iAFP-ET: A robust approach for accurate identification of antifungal peptides using extra tree classifier and multi-view fusion

Ali Raza[a], Ashfaq Ahmad[a*], Zafar Iqbal[a], Qadeer Yasin[a], Hamza Javed[a], Adil Shah[a], Sanya Chaudhary[a]

[a]Department of Computer Science, MY University, Islamabad, Pakistan

Abstract

Antifungal peptides (AFPs) have emerged as promising alternatives to conventional antifungal agents due to their broad-spectrum activity, low toxicity, and reduced propensity for resistance development. However, it is still a challenging task to quickly identify possible AFPs from large protein databases. The current antifungal therapies and medications are widely acknowledged as insufficient due to their associated adverse effects. To assess the effectiveness of AFPs in the human system, developing a reliable, intelligent model becomes imperative for the meticulous and precise identification of these peptides endowed with antifungal properties. Therefore, developing a machine learning framework is imperative to identify AFP effectively. In this paper, we present a novel approach for the identification of antifungal peptides using the Extra Tree Classifier with fusion features, including amino acid composition (AAC), dipeptide composition (DPC), and pseudo amino acid composition (PseAAC). Combining these hybrid features enhances classification accuracy and facilitates the efficient screening of potential AFP candidates. Through the implementation of a five-fold cross-validation strategy, the obtained results demonstrated the outstanding performance of our model, achieving an accuracy rate of 91.29% and an area under the curve (AUC) value of 0.96 in the identification of antifungal peptides on the training dataset. Our proposed model, named iAFP-ET, demonstrated exceptional performance, surpassing existing computational models and attaining the highest level of accuracy. The development of this model is expected to have a significant impact on research in academia, with an important contribution towards the growth of Proteomics and drug development.

Keywords: *Antifungal peptides, Extra Tree Classifier, fusion features, Amino acid composition, Dipeptide composition, Pseudo amino acid composition.*

**** Corresponding Author:*** *Dr. Ashfaq Ahmad, Assistant Professor, Computer Science Department, MY university Islamabad Pakistan.*

1. INTRODUCTION

Worldwide, infections caused by fungi are a severe health concern for people. When an invasive fungus appears on a specific area of the human body, the immune system finds it difficult to fight it off, resulting in a fungus sickness. According to contemporary figures, this chronic disease affects approximately 1.5 million deaths annually, almost three times more than malaria and nearly equal to the death rate of tuberculosis [1]. Over 150 million severe cases of fungal infection occur worldwide, according to a recent report from the World Health Organization (WHO) [2]. Over the past few years, numerous biochemical-based antifungals and treatments have been used. The four primary categories of antifungal agents that dominate the market are azole, which prevents the production of ergosterol: polyenes, which physically and chemically interact with the sterols in fungal membranes; Fluorinated pyrimidines interfere with pyrimidine metabolism, which prevents the creation of RNA and DNA, and echinocandins that prevent the formation of glucan. However, these approaches are insufficient due to their adverse drug responses, immunosuppressive narrow-spectrum activity, prolonged treatment process, and cross-resistance [3]. The medications used to treat fungi disorders can have serious side effects, including immune suppression, environmental contamination, and aggressive drug reactions [3, 4]. Developing efficient and accurate antifungal medications is essential for addressing these deficiencies Because of their high effectiveness and selectivity, and scientists recently discovered that peptide therapy-based medicines are safer than conventional ones. However, antimicrobial peptides (AMPs) are precisely the majority of peptide treatments depend on. AMPs have come to be considered a new medicine for treating and preventing fungal infections [5]. Anticancer [6], antiviral [7], antiparasitic [8], antibacterial [9], and antifungal [10] are among the classes of AMPs. In order to take into account the significance of AMPs, which has led to more than a hundred new drugs being developed so far, scientists have produced over 2300 active medicinal products. Due to their excellent selectivity, low microbial resistance, efficiency, and quick action, AMPs with antifungal properties have been considered a strong candidates [5]. As a result, the researchers used AFPs as an alternative to antibiotics while creating new drugs to treat fungal illnesses. To identify different bio-peptides, such as cell-penetrating peptides [11-13], anticancer peptides [14], antitubercular peptides [15], antimicrobial peptides [16], and antibacterial peptides [17], several studies have been proposed in the literature. In addition, numerous computational models for identifying antifungal peptides (AFPs) have been created. The ClassAMP method was developed by Joseph et al. to determine if protein samples

