DeepEnzyPred: A Bi-Layered Deep Learning Framework for prediction of Bacteriophage Enzymes and their Sub-Hydrolases Enzymes via Novel Multi Level- Multi Thresholds Feature Selection technique

Yu Wang
College of Information Science and Engineering
ZAHEER ULLAH KHAN (zaheerkhan@nuaa.edu.cn)
School of Computer Science and Technology, NUAA Nanjing, China https://orcid.org/0000-0003-2263-6109
Shaukat Ali
Islamia College Peshawar
Maqsood Hayat
Abdul Wali Khan University Mardan

Research article

Keywords: Bacteriophage enzymes, hydrolase, MLMT-SFS, Secondary Sequence feature, Deep Learning

Posted Date: November 9th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-72347/v1

DeepEnzyPred: A Bi-Layered Deep Learning Framework for prediction of Bacteriophage Enzymes and their Sub-Hydrolases Enzymes via Novel Multi Level- Multi Thresholds Feature Selection technique

Yu Wang[1], Zaheer Ullah Khan[2*] ,Shaukat Ali[3], Maqsood Hayat[4]

[1]College of Information Science and Engineering, Shandong Agricultural University, China
[2]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China
[3]Department of Computer Science, Islamic College Peshawar,Pakistan
[4]Department of Computer Science, Abdul Wali Khan University Mardan,Pakistan

Author's email addresses:
Yu Wang: yu.wang@sdau.edu.cn
Zaheer Ullah Khan: zaheerkhan@nuaa.edu.cn
Shaukat Ali: shaukatali@icp.edu.pk
Maqsood Hayat: m.hayat@awkum.edu.pk

Corresponding Author Name & Email: Zaheer Ullah Khan: zaheerkhan@nuaa.edu.cn

Abstract

Background

Bacteriophage or phage is a type of virus that replicates itself inside bacteria. It consist of genetic material surrounded by a protein structure. Bacteriophage plays a vital role in the domain of phage therapy and genetic engineering. Phage and hydrolases enzyme proteins have a significant impact on the cure of pathogenic bacterial infections and disease treatment. Accurate identification of bacteriophage proteins is important in the host subcellular localization for further understanding of the interaction between phage, hydrolases, and in designing antibacterial drugs. Looking at the significance of Bacteriophage proteins, besides wet laboratory-based methods several computational models have been developed so far. However, the performance was not considerable due to inefficient feature schemes, redundancy, noise, and lack of an intelligent learning engine. Therefore we have developed an anovative bi-layered model name DeepEnzyPred. A Hybrid feature vector was obtained via a novel Multi-Level Multi-Threshold subset feature selection (MLMT-SFS) algorithm. A two-dimensional convolutional neural network was adopted as a baseline classifier.

Results

A conductive hybrid feature was obtained via a serial combination of CTD and KSAACGP features. The optimum feature was selected via a Novel Multi-Level Multi-Threshold Subset Feature selection algorithm. Over 5-fold jackknife cross-validation, an accuracy of 91.6 %, Sensitivity of 63.39%, Specificity 95.72%, MCC of 0.6049, and ROC value of 0.8772 over Layer-1 were recorded respectively. Similarly, the underline model obtained an Accuracy of 96.05%, Sensitivity of 96.22%, Specificity of 95.91%, MCC of 0.9219, and ROC value of 0.9899 over layer-2 respectivily.

Conclusion

This paper presents a robust and effective classification model was developed for bacteriophage and their types. Primitive features were extracted via CTD and KSAACGP. A novel method (MLMT-SFS) was devised for yielding optimum hybrid feature space out of primitive features. The result drew over hybrid feature space and 2D-CNN shown an excellent classification. Based on the recorded results, we believe that the developed predictor will be a valuable resource for large scale discrimination of unknown Phage and hydrolase enzymes in particular and new antibacterial drug design in pharmaceutical companies in general.

Keywords: Bacteriophage enzymes; hydrolase, MLMT-SFS, Secondary Sequence feature, Deep Learning

1. Background

Bacteriophages are among the most widely recognized and diverse elements in the biosphere[1][2]. They are also known as Phage, which remains a natural enemy of the bacterium by still keeps specificity for pathogenic bacteria and beneficial flora. Phages replicate itself within the bacterium by injecting their genome into its cytoplasm. It is estimated that there are more than 10^{31} bacteriophages on the planet, more than every other organism on Earth, including bacteria, combined. Phage-coded hydrolyses is a key component of cleavage and helps fight bacterial pathogens, especially those that cannot be killed by antibiotics and chemicals. Research studies have shown[3][4], that phage was extensively used as a cure for those bacterial infections, which don't respond to anti-biotic[4]–[7].

