Send Orders for Reprints to reprints@benthamscience.net

Current Bioinformatics, xxxx, *xx,* x-x

RESEARCH ARTICLE

An Extended Feature Representation Technique for Predicting Sequenced-based Host-pathogen Protein-protein Interaction

Jerry Emmanuel[1,3], Itunuoluwa Isewon[1,2,3], Grace Olasehinde[3,4] and Jelili Oyelade[1,2,3,*]

[1]Department of Computer & Information Sciences, Covenant University, Ota 112104, Nigeria; [2]Covenant University Bioinformatics Research (CUBRe), Covenant University, Ota 112104, Nigeria; [3]Covenant Applied Informatics and Communication African Centre of Excellence (CApIC-ACE), Covenant University, Ota 112104, Nigeria; [4]Department of Biological Science, Covenant University, Ota 112104, Nigeria

ARTICLE HISTORY

Received: October 11, 2023
Revised: December 13, 2023
Accepted: December 20, 2023

DOI:
10.2174/0115748936286848240108074303

Abstract: ***Background:*** The use of machine learning models in sequence-based Protein-Protein Interaction prediction typically requires the conversion of amino acid sequences into feature vectors. From the literature, two approaches have been used to achieve this transformation. These are referred to as the Independent Protein Feature (IPF) and Merged Protein Feature (MPF) extraction methods. As observed, studies have predominantly adopted the IPF approach, while others preferred the MPF method, in which host and pathogen sequences are concatenated before feature encoding.

Objective: This presents the challenge of determining which approach should be adopted for improved HPPPI prediction. Therefore, this work introduces the Extended Protein Feature (EPF) method.

Methods: The proposed method combines the predictive capabilities of IPF and MPF, extracting essential features, handling multicollinearity, and removing features with zero importance. EPF, IPF, and MPF were tested using bacteria, parasite, virus, and plant HPPPI datasets and were deployed to machine learning models, including Random Forest (RF), Support Vector Machine (SVM), Multilayer Perceptron (MLP), Naïve Bayes (NB), Logistic Regression (LR), and Deep Forest (DF).

Results: The results indicated that MPF exhibited the lowest performance overall, whereas IPF performed better with decision tree-based models, such as RF and DF. In contrast, EPF demonstrated improved performance with SVM, LR, NB, and MLP and also yielded competitive results with DF and RF.

Conclusion: In conclusion, the EPF approach developed in this study exhibits substantial improvements in four out of the six models evaluated. This suggests that EPF offers competitiveness with IPF and is particularly well-suited for traditional machine learning models.

Keywords: Protein-Protein Interaction; Feature Representation; Host-Pathogen Interaction; Machine Learning; Protein Sequence; Feature Vectors.

1. INTRODUCTION

Proteins frequently interact with one another to facilitate diverse biochemical functions, culminating in the formation of protein complexes [1-4]. This phenomenon is referred to as protein-protein interactions (PPIs), where electrostatic or hydrophobic forces are employed to facilitate the binding of two or more proteins [5-8]. The identification of PPIs is crucial for understanding cellular functions and vital biological processes, such as protein function, disease pathogenesis, and drug development [9-12]. PPIs can either involve interactions between the proteins of the same organism or those of different organisms. These interactions are referred to as intra- or inter-species interactions, respectively [13, 14]. A pathogen (an organism that causes disease) and a host (the victim organism) can interact through a process known as host-pathogen interaction (HPI), also referred to as host-pathogen protein-protein interaction (HPPPI) [15-17]. Over time, the concept of HPPPI has contributed to the general understanding of the pathogenesis of infectious diseases, which in turn has led to the development of treatments for those diseases [18-20].

Despite the use of experimental methods to identify PPIs, the comprehensive set of PPIs in organisms remains limited [12, 21]. Some of the most commonly used experimental methods for detecting PPIs are fluorescence resonance energy transfer (FRET) [22], protein-fragment complementation

*Address correspondence to this author at the Department of Computer & Information Sciences, Covenant University, Ota 112104, Nigeria; Tel: +2348035755778; E-mail: ola.oyelade@covenantuniversity.edu.ng

1574-8936/xx $65.00+.00

assay (PFCA) [23], LUMIER [24], and yeast two-hybrid (Y2H) [25-27]. However, using these aforementioned experimental methods to detect HPPPIs is not scalable as the space of HPPPI is too large to be explored experimentally and to generate interactions with high confidence [8, 25, 29]. Machine learning, on the other hand, enables computers to learn and advance through the use of statistical techniques so that they may make more accurate predictions without being particularly programmed to do so [30]. To complement the effort being made in detecting HPPPI using the experimental methods, machine learning techniques, such as Random Forest (RF), Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), and Deep Neural Network (DNN) have been utilized over the years [15, 31-33]. Similarly, Symbolic Regression (SR), which is a new Genetic Programming model, has also been used for predicting disease-related PPIs [34, 35]. SR involves deriving a mathematical function that describes a given dataset [36]. While traditional regression methods rely on the pre-determined independent variable(s) and aim to create a fit model by adjusting various numerical coefficients, SR seeks to simultaneously discover both the parameters and equations. Hence, with SR, a mathematical expression that fits a given data is produced [36, 37].

Machine learning models require data for effective performance [38-40], and in recent times, there has been an increase in the number of interacting protein complexes and their respective sequence data. As a result, computational methods for HPPPI prediction using only the protein sequence information have attracted growing interest [41, 42]. In an experiment conducted by Christian Anfinsen several decades ago, he found that a protein's sequence determines its structural conformation, which in turn affects how the protein works and how it interacts with other proteins [43-45]. Also, several studies [46-49] have demonstrated that utilizing information from amino acid sequences is sufficient to discover highly potential HPPPIs that can also be experimentally validated. Furthermore, compared to other data types like gene expression and knockout phenotypes that might also be employed for HPPPI prediction, amino acid sequence information is the protein data type that is most easily accessible [50]. However, using protein sequences in machine learning requires that they are processed into feature vectors. To create a fixed-length vector that can be employed in standard prediction models, features from the input amino acid sequences are extracted using feature extraction methods, such as Pseudo Amino Acid Composition (PseAAC) [51], Moran Autocorrelation [15], and Conjoint Triads [52]. In addition, due to the single representation of protein sequences in techniques, such as dipeptide composition (DipC), PseAAC, and position-specific scoring matrix (PSSM), Wang and Liu [53] introduced a fusion of PSSM with DipC as well as PSSM with PseAAC to produce techniques for future representation called DipPSSM and PseAAPSSM respectively. Also, taking into consideration the distances between polynucleotide sequences using the concept of fuzzy polynucleotide space (FPS), Nieto *et al.* [54] examined varying methods of calculating distances between the nucleotide sequences as opposed to the method used in the original FPS.

There are a variety of approaches used by researchers in the application of feature encoding methods for generating feature vectors in training machine learning models for HPPPI prediction[15,55,56]. Several authors perform feature representation of the host and pathogen amino acid sequences independently before concatenating the resulting encoded features, while several other authors combine the amino acid sequences of both the host and the pathogen before using an encoding technique [16,33,57]. For reference purposes in this work, the two methods will be referred to as Independent Protein Feature (IPF) and Merged Protein Feature (MPF), respectively. Although studies have shown that the IPF strategy is the most widely applied, most recently, authors have adopted the use of MPF. This presents the challenge of what approach should be adopted for improved HPPPI prediction. Therefore, in this work, an Extended Protein Feature (EPF) method is presented. As a contribution, this study examines the strengths and weaknesses of these two methods and then presents a new and improved strategy that integrates their predictive abilities. In addition, sequel to the confirmation of the existence of important features and less or non-important features generated when both IPF and MPF are used, EPF ensures only important features that enhance model performance are extracted for subsequent training of machine learning models. EPF automatically checks for multicollinearity, which aids in the identification of redundant features. Finally, the method presented in this work adopts a generalized system approach whereby a combination of protein sequences from interacting and non-interacting organisms was used in training and evaluating the existing methods and the proposed methods as opposed to the most frequent approach of training models using data from a single interacting pair one at a time.

The rest of this paper is organized as follows: materials and methods, which contain the dataset description, the feature encoding methods used, the proposed methodology, results, discussion, and conclusion.

2. MATERIALS AND METHODS

This section describes the dataset used for this study, the feature representation methods, and the machine learning models used to train the extracted features. The section concludes with a description of the proposed system architecture as a whole.

2.1. Dataset

PPI computational prediction methods require two datasets: one of the known interacting protein pairs and one of the non-interacting protein pairs to enhance their ability to discern between the two [58-60]. The HPPPI datasets employed for this work were experimentally derived as reported by Dey *et al*. [57], Kösesoy *et al.* [61], Wuchty *et al*. [62], Gordon *et al.* [63], and Mukhtar *et al.* [64]. These datasets include the human-*Bacillus anthracis*, human-*Plasmodium*

falciparum, human-SARS-Cov2, *Arabidopsis-Pseudomonas syringae, Hyaloperonospora arabidopsidis,* and *Golovinomyces orontii*.

The interspecies datasets used by Kösesoy *et al*. [61] were employed for the human *Bacillus anthracis*. The dataset is made up of interactions between humans and *Bacillus anthracis*, which contains 2500 positive and 9500 negative interactions. The complete positive interactions were all used as each protein pair had already been verified by the authors [61]. To train a machine learning model, an equal number of positive and negative samples are required [65, 66]. Hence, a total of 2500 samples were randomly selected from the negative interaction dataset. Before that, a script was developed and used to ensure that none of the interacting positive pairs was present in the pool of negative pairs.