are likely to have antiviral, antibacterial, or antifungal action [18]. To numerically represent protein samples, a range of sequential and physical properties formulation approaches have been used, eliminating recursive features for selecting highly discriminative characteristics. Moreover, SVM and RF were used to assess the proposed scheme's capacity for learning. The same prediction approach was presented by Mousavizadegan et al. to distinguish the AFP samples from antimicrobial peptides [19]. To train the model, pseudo amino acid composition (PseAAC) & SVM were used. Later, Agrawal et al. employed descriptors for AFPs based on amino acid composition (AAC), dipeptide composition (DPC), binary profile, and split amino acid composition [20]. SVM performed well among the classifiers used to assess the retrieved vectors from training and independent datasets. The existing predictors used the conventional sequential techniques to obtain numerical descriptors derived from peptide samples, it was found after examining the existing AFPs models. However, there are several issues with these feature extraction techniques, especially for short amino acid sequences. Nonetheless, the abovementioned existing predictors' performance is insufficient and remains to be improved. In order to further improve the identification ability, we have proposed a novel model for identifying AFPs. We first encoded the samples with several features that included AAC, DPC, and PseAAC. In various biological problems, it has been demonstrated that multiple features can effectively discriminate between positive and negative cases. Second, we combine these features to enhance classification accuracy and facilitate the efficient screening of potential AFP candidates. Finally, we used the Extra Tree classifier to construct an identification model based on the Fusion features. The experiment results demonstrate that our proposed model is more efficient than current methods. The detailed flowchart of the proposed model is illustrated in Fig 1.

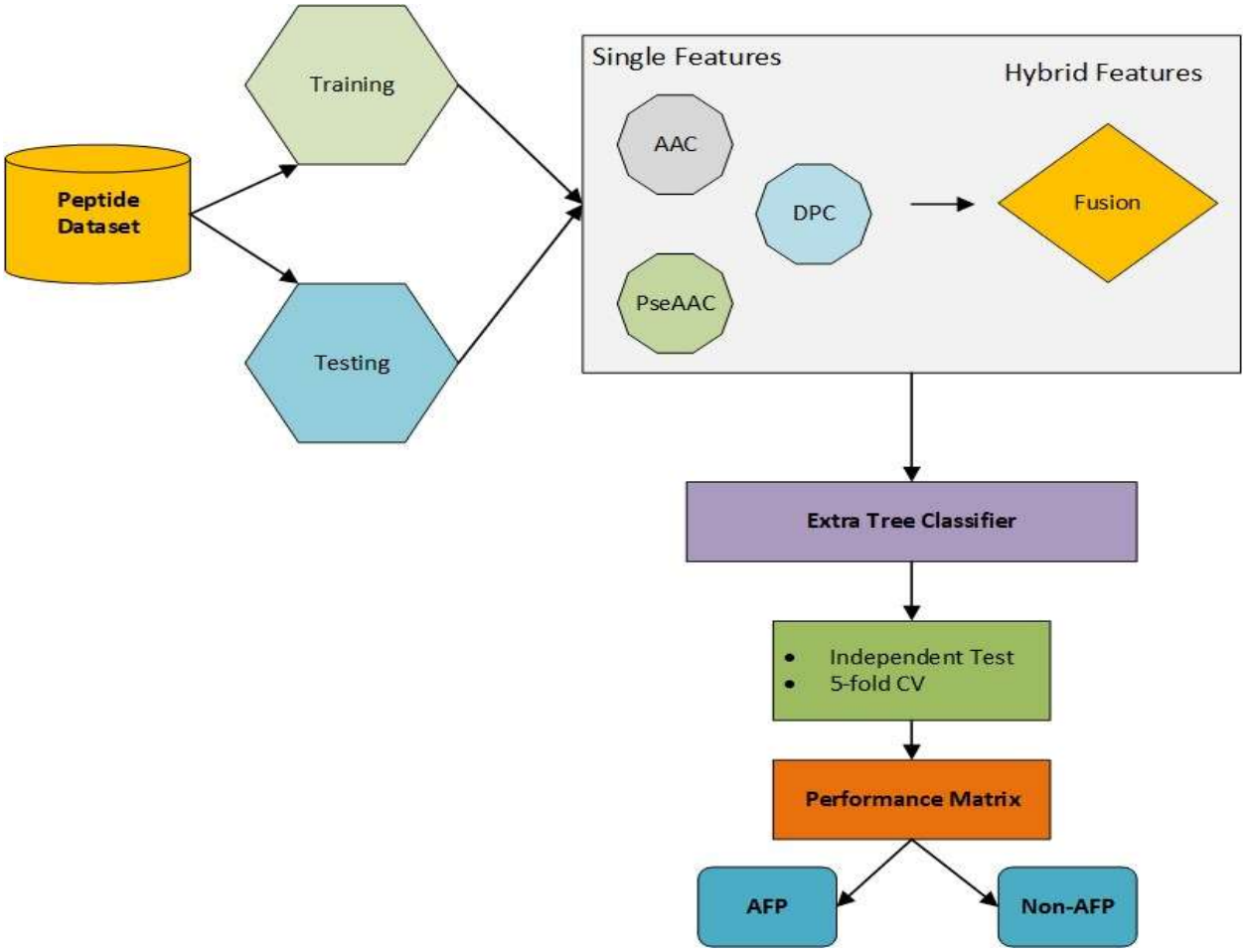

Fig. 1. A flowchart of the proposed iAFP-ET model.

