Data is downloaded from the two major databases FARRP and UniProtKB. The results of this model are validated with the help of self-consistency testing, independence testing and jackknife testing. The accuracy of self-consistency validation is 99.89%, while for the independence testing 74.23% accuracy produced and 10-fold cross validation produce 97.17%. The AllerPredictor model has the better capability to predict the allergen non allergen protein than the existing methods.

### REFERENCES

[1] G. Devereux, “The increase in the prevalence of asthma and allergy: Food for thought,” Nat. Rev. Immunol., vol. 6, no. 11, pp. 869–874, 2006.

[2] N. J. Stagg, H. N. Ghantous, G. S. Ladics, R. V. House, S. M. Gendel, and K. L. Hastings, “Workshop proceedings: Challenges and opportunities in evaluating protein allergenicity across biotechnology industries,” Int. J. Toxicol., vol. 32, no. 1, pp. 4–10, 2013.

[3] “Difference between antigen and antibody,” 2017. [Online]. Available: https://www.technologynetworks.com/immunology/articles/antigen-vs-antibody-what-are-the-differences-293550.

[4] S. Maurer-Stroh et al., “AllerCatPro-prediction of protein allergenicity potential from the protein sequence,” Bioinformatics, vol. 35, no. 17, pp. 3020–3027, 2019.

[5] Y. F. Gao, B. Q. Li, Y. D. Cai, K. Y. Feng, Z. D. Li, and Y. Jiang, “Prediction of active sites of enzymes by maximum relevance minimum redundancy (mRMR) feature selection,” Mol. Biosyst., vol. 9, no. 1, pp. 61–69, 2013.

[6] H. A. Sampson and N. York, “Allergy Clinical Immunology disorders,” pp. 717–728, 1976.

[7] G. S. Ladics et al., “Bioinformatics and the allergy assessment of agricultural biotechnology products: Industry practices and recommendations,” Regul. Toxicol. Pharmacol., vol. 60, no. 1, pp. 46–53, 2011.

[8] T. P. Chang et al., “No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析 Title,” Cem. Concr. Res., vol. 19, no. 4, pp. 645–655, 2005.

[9] F. Dall’Antonia, T. Pavkov-Keller, K. Zangger, and W. Keller, “Structure of allergens and structure-based epitope predictions,” Methods, vol. 66, no. 1, pp. 3–21, 2014.

[10] L. Zhang, Y. Huang, Z. Zou, Y. He, X. Chen, and A. Tao, “SORTALLER: Predicting allergens using an substantially optimized algorithm on allergen family featured peptides,” Bioinformatics, vol. 28, no. 16, pp. 2178–2179, 2012.

[11] C. Mooney, N. J. Haslam, T. A. Holton, G. Pollastri, and D. C. Shields, “PeptideLocator: prediction of bioactive peptides in protein sequences.,” Bioinformatics, vol. 29, no. 9, pp. 1120–1126, 2013.

[12] I. Dimitrov, L. Naneva, I. Doytchinova, and I. Bangov, “Systems biology AllergenFP : allergenicity prediction by descriptor fingerprints,” no. 2005, pp. 1–6, 2013.

[13] H. X. Dang and C. B. Lawrence, “Sequence analysis Allerdictor : fast allergen prediction using text classification techniques,” vol. 30, no. 8, pp. 1120–1128, 2014.

[14] X. Wang, W. Zhang, Q. Zhang, and G. Z. Li, “MultiP-SChlo: Multi-label protein sub chloroplast localization prediction with Chou’s pseudo amino acid composition and a novel multi-label classifier,” Bioinformatics, vol. 31, no. 16, pp. 2639–2645, 2015.

[15] S. S. Negi and W. Braun, “Cross-React: A new structural bioinformatics method for predicting allergen cross-reactivity,” Bioinformatics, vol. 33, no. 7, pp. 1014–1020, 2017.

[16] Z. H. Zhang, J. L. Y. Koh, G. L. Zhang, K. H. Choo, M. T. Tammi, and J. C. Tong, “AllerTool: A web server for predicting allergenicity and allergic cross-reactivity in proteins,” Bioinformatics, vol. 23, no. 4, pp. 504–506, 2007.

