ORIGINAL ARTICLE

Antifp SRF: Identifying Antifungal Peptides by Sequence Statistical Moments and Random Forest Classifier

Hina Asad

Authors' Affiliation:

Department of Computer Sciences, Abdul Wali Khan University Mardan, Pakistan

Article History:

Submitted: November 18, 2023
Revised: March 05, 2024
Accepted: March 11, 2024
Published online: March 30, 2024

Corresponding author(s):

Name: Hina Asad
Email: hiramehr12@gmail.com

A B S T R A C T

Numerous types of fungus are unhealthy and can lead to life-threatening conditions in people. A fungal infection that targets specific bodily regions and that the immune system is unable to fight off. Numerous studies have been conducted on the topic, but due to its negative side effects, medications, chemicals, and conventional techniques have not proven effective enough. Due to the rapidly invading fungus species, antimicrobial peptides having antifungal properties emerged as a new and powerful candidate due to their efficiency and selectivity. Many mathematical models are being developed for its identification, including Deep-AntiFP and AntiFP, the most recently. However, it has limited performance in accuracy, sensitivity, precision, and MCC. This study investigates an efficient and accurate computational model for the prediction of antifungal peptides (AFPs) using machine learning approaches. The model uses the statistical moment as feature extraction and Random Forest (RF) as a predictor to predict AFPs from non-AFPs. 10-fold cross-validation, independent tests, and self-consistency tests are used to validate the performance of the model. Three datasets are used for model training and its evaluation, namely Antifp Main, Antifp DS1, and Independent Main. The model achieved an accuracy of 95.73%, 97.01%, and 99.15% for Antifp Main, Antifp DS1, and Independent main respectively through independent testing; 97.4%, 97.3%, and 97.4% by 10-fold cross-validation and 99.74%, 99.87 and 99.83 by self-consistency testing. The proposed model has shown excellent results for all three datasets as compared to the existing models available in the literature.

Keywords: Antifungal peptides, Random Forest, Prediction accuracy, Computational modeling, Machine learning, Statistical moment
(MSC2020) Codes: 68T05 , 03B52,

1 | INTRODUCTION

Fungi are organisms that are eukaryotic, multicellular, and heterotrophic. They do not have chloroplast like mold, mushrooms, and yeast [1]. The main human fungal pathogens are Aspergillus, Pneumocystis, and Candida [2]. A lot of development in technologies has been done but still for human

health fungi is a serious issue [3]. In humans, fungi cause severe diseases, about 1.7 million deaths are caused by fungal infection every year according to a report [4]. Traditionally various plants were utilized to treat fungal infections because they produced protected, eco-friendly, and renewable antifungal drugs. The traditional methods have limited action, just toward bacteria and fungi; however, they lack specificity, tolerance, and effectiveness [5].

Because of the above-limited cure for treating the fungal infection, it has been investigated to replace the chemical-based drugs with the desired non-chemical alternative, which is found in the form of peptide-based therapeutics [6]. Peptide-based antimicrobials then emerged as a new and advanced factor in biopharmaceutics to treat fungal infection which can also act as an immune stimulator and suppressant [7]. These antimicrobial peptides were not precise towards fungal infections but were also effective in the chemoattraction of immune cells, cytokine, and histamine production and secretion [8]. Antimicrobial peptides show vast activities against protozoa, show resistance to bacteria, fungi, parasites, and viruses, and even kill cancer cells [9]. Because of its effectiveness for a broad range of species, the development of some specific antifungal peptides was sensed as necessary for the treatment of fungal infections unambiguously and effectively [3]. For the identification of biopeptides such as antimicrobial peptides, anticancer peptides, Antiviral peptides, and Antibacterial peptides traditional methods were used like molecular [10], non-molecular [11], and immunological methods. [12] Worked on Tanzanian plants which exhibit antifungal activity against many fungal species like candida glabrata, Cryptococcus neoformans, and candida krusei. These plants by testing on 38 families show 45% antifungal nature. The strongest AF activities were exhibited by the extracts of Turraea holstii Gurk and Clausena anisata Oliv [13].

Due to its enormous importance, many researchers have developed a computational model using data mining for its identification to overcome the limitations of traditional methods, which are considered time-consuming and costly. [14] Worked on AFPs classification, bioactivity, and its efficiency which emerged as a novel therapy for fungal infections. Also, the synthetic production and exploitation of AFPs were discussed. [15] Using Chou's pseudo amino acid composition with different machine learning algorithms to compare and predict the more accurate model for the prediction of AFPs. Five algorithms were used among which high-performance classifiers were SVM and Bagged-C4.5 shows 80% accuracy. [16] Worked on the screening of antifungal peptides by developing a prediction method using Chou's pseudo amino acid composition along with the SVM classifier for its binomial classification and achieved an accuracy of 94.76%. [17] Developed a model based on SVM as a classifier to predict the antifungal peptides, it shows an accuracy of 88.78% for the training dataset and 83.33%. for testing the dataset. They also investigated the datasets by using binary patterns which achieved 84.88% accuracy through the training dataset and 84.64 through the testing dataset. [18] Investigated the model based on

Naïve Bayes (NB) and decision trees as classifiers working on small data sets, NB gives a robust performance. Furthermore [3] used Support Vector Machine (SVM), Probabilistic neural network (PNN), and k-nearest neighbor (KNN) to examine their efficiency and accuracy in predicting antifungal peptides and achieved an accuracy of 94%.

There is still room for improvement as most of the existing models have used PseAAC (Pseudo Amino Acid Composition) which has some shortcomings in working on small peptide sequences. Most importantly, they all have low performance in terms of accuracy, sensitivity, precision, and MCC. Therefore, in this proposed research a more accurate and efficient model for antifungal peptide is developed, which provides high accuracy in prediction and other performance matrices as compared to the existing models. We use statistical moments for feature extraction and further examine machine learning classifiers: SVM, PNN, K-NN, RF, A-NN, and NB, in these algorithms the best operational model the RF gives the highest accuracy. Furthermore, testing methods like K-Fold cross-validation, independent testing, and self-consistency validation are used to evaluate the accuracy of the proposed model.

2 | METHODS AND MATERIALS

The proposed system was developed to classify antifungal peptides and non-antifungal peptides. The methodology of the proposed system is divided into the following five steps:

Collection of a valid Benchmark dataset.

1. Feature Extraction
2. Machine learning classifiers
3. Methods for evaluating the performance of classifiers.
4. Results and Discussions