2 MATERIALS AND METHODS

2.1. Dataset

To develop an intelligent prediction model, generating the appropriate benchmark data is of fundamental importance within the machine learning domain. Similarly, selecting an appropriate training data set substantially influences the performance of a computational predictor. Therefore, we have chosen to use the training dataset previously used by Agarwal et al., [20] Antifp_main, in recognition of the critical role of the benchmark dataset. The Antifp_main dataset encompasses two distinct classes: antifungal peptides (positive samples) and non-antifungal peptides (negative samples). It comprises a total of 2336 samples, with 1168 samples representing AFPs and an equal number representing non-AFPs. The AFP samples are sourced from the DRAMP database [21], whereas the non-AFPs are randomly generated from the Swiss-prot database [22]. In addition, the selected peptide samples are subject to further processing steps, such as eliminating duplication

sequences and invalid amino acids. On the other hand, we assess the generalization ability of our suggested model and address the issues related to over fitting by using an independent dataset previously utilized by Agarwal et al. 582 samples from the DRAMP database of AFP and non-AFP.

2.1 FEATURE REPRESENTATION

One of the difficulties in designing techniques for predicting sequences is to create relevant feature vectors from primary sequences. In this work, the amino acid composition descriptor (AAC), pseudo amino acid composition (PAAC), and dipeptide composition (DPC) were assessed as three frequently utilized characteristics for AFP prediction.

2.1.1. Amino acid composition (AAC)

AAC, or amino acid composition, is a technique used in bioinformatics to describe the structure of a protein. It is necessary to calculate the ratio between each of 20 naturally occurring amino acids in a protein sequence [23]. This method aims at providing a clear and fundamental understanding of the protein's structure that may then serve to predict its functions or structures, as described below. This approach has been described as one of the easiest methods to predict protein function and structure in literature, providing easy representation of proteins' sequence. The following equation 1 was used to calculate the composition of all 20 naturally occurring amino acids.

$$AAC_i = \frac{R_i}{L}$$

Where AAC_i the amino acid composition of residue type is i, R_i is the number of residues of type i in the protein sequence, L is the total number of residues in the protein sequence.

2.1.2. Pseudo amino acid composition (PseAAC)

Pseudo Amino Acid Composition (PseAAC) is a computational encodings utilize the hydrophobicity measurements suggested by Tanford [24], the hydrophilicity measurements suggested by Hopp and Woods [25], and the side chain mass measurements are the conventional ones to represent protein sequences as numerical vectors based on their amino acid composition and the correlation of different physicochemical properties of amino acids. The original

measurements are denoted as $H_1^0(t), H_2^0(t)$, and $M^0(t)$, where t represents each of the 20 inherent amino acids. These measurements are centralized and normalized in the subsequent manner.

$$P(t) = \frac{P^0(t) - \frac{1}{20}\sum_{i=1}^{20} P^0(i)}{\sqrt{\frac{\sum_{i=1}^{20}\left[P^0(i) - \frac{1}{20}\sum_{j=1}^{20} P^0(i)\right]^2}{20}}} t \in A$$

Within this context, the symbol A represents the group of the 20 natural amino acids. Correspondingly, the notation $P(t)$ conveys the centrality and standardization of the numerical value associated with any of the three characteristics (H_1, H_2, M) pertaining to the amino acid denoted by t. Consequently, the resulting variables are expressed as $H_1(t), H_2(t)$, and $M(t)$.

2.1.3. Dipeptide Composition (DPC)

The Dipeptide Composition is the frequency of each consecutive pair of amino acids in the sequence is calculated by the dipeptide composition method [23].

$$DPC(t,u) = \frac{N(t,u)}{N-1}, t,u \in A$$

Where A represents the collection of the 20 inherent amino acids, $N(t,u)$ indicates the frequency of occurrence of the amino acid pair t,u within the sequence, and N signifies the length of the sequence.

2.2 MACHINE LEARNING APPROACHES

In this paper, we utilized various machine learning classification algorithms to develop identification models containing Extra Tree Classifier, Random Forest, and KNN. The description of these methods is as follows.

2.2.1 Extra Tree Classifier

The Extra-Trees classifier generates an ensemble of undecorated decision trees using the conventional top-down methodology. This process involves considerable randomness in selecting attributes and cutting points when the nodes are split. In the most extreme case, it builds fully

randomized trees with structures independent of the output values from the training sample. The method, in two separate respects, differs from other ensembles' methods which rely on decision trees: firstly, it employs entirely random cut-point selection during node splitting, and secondly, it employs the entire training sample (rather than bootstrap replicas) to grow the trees. The final prediction is established by aggregating the predictions of all the trees through majority voting. The underlying concept of the Extra-Trees classifier is that the comprehensive randomization of both cut-point and attribute selection, coupled with ensemble averaging, effectively reduces variance compared to the weaker randomization strategies employed by alternative methods. In order to reduce the incidence of bias, original training samples shall be used instead of bootstrap replicas. The computational efficiency of this algorithm is one of its notable strengths [26]. The Extra Trees algorithm found extensive and diverse applications in the literature, just as with other algorithms. Recent examples include land cover classification utilizing Extremely Randomized Trees [27] and a multi-layer intrusion detection system incorporating Extra Trees feature selection, an ensemble of extreme learning machines, and softmax aggregation [28].