Malaria, a lethal infectious disease caused by *Plasmodium falciparum*, affects a remarkably large part of the world's population [67-71]. As a result of this, the human-*Plasmodium falciparum* interaction dataset that was used by Wuchty *et al.* [62] will be used in this paper for the human-parasite interaction. The initial dataset was comprised of 1,112 positive interactions and 1136 negative interactions. The files include the identities of human and parasite proteins. Following the pre-processing stage, a total of 727 positive and negative samples were utilized. To eliminate biases, the proposed model will be trained using the same amount of positive and negative examples. Hence, the model was trained using a total of 1454 interactions.

Gordon *et al.* [63]'s human-SAR-CoV-2 interaction dataset was used for the human-virus dataset. Affinity-purification mass spectrometry (AP-MS) was used to prepare the dataset, which includes interactions between new coronavirus proteins and human proteins. The database had 332 unique interactions between 332 human proteins, four structural coronavirus proteins, and 20 coronavirus proteins that are not needed for the virus to work. As described by the authors, positive training and testing datasets were created using these 332 human-SAR-CoV-2 proteins that had undergone experimental validation. The negative samples for this category of the dataset were created by choosing human proteins from the HPRD database release 9 that are absent from the positive dataset and have a low degree in the human PPI network, as explained by Dey *et al.*[57], who also used the 332 positive samples in conducting their research.

Positive datasets consisting of 459 Arabidopsis (Ara)-pathogen interactions with three different pathogens were compiled from published literature. For the bacterial pathogen *Pseudomonas syringae* (Psy), 104 Ara-Psy PPIs were found. These PPIs involved 60 proteins from Arabidopsis and 38 from Psy's effectors [64]. 233 Ara-Hpa PPIs were retrieved for the oomycete *Hyaloperonospora arabidopsidis* (Hpa), which involved 122 proteins from Arabidopsis and 64 effectors from Hpa [66]. We retrieved 122 Ara - Gor HPPPI from Weßling *et al.* [72] for the fungal pathogen *Golovinomyces orontii* (Gor), which contained 60 Arabidopsis proteins and 46 Gor effectors. Negative samples were analyzed by retrieving PPIs involving 2639 Arabidopsis proteins from the TAIR database. As a result of the preprocessing, however, 444 positive and negative samples were ultimately used.

Moreover, consequent to the unbalanced nature of interacting and non-interacting PPI datasets, as evident from the aforementioned data sources and in a biologically relevant scenario, the techniques discussed in this work were also trained and evaluated using a complete set of the unbalance datasets.

In the data preparation and preprocessing stage, the interacting host-pathogen protein pairs that were identified and retrieved were processed and prepared to fit into machine learning models. During this process, any pair with either the host or pathogen protein sequence or both missing was removed from the interaction and non-interaction list. Similarly, before feature encoding was performed, protein pairs with any invalid amino acid code were filtered out. This further reduced the host-pathogen protein pairs. The final valid number of balanced and unbalanced datasets used in this work is presented in Table **1**. Finally, the preprocessed datasets were combined to create a comprehensive dataset consisting of various host-pathogen pairs from different organisms (specifically, four host-pathogen pairs in this instance). This resulted in a total of 8006 pairs for the balanced dataset and 16488 pairs for the unbalanced dataset, as shown in Table **1**, including both interacting and non-interacting pairs. The protein sequences of these datasets in fasta format can be found at https://github.com/JerrySteam/extended-protein-feature/tree/master/Dataset. The pre-processed data were then utilized for training and assessing machine learning models to develop a more generalized HPPPI prediction model that incorporates an enhanced feature extraction approach.

2.2. Extended Protein Feature Representation

A feature is a fundamental variable used to express and capture the information included in data while doing data-driven tasks like knowledge discovery and machine learning [73]. The success of workflows in data science is directly correlated to the quality of the features that are engineered. The use of feature representation methods is crucial for the creation of computational models for HPPPI prediction [74]. The Composition of k-spaced Amino Acid Pairs (CKSAAP), which is equivalent to the Amino Acid Pairs (AAP) when the value of k is 0, is used as the primary feature representation of sequence information in this study due to its high information content and protein-specificity [61, 75].

Given the substantial number of feature vectors generated by AAP (400 per protein sequence), we also incorporate Amino Acid Composition (AAC)[1-4]. AAC in this work produces only 20 feature vectors for each protein sequence. This represents a significant reduction compared to the 400 vectors generated by AAP. The Amino Acid Composition encoding calculates the frequency of each amino acid type in a protein or peptide sequence [75]. The frequencies of all 20 natural amino acids (ACDEFGHIKLMNPQRSTVWY) can be calculated as:

Table 1. Host-pathogen protein-protein Interaction datasets showing the number of interacting and noninteracting pairs.

	Balance		Unbalance	
Host-Pathogen	**Interacting**	**Non-Interacting**	**Interacting**	**Non-Interacting**
Human-*Bacillus anthracis*	2500	2500	2500	8491
Human-*Plasmodium falciparum*	727	727	738	727
Human-SAR-CoV-2	332	332	332	617
Arabidopsis - *Pseudomonas syringae, Hyaloperonospora arabidopsidis, Golovinomyces orontii*	444	444	444	2639
TOTAL	4003	4003	4014	12474

$$f(t) = \frac{N(t)}{N}, t \in \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\} \quad (1)$$

where *N(t)* is the number of amino acid type *t*, while *N* is the length of a protein or peptide sequence.

The AAP feature encoding calculates the frequency of amino acid pairs. Pairing the 20 natural amino acids given previously produces 400 residue pairs, and the feature vectors can be calculated as:

$$f(400) = \frac{N_{AA}}{N_{total}}, \frac{N_{AC}}{N_{total}}, \frac{N_{AD}}{N_{total}} \ldots, \frac{N_{YY}}{N_{total}} \quad (2)$$

Where $N_{AA} \ldots N_{YY}$ are the number of times the residual pairs appear in a peptide or protein sequence while N_{total} is the number of amino acids in the protein or peptide sequence.

Now, as an improvement on the two approaches of applying feature representation as discussed in section 1 of this work, this proposed system applies feature representation primarily at three different points as displayed in the proposed system framework in Fig. (**1**). Firstly, the feature encoding method (AAP or AAC in this case) is applied separately on the host and pathogen protein sequence, after which the produced feature vectors are concatenated. For example, if a protein sequence is encoded using the AAP method, this first component generates 400-dimensional feature vectors for each of the host and pathogen protein sequences. Concatenating the vectors produces an 800-dimension feature vector. In the second component of this work, the host and pathogen amino acid sequences are first concatenated before an AAP method is applied. This component generates just a 400-dimensional vector space. At this point, the respective feature vectors generated by the two different components are then concatenated, thereby generating a 1200-dimension feature vector. Similarly, if AAC is employed, a 60-dimension feature vector is generated for each protein pair.

2.2.1. Feature Reduction with Feature Importance

When analyzing large datasets with thousands of variables, feature selection is a crucial step. It selects a smaller collection of relevant features for classification. In this work, five methods of determining feature importance were evaluated and compared to determine the most suitable feature importance method for the developed approach. These methods include:

a. Impurity-based feature importances: This method is based on the mean decrease in impurity of decision trees, and it measures the importance of each feature by calculating the reduction in impurity (usually measured by Gini impurity or entropy) that results from splitting the data based on that feature. The higher the value, the more important the feature [76]. The Gini impurity of a node is calculated as:

$$G(N) = 1 - \sum_{i=1}^{c} (p_i)^2 \quad (3)$$

Where c is the number of classes in the dataset, and p_i is the probability of class i in node N. To calculate the feature importance using Gini impurity, the decrease in impurity over all nodes is evaluated where the feature is used for splitting, weighted by the number of data points in each node.

b. Permutation-based feature importances: This method calculates the feature importance by randomly permuting the values of each feature and measuring the decrease in model performance. This method is computationally expensive [77].

c. Decision tree-based feature importance: This method is similar to the impurity-based feature importance method, but it uses only a single decision tree instead of a forest of trees [78, 79].

d. Logistic Regression-based feature importance: This method measures the importance of each feature by calculating the absolute value of the coefficients of the logistic regression model [80].

e. Feature selection with information gain: This method uses mutual information to measure the dependence between each feature and the target variable, and it selects the features with the highest information gain [81, 82]. Information gain is calculated as the decrease in entropy after the dataset is split on a feature. The entropy of a node in a classification task is calculated as follows:

$$Entropy(N) = -\sum_{i=1}^{c} p_i log_2(p_i) \quad (4)$$

Where c is the number of classes and p_i is the probability of class i in node N. Information gain is then calculated as the entropy of the parent node minus the weighted average of the entropies of the child nodes after a split. The feature im-