Overconsumption of antibiotics is the most important factor leading to antibiotic resistance all over the world. Some drug-resistant viruses cannot be effectively controlled due to the abuse of antibiotics. This problem can be solved by phage hydrolysis therapy, which breaks down the host-virus during the release of the offspring of the bacteriophage[10], [11]. Besides these, they are also used as food safety tools to reduce bacterial contamination. As a result, there is an increasing demand in public health domain for the rapid detection of phages and hydrolytic enzymes [8][9].

Phage therapy has some advantages over antibiotic therapy, as follows: cost-effectiveness, facile isolation, and purification from the living environment, abundance in the environment, strong specific effects on bacteria, and low side effects. Along with study of phage hydrolysis enzyme got momentum, in studying host cell lysis activated by hydrolysase, it was found that calcium can regulate bacteriotic lysis induced by phage[12]. Therefore, the correct identification of hydrolytic enzymes encoded in phage has become an important research topic. Although various technological wet-lab techniques such as mass spectrometry have been developed to annotate the phage proteins from sequence data but these biochemical experimental techniques are overpriced and time-consuming. The computational methods provide the best opportunity to study and analyze the phage hydrolysis enzymes in contrast to biochemical-based methods[13], [14].

Phylogenetic analysis or similarity search could find relative conservation of motifs among related species[4], [8], [15], [16], but in the case of phage open reading frame (ORF) which varies greatly, with more than 70% of the base sequence syllables in GenBank are then unable to find similar genes with desired annotation function [4], [5], [9].

With the accumulation of more post-genome data, several models have been developed for discrimination of the functions of phage proteins. Reed et al, proposed a model to predict the three-dimensional structure of t-even phage-type tail fibrin. Their finding were more likely consistent with electron microscopic data[17]. Over computational approach, several models have been developed for phage T7[13], [18]–[21]. Recently, the viral protein encoded by phage was studied by Feng et al using simple Bayes algorithms combined with first-level sequence information[16]. They have recorded a 85.02% of ACC from overall ACC of (79.15%) using the feature selection approach. Recently, Hong- Fi li et al[3] have developed an excellent computational model, by combining multiple feature vectors. They have obtained 85.1% of Accuracy, 88% of Specificity, and 83% of sensitivity for the prediction of Phage enzymes and 94.3% of Acc, 93% of Specificity, and 96% of Sensitivity for the further discrimination of phage

hydrolase enzymes. They have used ANOVA as a feature selection strategy with an SVM as a baseline classifier.

Successively, each predictor brought a significant improvement in different classification metrics but, these models lack optimum classification power to correctly predict phage enzymes and its hydrolase encoded phage. This makes the model unable to get a true generalization power over unseen data. To fill this research gap, we have developed a more robust and intelligent computational predictor, in which optimum features were obtained via a novel multi-threshold values feature selection algorithm. The proposed model is simulated via a 2-dimensional convolutional neural network build in python Keras library. For model evaluation and generalization testing, a 5-fold rigorous cross-validation was used. Empirical results shown that the proposed technique outperformed with existing phage enzymes prediction tools due to (i) new feature fusion scheme via feature selection, and (ii) state-of-the-art deep learning algorithm. We believe that the proposed model will be a very handy tool in the field of biological research, academia, and the applied drug design industry by further providing additional future insight in the field of computational proteomics.

The main contributions presented in this article are as follows:

1) A new Bi-layer computational method for identifying phage enzymes are proposed, the first layer identifies the phage and non-phage enzymes, and the second layer identifies the further types of hydrolase phage enzymes. To the best of our knowledge, the bi-layered fasion approach model was first devised. The proposed model is not only cheap and computationally fast but also reliable than wet laboratory methods.

2) An automatic feature extraction and selection scheme are proposed to obtain the most conductive feature information out of the base feature vector.

3) A Novel Multi-Level Multi-threshold subset feature selection (MLMT-SFS) model is proposed for selecting the best features for establishing a reliable model.