2.2.2 Random Forest

Random forest, proposed by [29], is acknowledged as one of the most robust ensemble learning techniques. The random forest has become widely accepted in the field of bioinformatics because of its exceptional efficiency. This flexible algorithm can address both regression and classification tasks. Random forests use a random feature selection approach to build a number of decision trees, ranging from hundreds to thousands, Akbar et al., [30] to address the problem. A voting mechanism shall be used to determine the final identification outcome of these Decision Trees. In this study, the random forest algorithm is sourced from WEKA [31], with all parameters set to their default values.

2.2.3 K-nearest Neighbor

The K-nearest Neighbor (K-NN) algorithm is highly regarded in pattern recognition and classification domains due to its exceptional performance, straightforward interpretability, and inherent simplicity. Compared to a number of alternative mechanisms in the area of machine learning, it consistently delivers competitive results despite its simplicity. KNN, being a non-parametric classification algorithm, distinguishes itself as an instance-based learner [13]. Within the K-NN classification process, instances are assigned to categories based on their proximity to

the K nearest instances in the feature space, as depicted in Figure 5 [32]. The KNN classifier will begin by estimating the Euclidian distance between a testing case and all training cases. The K-NN classifier commences by calculating the Euclidean distance between the testing and training instances. After that, only K instances from the feature space shall be selected closest to a test instance [33].

Consequently, the class that occurs most frequently among these K instances is assigned to the testing instance. In the event of a tie, the class assignment is randomized. In order to ensure a balanced decision-making process, an odd value is usually selected for K.

2.3 EVALUATION METRICS

The performance of a prediction method was evaluated using seven different metrics: Sensitivity (Sn), specificity (Sp), accuracy (Acc), and Matthews correlation coefficient (MCC) are commonly used metrics in the evaluation of a predictive model [15, 34-39].

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sen = \frac{TP}{TP + FN}$$

$$Spe = \frac{TN}{TN + FP}$$

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

The metrics Sensitivity, Specificity, Accuracy, and Matthews Correlation Coefficient (MCC) are utilized to evaluate the performance of a predictive model. True positives (TP) represent the cases where the model correctly identifies a positive outcome, while true negatives (TN) represent cases where the model correctly identifies a negative outcome. False positives (FP) are instances in which the model predicts a positive outcome inaccurately, and false negatives (FN) are instances in which the model is unable to detect a positive outcome [38]. When all aspects of the prediction approach are taken into account, MCC is one of the most complete and comprehensive metrics. It ranges from -1 (worst) to 1 (best), with a score of -1 signifying complete discrepancy between the

forecast and observation and a value of 0 signifying that the prediction method is no better than a guess at random.

3 RESULTS AND DISCUSSION

The results showed that the Extra Tree Classifier method achieved excellent accuracy in terms of Sen, Spe and MCC compared to traditional methods. The results further demonstrated that using an extra tree classifier increases the robustness of this method and decreases over fitting.

3.1. Performance evaluation

The study assessed the efficiency of different types of feature encoding using machine learning classifiers like Extra Tree, Random Forest, and k nearness neighbor. Single-encoding models, such as AAC, DPC, and PseAAC, and feature fusion models were evaluated. The results, presented in Tables 1 and 2, showed that the accuracies of the feature fusion models were consistently higher than those of the single-encoding models, indicating that combining multiple sources of information effectively achieved better results. The evaluation was conducted using a five-fold cross-validation test

3.2. Results of classifiers using single feature encoder

The results of the classifiers using each feature set are presented in Tables 1 and 2 for both datasets. Random Forest (RF) achieved accuracies of 86.29%, 87.49%, and 87.26% on AAC, DPC, and PseAAC, respectively. These results demonstrate that the DPC features were particularly informative and performed well. The AAC and PseAAC features generated approximately the same results. K-nearest neighbor (KNN) achieved poor accuracies than RF when using AAC, DPC, and PseAAC features. The best performance was achieved by Extra Tree Classifier using all of the feature extractors. On AAC, Extra Tree Classifier achieved 1.84% and 3.42% higher accuracies than RF and KNN, respectively.

Similarly, Extra Tree Classifier on DPC improved the accuracies by 1.2% and 11.05% compared to RF and KNN, respectively. The better results were achieved with PseAAC with Extra Tree Classifier. Ultimately, the best performance was achieved by Extra Tree Classifier, demonstrating its effectiveness in accurately identifying antifungal peptides.