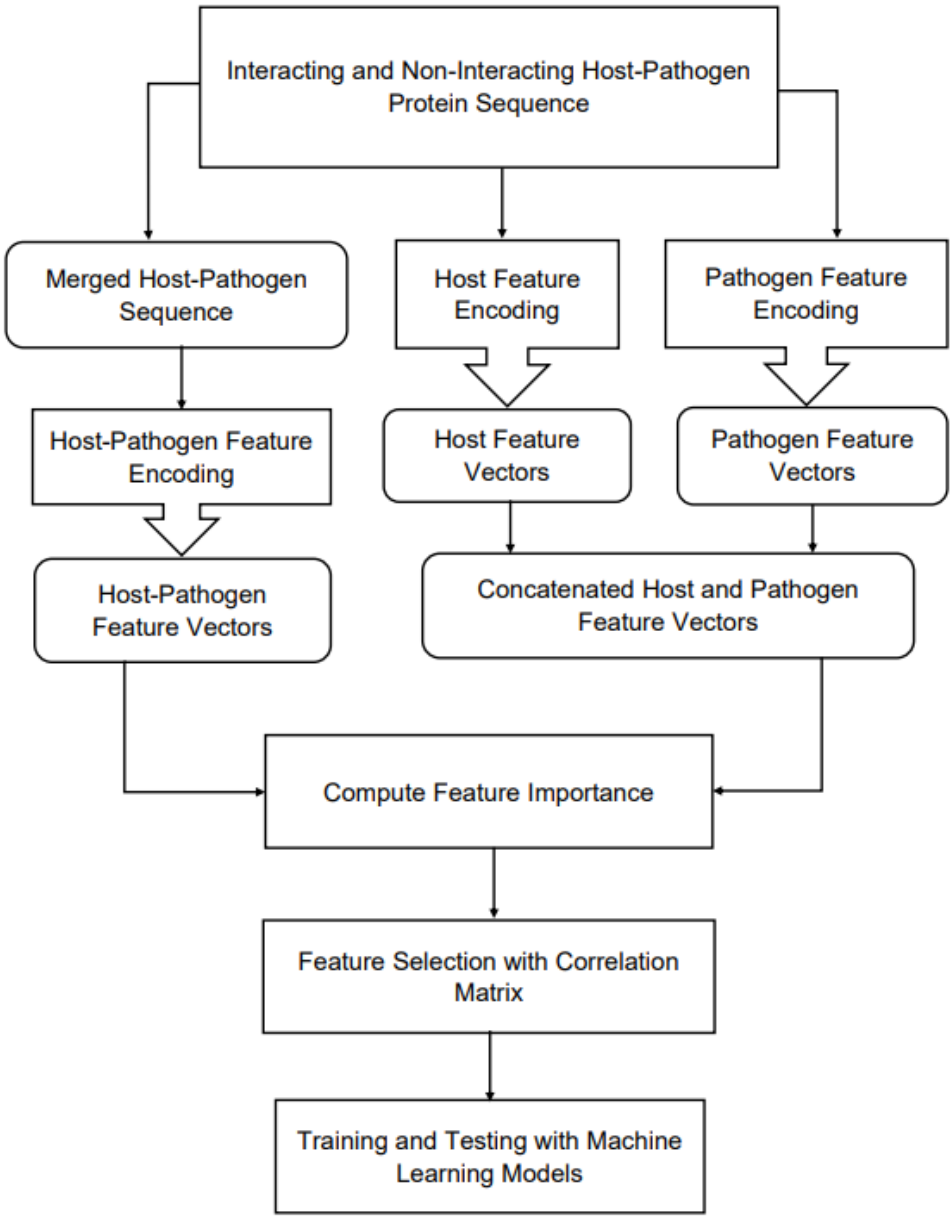

Fig. (1). Extended Protein Feature representation system framework. The proposed system primarily applies feature representation at three different points. The importance of the resulting feature vectors is determined, and features with multicollinearity are identified and eliminated to reduce redundancy.

portance using information gain can be computed by summing the information gains over all nodes where the feature is used for splitting.

2.2.2. Handling Multicollinearity Using Correlation Matrix

Feature selection using a correlation matrix involves computing the pairwise correlation coefficients between all pairs of features in the data matrix and selecting a subset of features based on their correlation with the target variable and their correlation with other features. The correlation coefficient between two features i and j is defined as the Pearson correlation coefficient:

$$r_{i,j} = \frac{(1/(n-1)) * Sum\left(\left(x_{i,k} - mean(x_i)\right) * (x_{j,k} - mean(x_j))\right)}{std(x_i) * std(x_j)} \quad (5)$$

where n is the number of samples, $x_{i,k}$ and $x_{j,k}$ are the values of features i and j for sample k, and $mean(x_i)$ and $std(x_i)$ are the mean and standard deviation of feature i across all samples. The correlation coefficient ranges from -1 to 1, where a value of 1 indicates a perfect positive correlation (that is, the two features increase or decrease together), a value of -1 indicates a perfect negative correlation (the two features have opposite effects), and a value of 0 indicates no correlation. In this work, the correlation threshold was set at 0.8; hence, highly correlated features are discarded.

As described in Fig. (**1**) and Algorithm **1**, variable importance with the methods described above and reduction with correlation matrix was employed to find and eliminate irrelevant variables, specifically, variables with zero im-

Algorithm 1: Extended Protein Feature algorithm

1. **Inputs**:
   Interacting host protein sequence
   Interacting pathogen protein sequence
   Noninteracting host protein sequence
   Noninteracting pathogen protein sequence
2. **Output**:
   Model prediction performance (Accuracy, sensitivity, specificity, precision, F1, Score, MCC, AUROC)
3. **Begin**
4. IPF = Compute independent host and pathogen protein sequence feature encoding
5. MPF = Compute merged host-pathogen protein sequence feature encoding
6. Combine IPF and MPF
7. HPI_data = Combine and add labels to interacting and noninteracting data
8. HPI_imp_features = Compute feature importance (HPI_data)
9. HPI_imp_features = select features with HPI_imp_features != 0.00
10. HPI_data_cm = compute Correlation_Matrix (HPI_imp_features) using Eqn 1
11. HPI_data = discard highly correlated features from HPI_data_cm (>0.8)
12. **for** model in MLModels[RF, SVM, MLP, NB, LR, DF]
13. HPI_train = split data into training and test set (HPI_data) using 10-fold StratefiedGroupCV
14. HPI_test = split data into training and test set (HPI_data) using 10-fold StratefiedGroupCV
15. Train model (HPI_train) using 9 folds
16. Test model (HPI_test) using 1 fold
17. Repeat 13 to 16 until all folds are used for training and testing.
18. **return** Compute model accuracy, sensitivity, specificity, precision, F1 Score, MCC, AUROC
19. **end for**
20. **End**

portance. Subsequently, highly correlated features resulting from the correlation matrix were further disengaged. As would be seen in subsequent sections of this work, the application of these techniques further improved the extended method that is proposed.

Furthermore, the final feature vectors obtained are used in training machines learning models, such as Logistic Regression (LR), Support Vector Machine (SVM), Multilayer Perceptron (MLP), Naïve Bayes (NB), Random Forest (RF), and Deep Forest (DF) for eventual HPPPI prediction.

To evaluate the performance of machine learning models on a dataset, it is important to use a robust evaluation technique. One such technique is called cross-validation. In this particular case, a 10-fold cross-validation technique was used to evaluate the performance of the machine learning models. This technique involves splitting the dataset into 10 equally sized parts, or "folds". The model is then trained on 9 of the folds and evaluated on the remaining fold. This process is repeated 10 times, with each fold used exactly once as the evaluation set. Using this method, the performance of the machine learning models can be evaluated more robustly, as it ensures that each data point is used for both training and testing. By evaluating the performance of the models on multiple splits of the data, it is also possible to gain a better understanding of how well the models can generalize to new data.

However, due to the nature of host-pathogen interaction data, random splitting of host-pathogen pairs during cross-validation can lead to accuracy inflation. To handle this problem, we introduced a Stratified Group 10-fold cross-validation (cv). This method creates folds taking into consideration the dataset groups and also preserves the percentage of samples for each class as much as possible in each fold, given the constraint of non-overlapping groups between folds. To achieve this, at the data preparation stage, we constructed a function to ensure that host-pathogen protein pairs were grouped based on the host protein sequences and assigned group IDs, with group ID starting from 1. We had a total of 10979 group IDs from 16488 host-pathogen protein pairs (interacting and non-interacting). So instead of splitting the host-pathogen pairs randomly, the Stratified Group 10-fold cv method ensures that any host-pathogen protein pair in the same group cannot appear in the training and validation set at the same time during cross-validation. So this implies that it is not possible for the same host protein (interacting with different pathogen proteins and having the same group ID) to appear in both training and validation.

3. RESULTS

In this section, the results of the proposed system's predictive abilities are presented. Also, the two approaches discovered from the literature (IPF and MPF), as used by other

researchers, were evaluated and benchmarked using the same dataset and the same set of predictors. The results obtained were compared and discussed among the three different approaches. To ascertain if there is a performance difference based on the dataset size, we conducted experiments on each of the methods being evaluated using different dataset sizes comprising 10%, 25%, 50%, and 75% of the original dataset size. To ensure that the right proportion of the host-pathogen pair was used in training the models, equal fractions of the interacting and non-interacting dataset sizes were obtained and used in model training. According to the results obtained, 10% of the original dataset presented the lowest performance as compared to the remaining dataset sizes.

Interestingly, increasing the dataset sizes provides a corresponding increase in model performance across RF, SVM, MLP, and DF, while NB and LR had their performances maintained regardless of the size of the data. However, model performance increases across the board with an increase in the number of feature vectors from 20 (AAC) features to 400 (AAP) features (see Figs. **2**-**9**). The result presented in this section is obtained using the complete dataset. Accuracy, Sensitivity, Specificity, F1 score, Precision, Matthews Correlation Coefficient (MCC), and Area Under ROC (AUROC) were used in measuring the predictive performance of the approaches under consideration [61].

3.1. Evaluating the Feature Importance of the Proposed Method

In Section 2.2.1, it was highlighted that each feature importance method was applied to the features generated by EPF. The essence of conducting this process was to show that the features generated by the MPF and IPF, to a good extent describe any protein pair of interest. This process also shows the various levels of importance for each feature. The process also helped identify features with zero or close to zero importance. Ultimately, using multiple methods helps determine which of the methods is most advantageous to the proposed approach in this work. In addition, as shown in Figs. (**2** and **3**) for example, using the approaches helps to discover if there are overlapping features from both IPF and MPF. From the images, it was discovered that there were overlapping features; hence there was a need for checking multicollinearity and eliminating redundant features. Figs. (**2** and **3**) demonstrate that features generated using IPF (blue) and MPF (orange) contributed significantly to the outcome, regardless of the feature importance method, except for the permutation-based method, which indicated that all or majority of the features had zero importance, despite its computational complexity. To determine the most appropriate feature importance method, the five methods were evaluated on EPF using six machine learning models. As depicted in Table **2**, a score of 1 is assigned to a method that exhibits the highest predictive accuracy amongst the five methods, whereas a score of 0.5 is assigned to the second-best performing method while 0 is assigned to others. Having obtained the highest cumulative score, information gain was selected as the most favorable method to proceed with. Henceforth, the following discussions on EPF results are grounded on the feature selection conducted *via* information gain. Additionally, the features that exhibited an importance value of less than or equal to 0.00 were eliminated based on the analysis of feature importances.