4) A Two Dimensional Convolutional Neural Network has been used as baseline classification algorithms

The organization of the paper is as follows. In Section 1 detail background study and literature have been discussed. Section 2 discussed the benchmark dataset, feature extraction, and selection technique, and evaluation metrics are discussed. Section 3 plot a detailed picture of results and discussion. Finally, Section 4 makes a summary of this paper, conclusion, and directions for future work.

2. Methods

2.1. Benchmark Dataset

Constructing a reliable benchmark dataset could guarantee the reliability of the proposed computational model [22]–[27]. In this work, samples were gained from previous studies of [3], [20] studies [3], [20], which were rigorously screened through the following three steps:

(1) Phage proteins have been annotated by the standard operating procedure for UniProt manual curation (Swiss-Prot);
(2) Protein sequences samples containing illegal characters were deleted;
(3) Sequence identity in the dataset must be less than 30%, which was implemented by CD-HIT (Fu et al., 2012) software. Consequently, the definitive benchmark dataset contains 255 phage proteins, of which 124 proteins belong to phage enzymes (positive samples of set 1), and the remaining 131 are phage non-enzymes (negative samples of set 1). Furthermore, 124 phage enzymes are divided into 69 hydrolases (positive samples of set 2) and 55 non-hydrolases (negative samples of set 2), respectively. The following calculations are all based on these data.

2.2. Proposed Model framework

Diagrammatic representation enhances the readability of the complex relationship of any computational model inter-processes. For ease of understanding the schematic flow of the proposed model is illustrated in Figure 1 below.