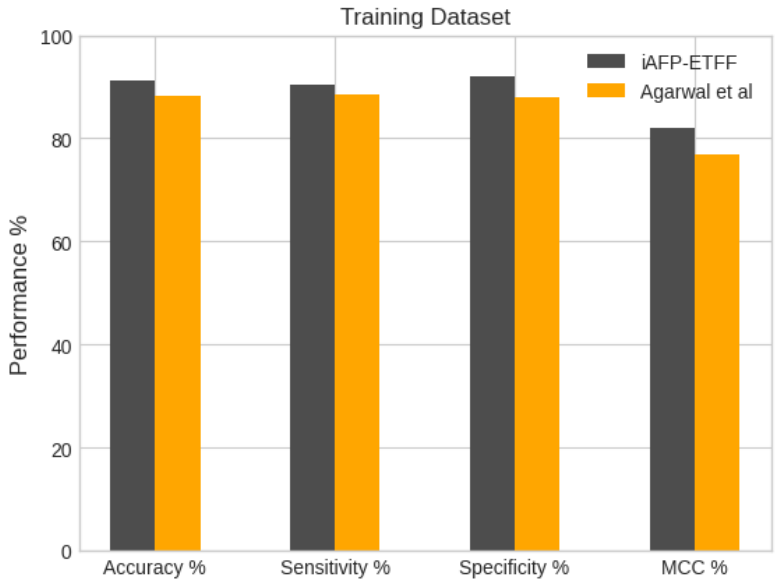

Fig. 2. Performance comparison of iAFP-ET with existing model using Training Dataset

3.3. Performance of classifiers with Hybrid features

Previous research has suggested that combining different features can improve prediction models. In this study, the researchers combined features from various descriptors in different series combinations and evaluated the performance using different classifiers. They found that the integrated feature set resulted in better predictions for AFP. Specifically, RF, KNN, and Extra Tree Classifier further improved the performance with AAC+DPC+PseAAC. Additionally, the analysis of the prediction results with all classifiers using a "fusion set" demonstrated remarkable performance with all assessment indexes. Extra Tree Classifier achieved the highest accuracy, sensitivity, specificity, and MCC among all classifiers. The results suggest that using a fusion feature set of "All feature set" is highly effective in identifying AFPs. The details of the fusion set results are given in Tables 1 and 2.

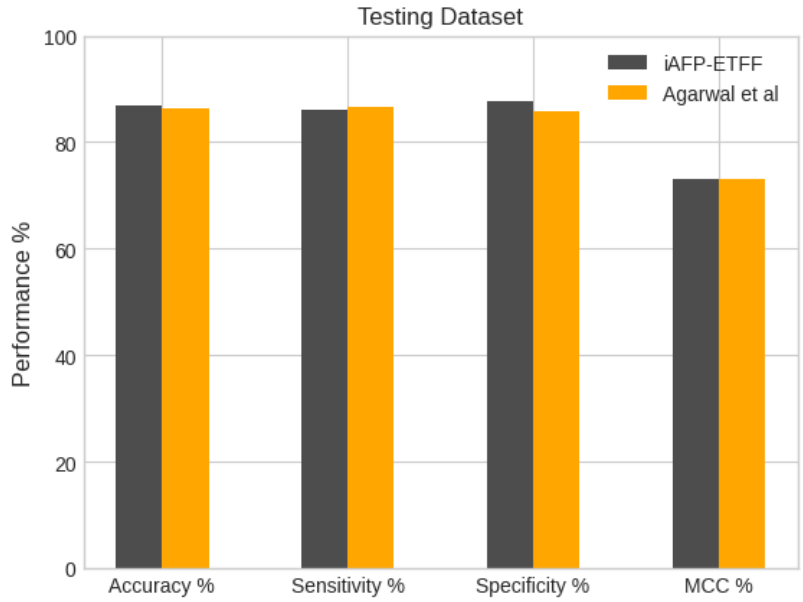

Fig. 3. Performance comparison of iAFP-ET with existing model using Testing Dataset

Table 1. Prediction Analysis of Classifiers algorithms using Training dataset.

Encoding Method	Classifier	Acc (%)	Sn (%)	Sp (%)	MCC	AUC
AAC	RF	86.29	85.34	87.24	0.72	0.9
	ET	88.13	87.40	88.86	0.76	0.94
	KNN	84.71	80.11	89.29	0.69	0.91
DPC	RF	87.49	87.40	87.58	0.74	0.93
	ET	88.69	89.11	88.27	0.77	0.94
	KNN	77.64	74.12	81.16	0.55	0.85
PAAC	RF	87.26	86.44	88.08	0.74	0.94
	ET	89.41	89.70	89.11	0.78	0.95
	KNN	85.38	80.78	89.97	0.71	0.92
Fusion-Vector	RF	90.95	89.27	92.63	0.81	0.96
	ET	91.29	90.48	92.11	0.82	0.96
	KNN	85.38	80.78	89.97	0.71	0.92