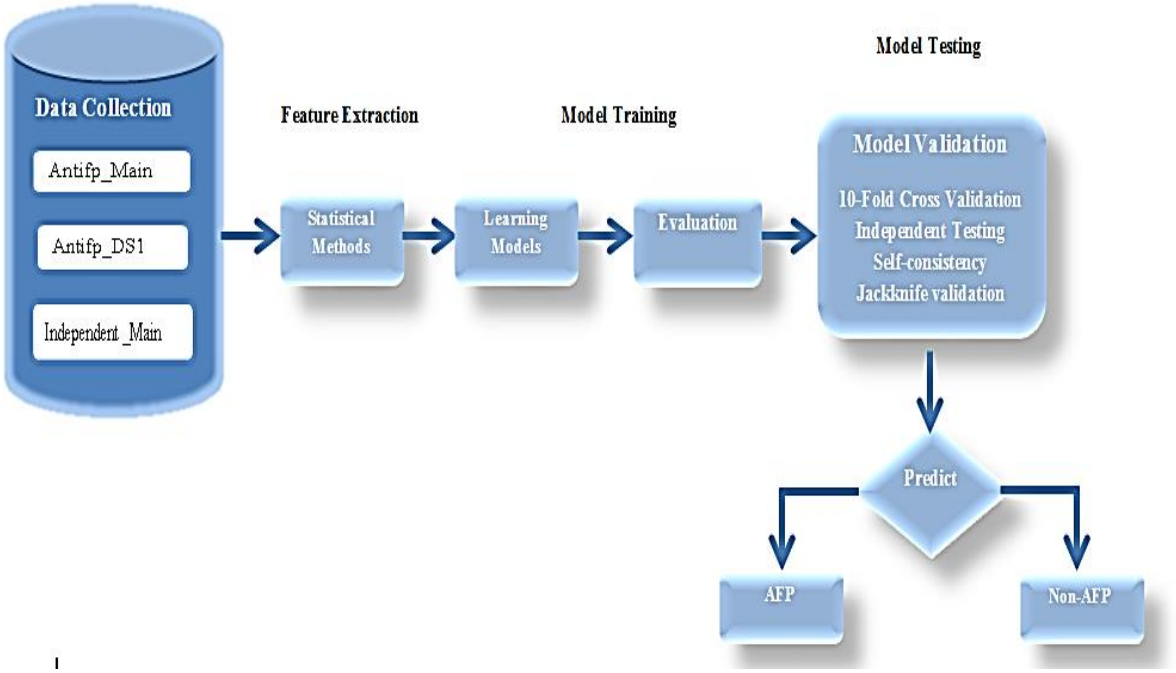

Figure 1: Proposed Model

2.1 | Benchmark Dataset

An appropriate benchmark dataset is essential for the development of good computational model. The dataset is retrieved from the Swiss-Prot and the DRAM, further, it is divided into the following three groups namely Antifp_Main, Antifp_DS1, and Independent_Main as described below and classified in Table 1 used by [3].

Antifp Main

This dataset consists of 2336 peptide sequences, among which 1168 are negative (non-antifungal peptides) samples of antimicrobial peptides other than antifungal and random peptides collected from Swiss-Port, and 1168 are positive samples which are exclusive Antifungal peptides collected from DRAMP.

Antifp DS1

The dataset contains 2336 peptide sequences, among which 1168 are negative (non-antifungal peptides) samples which are antimicrobial peptides other than antifungal and 1168 are positive samples which are exclusive Antifungal peptides.

Independent Main

This dataset consists of 582 peptide sequences where 291 are positive peptides (Antifungal) collected from DRAMP and 291 are negative peptide (non-antifungal) sequences collected from Swiss-Prot.

Table 1: Datasets Classification Summary

Datasets	No of Peptide	Database	Peptide Sequences Nature	Model Prediction
Antifp_Main	1168 AFPs (Positive)	DRAMP	exclusive Antifungal peptides	0
	1168 Non-AFPs (Negative)	Swiss-Port	antimicrobial peptides other than antifungal and random peptides	1
Antifp_DS1:	1168 AFPs (Positive)	DRAMP	exclusive Antifungal peptides	0
	1168 Non-AFPs (Negative)	Swiss-Port	Antimicrobial peptides other than antifungal	1
Independent _Main:	291 AFPs (Positive)	DRAMP	exclusive Antifungal peptides	0
	291 Non-AFPs (Negative)	Swiss-Port	antimicrobial peptides other than antifungal and randomly collected	1

2.1 | Feature Vector Construction

Now a day biological data and sequences are growing wide in almost every phase of life. Biological analysis has become an essential part in terms of solving biological problems. The hardest task for one is to express the biological data and sequences into a vector or discrete form without dropping sequence pattern information and its characteristics or components for final analysis. Vector formulation is an important task because all computer-based models such as RF (Random Forest Algorithm) [19], KNN [20], and SVM (Support Vector Machine) [21] use the vector dataset for prediction and evaluation. However, to avoid the problem of data loss during the formulation, many experts and researchers have used the world-wide famous feature extraction methods like Amino Acid Composition (AAC) and Pseudo Amino Acid Composition (PseAAC) [22] statistical moments. This proposed model has been developed by using statistical moment [23] which consists of the raw moments, Hahn and central moments, and other moments are discussed in detail as stated below.

2.1.1 | Statistical Moment Calculation

The same results and conclusions are provided by the central moments. Next to the centroid of data, these moments are calculated consequently which are incessant to their location according to the centroid but on the other hand, are scale variants. Hahn polynomials are consequent from Hahn moments so; these are neither location nor scale variants. The major goal and purpose of the calculation of statistical moment is to find the size of the dataset, and the composition and position of protein sequences. This method is used to extract different features of protein.

Many moments have been explained by mathematicians and statisticians based on polynomials and distribution functions, definitely well-known [24]. In these observations, the raw moment, central moment [25], and Hahn moments are calculated [26]. Further, the two-dimensional matrix of order n x n is based on the protein sequence P which is expressed as below:

$$P = \{\alpha_1, \alpha_2 \ldots\ldots \alpha_k\} \qquad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where $\boldsymbol{\alpha i}$ is the $\mathbf{i^{th}}$ amino acid remains or residues parts in prime sequence containing j residues, further,

$$n = \lceil \sqrt{k} \rceil \qquad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

2.1.2| Determination of Position Relative Incidence Matrix (PRIM)

The order of protein sequences represents significant information for the establishment of any mathematical form or model. As earlier articulated, one of the very important physical properties of the protein is the relative placement of amino acid residues. Consequently, relative position is important in perceiving how amino acids are interconnected in the polypeptide chain. To quantize the relative position of the specific protein sequence is very important [27]. Therefore, the Position Relative Incidence Matrix (PRIM) is constructed and is a 20 by 20 matrix. Using this matrix the features that are related to the relative position of amino acid residue are calculated,

$$\epsilon_{PRIM} = \begin{bmatrix} \varepsilon_{1\to1}, & \varepsilon_{1\to2}, \ldots & \varepsilon_{1\to j}, \ldots & \varepsilon_{1\to20} \\ \varepsilon_{2\to1}, & \varepsilon_{2\to2}, \ldots & \varepsilon_{2\to j}, \ldots & \varepsilon_{2\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \varepsilon_{i\to1}, & \varepsilon_{i\to2}, \ldots & \varepsilon_{i\to j}, \ldots & \varepsilon_{i\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \varepsilon_{n\to1}, & \varepsilon_{n\to2}, \ldots & \varepsilon_{n\to j}, \ldots & \varepsilon_{n\to20} \end{bmatrix} \qquad \ldots\ldots\ldots\ldots\ldots\ldots (13)$$

2.1.3| Determination of Reverse Position Relative Incidence Matrix (PRIM)

Some primary sequence of protein has incomprehensible unknown features. Our work is to find this hidden quality of protein with homologous ambiguities in protein sequence and for this purpose, we implemented the Reverse Position Relative Incidence Matrix (RPRIM) which produces a similar matrix (20x20 dimensions) of 400 coefficients as PRIM but for the reversed sequence [19] or order. RPRIM can be defined as,

$$\epsilon_{RPRIM} = \begin{bmatrix} \varrho_{1\to1}, & \varrho_{1\to2}, \ldots & \varrho_{1\to j}, \ldots & \varrho_{1\to20} \\ \varrho_{2\to1}, & \varrho_{2\to2}, \ldots & \varrho_{2\to j}, \ldots & \varrho_{2\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \varrho_{i\to1}, & \varrho_{i\to2}, \ldots & \varrho_{i\to j}, \ldots & \varrho_{i\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \varrho_{n\to1}, & \varrho_{n\to2}, \ldots & \varrho_{n\to j}, \ldots & \varrho_{n\to20} \end{bmatrix} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (14)$$

2.1.5| Generating Accumulative Absolute Position Incidence Vector (AAPIV)

AAPIV embraces the cumulative recurrence event of amino acids. Further, data regarding the creation of the protein is held in this. This filters the information regarding the relative positioning of amino acid residues. To extract compositional information from sequences, the frequency matrix is calculated, but now the problem came that it did not provide in sequence the relative positions of the residues, hence, the cumulative absolute impact vector of the position (AAPIV) of the length twenty 20 elements is calculated.