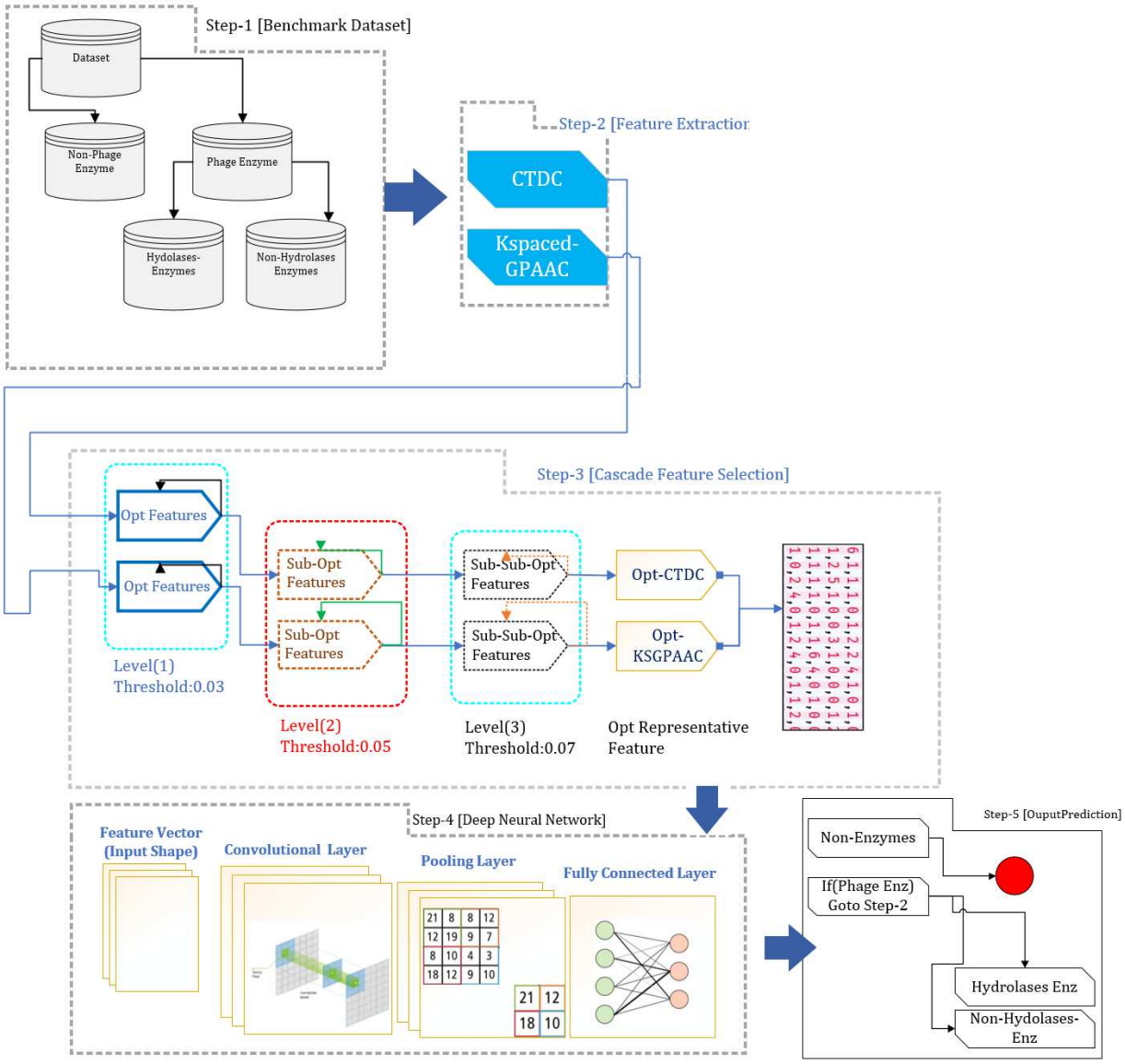

Figure.1. Schematic Workflow Model of the proposed model. Step (1): Training and Testing dataset, Step (2): Feature Extraction, Step (3): MLMT-SFS and Step (4): Classifier Evaluation and Prediction of Phage Enzymes and their sub-types)

2.3. Feature Extraction Technique

2.4. CTD(Composition Transition Distribution)

The global feature holds decisive information and has an effective contribution to the prediction performance of a predictor. Considering this, we use the Composition, Transition, and Distribution (CTD) algorithm [28]. In this approach, the 20 amino acids are split into three groups such as neutral, hydrophobicity, and polar according to the seven physicochemical properties including polarity, solvent accessibility, charge, polarizability, vander waals volume, secondary structure, and hydrophobicity. The complete list is given in Table 1.

Table 1. The distribution of 20 amino acids in three groups

Physicochemical properties	Group I	Group II	Group III
Polarity	DEKNQR	AGHPSTY	CFILMVW
Solvent accessibility	ACDGPST	EILNQV	FHKMRWY
Charge	CFILMVWY	AGPST	DEHKNQR
Polarizability	ADGST	CEILNPQV	FHKMRWY
Vander waals volume	KR	DE	ADFGHILMNPQSTVWY
Secondary structure	AEHKLMQR	CFITVWY	DGNPS
Hydrophobicity	ACFGILVW	DEKNQR	HMPSTY

The Composition (C) defines the global percent composition of each unit and noted as:

$$\left(\frac{G_1}{L}, \frac{G_2}{L}, \frac{G_3}{L}\right) \tag{1}$$

here G_i, $i \in \{1, 2, 3\}$, indicates the amino acids to group i of the protein sequence with length L.
The Transition (T) describes the percent frequency of amino acids in one group followed by amino acids from the other group and calculated by Eq.2

$$\left(\frac{G_{a_1a_2} + G_{a_2a_1}}{L}, \frac{G_{a_1a_3} + G_{a_3a_1}}{L}, \frac{G_{a_2a_3} + G_{a_3a_2}}{L}\right) \tag{2}$$

Here a_i, $i \in \{1, 2, 3\}$, shows the number of one group and $G_{a_ia_j}$ represents the number of dipeptides in form of a_ia_j.
The Distribution (D) calculates the corresponding positions of the first, 25%, 50%, 75% and 100% of the amino acids in a group, which is described as:

$$\left(\frac{G_{11}}{L}, \frac{G_{12}}{L}, \dots, \frac{G_{15}}{L}, \frac{G_{21}}{L}, \frac{G_{22}}{L}, \dots, \frac{G_{25}}{L}, \frac{G_{31}}{L}, \frac{G_{32}}{L}, \dots, \frac{G_{35}}{L}\right) \tag{3}$$

here G_{i1} , G_{i2}, G_{i3}, G_{i4} and G_{i5} represent the chain length of first, 25%, 50%, 75% and 100% of the amino acids in i group respectively are located. According to Table 1, a 147-dimension CTD feature vector against each protein sequence is generated.

2.5. KSAACGP (K-spaced Amino Acid Composition Group Pairs)

The composition of the k-spaced amino acid feature descriptor is a variant of CKSAAP[24], [29] which calculated the cumulative frequency of amino acid by k-separated ($k \in W = \{0,1,2,3,4,5\}$) amino acids residue[30]. In KSAAGP feature descriptor, five group pairs of native 20 amino acid have been taken according to their physicochemical properties e.g., molecular size, hydrophobicity, and charge[31]. Among these five groups, g1: aliphatic group, g2: aromatics group, g3: positive charge group, g4: negative charged group, and g5: uncharged group. These groups further inherit following amino acids i.e. $g_1 \in \{G, A, V, L, M, I\}$, $g_2 \in \{F, Y, W\}$, $g_3 \in \{K, R, H\}$, $g_4 \in \{D, E\}$ and $g_5 \in \{S, T, C, P, N, Q\}$.