3.2. Performance Evaluation of Models on Independent Protein Features

Results obtained from this research were presented using heatmaps. The darker areas of the heatmap represent higher values, while the lower values are depicted using more lighter color display. With regards to the result, Fig. (**4**) presents the performance of IPF when AAC was used as the feature encoding method. The accuracy values show DF has the highest accuracy with 85.51%, followed by RF at 82.62%, MLP at 68.78%, SVM at 67.63%, LR at 60.86%, and NB at 60.10%. The sensitivity values, which represent the proportion of actual positives that are correctly identified by the model, are highest for RF with 86.22%, followed closely by DF with 84.77%. NB has the lowest sensitivity, with 68.86%.

The specificity values, which represent the proportion of actual negatives that are correctly identified by the model, are highest for DF with 86.27%, followed by RF with 79.02%. NB has the lowest specificity, with 44.91%. Looking at the precision values, which represent the proportion of positive predictions that are true positives, DF has the highest value with 86.79%, followed by RF with 81.66%. NB has the lowest precision, with 59.34%. The F1 score values, which balance both precision and sensitivity, are highest for DF with 85.59%, followed by RF with 83.55%. LR has the lowest F1 score, with 64.23%. MCC values, which indicate the quality of the classification, are highest for DF with 71.33%, followed by RF with 65.84%. NB has the lowest MCC, with 19.21%. Finally, AUROC values, which indicate the model's ability to discriminate between positive and negative classes, are highest for DF with 85.52%, followed by RF with 82.62%.

Overall, based on these performance metrics, it appears DF performs the best in this classification task, while the naive Bayes model (NB) performs the worst.

Fig. (**5**), similar to Fig. (**4**), illustrates that various machine learning algorithms exhibit varying levels of classification accuracy. RF and DF are the most precise and accurate models, whereas the other models tested demonstrate lower accuracy and precision, with differing levels of sensitivity and specificity. Interestingly, when AAP is used as the encoding method for IPF, the model appears to perform better than when using AAC. This is evidenced by the superior accuracy rates achieved by DF (88.17%) and RF (83.38%) using AAP, compared to DF (85.51%) and RF (82.62%) using ACC.

3.3. Performance Evaluation on Merged Protein Features

As depicted in Fig. (**6**), the Deep Forest (DF) model has the highest accuracy rate and AUROC of 76.58%, indicating that it performs the best overall. It also has a high specificity

Fig. (2). EPF Feature importance with AAP. Feature importance based on information gain showing 1200 features generated using EPF with AAP as feature encoding method. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

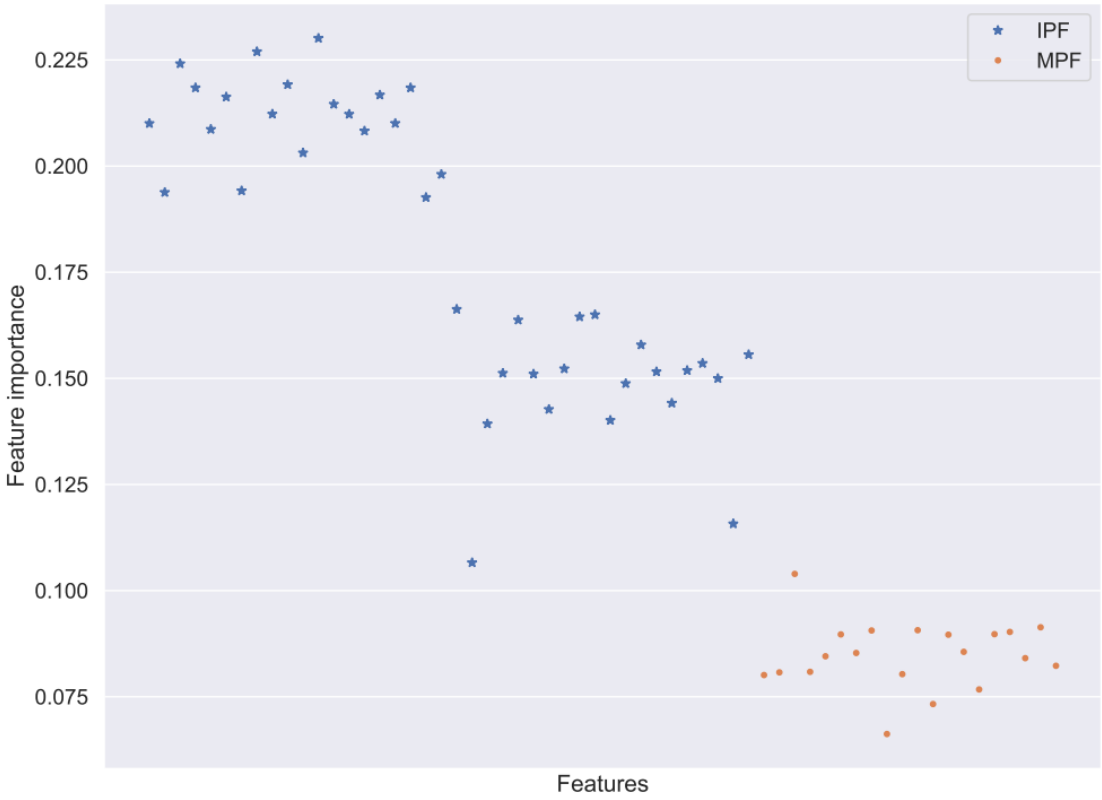

Fig. (3). EPF Feature importance with AAC. Feature importance based on information gain showing 60 features generated using EPF with AAC as feature encoding method. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

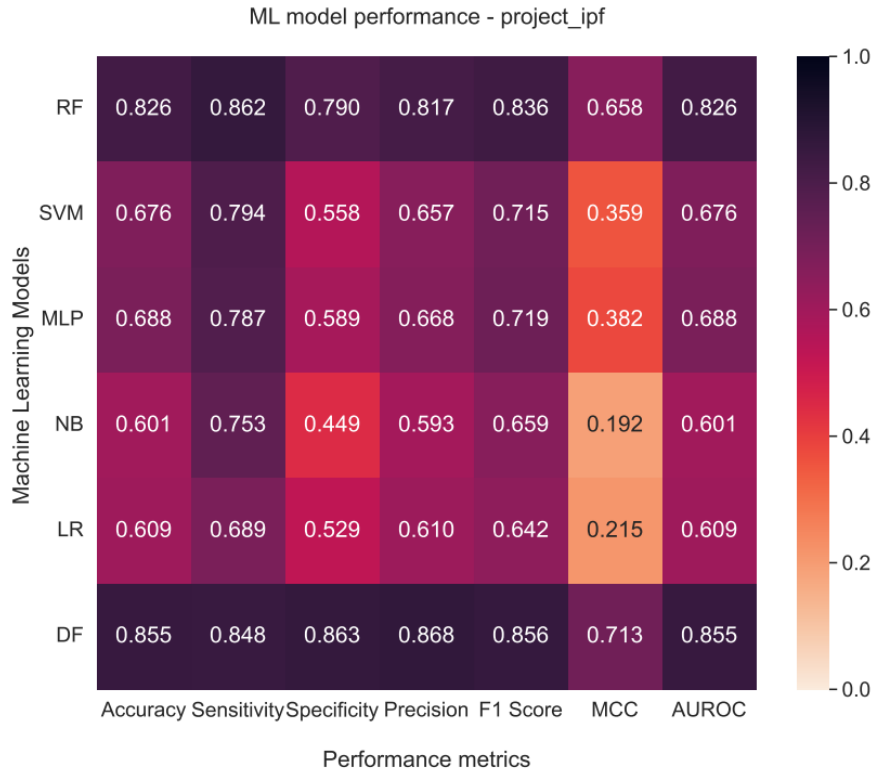

Fig. (4). IPF predictive performance with AAC. Performance evaluation of models on Independent Protein Feature (IPF) using AAC as feature encoding method. Models are random forest (RF), support vector machine (SVM), multilayer perceptron (MLP), naive Bayes (NB), logistic regression (LR), and decision tree (DF). (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Fig. (5). IPF predictive performance with AAP. Performance evaluation of models on Independent Protein Feature (IPF) using AAP as feature encoding method. Models are random forest (RF), support vector machine (SVM), multilayer perceptron (MLP), naive Bayes (NB), logistic regression (LR), and decision tree (DF). (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Table 2. Evaluating the feature importance methods on EPF.