2.2 | Machine Learning Classifiers

To identify and classify the antifungal peptides and non-antifungal peptides, we used machine learning classification algorithms. Artificial Neural Network and Random Forest are used to analyze the data set. Among these classification algorithms, the random forest classifier provides the highest accuracy

2.2.1| Prediction Model

Random forest is the technique that is more efficient to operate large data sets. Random forest is a robust machine learning algorithm consisting of many decision trees and performs various tasks, including regression and classification [28]. It is flexible and easy to use. A random forest classifier is initially proposed by [29]. It contains many decision trees and trains each tree on a different sample of observation, so it gives high accuracy. Random forest alleviates the issue of overfitting by an average of the prediction result given by different decision trees.

The following steps are used to complete the random forest algorithm.

1. Random forest algorithms work on random samples from the given dataset.
2. A decision tree will be created by the algorithm for each sample then these decision trees will give us prediction results.
3. For the predicted result voting will be done. The mode will be used for the classification problem and the Mean for the regression problem.
4. In the last step, the majority voting result will be selected by random forest as the final prediction.

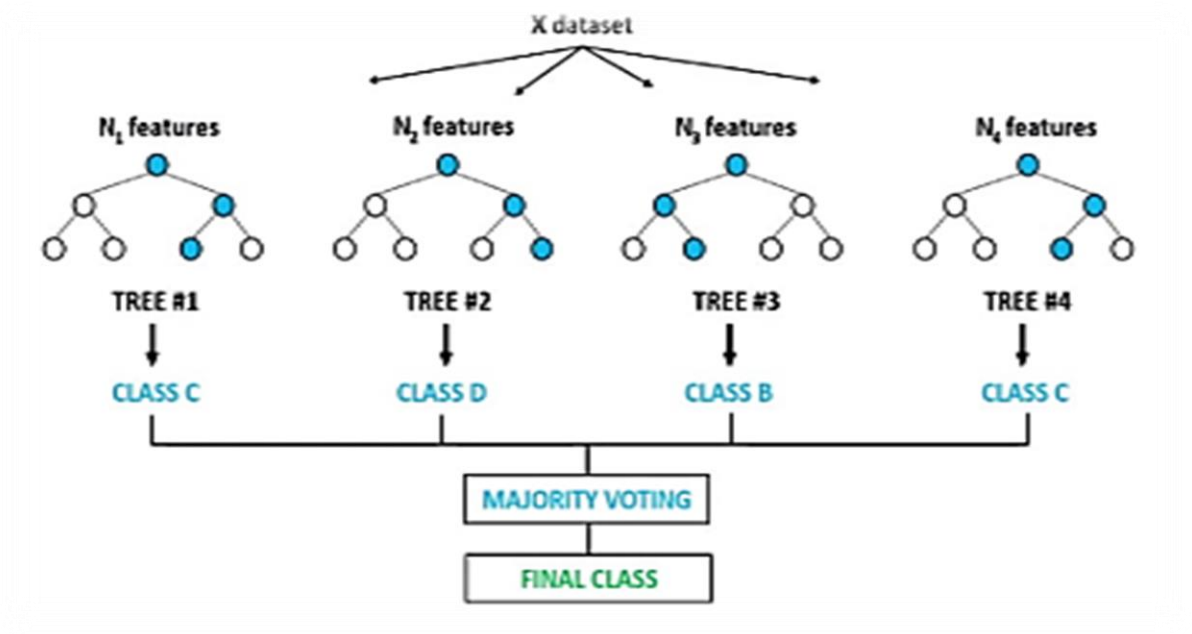

Figure 2: Random Forest Classifier

3 | PERFORMANCE METRICS

Machine learning uses various evaluation metrics to measure classifier performance. In this study, the classifiers are assessed based on the following measurement metrics: Accuracy (ACC), Sensitivity (SEN), Specificity (SPC), and Mathews Correlation Coefficient (MCC). The following formulas are commonly used to calculate the parameters discussed above.

$$\text{ACC} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{SEN} = \frac{TP}{TP+FN}$$

$$\text{SPC} = \frac{TN}{TN+FP}$$

$$\text{MCC} = \frac{(TP \text{ X } TN - FP \text{ X } FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

4 | RESULTS AND DISCUSSION

4.1. | Testing Methods

In machine learning, we evaluate intelligent model outcomes by using several cross-validation Techniques. In this research K-fold cross-validation, independent test, self-consistency, and jackknife test are used to check the success rates of our proposed model.

4.1.1 | K-fold Cross-validation

To evaluate the performance of the proposed model, we use a 10 10-fold cross-validation test. Herein 10 folds sequences in datasets were split randomly into 10 different folds uniformly. Further model is trained for (10-i) folds and test for the remaining i^{th} fold where i=1, 2, 3,..,10. Similarly, the process is

repeated for each i, and lastly, the overall performance is calculated by averaging the results of all the 10 folds, as described mathematically as follows.

A is the combination of all samples in a dataset which may belong to a positive class or negative class:

$$A = \{a_1, a_2, a_3 \dots a_n\}$$

Here Si represents the comparable size of the independent k sets of datasets such that

$$|S_i| \cong |S_j|$$

and

$$\bigcup_{i=1}^{k} S_i = A$$

also

$$\bigcap_{i=1}^{k} S_i = \emptyset$$

Further, training is performed on the set {S1, S2, S3,…, Si-1, Si+1,...Sk} whereas testing on the remaining one set Si, where, i= 1,2,3,4,….k which gives computation of Ri, in the ith iteration of cross-validation. Then the average is taken of all the k values which are represented by Ra.

$$R_a = \frac{\sum_{i=1}^{k} R_i}{k}$$

The performance measures can be calculated when these iterations are, once done.

Using K folds cross-validation K=10, the results will be calculated by the RF predictor shown in the Table below.

Table 2: 10-Fold Cross Validation

K-Folds	Antifp_Main				Antifp_DS1				Independent _Main			
	Acc	**Spe**	**Sen**	**MCC**	**Acc**	**Spe**	**Sen**	**MCC**	**Acc**	**Spe**	**Sen**	**MCC**
Fold 1	82.5	83.8	81.2	0.6	82.1	83.8	80.3	0.6	86.4	83.3	89.7	0.7
Fold 2	99.2	99.2	99.2	1.0	100.0	100.0	100.0	1.0	98.3	100.0	96.7	1.0
Fold 3	99.2	99.2	99.2	1.0	99.2	100.0	98.3	1.0	98.3	100.0	96.6	1.0
Fold 4	98.7	98.3	99.2	1.0	99.6	100.0	99.2	1.0	98.3	96.6	100.0	1.0
Fold 5	98.7	97.4	100.0	1.0	98.3	99.2	97.4	1.0	98.3	100.0	96.6	1.0
Fold 6	99.2	99.2	99.2	1.0	97.9	97.4	98.3	1.0	100.0	100.0	100.0	1.0
Fold 7	99.1	99.1	99.2	1.0	99.1	98.3	100.0	1.0	100.0	100.0	100.0	1.0
Fold 8	99.1	99.1	99.2	1.0	99.1	98.3	100.0	1.0	100.0	100.0	100.0	1.0
Fold 9	98.7	99.2	98.3	1.0	98.7	99.2	98.3	1.0	96.6	100.0	93.1	0.9
Fold 10	99.1	98.3	100.0	1.0	99.6	99.2	100.0	1.0	98.3	96.6	100.0	1.0
AVG	97.4	97.3	97.4	0.9	97.3	97.5	97.2	0.9	97.4	97.6	97.3	0.9