2.6. Feature Selection

To remove and make the feature space optimally fit for the prediction of unseen data an excellent feature selection strategy is employed. As it is evident from several research studies [24], [27], [32]–[34], that feature selection plays a prominent role in building a reliable computational model. The optimum selected feature prevents the model from the curse of dimensionality, avoids overfitting, reduces training time, and enhanced model generalizability. A new scheme of feature selection is discussed in preceding section.

2.6.1. Multi-Level Multi Threshold Feature Subset Selection

One challenging task in computational biological problems is the formulation of the biological sequence via a strong mathematical equation because of a machine learning model unable to process raw biological sequence [25]–[27], [35]–[37]. Therefore, these protein sequences are represented in a more compact and mathematical discrete representation. Studies have shown that extracted primitive feature vector often contains redundant, vague, and irreverent features, which not only mislead the prediction of a classifier but also cause the curse of dimensionality, which eventually, leads to overfitting/underfitting. Various studies [38]–[40] indicate that a single representative feature fails to represent significant feature information concealed in a protein sequence. Hence by coping with such situation, the concept of fusion is used, in which hybrid feature is obtained from a multiple selected feature. To reduce the possibility of the curse of dimensionality and model over-fitting [39], [41]–[46], we have implemented a novel Multi-Level Multi-Threshold feature subset selection to select only the most favorable feature for building the model. In MLMT-SFS, the proposed feature selection model operates over a set of three threshold values, i.e. 0.03, 0.05, and 0.07. Iteratively, the model distilled and obtained optimum features by applying these three sets of threshold values through a cascading approach. A score function is used by calculating the weightage value of each feature. Then only those features are selected by having a score value greater than the 0.03 threshold value. Again the obtained feature space is run through threshold value of 0.05 and 0.07 to get the most optimum feature space. The proposed model schematic diagram is given in Figure.1

2.7. Model Architecture

The proposed methodology has been simulated via different classification algorithms i.e., Multilayer Perceptron (MLP)[25], Support Vector Machines (SVM)[27], DT (Decision Tree), RF (Random Forest)[47][48][49] and 2D-CNN. Deep learning attained huge and considerable

attention concerning its implication in the field of computational genomics and proteomics[50]–[53]. We have built 2D-CNN model with a Keras framework (http://www.keras.io)

For a baseline chosen hybrid feature space, the following tuned parameters were used for building the convolutional neural network model. A CNN layer was instantiated with a 2-Dimensional(2D) Convolutional layer of 32 filters, a kernel shape of (3,3), 2D Zero Padding with an input shape of (1,84,84) and activation function 'relu'. A 2D-MaxPooling layer with consistent stride shape of (2,2), dim ordering 'th' and ZeroPadding2D with a shape of (1,1) was adopted on each following 2D Dense layer instantiation. The second 2D-CNN layer with 64 filters, of shape (3,3) kernel and activation function 'relu' and third layer was instantiated with a filter size of 128, kernel shape of (3,3,) and activation function 'relu' was adopted.

Final flatten and output dense layer with a filter size of 128 , nb_classes of 2 and activation squash function 'sigmoid' was used. Binary Cross Entropy was used as a loss function. Adadelta, obtained best success result as optimization parameter. To provent model from overfitting and underfitting , a drop out of 0.3 as regularization technique was used.

2.8. Cross validation

Model evaluation is assessed over rigorous CV (cross-validation) in order to find out the best generalization parameter of the model[26], [54]. Different types of CV like a jackknife, subsampling (K-fold), and independent dataset test are employed. In CV, jackknife always could generate unique results. To assess the efficiency of the novel predictor, we utilized 5-fold jackknife Cross validation methodology.