	Information Gain	Logistic Regression	Decision Tree	Mean Decrease Impurity (RF)	Permutation-based (RF)
RF	1	0.5	0	0.5	0
SVM	1	0.5	0	0.5	0
MLP	0.5	0.5	1	0.5	0
NB	0.5	1	0	1	0
LR	0.5	1	0	1	0
DF	0.5	0	1	0	0

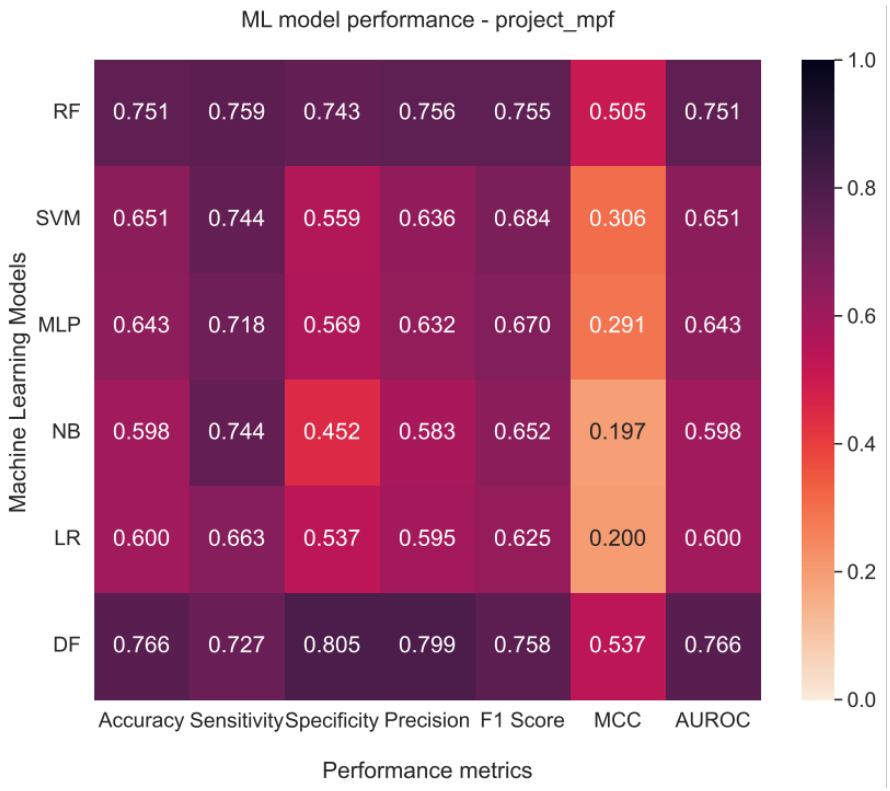

Fig. (6). Merged Protein Feature (MPF) predictive performance with AAC. Performance evaluation of models on MPF using AAC as feature encoding method. Models are random forest (RF), support vector machine (SVM), multilayer perceptron (MLP), naive Bayes (NB), logistic regression (LR), and decision tree (DF). (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

of 80.5% and a slightly lower sensitivity of 72.7%, indicating that it is good at correctly identifying negative instances but may miss some positive instances. Additionally, DF has a high precision rate of 79.9%. The Random Forest (RF) model also performs well, with an accuracy rate of 75.1% and an AUROC of 75.1%. It has a higher sensitivity of 75.9% and a slightly lower specificity of 74.3%, indicating that it is good at correctly identifying positive instances but may misclassify some negative instances. RF also has a high precision rate of 75.6. SVM, MLP, NB, and LR models all have lower levels of accuracy and precision, with varying levels of sensitivity and specificity.

Furthermore, the highest performing models based on most of the metrics used when MPF features are encoded using AAP are DF and MLP, with DF having the highest accuracy, specificity, precision, F1 Score, MCC, and AUROC as shown in Fig. (**7**). However, MLP has the highest sensitivity rate. This is in line with what was observed in the literature and several reviews that were carried out. MLP continues to outperform other models in terms of its ability to predict more interacting proteins when most of the dataset and features are put into perspective. RF, SVM, NB, and LR have lower levels of performance across most of the metrics.

3.4. Performance Evaluation of the Proposed System

Figs. (**8** and **9**) depict the performance of the machine learning model on feature vectors generated by the method developed in this study, as discussed in Section 3, utilizing AAC and AAP as the feature encoding methods, respectively. Using AAC, the highest accuracy rate is achieved by DF at 84.11%, followed by RF at 81.30%, MLP at 71.44%, SVM at 68.34%, LR at 61.40%, and NB at 60.12%. Therefore, DF and RF are the highest-performing models in terms

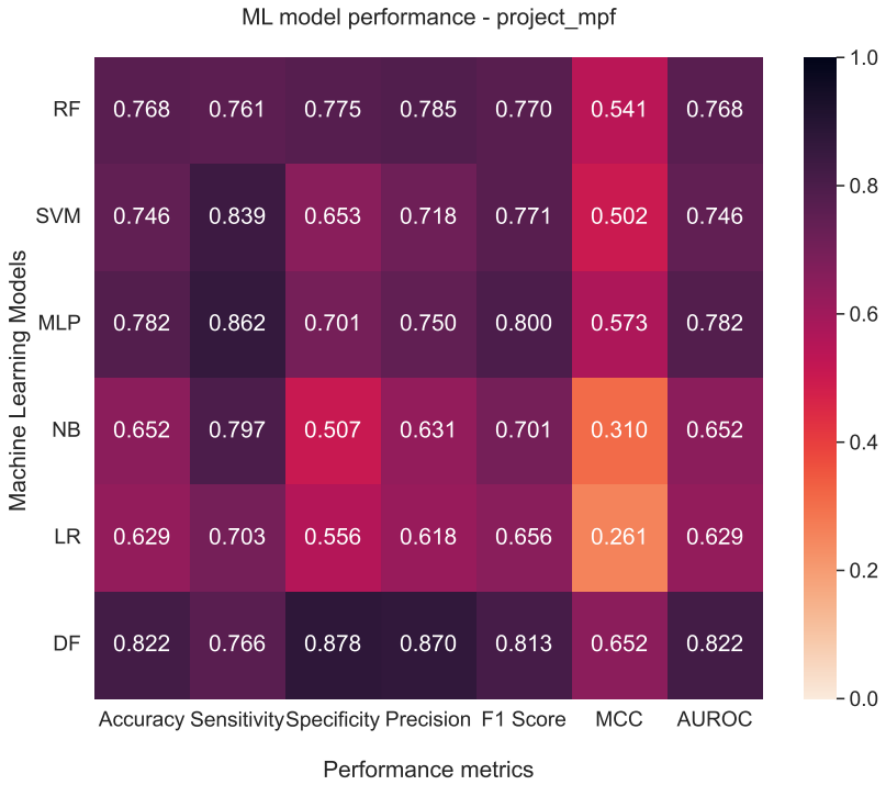

Fig. (7). Merged Protein Feature (MPF) predictive performance with AAP. Performance evaluation of models on Merged Protein Feature (MPF) using AAP as feature encoding method. Models are random forest (RF), support vector machine (SVM), multilayer perceptron (MLP), naive Bayes (NB), logistic regression (LR), and decision tree (DF). (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

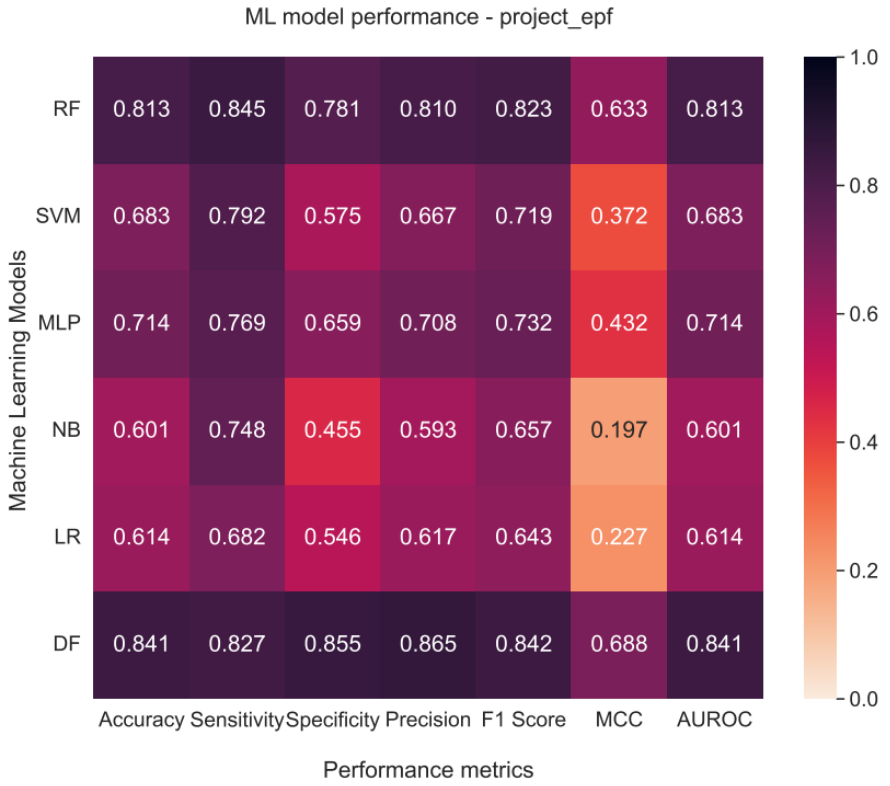

Fig. (8). Extended Protein Feature (EPF) predictive performance with AAC. Performance evaluation of models on EPF using AAC as feature encoding method. Models are random forest (RF), support vector machine (SVM), multilayer perceptron (MLP), naive Bayes (NB), logistic regression (LR), decision tree (DF). (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Fig. (9). Extended Protein Feature (EPF) predictive performance with AAP. Performance evaluation of models on EPF using AAP as feature encoding method. Models are random forest (RF), support vector machine (SVM), multilayer perceptron (MLP), naive Bayes (NB), logistic regression (LR), decision tree (DF). (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

of accuracy. DF also has the highest sensitivity rate at 82.74%, followed by RF at 84.54%, MLP at 76.95%, NB at 74.78%, SVM at 79.17%, and LR at 68.18%. Therefore, DF and RF are the highest-performing models in terms of sensitivity. Similarly, DF has the highest specificity rate at 85.49%, followed by RF at 78.07%, MLP at 65.93%, SVM at 57.52%, LR at 54.62%, and NB at 45.48%. Therefore, DF presents the highest-performing model in terms of specificity.

Following the same trend but with an improved performance than using AAC, the EPF approach with the AAP encoding method displays DF as the best model with 87.25% accuracy, followed by MLP at 81.38%, RF at 81.28%, SVM at 77.84%, LR at 68.72%, and NB at 67.27%. Similar to the result of IPF and MPF, MLP achieved the highest sensitivity with 89.86% performance. However, DF shows much better strength than other models with regard to the remaining metrics. Again, and notably so, NB displays the lowest predictive performance across the majority of metrics.