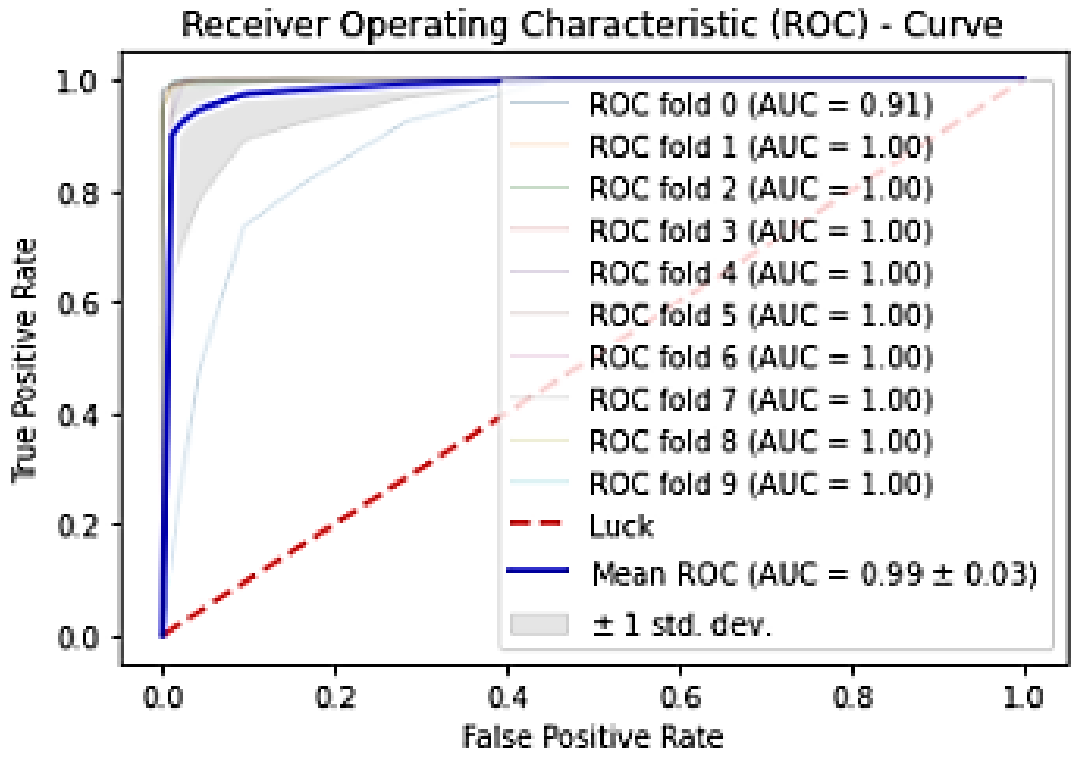

Figure 3: 10-Fold Cross Validation ROC Curve Antifp_Main

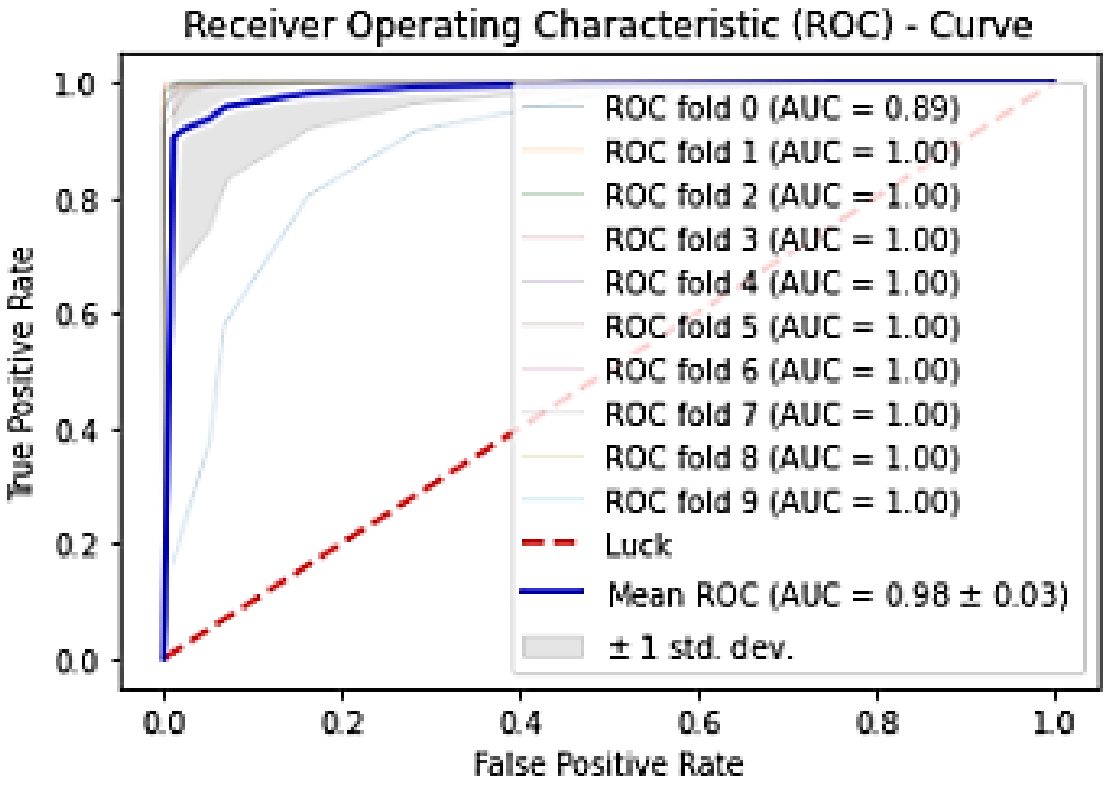

Figure 4: 10-Fold Cross Validation ROC Antifp_DS1

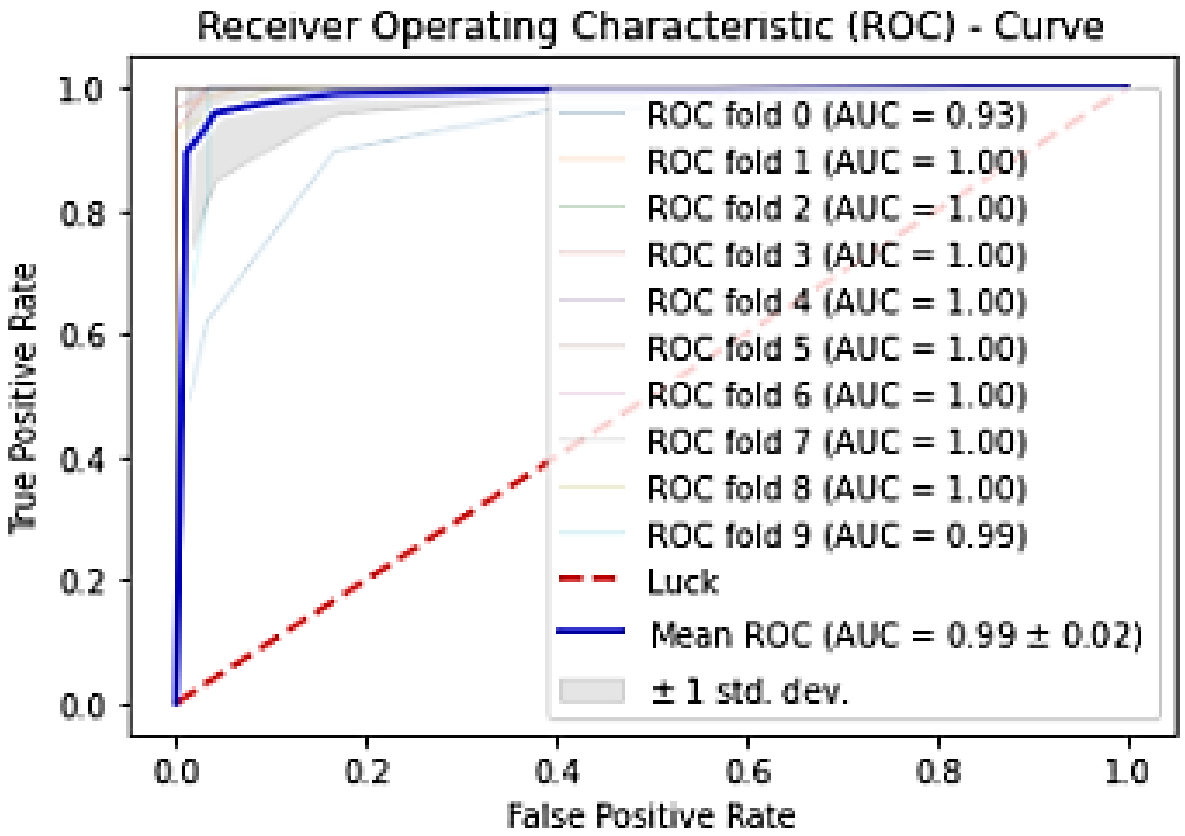

Figure 5: 10-Fold Cross Validation ROC Independent _Main

4.1.2 | Independent Test

In this cross-validation method, the predictor's performance is checked against the said criteria [32]. Dataset split into two parts one is training and the other is testing. Training data is used to train the predictor and the remaining part is used for testing. In this research, we split our data into two parts, 80% used for training and 20% used for testing. RF predictor calculates the accuracy of 95.2% on training data and 95.7% on the test data, as shown in Fig 2.

Table 3: Independent Testing Results

Independent Testing Results	Antifp_Main				Antifp_DS1				Independent _Main			
	Acc	Spe	Sen	MCC	Acc	Spe	Sen	MCC	Acc	Spe	Sen	MCC
Training (80%)	95.2	1.0	1.0	1.0	96.4	1.0	1.0	1.0	98.7	1.0	1.0	1.0
Testing (20%)	95.7	1.0	1.0	1.0	97.0	1.0	1.0	1.0	99.2	1.0	1.0	1.0