2.9. Model evaluation Metrics

To measure the quality of a model, two things are kept under consideration, i) Quantitative measure metrics and ii) cross-validation test (further discussed in Section 2.11)[55]. Different statistical model evaluation metrics were exercised to quantify the robustness and authenticity of the model[56], [57]. Accuracy is a well-known statistical metrics used as a classifier correctness measure tool, sensitivity or recall measure true positive rate, specificity measure the true negative rate, MCC measures the model stability and ROC curve measure the overall model performance and authenticity in respect of just making random judging. F-measure is the harmonic mean of sensitivity and specificity operating in a range of 0 and 1. Cohens-kappa statistics, measure the inter-observer agreement, the reliability of the system. Average precision is a singular value metric to evaluate the model prediction.

$$\begin{cases} Sn = 1 - \frac{N_-^+}{N^+} & 0 \le Sn \le 1 \\ Sp = 1 - \frac{N_+^-}{N^-} & 0 \le Sp \le 1 \\ Acc = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le Acc \le 1 \\ MCC = \frac{1 - \left(\frac{N_-^+ + N_+^-}{N^+ + N^-}\right)}{\sqrt{\left(1 + \frac{N_+^- - N_-^+}{N^+}\right)\left(1 + \frac{N_-^+ - N_+^-}{N^-}\right)}} & -1 \le MCC \le 1 \\ FScore = (2/recall^{-1} + precision^{-1}) & 0 \le F1 \le 1 \\ kappa = 1 - \frac{1-p_0}{1-p_e} & 0 \le kappa \le 1 \end{cases} \quad (4)$$

In Eq. 9, Sn refers to sensitivity, Sp to specificity, Acc to accuracy, and MCC to Mathew correlation coefficient. We have incorporated another few metrics, Log-loss (LL), Gini-index (GI), and Normalized Gini-Index (NGI) for evaluating the imbalance issue measure. Log-loss values approaching zero shows an optimum model. Similarly, Gini-index values are the area below line of perfect quality minus the area below the Lorenz curve divided by the area below the perfect quality line. The lower value of Gini-index shows the perfect distribution of model prediction towards each target class. In Eq.4 N^+ represents positive observational samples, N^- is the set of the total number of negative samples investigated. Whereas N_+^- and N_-^+ are termed as a number of negative samples predicted incorrectly as positive, and the number of positive samples incorrectly predicted as negative, respectively.

3. Result and Discussion

3.1. Simulation analysis over primitive feature space

The proposed model has been simulated over individual feature spaces, which are given in in Table .2. Over Layer-1, the proposed model over CTD feature space yielded a 91.80% of Accuracy (ACC), 69.44% of Sensitivity (Sn), 95.00% of Specificity (Sp), 0.681 of f-score, and 0.8643 of ROC curve. Similarly over KSGPAAC feature space the model obtained a figure of 91.93%, 64.25%, 96.00%, 0.6662, 0.6257, 0.6211, 0.7367, 0.911, 0.3017, 0.3585 and 0.822 as ACC, Sn, Sp, F-Score, MCC, Kappa, APR, ROC, LL, Gini Index, and NGI(Normalized Gini) respectively.

Over layer-2 the proposed model obtained 91.32% of Accuracy, 60.65% of Sensitivity and 95.82% of Specificity over CTD feature space and 93.82% of ACC, 73% of Sensitivity and 96.92% of specificity over KSGPAAC feature space respectively. These Layer-2 detail results of discrimination of Phage Hydrolases enzymes are given in Table.3

Table 2. The success rate of different classifiers over CTD+KSGPAAC Feature on Layer-1

Feature	Classifier	ACC	Sn	Sp	F-Score	MCC	Kappa	APR	ROC	LL	GI	NGI
CTDC	SVM	87.01	0.960	99.57	0.0185	0.0258	0.0092	0.1521	0.593	0.8899	0.0899	0.206
	MLP	80.39	18.26	89.46	0.1919	0.0808	0.0807	0.1874	0.6099	0.9901	0.0959	0.2199
	DT	77.57	20.19	85.95	0.1867	0.0577	0.0575	0.1368	0.5257	7.7536	0.0792	0.1815
	2DCNN	91.80	69.44	95.00	0.681	0.6361	0.6348	0647	0.8643	0.6455	0.3178	0.7286
KSGPAAC	SVM	64.21	65.38	64.04	0.3178	0.2006	0.1546	0.2351	0.6627	0.635	0.1419	0.3253
	MLP	80.88	22.11	89.46	0.2277	0.1188	0.1187	0.1945	0.6002	1.2206	0.0875	0.2005

	DT	76.47	19.23	84.83	0.1724	0.0372	0.037	0.133	0.5203	8.1269	0.0481	0.1102
	2DCNN	91.93	64.25	96.00	0.6662	0.6257	0.6211	0.7367	0.911	0.3017	0.3585	0.822