3.5. Comparative Analysis of the Evaluated Approaches

A summarized performance of IPF, MPF, and EPF using six machine learning models is displayed in Figs. (**10** and **11**) using low (AAC) and high (AAP) dimensional feature vectors. As seen from the Figure and the previous figures presented, the three approaches follow a similar trend in terms of the machine learning model performance, although with a variety of values. For instance, the best-performing model across the 3 approaches is DF, and the least-performing is NB. However, contrasting these approaches, there is a considerable improvement of SVM, MLP, NB, and LR models when the EPF method is used for feature representation of host-pathogen protein sequences. In addition, EPF and IPF generally outperform MPF across all the models tested irrespective of the feature encoding method employed, as shown in Figs. (**10** and **11**). However, IPF presented better performance than EPF and MPF consistently across the decision tree models (RF and DF), thereby presenting room for further improvement of the proposed method. Furthermore, Table **3** presents the confidence intervals for each of the machine learning models across the three approaches. The accuracy confidence interval was calculated using the t-distribution with a confidence level of 95%. This implies that with 95% confidence, the true value of the different accuracies obtained in this work falls within the range provided in Table **3**. The significance of the small values obtained means that the accuracy produced has a more precise estimate. Smaller confidence intervals suggest that the range of possible values has been narrowed.

Furthermore, to optimize the performance of these techniques, hyperparameters were tuned using the GridSearchCV library in Python. The results of the experiment are presented in Supplementary Figs. (**S1** and **S2**), and it was found that there was no significant difference between the results obtained using EPF and IPF with GridSearchCV hyperparameter tuning. Supplementary File **S3** provides more detailed

Table 3. Accuracy confidence interval (CI).

	IPF Accuracy CI (+/-)	MPF Accuracy CI (+/-)	EPF Accuracy CI (+/-)
RF	0.046	0.047	0.045
SVM	0.053	0.052	0.057
MLP	0.019	0.036	0.025
NB	0.081	0.066	0.078
LR	0.051	0.036	0.044
DF	0.028	0.028	0.033

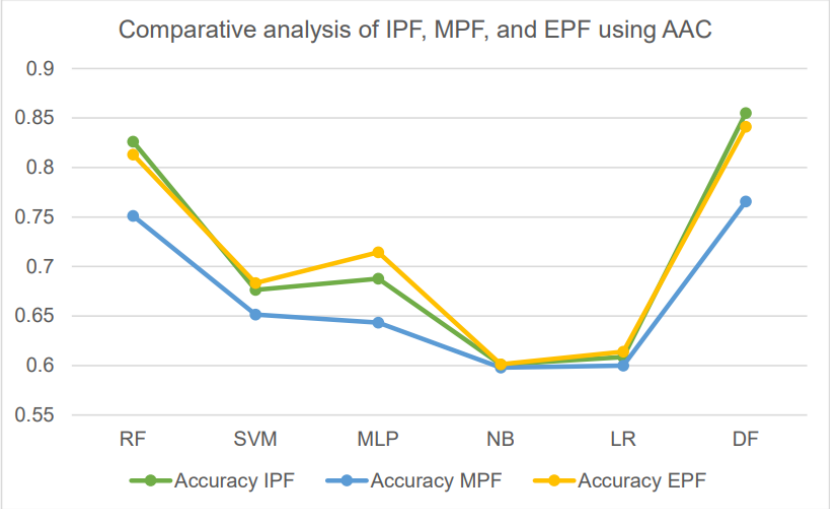

Fig. (10). Comparative analysis of the evaluated approaches using Amino Acid Composition. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

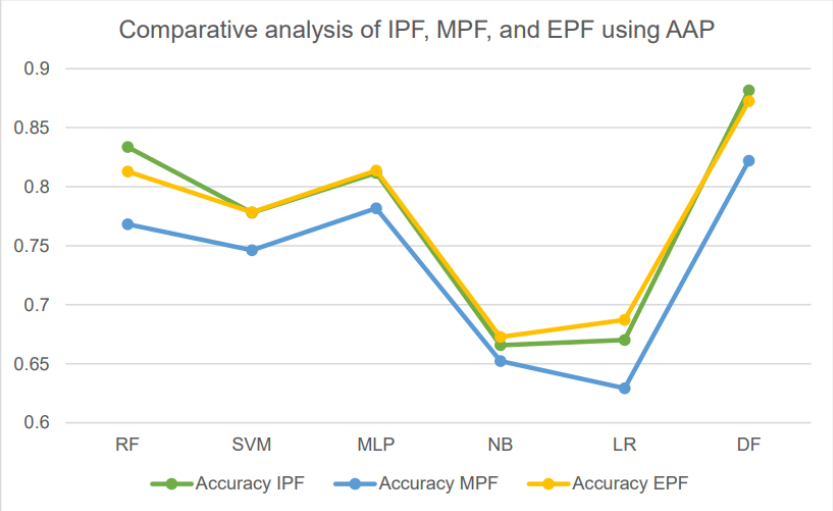

Fig. (11). Comparative analysis of the evaluated approaches using Amino Acid Pairs. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

information about the specific hyperparameters used for the experiment.

4. DISCUSSION

Overall, our comparative analysis revealed that the EPF approach was the most effective in terms of prediction accuracy, outperforming other approaches in four out of the six machine learning models tested (SVM, MLP, NB, and LR). This consistent performance was observed regardless of the feature encoding method used, whether it was low or high dimensional. These results indicate that the EPF approach is well-suited for accurately representing the features of protein

sequences, making it a promising method for predicting HPPPI.

On the other hand, the IPF approach produced better prediction results for the decision tree-based models employed (DF and RF), suggesting that it is better suited for decision tree models. This observation was consistent with the findings reported in previous subsections (see Figs. **4**, **5**, **10**, and **11**), further supporting the adoption of IPF for feature representation of protein sequences.

MPF approach presented the least performance across the majority of the metrics under consideration. This suggests that the MPF method may not adequately collect the necessary data about protein sequences for use in future predictions and may not be the most appropriate approach for predicting HPPPI.

Although the NB and LR models did not perform well in predicting HPPPI, the proposed approach was able to achieve some level of improvement in the model performance. This indicates that the adoption of the proposed approach would greatly impact the accurate prediction of HPPPI regardless of the dataset type, dataset size, and feature encoding method used.

Furthermore, performing a 10-fold cross-validation training using all the earlier mentioned models with an unbalanced dataset presented a significant increase in performance accuracy across all the models (with NB and LR performing lowest), as shown in Supplementary Files S4, S5, and S6. However, we observed that due to the unbalanced nature of the training datasets where the number of negative pairs (non-interacting) outweighed the positive (interacting) pairs, the models were more tilted towards predicting more non-interacting pairs more than the interacting ones with the highest sensitivity of 80.3% as compared to 99.6% specificity using the EPF approach.

Overall, our findings demonstrate the effectiveness of the EPF approach for predicting HPPPI and highlight the importance of carefully selecting feature encoding methods when developing predictive models for complex biological problems like HPPPI.

CONCLUSION

This study explores the effectiveness of existing approaches, namely Independent Protein Features (IPF) and Merged Protein Features (MPF), alongside the novel EPF method. We employ datasets encompassing host-pathogen interactions involving bacteria, parasites, viruses, and plants. These datasets are integrated into machine learning models, such as Random Forest (RF), Support Vector Machine (SVM), Multilayer Perceptron (MLP), Naïve Bayes (NB), Logistic Regression (LR), and Deep Forest (DF). Our experimental results provide compelling evidence in favor of the widespread adoption of IPF over MPF due to its superior capability to represent the intricate compositional aspects of protein sequences, thereby enhancing machine learning predictions. Conversely, the EPF approach developed in this study exhibits substantial improvements in four out of the six models evaluated. This suggests that EPF offers competitiveness with IPF and is particularly well-suited for traditional machine learning models, including Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), and Multilayer Perceptron (MLP). It is worth noting that IPF outperforms EPF in decision tree-based models, such as Random Forest (RF) and Deep Forest (DF), a consistent trend observed across all three approaches, irrespective of the encoding method employed, be it amino acid composition or amino acid pairs. Furthermore, to enhance the generalizability of our approach, we conducted training and testing on a diverse range of host-pathogen pairs. Robustness was ensured through a cross-validation technique involving a ten-fold split, with nine folds allocated for training and one for testing, repeated iteratively until all folds were employed for testing.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No animals/humans were used in this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The source code for this article and its corresponding datasets is available in the GitHub repository at https://github.com/JerrySteam/extended-protein-feature.

FUNDING

This research was funded by Covenant Applied Informatics and Communication African Centre of Excellence (CApIC-ACE) funded by the World Bank Project.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's website along with the published article.