[17] M. J. Abramson, R. M. Puy, and J. M. Weiner, “Allergen immunotherapy for asthma,” Cochrane Database Syst. Rev., no. 4, 2003.

[18] R. Wang, H. Guo, L. S. Davis, and Q. Dai, “Covariance Discriminative Learning: A Natural and Efficient Approach to Image Set Classification Institute for Advanced Computer Studies , University of Maryland , College Park , MD , 20742,” Comput. Vis. Pattern Recognit., pp. 2496–2503, 2012.

[19] P. Pavlidis, I. Wapinski, and W. S. Noble, “Support vector machine classification on the web,” Bioinformatics, vol. 20, no. 4, pp. 586–587, 2004.

[20] H. He, W. Graco, and X. Yao, “Application of Genetic Algorithm and

k-Nearest," Knowl. Acquis., pp. 74–81, 1999.

[21] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," Comput. Stat. Data Anal., vol. 52, no. 4, pp. 2249–2260, 2008.

[22] W. Hussain, Y. D. Khan, N. Rasool, S. A. Khan, and K. C. Chou, "SPrenylC-PseAAC: A sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-prenylation sites in proteins," J. Theor. Biol., vol. 468, pp. 1–11, 2019.

[23] K.-C. Chou, "An Unprecedented Revolution in Medicinal Chemistry Driven by the Progress of Biological Science."

[24] D. S. Cao, Q. S. Xu, and Y. Z. Liang, "Propy: A tool to generate various modes of Chou's PseAAC," Bioinformatics, vol. 29, no. 7, pp. 960–962, 2013.

[25] S. Iqbal, L. M. Kiah, M. Hussain, M. K. Khan, and K. Raymond, "Author ' s Accepted Manuscript On Cloud Security Attacks : A Taxonomy and Intrusion Detection and Prevention as a Service Reference : On Cloud Security Attacks : A Taxonomy and Intrusion Detection and Prevention as a Service," J. Netw. Comput. Appl., 2016.

[26] P. Du, S. Gu, and Y. Jiao, "PseAAC-General: Fast building various modes of general form of chou's pseudo-amino acid composition for large-scale protein datasets," Int. J. Mol. Sci., vol. 15, no. 3, pp. 3495–3506, 2014.

[27] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K. C. Chou, "Pse-in-One: A web server for generating various modes of pseudo components of DNA, RNA, and protein sequences," Nucleic Acids Res., vol. 43, no. W1, pp. W65–W71, 2015.

[28] P. P. In, "Protein phosphorylation in prokaryotes.," Biochimie, vol. 71, no. 9–10, pp. 987–1105, 1989.

[29] [29] K. Yamaoka, T. Nakagawa, and T. Uno, "Statistical moments in pharmacokinetics," J. Pharmacokinet. Biopharm., vol. 6, no. 6, pp. 547–558, 1978.

[30] A. Winkelbauer, "Moments and Absolute Moments of the Normal Distribution," no. 2, pp. 1–4, 2012.

[31] M. Knüppel, "Evaluating the Calibration of Multi-Step-Ahead Density Forecasts Using Raw Moments," J. Bus. Econ. Stat., vol. 33, no. 2, pp. 270–281, 2015.

[32] R. Mukundan, S. H. Ong, and P. A. Lee, "Image analysis by Tchebichef moments," IEEE Trans. Image Process., vol. 10, no. 9, pp. 1357–1364, 2001.

[33] P. Cohen, "The origins of protein phosphorylation," Nat. Cell Biol., vol. 4, no. 5, 2002.

[34] K. Fatima, P. School of Systems & Technology, University of Management & Technology, Lahore, and ; Sumbal Nawaz ; Sobia Mehrban, "Biometric Authentication in Health Care Sector: A Survey," 2019.

[35] X. Xiao, H. X. Ye, Z. Liu, J. H. Jia, and K. C. Chou, "iROS-gPseKNC: Predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition," Oncotarget, vol. 7, no. 23, pp. 34180–34189, 2016.

[36] J. V. Olsen et al., "Global, In Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks," Cell, vol. 127, no. 3, pp. 635–648, 2006.

[37] D. H. P. & T. W. M. Jung-Min Kee, Rob C Oslund, "A pan-specific antibody for direct detection of protein histidine phosphorylation."