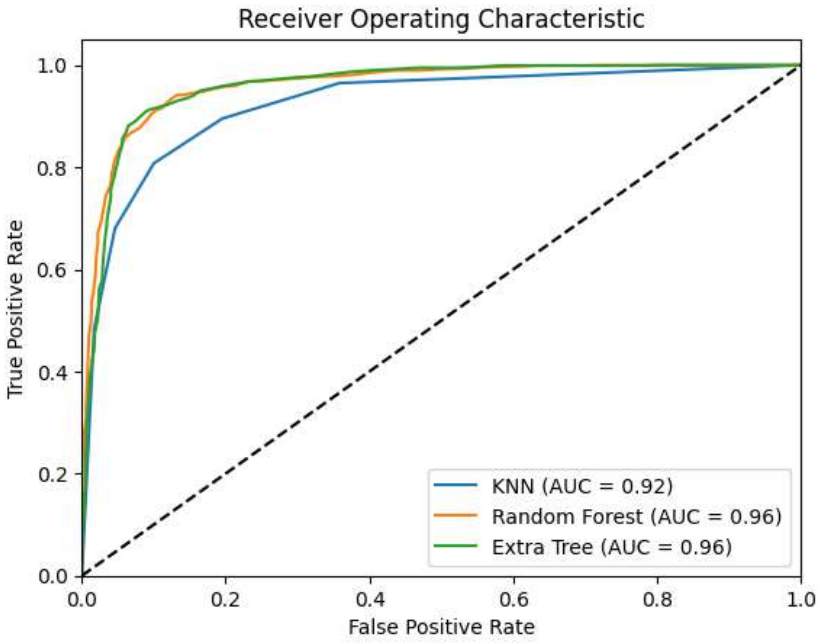

Fig. 4. ROC Analysis of Training dataset

Table 2. Prediction Analysis of Classifiers algorithms using Test dataset.

Encoding Method	Classifier	Acc (%)	Sn (%)	Sp (%)	MCC	AUC
AAC	RF	81.92	81.03	82.81	0.63	0.89
	ET	83.65	83.44	84.19	0.67	0.91
	KNN	79.33	84.48	74.22	0.59	0.87
DPC	RF	82.26	81.37	83.16	0.64	0.90
	ET	85.53	83.10	87.97	0.71	0.92
	KNN	73.83	82.75	64.94	0.48	0.82
PAAC	RF	84.84	84.48	85.22	0.69	0.91
	ET	86.22	85.17	87.28	0.72	0.93
	KNN	79.02	83.10	74.91	0.58	0.88
Fusion-Vector	RF	85.02	82.41	87.62	0.70	0.91
	ET	86.91	86.20	87.62	0.73	0.93
	KNN	80.03	79.31	80.75	0.60	0.88

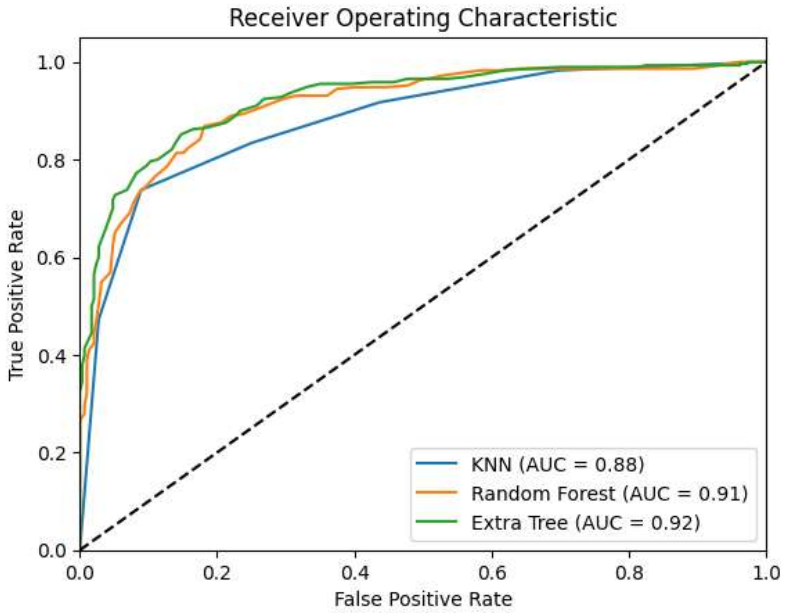

Fig. 5. ROC Analysis of Test dataset

3.4. Comparison with existing models

We compared our Proposed Model, iAFP-ET, with existing predictors such as Agrawal et al. [20] on both training and independent datasets. The results in Table 3 showed that iAFP-ET achieved higher accuracy, sensitivity, specificity, and MCC compared to the other methods, such as Agrawal et al. [20]. The study also evaluated the generalization ability of iAFP-ET by testing it on the testing dataset, where it again outperformed previous methods in the literature. The performance of iAFP-ET was compared to existing predictors, such as Agrawal et al. [20], on the testing dataset. The results showed that iAFP-ET outperformed Agrawal et al. [20] by achieving higher accuracy, sensitivity, specificity, and Matthews correlation coefficient (MCC). Besides, the other performance metrics, the Area under Curve is evaluating the prediction performance of a classification model, AUC is still considered to be effective. Where a model with higher AUC values is considered more effective than a lower AUC. Therefore, the AUC values for all used datasets shown in figure 4 and 5 shall be calculated when measuring the relevance of a statistical predictor. Whereas, Extra Tree Classifier For the antifp_main, and independent dataset it calculated the highest AUC values of 0.96 and 0.92. These comparisons are visualized in Figures 2 and 3.

Table 3. Performance Comparison of iAFP-ET with existing models.