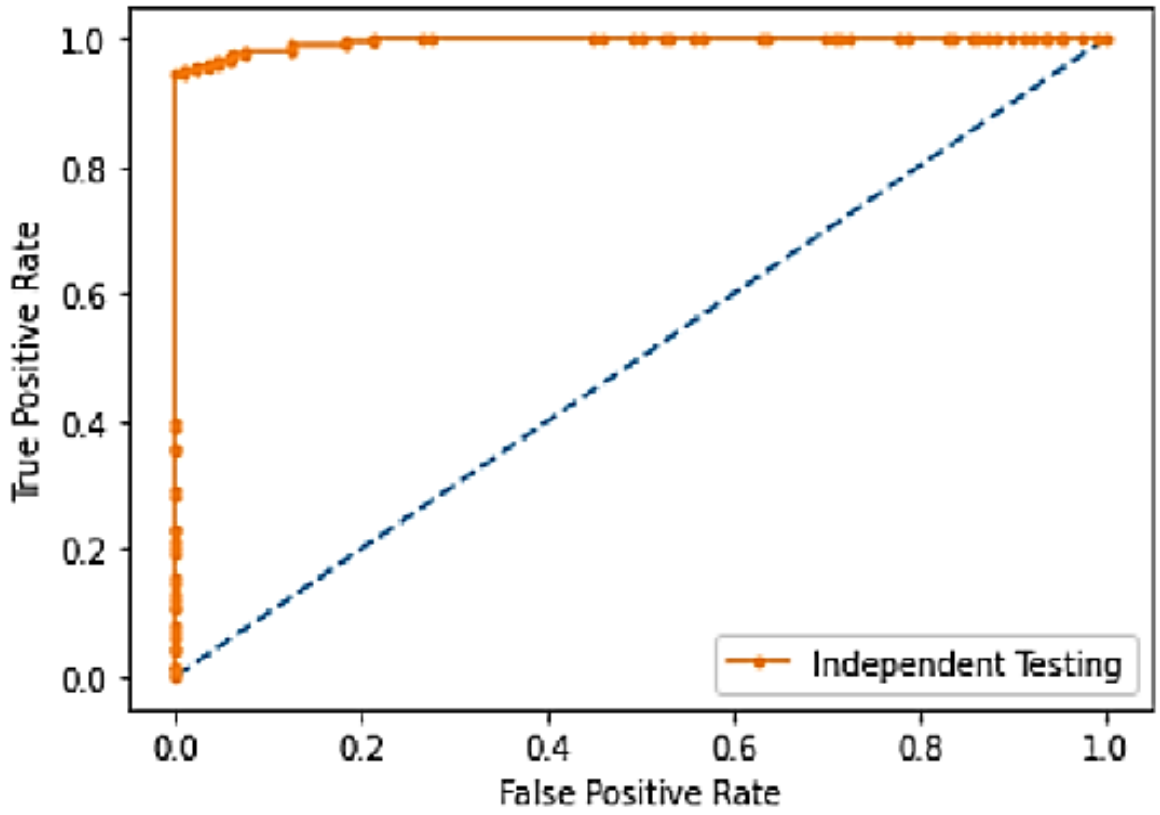

Figure 6: Independent Testing ROC Curve Antifp_Main

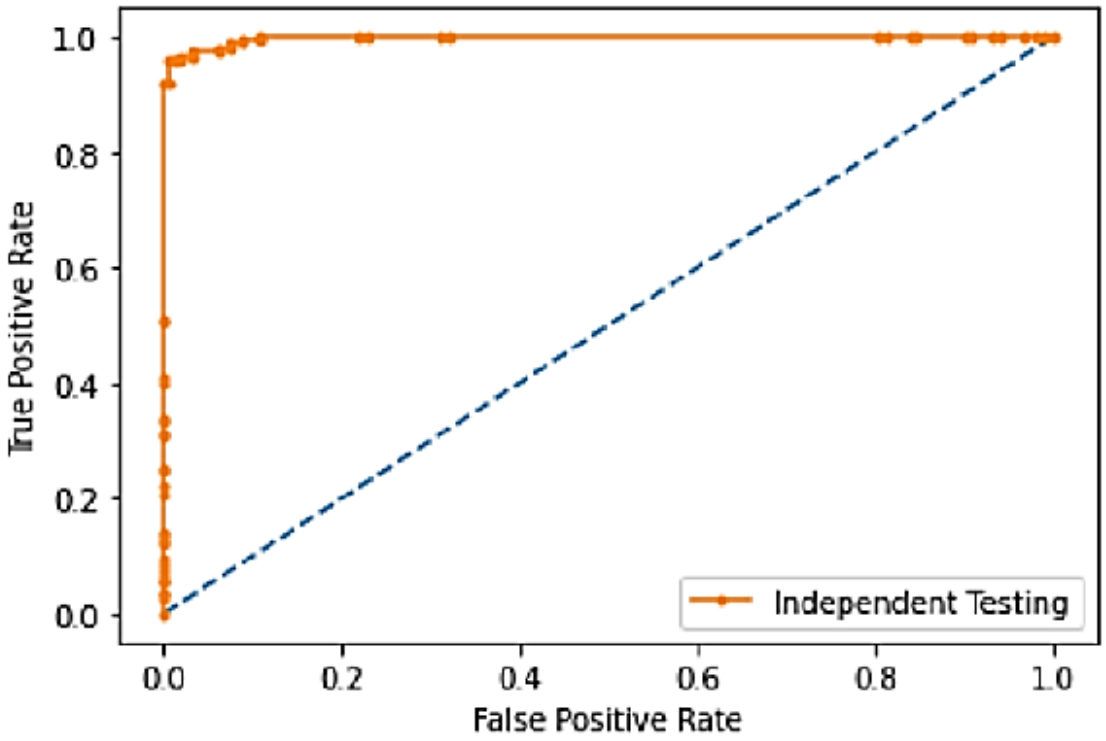

Figure 7: Independent Testing ROC Antifp_DS1

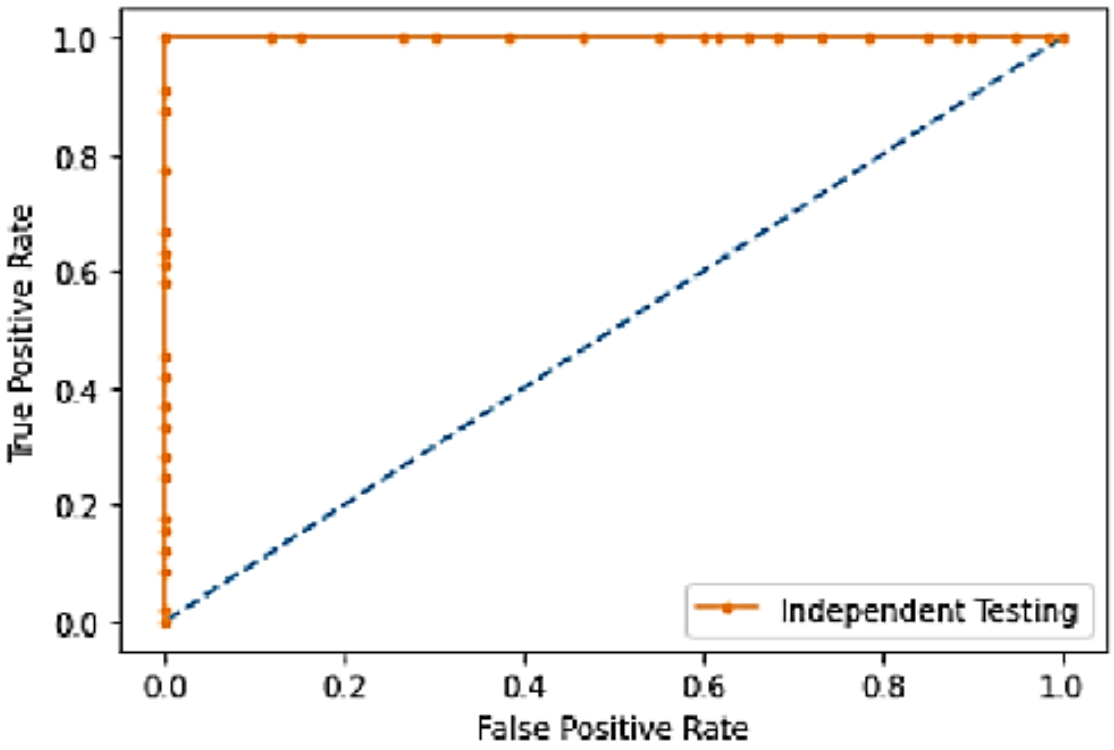

Figure 8: Independent Testing ROC Independent _Main

4.1.3 | Self-consistency

In self-consistency, the same dataset is used for training and testing the model. With the help of decision rules, the training dataset and the characteristics of the dataset will be predicted, in a self-consistency test. In this, we achieved the accuracy, specificity, sensitivity and MCC are 99.7%, 1.0%, 1.0%, and 1.0%, respectively. To evaluate the model's performance Area under the Curve (AUC) and Receiver Operating Characteristics (ROC) [30] are used, whereas AUC shows the predictor performance and the ROC curve is a two-dimensional graph in which true positive rate (TPR) is plotted on the Y-axis and the false positive rate (FPR) is plotted on X-axis. ROC curve shows the compromise between sensitivity (TPR) and specificity (1-TPR). ROC curves are valuable to visualize and indicate better performance when the classifier gives the curve closer to the upper left corner.

Table 4: Self-Consistency Results

Self-Consistency Results	Antifp_Main				Antifp_DS1				Independent _Main			
	Acc	Spe	Sen	MCC	Acc	Spe	Sen	MCC	Acc	Spe	Sen	MCC
	99.7	1.0	1.0	1.0	99.9	1.0	1.0	1.0	99.8	1.0	1.0	1.0