REFERENCES

[1] Zhang B, Li J, Quan L, Chen Y, Lü Q. Sequence-based prediction of protein-protein interaction sites by simplified long short-term memory network. Neurocomputing 2019; 357: 86-100. http://dx.doi.org/10.1016/j.neucom.2019.05.013

[2] Ziegler SJ, Mallinson SJB, St John PC, Bomble YJ. Advances in integrative structural biology: Towards understanding protein complexes in their cellular context. Comput Struct Biotechnol J 2020; 19: 214-25. http://dx.doi.org/10.1016/j.csbj.2020.11.052 PMID: 33425253

[3] Richards AL, Eckhardt M, Krogan NJ. Mass spectrometry-based protein-protein interaction networks for the study of human diseases. Mol Syst Biol 2021; 17(1): e8792. http://dx.doi.org/10.15252/msb.20188792 PMID: 33434350

[4] Meldal BHM, Perfetto L, Combe C, *et al.* Complex portal 2022: New curation frontiers. Nucleic Acids Res 2022; 50(D1): D578-86. http://dx.doi.org/10.1093/nar/gkab991 PMID: 34718729

[5] Khatun MS, Shoombuatong W, Hasan MM, Kurata H. Evolution of sequence-based bioinformatics tools for protein-protein interaction prediction. Curr Genomics 2020; 21(6): 454-63. http://dx.doi.org/10.2174/1389202921999200625103936 PMID: 33093807

[6] Marchand A, Van Hall-Beauvais AK, Correia BE. Computational design of novel protein-protein interactions - An overview on methodological approaches and applications. Curr Opin Struct Biol 2022; 74: 102370. http://dx.doi.org/10.1016/j.sbi.2022.102370 PMID: 35405427

[7] Balasubramanian K, Gupta SP. Quantum molecular dynamics, topological, group theoretical and graph theoretical studies of protein-protein interactions. Curr Top Med Chem 2019; 19(6): 426-43. http://dx.doi.org/10.2174/1568026619666190304152704 PMID: 30836919

[8] Heifetz A, Sladek V, Townsend-Nicholson A, Fedorov DG. Characterizing protein-protein interactions with the fragment molecular orbital method. Methods Mol Biol 2020; 2114: 187-205. http://dx.doi.org/10.1007/978-1-0716-0282-9_13 PMID: 32016895

[9] Yugandhar K, Gupta S, Yu H. Inferring protein-protein interaction networks from mass spectrometry-based proteomic approaches: A mini-review. Comput Struct Biotechnol J 2019; 17: 805-11. http://dx.doi.org/10.1016/j.csbj.2019.05.007 PMID: 31316724

[10] Sun T, Zhou B, Lai L, Pei J. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC Bioinformatics 2017; 18(1): 277. http://dx.doi.org/10.1186/s12859-017-1700-2 PMID: 28545462

[11] Rosa S, Bertaso C, Pesaresi P, Masiero S, Tagliani A. Synthetic protein circuits and devices based on reversible protein-protein interactions: An overview. Life 2021; 11(11): 1171. http://dx.doi.org/10.3390/life11111171 PMID: 34833047

[12] Murakami Y, Mizuguchi K. Recent developments of sequence-based prediction of protein-protein interactions. Biophys Rev 2022; 14(6): 1393-411. http://dx.doi.org/10.1007/s12551-022-01038-1 PMID: 36589735

[13] Nussbaumer T. Host_microbe_PPI - R package to analyse intra-species and inter-species protein-protein interactions in the model plant arabidopsis thaliana. bioRxiv 2019; 551275. http://dx.doi.org/10.1101/551275

[14] Dick K, Samanfar B, Barnes B, *et al.* PIPE4: Fast PPI predictor for comprehensive inter- and cross-species interactomes. Scientific Reports 2020; 10(1): 1-15. http://dx.doi.org/10.1038/s41598-019-56895-w

[15] Sunggawa MI, Bustamam A, Siswantining T. Sequence-based prediction of pathogen-host interaction using an ensemble learning classifier and moran autocorrelation feature encoding method. TURCOMAT 2021; 12(14): 598-605.

[16] Ghedira K, Hamdi Y, El Béji A, Othman H. An integrative computational approach for the prediction of human-*plasmodium* protein-protein interactions. BioMed Res Int 2020; 2020: 1-11. http://dx.doi.org/10.1155/2020/2082540 PMID: 33426052

[17] Chen H, Guo W, Shen J, Wang L, Song J. Structural principles analysis of host-pathogen protein-protein interactions: A structural bioinformatics survey. IEEE Access 2018; 6: 11760-71. http://dx.doi.org/10.1109/ACCESS.2018.2807881

[18] Sironi M, Cagliani R, Forni D, Clerici M. Evolutionary insights into host-pathogen interactions from mammalian sequence data. Nat Rev Genet 2015; 16(4): 224-36. http://dx.doi.org/10.1038/nrg3905 PMID: 25783448

[19] Engering A, Hogerwerf L, Slingenbergh J. Pathogen-host-environment interplay and disease emergence. Emerg Microbes Infect 2012; 2013: 2. http://dx.doi.org/10.1038/emi.2013.5 PMID: 26038452

[20] Steps E, Causation DD. Chapter 4-biomedical research chapter 4-lesson 4 host-pathogen interactions. Biomed Res 2012; •••: 123-8.

[21] Chen H, Shen J, Wang L, Song J. Towards data analytics of pathogen-host protein-protein interaction: A survey. Proceedings - 2016 IEEE International Congress on Big Data, BigData Congress . 377-88. http://dx.doi.org/10.1109/BigDataCongress.2016.60

[22] Kaur A, Kaur P, Ahuja S. Förster resonance energy transfer (FRET) and applications thereof. Anal Methods 2020; 12(46): 5532-50. http://dx.doi.org/10.1039/D0AY01961E PMID: 33210685

[23] Chrétien AÈ, Gagnon-Arsenault I, Dubé AK, *et al.* Extended linkers improve the detection of protein-protein interactions (PPIs) by dihydrofolate reductase protein-fragment complementation assay (DHFR PCA) in living cells. Mol Cell Proteomics 2018; 17(2): 373-83. http://dx.doi.org/10.1074/mcp.TIR117.000385 PMID: 29203496

[24] Pichlerova K, Hanes J. Technologies for the identification and validation of protein-protein interactions. Gen Physiol Biophys 2021; 40(6): 495-522. http://dx.doi.org/10.4149/gpb_2021035 PMID: 34897023

[25] Ho Y, Gruhler A, Heilbut A, *et al.* Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 2002; 415(6868): 180-3. http://dx.doi.org/10.1038/415180a PMID: 11805837

[26] Formstecher E, Aresta S, Collura V, *et al.* Protein interaction mapping: A *drosophila* case study. Genome Res 2005; 15(3): 376-84. http://dx.doi.org/10.1101/gr.2659105 PMID: 15710747

[27] Nicod C, Banaei-Esfahani A, Collins BC. Elucidation of host-pathogen protein-protein interactions to uncover mechanisms of host cell rewiring. Curr Opin Microbiol 2017; 39: 7-15. http://dx.doi.org/10.1016/j.mib.2017.07.005 PMID: 28806587

[28] Khorsand B, Savadi A, Zahiri J, Naghibzadeh M. Alpha influenza virus infiltration prediction using virus-human protein-protein interaction network. Math Biosci Eng 2020; 17(4): 3109-29. http://dx.doi.org/10.3934/mbe.2020176 PMID: 32987519

[29] Chen H, Li F, Wang L, *et al.* Systematic evaluation of machine learning methods for identifying human-pathogen protein-protein interactions. Brief Bioinform 2021; 22(3): bbaa068. http://dx.doi.org/10.1093/bib/bbaa068 PMID: 32459334

[30] Aromolaran O, Aromolaran D, Isewon I, Oyelade J. Machine learning approach to gene essentiality prediction: A review. In: Briefings in Bioinformatics. Oxford University Press 2021. http://dx.doi.org/10.1093/bib/bbab128

[31] Loaiza C. Prediction of host-pathogen protein-protein interactions using. In: Student Research Symposium. Utah State University 2019.

[32] Brierley L, Fowler A. Predicting the animal hosts of coronaviruses from compositional biases of spike protein and whole genome sequences through machine learning. PLoS Pathog 2021; 17(4): e1009149. http://dx.doi.org/10.1371/journal.ppat.1009149 PMID: 33878118

[33] Prasasty VD, Hutagalung RA, Gunadi R, *et al.* Prediction of human-Streptococcus pneumoniae protein-protein interactions using logistic regression. Comput Biol Chem 2021; 92(March): 107492. http://dx.doi.org/10.1016/j.compbiolchem.2021.107492 PMID: 33964803

[34] Vyas R, Bapat S, Goel P, Karthikeyan M, Tambe SS, Kulkarni BD. Application of genetic programming (GP) formalism for building disease predictive models from protein-protein interactions (PPI) data. IEEE/ACM Trans Comput Biol Bioinformatics 2018; 15(1): 27-37. http://dx.doi.org/10.1109/TCBB.2016.2621042 PMID: 28113781

[35] Taha K. Employing Machine Learning Techniques to Detect Protein-Protein Interaction: A Survey, Experimental, and Comparative Evaluations. bioRxiv 2023; 2023.08.22.554321. http://dx.doi.org/10.1101/2023.08.22.554321

[36] Angelis D, Sofos F, Karakasidis TE. Artificial intelligence in physical sciences: Symbolic regression trends and perspectives. Arch Comput Methods Eng 2023; 30(6): 3845-65. http://dx.doi.org/10.1007/s11831-023-09922-z PMID: 37359747

[37] Papastamatiou K, Sofos F, Karakasidis TE. Machine learning symbolic equations for diffusion with physics-based descriptions. AIP Adv 2022; 12(2): 025004. http://dx.doi.org/10.1063/5.0082147

[38] Paturi UMR, Cheruku S. Application and performance of machine learning techniques in manufacturing sector from the past two decades: A review. Mater Today Proc 2021; 38: 2392-401. http://dx.doi.org/10.1016/j.matpr.2020.07.209