Method	Dataset	Acc (%)	Sn (%)	Sp (%)	MCC
iAFP-ET	Training	**91.29**	**90.48**	**92.11**	**0.82**
Agrawal el al. [20]		88.27	88.61	87.93	0.77
iAFP-ET	Testing	**86.91**	86.20	87.62	0.73
Agrawal el al. [20]		86.25	86.60	85.91	0.73

4. CONCLUSIONS

This study developed a new predictor called iAFP-ET to identify antifungal peptides. Antifungal peptides are diverse and complex, making it difficult to identify their unique features. To overcome this challenge, we used three different feature extraction techniques, AAC, DPC, and PseAAC, to identify the most important information. In addition, a fusion feature of different feature vectors is used as an alternative measure to overcome limitations on using individual feature strategies and then trained different models using Extra Tree Classifier (ETC), Random Forest (RF), and K nearest neighbor (KNN) classification algorithms. Notably, the incorporation of fusion features in conjunction with an Extra Tree Classifier (ETC) surpassed the performance of existing models documented in the literature. After comparing the performance of all models, it was found that iAFP-ET performed the best. The primary reason for the remarkable success of the current study can be attributed to the effective feature coding approaches and the use of appropriate classification algorithms. Consequently, using iAFP-ET is expected to be highly regarded as a valuable tool within the research community.

REFERENCES

[1] F. Bongomin, S. Gago, R. O. Oladele, and D. W. Denning, "Global and multi-national prevalence of fungal diseases—estimate precision," *Journal of fungi,* vol. 3, no. 4, p. 57, 2017.

[2] M. C. Fisher, N. J. Hawkins, D. Sanglard, and S. J. Gurr, "Worldwide emergence of resistance to antifungal drugs challenges human health and food security," *Science,* vol. 360, no. 6390, pp. 739-742, 2018.

[3] D. Sanglard, "Emerging threats in antifungal-resistant fungal pathogens," *Frontiers in medicine,* vol. 3, p. 11, 2016.

[4] R. Capita and C. Alonso-Calleja, "Antibiotic-resistant bacteria: a challenge for the food industry," *Critical reviews in food science and nutrition,* vol. 53, no. 1, pp. 11-48, 2013.

[5] M. Fernández de Ullivarri, S. Arbulu, E. Garcia-Gutierrez, and P. D. Cotter, "Antifungal peptides as therapeutic agents," *Frontiers in Cellular and Infection Microbiology,* vol. 10, p. 105, 2020.

[6] D. Gaspar, A. S. Veiga, and M. A. Castanho, "From antimicrobial to anticancer peptides. A review," *Frontiers in microbiology,* vol. 4, p. 294, 2013.

[7] M. Feng, S. Fei, J. Xia, V. Labropoulou, L. Swevers, and J. Sun, "Antimicrobial peptides as potential antiviral factors in insect antiviral immune response," *Frontiers in immunology,* vol. 11, p. 2030, 2020.

[8] F. Iordache, M. Ionita, L. I. Mitrea, C. Fafaneata, and A. Pop, "Antimicrobial and antiparasitic activity of lectins," *Current pharmaceutical biotechnology,* vol. 16, no. 2, pp. 152-161, 2015.

[9] P. Kosikowska and A. Lesner, "Antimicrobial peptides (AMPs) as drug candidates: a patent review (2003–2015)," *Expert opinion on therapeutic patents,* vol. 26, no. 6, pp. 689-702, 2016.

[10] J. Wang *et al.*, "Antimicrobial peptides: Promising alternatives in the post feeding antibiotic era," *Medicinal research reviews,* vol. 39, no. 3, pp. 831-859, 2019.

[11] M. Arif, S. Ahmad, F. Ali, G. Fang, M. Li, and D.-J. Yu, "TargetCPP: accurate prediction of cell-penetrating peptides from optimized multi-scale features using gradient boost decision tree," *Journal of computer-aided molecular design,* vol. 34, pp. 841-856, 2020.

[12] B. Manavalan, S. Subramaniyam, T. H. Shin, M. O. Kim, and G. Lee, "Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy," *Journal of proteome research,* vol. 17, no. 8, pp. 2715-2726, 2018.

[13] P. Pandey, V. Patel, N. V. George, and S. S. Mallajosyula, "KELM-CPPpred: kernel extreme learning machine based prediction model for cell-penetrating peptides," *Journal of proteome research,* vol. 17, no. 9, pp. 3214-3222, 2018.

[14] S. Akbar, A. U. Rahman, M. Hayat, and M. Sohail, "cACP: Classifying anticancer peptides using discriminative intelligent model via Chou's 5-step rules and general pseudo components," *Chemometrics and Intelligent Laboratory Systems,* vol. 196, p. 103912, 2020.

[15] S. Akbar, A. Ahmad, M. Hayat, A. U. Rehman, S. Khan, and F. Ali, "iAtbP-Hyb-EnC: Prediction of antitubercular peptides via heterogeneous feature representation and genetic algorithm based ensemble learning model," *Computers in Biology and Medicine,* vol. 137, p. 104778, 2021.

[16] S. Basith, B. Manavalan, T. Hwan Shin, and G. Lee, "Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening," *Medicinal research reviews,* vol. 40, no. 4, pp. 1276-1314, 2020.