[38] J. Mok and M. Snyder, Global Analysis of Phosphoregulatory Networks, Second Edi., vol. 2. Elsevier Inc., 2010.

[39] F. Takaiwa, "Seed-based oral vaccines as allergen-specific immunotherapies," Hum. Vaccin., vol. 7, no. 3, pp. 357–366, 2011.

[40] S. Muhammad Aizaz Akmal, Methodology, Software, Validation, 1 Nouman Rasool, Conceptualization, Data curation, 2 and Yaser Daanial Khan, "Prediction of N-linked glycosylation sites using position relative features and statistical moments."

[41] A. links open overlay panelWaqarHussainaYaser D. AfzalKhancfKuo-ChenChoud, "SPalmitoylC-PseAAC: A sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-palmitoylation sites in proteins."

[42] P. V. Attwood, P. G. Besant, and M. J. Piggott, "Focus on phosphoaspartate and phosphoglutamate," Amino Acids, vol. 40, no. 4, pp. 1035–1051, 2011.

[43] P. G. Besant and P. V. Attwood, "Mammalian histidine kinases," Biochim. Biophys. Acta - Proteins Proteomics, vol. 1754, no. 1–2, pp. 281–290, 2005.

[44] J. Albert et al., "Implementation of the Random Forest method for the Imaging Atmospheric Cherenkov Telescope MAGIC," Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip., vol. 588, no. 3, pp. 424–432, 2008.

[45] H. Da Huang, T. Y. Lee, S. W. Tzeng, and J. T. Horng, "KinasePhos: A web tool for identifying protein kinase-specific phosphorylation sites," Nucleic Acids Res., vol. 33, no. SUPPL. 2, pp. 226–229, 2005.

[46] 5 Alvaro Martinez Barrio, 1, 5 Daniel Soeria-Atmadja, 2, 4 Anders Nistér, 1 Mats G. Gustafsson, 3, 4 Ulf Hammerling, 2,* and Erik Bongcam-Rudloff1, "EVALLER: a web server for in silico assessment of potential protein allergenicity."

[47] K. C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," J. Theor. Biol., vol. 273, no. 1, pp. 236–247, 2011.

[48] L. Jiang, J. Zhang, P. Xuan, and Q. Zou, "BP Neural Network Could Help Improve Pre-MiRNA Identification in Various Species," Biomed Res. Int., vol. 2016, 2016.

[49] Saeed, S.; Mahmood, M. K.; Khan, Y. D., An exposition of facial expression recognition techniques. Neural Computing and Applications 2018, 29 (9), 425-443.

[50] Butt, A. H.; Khan, Y. D., CanLect-Pred: A cancer therapeutics tool for prediction of target cancerlectins using experiential annotated proteomic sequences. IEEE Access 2019, 8, 9520-9531.

[51] Amanat, S.; Ashraf, A.; Hussain, W.; Rasool, N.; Khan, Y. D., Identification of lysine carboxylation sites in proteins by integrating statistical moments and position relative features via general PseAAC. Current Bioinformatics 2020, 15 (5), 396-407.

[52] Ilyas, S., Hussain, W., Ashraf, A., Khan, Y. D., Khan, S. A., & Chou, K. C. (2019). iMethylK-PseAAC: Improving accuracy of lysine methylation sites identification by incorporating statistical moments and position relative features into general PseAAC via Chou's 5-steps rule. Current Genomics, 20(4), 275-292.

[53] Hussain, W.; Rasool, N.; Khan, Y. D., A Sequence-Based Predictor of Zika Virus Proteins Developed by Integration of PseAAC and Statistical Moments. Combinatorial chemistry & high throughput

screening 2020, 23 (8), 797-804.

[54] Khan, Y. D.; Alzahrani, E.; Alghamdi, W.; Ullah, M. Z., Sequence-based Identification of Allergen Proteins Developed by Integration of PseAAC and Statistical Moments via 5-Step Rule. Current Bioinformatics 2020, 15 (9), 1046-1055.

[55] Mahmood, M. K.; Ehsan, A.; Khan, Y. D.; Chou, K.-C., iHyd-LysSite (EPSV): Identifying Hydroxylysine Sites in Protein Using Statistical Formulation by Extracting Enhanced Position and Sequence Variant Feature Technique. Current Genomics 2020, 21 (7), 536-545.