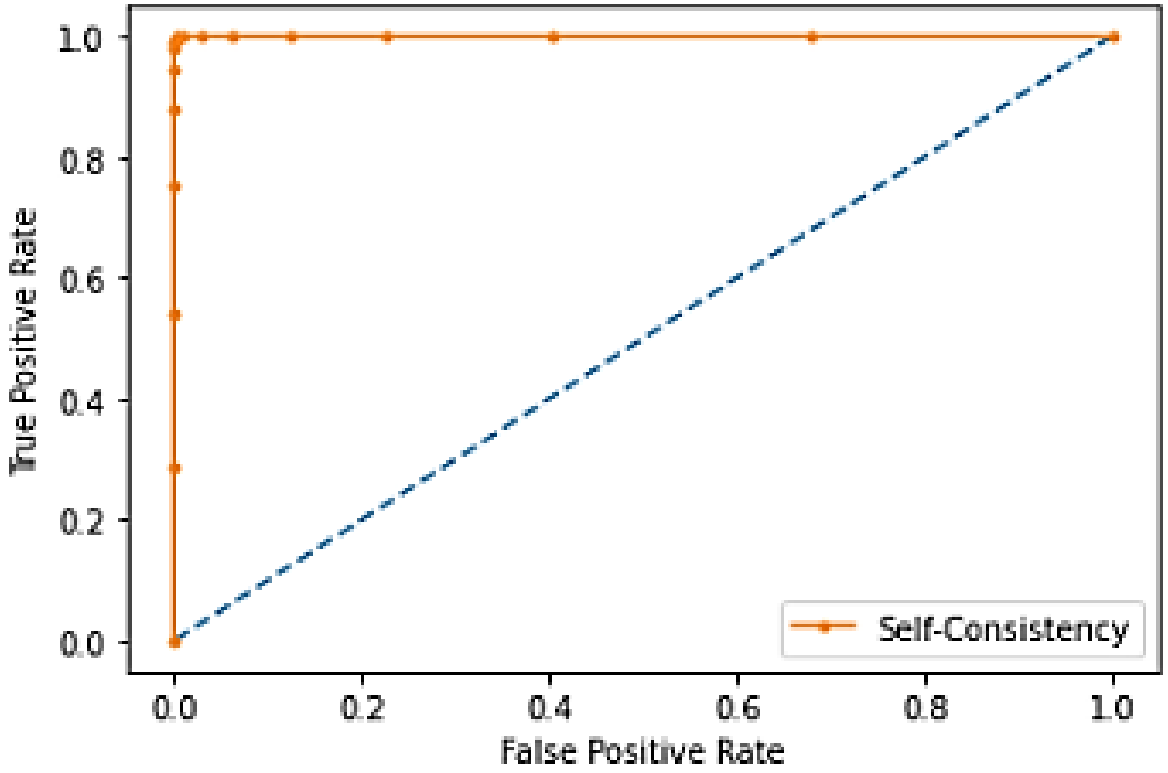

Figure 9: Self-Consistency ROC Curve Antifp_Main

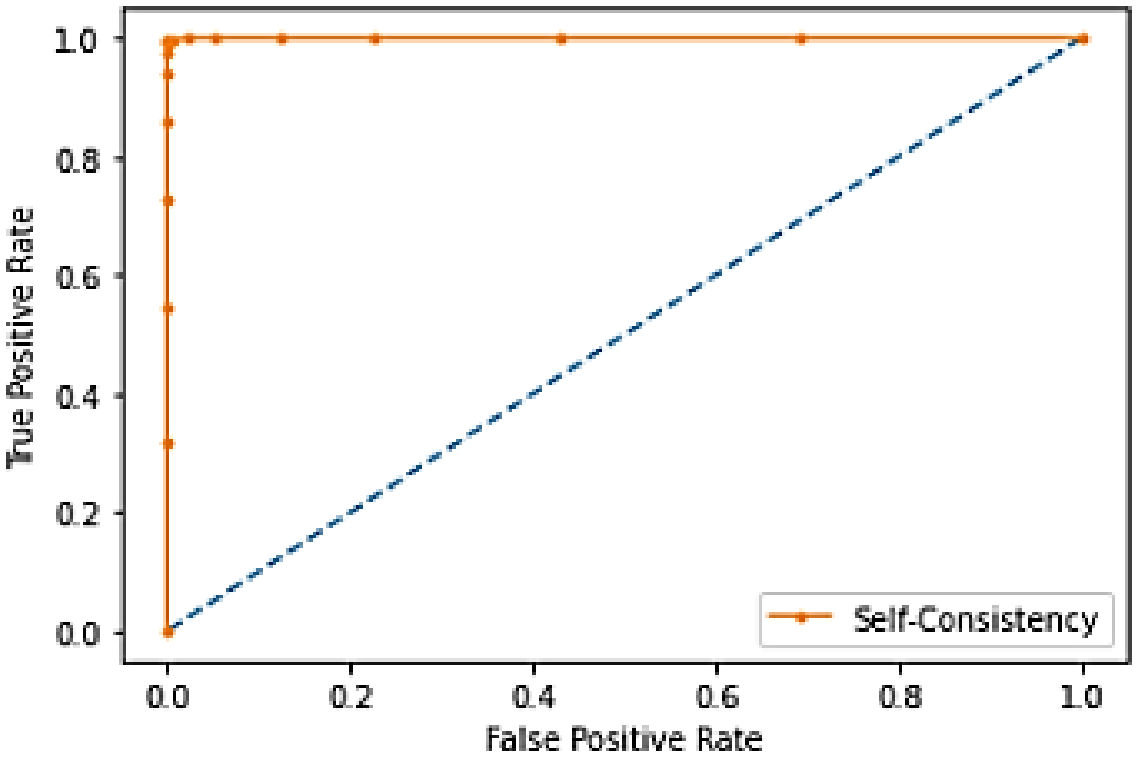

Figure 10: Self-Consistency ROC Antifp_DS1

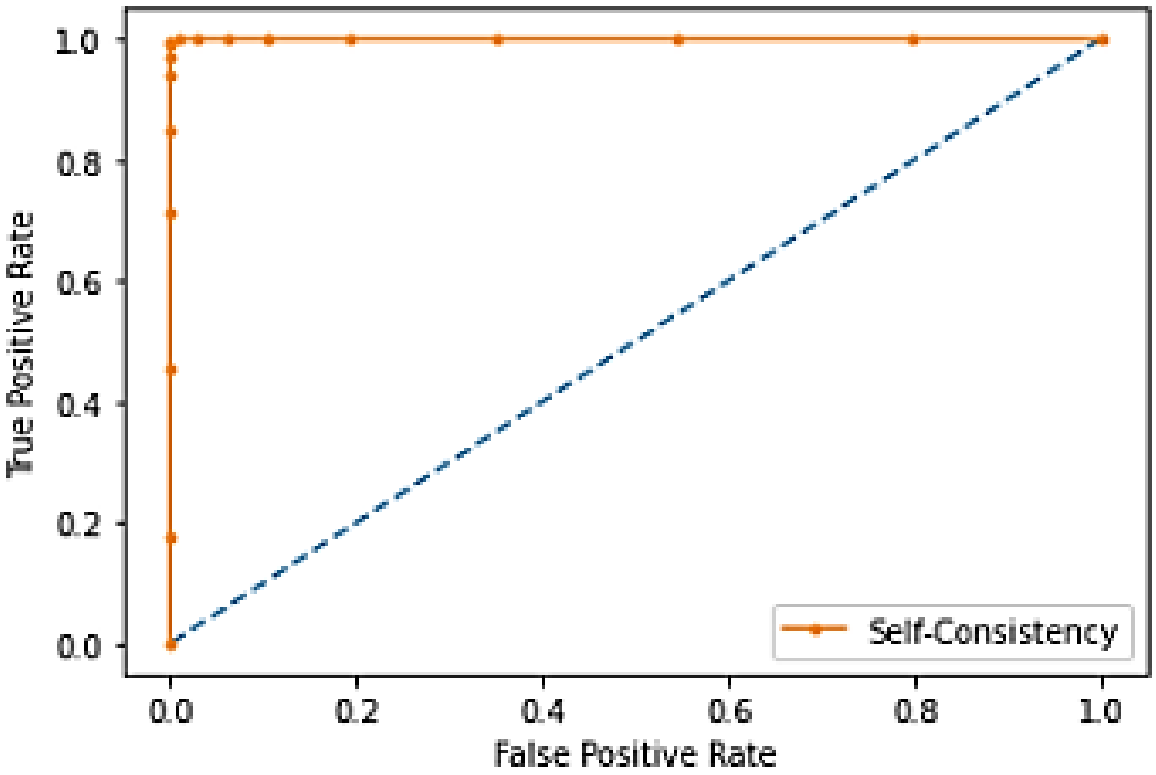

Figure 11: Self-Consistency ROC Independent _Main

4.2 | Comparison of Proposed Model Performance with Existing Models

The performance of the proposed model and existing models are provided in Table 5. Based on the same datasets we compared our proposed model with [17] and [3]. However, [17] used Amino acid composition, Split Amino acid composition, and dipeptide composition, and [3]CPP and QSO were used for feature extraction. On the other hand, our proposed model uses statistical moment for feature extraction and Random Forest as a classifier. After feature extraction of the Antifp_main dataset, [17] has an accuracy of 88% on the SVM classifier and [3] has 94% accuracy via DNN, whereas our proposed model has achieved 97.4% accuracy with Random Forest. By using Antifp_DS1 [17] achieved 86.26% accuracy; [3] reported 91.02% accuracy and our proposed model achieved a 97.3% success rate. In the end, the independent dataset achieved 97.4% accuracy which is higher than the existing models.

Table 5: Comparison of the proposed model with the existing model

Dataset	Prediction Model	Accuracy%	Sensitivity	Specificity	MCC
Antifp_Main	Agrawal et.al [19]	88.27	88.61	87.93	0.77
	A. Ahmad et.al [10]	94.23	94.85	93.97	0.89
	Proposed Model for Antifp	97.4	97.3	97.4	0.9
Antifp_DS1	Agrawal et.al [19]	86.26	86.90	85.62	0.73