[39] Dogan A, Birant D. Machine learning and data mining in manufacturing. Expert Syst Appl 2021; 166(166): 114060. http://dx.doi.org/10.1016/j.eswa.2020.114060
[40] Nguyen QH, Ly HB, Ho LS, *et al.* Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. Math Probl Eng 2021; 2021: 1-15. http://dx.doi.org/10.1155/2021/4832864
[41] Yang L, Xia JF, Gui J. Prediction of protein-protein interactions from protein sequence using local descriptors. Protein Pept Lett 2010; 17(9): 1085-90. http://dx.doi.org/10.2174/092986610791760306 PMID: 20509850
[42] Bell E W, Schwartz J H, Freddolino P L, Zhang Y. PEPPI: Whole-proteome protein-protein interaction prediction through structure and sequence similarity, functional association, and machine learning. J Mol Biol 2022; 167530. http://dx.doi.org/10.1016/j.jmb.2022.167530
[43] Dong TN, Brogden G, Gerold G, Khosla M. A multitask transfer learning framework for the prediction of virus-human protein-protein interactions. BMC Bioinformatics 2021; 22(1): 572. http://dx.doi.org/10.1186/s12859-021-04484-y PMID: 34837942
[44] labxchange Anfinsen's Experiment Shows That Primary Structure Determines Protein Conformation - LabXchange Available from: https://www.labxchange.org/library/items/lb:LabXchange:e17fa649:html:1 (accessed 2023-12-11).
[45] ANFINSEN C B, HABER E, SELA M, WHITE J. F H . The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc Natl Acad Sci 1961; 47(9): 28. http://dx.doi.org/10.1063/1.3066543
[46] Charih F, Biggar KK, Green JR. Assessing sequence-based protein-protein interaction predictors for use in therapeutic peptide engineering. Sci Rep 2022; 12(1): 9610. http://dx.doi.org/10.1038/s41598-022-13227-9 PMID: 35688894
[47] Göktepe YE, Kodaz H. Prediction of protein-protein interactions using an effective sequence based combined method. Neurocomputing 2018; 303: 68-74. http://dx.doi.org/10.1016/j.neucom.2018.03.062
[48] Chen M, Ju CJT, Zhou G, *et al.* Multifaceted protein-protein interaction prediction based on Siamese residual RCNN. Bioinformatics 2019; 35(14): i305-14. http://dx.doi.org/10.1093/bioinformatics/btz328 PMID: 31510705
[49] Liu L, Zhu X, Ma Y, *et al.* Combining sequence and network information to enhance protein-protein interaction prediction. BMC Bioinformatics 2020; 21(S16)(Suppl. 16): 537. http://dx.doi.org/10.1186/s12859-020-03896-6 PMID: 33323120
[50] Dyer MD, Murali TM, Sobral BW. Supervised learning and prediction of physical interactions between human and HIV proteins. Infect Genet Evol 2011; 11(5): 917-23. http://dx.doi.org/10.1016/j.meegid.2011.02.022 PMID: 21382517
[51] Chen C, Zhang Q, Ma Q, Yu B. LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion. Chemom Intell Lab Syst 2019; 191(May): 54-64. http://dx.doi.org/10.1016/j.chemolab.2019.06.003
[52] Afify HM, Zanaty MS. Computational predictions for protein sequences of COVID-19 virus *via* machine learning algorithms. Med Biol Eng Comput 2021; 59(9): 1723-34. http://dx.doi.org/10.1007/s11517-021-02412-z PMID: 34291385
[53] Wang S, Liu S. Protein sub-nuclear localization based on effective fusion representations and dimension reduction algorithm LDA. Int J Mol Sci 2015; 16(12): 30343-61. http://dx.doi.org/10.3390/ijms161226237 PMID: 26703574
[54] Nieto JJ, Torres A, Georgiou DN, Karakasidis TE. Fuzzy polynucleotide spaces and metrics. Bull Math Biol 2006; 68(3): 703-25. http://dx.doi.org/10.1007/s11538-005-9020-5 PMID: 16794951
[55] Jha K, Saha S, Tanveer M. Prediction of protein-protein interactions using stacked auto-encoder. Trans Emerg Telecommun Technol 2020; 2021(November): 1-13. http://dx.doi.org/10.1002/ett.4256
[56] Kösesoy İ, Gök M, Öz C. A new sequence based encoding for prediction of host - pathogen protein interactions. Comput Biol Chem 2019; 78: 170-7. http://dx.doi.org/10.1016/j.compbiolchem.2018.12.001
[57] Dey L, Chakraborty S, Mukhopadhyay A. Machine learning techniques for sequence-based prediction of viral-host interactions between SARS-CoV-2 and human proteins. Biomed J 2020; 43(5): 438-50. http://dx.doi.org/10.1016/j.bj.2020.08.003 PMID: 33036956
[58] Ding Z, Kihara D. Computational methods for predicting protein-protein interactions using various protein features. Curr Protoc Protein Sci 2018; 93(1): e62. http://dx.doi.org/10.1002/cpps.62 PMID: 29927082
[59] Zhang L, Yu G, Guo M, Wang J. Predicting protein-protein interactions using high-quality non-interacting pairs. BMC Bioinformatics 2018; 19(S19)(Suppl. 19): 525. http://dx.doi.org/10.1186/s12859-018-2525-3 PMID: 30598096
[60] Soyemi J, Isewon I, Oyelade J, Adebiyi E. Inter-species/host-parasite protein interaction predictions reviewed. Curr Bioinform 2018; 13(4): 396-406. http://dx.doi.org/10.2174/1574893613666180108155851 PMID: 31496926
[61] Kösesoy İ, Gök M, Kahveci T. Prediction of host-pathogen protein interactions by extended network model. Turk J Biol 2021; 45(2): 138-48. http://dx.doi.org/10.3906/biy-2009-4 PMID: 33907496
[62] Wuchty S. Computational prediction of host-parasite protein interactions between P. falciparum and H. sapiens. PLoS One 2011; 6(11): e26960. http://dx.doi.org/10.1371/journal.pone.0026960 PMID: 22114664
[63] Gordon DE, Jang GM, Bouhaddou M, *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature 2020; 583(7816): 459-68. http://dx.doi.org/10.1038/s41586-020-2286-9 PMID: 32353859
[64] Mukhtar MS, Carvunis AR, Dreze M, *et al.* Independently evolved virulence effectors converge onto hubs in a plant immune system network. Science 2011; 333(6042): 596-601. http://dx.doi.org/10.1126/science.1203659 PMID: 21798943
[65] Pacal I, Karaman A, Karaboga D, *et al.* An efficient real-time colonic polyp detection with YOLO algorithms trained by using negative samples and large datasets. Comput Biol Med 2022; 141(141): 105031. http://dx.doi.org/10.1016/j.compbiomed.2021.105031 PMID: 34802713
[66] Najm M, Azencott CA, Playe B, Stoven V. Drug target identification with machine learning: How to choose negative examples. Int J Mol Sci 2021; 22(10): 5118. http://dx.doi.org/10.3390/ijms22105118 PMID: 34066072
[67] Oyelade J, Isewon I, Rotimi S, Okunoren I. Modeling of the glycolysis pathway in *plasmodium falciparum* using petri nets. Bioinform Biol Insights 2016; 10: BBI.S37296. http://dx.doi.org/10.4137/BBI.S37296 PMID: 27199550
[68] Soyemi J, Isewon I, Oyelade J, Adebiyi E. Functional enrichment of human protein complexes in malaria parasites. Proceedings of the IEEE International Conference on Computing, Networking and Informatics, ICCNI 2017. 1-6. http://dx.doi.org/10.1109/ICCNI.2017.8123791
[69] Impact of Malaria Worldwide. Centers for Disease Control and Prevention 2020.
[70] Mayo C. Malaria transmission cycle Available from: https://www.mayoclinic.org/diseases-conditions/malaria/multimedia/malaria-transmission-cycle/img-20006373
[71] Tracking progress against malaria World Malaria Report 2021.
[72] Weßling R, Epple P, Altmann S, *et al.* Convergent targeting of a common host protein-network by pathogen effectors from three kingdoms of life. Cell Host Microbe 2014; 16(3): 364-75. http://dx.doi.org/10.1016/j.chom.2014.08.004 PMID: 25211078
[73] Agany D D M, Pietri J E, Gnimpieba E Z. Assessment of vector-host-pathogen relationships using data mining and machine learning. Comput Struct Biotechnol J 2020; 1704-21. http://dx.doi.org/10.1016/j.csbj.2020.06.031
[74] Chen H, Shen J, Wang L, Chi CH. APEX2S: A two-layer machine learning model for discovery of host-pathogen protein-protein interactions on cloud-based multiomics data. Concurr Comput 2020; 32(23): e5846. http://dx.doi.org/10.1002/cpe.5846
[75] Chen Z, Zhao P, Li F, *et al. iFeature* : A python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics 2018; 34(14): 2499-502. http://dx.doi.org/10.1093/bioinformatics/bty140 PMID: 29528364

[76] Hjerpe A. Degree project in the field of technology computing random forests variable importance measures (VIM) on mixed continuous and categorical data computing random forests variable importance measures (VIM) on mixed numerical and categorical data beräknin. No. Vim 2016.

[77] Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: A corrected feature importance measure. Bioinformatics 2010; 26(10): 1340-7. http://dx.doi.org/10.1093/bioinformatics/btq134 PMID: 20385727

[78] Grąbczewski K, Jankowski N. Feature selection with decision tree criterion. HIS 2005: Fifth International Conference on Hybrid Intelligent Systems. 212-7. http://dx.doi.org/10.1109/ICHIS.2005.43

[79] Kazemitabar SJ, Amini AA, Bloniarz A, Talwalkar A. Variable importance using decision trees. In: Adv Neural Inf Process Syst. 2017; pp. 426-35.

[80] Cheng Q, Varshney PK, Arora MK. Logistic regression for feature selection and soft classification of remote sensing data. IEEE Geosci Remote Sens Lett 2006; 3(4): 491-4. http://dx.doi.org/10.1109/LGRS.2006.877949

[81] Azhagusundari B, Thanamani AS. Feature selection based on information gain. Int J Innov Technol Explor Eng 2013; 2(2): 18-21.

[82] Shaltout NA, El-Hefnawi M, Rafea A, Moustafa A. Information gain as a feature selection method for the efficient classification of influenza based on viral hosts. Lect Notes Eng Comput Sci 2014; 1(July): 625-31.