[17] E. Khaledian and S. L. Broschat, "Sequence-Based Discovery of Antibacterial Peptides Using Ensemble Gradient Boosting," in *Proceedings*, 2020, vol. 66, no. 1: MDPI, p. 6.

[18] S. Joseph, S. Karnik, P. Nilawe, V. K. Jayaraman, and S. Idicula-Thomas, "ClassAMP: a prediction tool for classification of antimicrobial peptides," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 5, pp. 1535-1538, 2012.

[19] M. Mousavizadegan and H. Mohabatkar, "Computational prediction of antifungal peptides via Chou's PseAAC and SVM," *Journal of bioinformatics and computational biology*, vol. 16, no. 04, p. 1850016, 2018.

[20] P. Agrawal, S. Bhalla, K. Chaudhary, R. Kumar, M. Sharma, and G. P. Raghava, "In silico approach for prediction of antifungal peptides," *Frontiers in microbiology*, vol. 9, p. 323, 2018.

[21] L. Fan *et al.*, "DRAMP: a comprehensive data repository of antimicrobial peptides," *Scientific reports*, vol. 6, no. 1, p. 24482, 2016.

[22] E. Boutet *et al.*, "UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view," *Plant bioinformatics: methods and protocols*, pp. 23-54, 2016.

[23] M. Bhasin and G. P. Raghava, "Classification of nuclear receptors based on amino acid composition and dipeptide composition," *Journal of Biological Chemistry*, vol. 279, no. 22, pp. 23262-23266, 2004.

[24] C. Tanford, "Contribution of hydrophobic interactions to the stability of the globular conformation of proteins," *Journal of the American Chemical Society*, vol. 84, no. 22, pp. 4240-4247, 1962.

[25] T. P. Hopp and K. R. Woods, "Prediction of protein antigenic determinants from amino acid sequences," *Proceedings of the National Academy of Sciences*, vol. 78, no. 6, pp. 3824-3828, 1981.

[26] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, pp. 3-42, 2006.

[27] A. Zafari, R. Zurita-Milla, and E. Izquierdo-Verdiguier, "Land cover classification using extremely randomized trees: A kernel perspective," *IEEE geoscience and remote sensing letters*, vol. 17, no. 10, pp. 1702-1706, 2019.

[28] J. Sharma, C. Giri, O.-C. Granmo, and M. Goodwin, "Multi-layer intrusion detection system with ExtraTrees feature selection, extreme learning machine ensemble, and softmax aggregation," *EURASIP Journal on Information Security*, vol. 2019, no. 1, pp. 1-16, 2019.

[29] L. Breiman, "Random Forests. Statistics Department," *University of California, Berkeley, CA*, vol. 4720, 2001.

[30] S. Akbar, M. Hayat, M. Tahir, and K. T. Chong, "cACP-2LFS: classification of anticancer peptides using sequential discriminative model of KSAAP and two-level feature selection approach," *IEEE Access*, vol. 8, pp. 131939-131948, 2020.

[31] G. H. B. P. P. Reutemann, I. H. W. M. Hall, E. Frank, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10-18, 2009.

[32] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of theoretical biology*, vol. 273, no. 1, pp. 236-247, 2011.

[33] A. Khan, M. Khan, and T.-S. Choi, "Proximity based GPCRs prediction in transform domain," *Biochemical and biophysical research communications*, vol. 371, no. 3, pp. 411-415, 2008.

[34] A. Ahmad, S. Akbar, M. Tahir, M. Hayat, and F. Ali, "iAFPs-EnC-GA: identifying antifungal peptides using sequential and evolutionary descriptors based multi-information fusion and

ensemble learning approach," *Chemometrics and Intelligent Laboratory Systems*, vol. 222, p. 104516, 2022.

[35] A. Ahmad, S. Akbar, M. Hayat, F. Ali, S. Khan, and M. Sohail, "Identification of antioxidant proteins using a discriminative intelligent model of k-space amino acid pairs based descriptors incorporating with ensemble feature selection," *Biocybernetics and Biomedical Engineering*, vol. 42, no. 2, pp. 727-735, 2022.

[36] S. Akbar, F. Ali, M. Hayat, A. Ahmad, S. Khan, and S. Gul, "Prediction of Antiviral peptides using transform evolutionary & SHAP analysis based descriptors by incorporation with ensemble learning strategy," *Chemometrics and Intelligent Laboratory Systems*, vol. 230, p. 104682, 2022.

[37] S. Akbar, M. Hayat, M. Tahir, S. Khan, and F. K. Alarfaj, "cACP-DeepGram: classification of anticancer peptides via deep neural network and skip-gram-based word embedding model," *Artificial Intelligence in Medicine*, vol. 131, p. 102349, 2022.

[38] 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, 2021.

[39] S. Akbar, S. Khan, F. Ali, M. Hayat, M. Qasim, and S. Gul, "iHBP-DeepPSSM: Identifying hormone binding proteins using PsePSSM based evolutionary features and deep learning approach," *Chemometrics and Intelligent Laboratory Systems*, vol. 204, p. 104103, 2020.