	A. Ahmad et.al [10]	91.02	91.19	88.72	0.83
	Proposed Model for Antifp	97.3	97.5	97.2	0.9
Independent Dataset	Agrawal et.al [19]	86.25	86.6	85.91	0.73
	A. Ahmad et.al [10]	89.08	86.64	87.79	0.78
	Proposed Model for Antifp	97.4	97.6	97.3	0.9

5 | CONCLUSION

In this research, we developed an effective computational model to predict antifungal peptides. Statistical moments were applied for the feature extraction of AFP's sequences. For the identification of AFP's and Non- AFP's Random Forest algorithm is used. Furthermore, for the model validation, we used several techniques such as K-Fold cross-validation, Independent Testing, and Self-Consistency Testing. The results obtained from the Antifp main dataset using K-Fold cross-validation accuracy is 97.4%, Independent Testing is 95.73%, and Self-Consistency has 99.74% accuracy. Antifp_DS1 obtained 97.3% over K-Fold cross-validation, 97.01% over Independent Testing, and 99.87% over Self-Consistency Testing. In the end, independent main dataset obtained 97.4% results through K-Fold cross-validation, 99.15% through Independent Testing, and 99.83% through Self-Consistency Testing. So our proposed model gives higher accuracy than existing models that's why our model is best for the prediction of AFPs and Non-AFP

REFERENCES

[1] K. Kavanagh, Ed., *Fungi: Biology and Applications*, 1st ed. Wiley, 2011. doi: 10.1002/9781119976950.

[2] D. Sanglard, "Emerging Threats in Antifungal-Resistant Fungal Pathogens," *Front. Med.*, vol. 3, Mar. 2016, doi: 10.3389/fmed.2016.00011.