[56] Naseer, S.; Hussain, W.; Khan, Y. D.; Rasool, N., IPhosS (Deep)-PseAAC: Identify phosphoserine sites in proteins using deep learning on general pseudo amino acid compositions via modified 5-Steps rule. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2020.

[57] Naseer, S.; Hussain, W.; Khan, Y. D.; Rasool, N., Sequence-based identification of arginine amidation sites in proteins using deep representations of proteins and PseAAC. Current Bioinformatics 2020, 15 (8), 937-948.

[58] Shah, A. A.; Khan, Y. D., Identification of 4-carboxyglutamate residue sites based on position based statistical feature and multiple classification. Scientific Reports 2020, 10 (1), 1-10.

[59] Awais, M.; Hussain, W.; Rasool, N.; Khan, Y. D., iTSP-PseAAC: Identifying Tumor Suppressor Proteins by Using Fully Connected Neural Network and PseAAC. Current Bioinformatics 2021, 16 (5), 700-709.

[60] Hussain, W.; Rasool, N.; Khan, Y. D., Insights into Machine Learning-based approaches for Virtual Screening in Drug Discovery: Existing strategies and streamlining through FP-CADD. Current Drug Discovery Technologies 2021, 18 (4), 463-472.

[61] Khan, Y. D.; Khan, N. S.; Naseer, S.; Butt, A. H., iSUMOK-PseAAC: prediction of lysine sumoylation sites using statistical moments and Chou's PseAAC. PeerJ 2021, 9, e11581.

[62] Malebary, S. J.; Khan, R.; Khan, Y. D., ProtoPred: Advancing Oncological Research Through Identification of Proto-Oncogene Proteins. IEEE Access 2021, 9, 68788-68797.

[63] Malebary, S. J.; Khan, Y. D., Evaluating machine learning methodologies for identification of cancer driver genes. Scientific reports 2021, 11 (1), 1-13.

[64] Malebary, S. J.; Khan, Y. D., Identification of Antimicrobial Peptides Using Chou's 5 Step Rule. CMC-COMPUTERS MATERIALS & CONTINUA 2021, 67 (3), 2863-2881.

[65] Naseer, S.; Ali, R. F.; Khan, Y. D.; Dominic, P., iGluK-Deep: computational identification of lysine glutarylation sites using deep neural networks with general pseudo amino acid compositions. Journal of Biomolecular Structure and Dynamics 2021, 1-14.

[66] Naseer, S.; Hussain, W.; Khan, Y. D.; Rasool, N., NPalmitoylDeep-PseAAC: A Predictor of N-Palmitoylation Sites in Proteins Using Deep Representations of Proteins and PseAAC via Modified 5-Steps Rule. Current Bioinformatics 2021, 16 (2), 294-305.

[67] Naseer, S.; Hussain, W.; Khan, Y. D.; Rasool, N., Optimization of serine phosphorylation prediction in proteins by comparing human engineered features and deep representations. Analytical Biochemistry 2021, 615, 114069.

[68] Khanum, S., Ashraf, M. A., Karim, A., Shoaib, B., Khan, M. A., Naqvi, R. A., ... & Alswaitti, M. Gly-LysPred: Identification of Lysine Glycation Sites in Protein Using Position Relative Features and Statistical Moments via Chou's 5 Step Rule.

[69] Lv, H., Dao, F. Y., Zhang, D., Yang, H., & Lin, H. (2021). Advances in mapping the epigenetic modifications of 5-methylcytosine (5mC), N6-methyladenine (6mA), and N4-methylcytosine (4mC). Biotechnology and Bioengineering.

[70] Zulfiqar, H., Sun, Z. J., Huang, Q. L., Yuan, S. S., Lv, H., Dao, F. Y., ... & Li, Y. W. (2021). Deep-4mCW2V: A sequence-based predictor to identify N4-methylcytosine sites in Escherichia coli. Methods.

[71] Liu, Y., Wang, X., & Liu, B. (2019). A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. Briefings in bioinformatics, 20(1), 330-346.

[72] Zhang, D., Xu, Z. C., Su, W., Yang, Y. H., Lv, H., Yang, H., & Lin, H. (2021). iCarPS: a computational tool for identifying protein carbonylation sites by novel encoded features. Bioinformatics, 37(2), 171-177.