[3] A. Ahmad *et al.*, "Deep-AntiFP: Prediction of antifungal peptides using distanct multi-informative features incorporating with deep neural networks," *Chemometrics and Intelligent Laboratory Systems*, vol. 208, p. 104214, Jan. 2021, doi: 10.1016/j.chemolab.2020.104214.

[4] F. Bongomin, S. Gago, R. Oladele, and D. Denning, "Global and Multi-National Prevalence of Fungal Diseases—Estimate Precision," *JoF*, vol. 3, no. 4, p. 57, Oct. 2017, doi: 10.3390/jof3040057.

[5] P. M. Kilima, I. Ostermayer, M. Shija, M. M. Wolff, and P. J. Evans, "Drug utilization, prescribing habits and patients in City Council Health Facilities, Dar es Salaam, Tanzania," *DUHP, Swiss Tropical Institute, Basel*, vol. 19, 1993.

[6] P. Kapoor, H. Singh, A. Gautam, K. Chaudhary, R. Kumar, and G. P. S. Raghava, "TumorHoPe: A Database of Tumor Homing Peptides," *PLoS ONE*, vol. 7, no. 4, p. e35187, Apr. 2012, doi: 10.1371/journal.pone.0035187.

[7] A. M. Nicola *et al.*, "Antifungal drugs: New insights in research & development," *Pharmacology & Therapeutics*, vol. 195, pp. 21–38, Mar. 2019, doi: 10.1016/j.pharmthera.2018.10.008.

[8] K. L. Brown and R. E. Hancock, "Cationic host defense (antimicrobial) peptides," *Current Opinion in Immunology*, vol. 18, no. 1, pp. 24–30, Feb. 2006, doi: 10.1016/j.coi.2005.11.004.

[9] R. E. W. Hancock and D. S. Chapple, "Peptide Antibiotics," *Antimicrob Agents Chemother*, vol. 43, no. 6, pp. 1317–1323, Jun. 1999, doi: 10.1128/AAC.43.6.1317.

[10] K. Kuba, "[Molecular and immunological methods applied in diagnosis of mycoses]," *Wiad Parazytol*, vol. 54, no. 3, pp. 187–197, 2008.

[11] M. Arvanitis, T. Anagnostou, B. B. Fuchs, A. M. Caliendo, and E. Mylonakis, "Molecular and Nonmolecular Diagnostic Methods for Invasive Fungal Infections," *Clin Microbiol Rev*, vol. 27, no. 3, pp. 490–526, Jul. 2014, doi: 10.1128/CMR.00091-13.

[12] O. J. M. Hamza *et al.*, "Antifungal activity of some Tanzanian plants used traditionally for the treatment of fungal infections," *Journal of Ethnopharmacology*, vol. 108, no. 1, pp. 124–132, Nov. 2006, doi: 10.1016/j.jep.2006.04.026.

[13] P. D. Khot and D. N. Fredricks, "PCR-based diagnosis of human fungal infections," *Expert Review of Anti-infective Therapy*, vol. 7, no. 10, pp. 1201–1221, Dec. 2009, doi: 10.1586/eri.09.104.

[14] M. Fernández De Ullivarri, S. Arbulu, E. Garcia-Gutierrez, and P. D. Cotter, "Antifungal Peptides as Therapeutic Agents," *Front. Cell. Infect. Microbiol.*, vol. 10, p. 105, Mar. 2020, doi: 10.3389/fcimb.2020.00105.

[15] M. Mousavizadegan and H. Mohabatkar, "An Evaluation on Different Machine Learning Algorithms for Classification and Prediction of Antifungal Peptides," *MC*, vol. 12, no. 8, pp. 795–800, Oct. 2016, doi: 10.2174/1573406412666160229150823.

[16] M. Mousavizadegan and H. Mohabatkar, "Computational prediction of antifungal peptides via Chou's PseAAC and SVM," *J. Bioinform. Comput. Biol.*, vol. 16, no. 04, p. 1850016, Aug. 2018, doi: 10.1142/S0219720018500166.

[17] P. Agrawal, S. Bhalla, K. Chaudhary, R. Kumar, M. Sharma, and G. P. S. Raghava, "In Silico Approach for Prediction of Antifungal Peptides," *Front. Microbiol.*, vol. 9, p. 323, Feb. 2018, doi: 10.3389/fmicb.2018.00323.

[18] J. Sperschneider, P. N. Dodds, D. M. Gardiner, K. B. Singh, and J. M. Taylor, "Improved prediction of fungal effector proteins from secretomes with EffectorP 2.0," *Molecular Plant Pathology*, vol. 19, no. 9, pp. 2094–2110, Sep. 2018, doi: 10.1111/mpp.12682.

[19] S. Khanum *et al.*, "Gly-LysPred: Identification of Lysine Glycation Sites in Protein Using Position Relative Features and Statistical Moments Via Chou's 5 Step Rule," *Computers, Materials & Continua*, vol. 66, no. 2, pp. 2165–2181, 2021, doi: 10.32604/cmc.2020.013646.

[20] S. Mishra, C. N. Bhende, and B. K. Panigrahi, "Detection and Classification of Power Quality Disturbances Using S-Transform and Probabilistic Neural Network," *IEEE Trans. Power Delivery*, vol. 23, no. 1, pp. 280–287, Jan. 2008, doi: 10.1109/TPWRD.2007.911125.

[21] "An Introduction to Support Vector Machines (SVM)." Accessed: Feb. 01, 2021. [Online]. Available: https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/

[22] I. Limongelli, S. Marini, and R. Bellazzi, "PaPI: Pseudo amino acid composition to score human protein-coding variants," *BMC Bioinformatics*, vol. 16, no. 1, p. 123, Dec. 2015, doi: 10.1186/s12859-015-0554-8.

[23] "statistical moment - Google Search." Accessed: Feb. 02, 2021. [Online]. Available: https://www.google.com/search?q=statistical+moment&oq=statistical+moment&aqs=chrome..69i57j0l5j0i20i263i395j69i60.5630j1j7&sourceid=chrome&ie=UTF-8

[24] Z. Ju and S.-Y. Wang, "Prediction of citrullination sites by incorporating k-spaced amino acid pairs into Chou's general pseudo amino acid composition," *Gene*, vol. 664, pp. 78–83, Jul. 2018, doi: 10.1016/j.gene.2018.04.055.

[25] Y. D. Khan, N. Rasool, W. Hussain, S. A. Khan, and K.-C. Chou, "iPhosY-PseAAC: identify phosphotyrosine sites by incorporating sequence statistical moments into PseAAC," *Mol Biol Rep*, vol. 45, no. 6, pp. 2501–2509, Dec. 2018, doi: 10.1007/s11033-018-4417-z.

[26] Y. D. Khan, S. A. Khan, F. Ahmad, and S. Islam, "Iris recognition using image moments and k-Means algorithm," *The Scientific World Journal*, vol. 2014, 2014, doi: 10.1155/2014/723595.

[27] M. A. Akmal, N. Rasool, and Y. D. Khan, "Prediction of N-linked glycosylation sites using position relative features and statistical moments," *PLoS ONE*, vol. 12, no. 8, p. e0181966, Aug. 2017, doi: 10.1371/journal.pone.0181966.

[28] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How Many Trees in a Random Forest?," in *Machine Learning and Data Mining in Pattern Recognition*, vol. 7376, P. Perner, Ed., in Lecture Notes in Computer Science, vol. 7376. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 154–168. doi: 10.1007/978-3-642-31537-4_13.

[29] Tin Kam Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998, doi: 10.1109/34.709601.

[30] D. K. McClish, "Analyzing a Portion of the ROC Curve," *Med Decis Making*, vol. 9, no. 3, pp. 190–195, Aug. 1989, doi: 10.1177/0272989X8900900307.