ACC: Accuracy, Sn. Sensitivity, Sp. Specificity, F-Score, MCC. Mathew Correlation coefficient, Kappa. Cohen Kappa Statistics, APR. Average Precision Recall, ROC. Receiver Operating Characteristics curve, LL. Log Loss, GI. Gini Index, NGI. Normalized Gini Index. SVM. Support Vector Machine, RF. Random Forest, MLP. Multi-Layer Perceptron. DT. Decision Tree, LR. Linear Regression, 2DCNN. 2-Dimensional Convolutional neural Network

Table 3. The success rate of different classifiers over CTD+KSGPAAC Feature on Layer-2

Feature	Classifier	ACC	Sn	Sp	F-Score	MCC	Kappa	APR	ROC	LL	GI	NGI
CTDC	SVM	83.08	30.76	90.73	0.3168	0.2206	0.2204	0.2788	0.6979	0.6183	0.1727	0.3958
	MLP	76.77	24.76	84.43	0.2149	0.0824	0.0814	0.1436	0.5454	8.0234	0.1103	0.2531
	DT	76.03	48.57	80.08	0.3423	0.2258	0.2111	0.2256	0.6607	0.7321	0.14	0.3213
	2DCNN	91.32	60.65	95.82	0.6241	0.5795	0.5762	0.6052	0.8686	0.5009	0.3215	0.7372
KSGPAAC	SVM	64.54	67.61	64.09	0.3287	0.2166	0.1668	0.2854	0.6936	0.6289	0.1687	0.3871
	MLP	79.77	32.69	86.65	0.2918	0.1769	0.1755	0.2161	0.6473	1.2213	0.1285	0.2946
	DT	78.48	30.47	85.55	0.2667	0.1444	0.1429	0.1615	0.5787	7.434	0.13	0.2982
	2DCNN	**93.82**	**73.0**	**96.92**	**0.7513**	**0.718**	**0.7165**	**0.8136**	**0.9379**	**0.2729**	**0.3819**	**0.8757**

3.2. Simulation analysis over Fuse Feature space without Feature selection

In a fair comparison, the proposed model was simulated over hybrid feature space without employing a feature selection algorithm. It can be observed from the listings of Table.4, that the proposed model drastically, degrade classification performance due to the presence of noise and outliers in the feature space. Over layer-1 the model obtained an ACC of 87.12%, Sn of 31.87%, Sp of 95.24%, and 0.7980 of ROC value. Successively, over Layer-2 the model yielded an accuracy of 85.60%, Sn of 52.20%, Sp of 90.48%, and 0.8385 of ROC value. The aforementioned detailed simulation results are given in Table.4.

Table 4 Success rate of different classifiers over Fuse Feature without Feature Selection

Layers	Classifier	Acc%	Sn%	Sp%	F-Score	MCC	Kappa	APR	ROC	LL	GI	NGI
Layer-1	SVM	87.01	0.960	99.57	0.0185	0.0258	0.0092	0.1521	0.593	0.8899	0.0899	0.206
	MLP	87.41	2.85	99.86	0.055	0.13	0.046	0.3161	0.733	0.342	0.203	0.466
	DT	78.55	25.96	86.24	0.2358	0.113	0.112	0.1513	0.564	7.368	0.132	0.304
	2D-CNN	**87.12**	**31.78**	**95.24**	**0.3546**	**0.308**	**0.295**	**0.419**	**0.796**	**0.429**	**0.258**	**0.592**
Layer-2	SVM	61.64	76.92	59.41	0.3383	0.2434	0.174	0.2595	0.7202	0.6451	0.1922	0.4405
	MLP	73.16	53.84	75.98	0.3384	0.222	0.1982	0.2453	0.6998	0.5472	0.1743	0.3995
	DT	76.34	28.84	83.28	0.2372	0.1047	0.1024	0.1495	0.5647	8.1277	0.3325	0.7622
	2D-CNN	**85.60**	**52.20**	**90.48**	**0.483**	**0.4017**	**0.4**	**0.4817**	**0.8385**	**0.4453**	**0.2952**	**0.6769**

3.3. Simulation Analysis over MLMT-SFS based Hybrid Feature space

Feature selected via novel Multi Level Multi Threshold Sequential Feature Selection (MLMT-SFS) grab great boost in the classification power of the proposed model. We have observed that, over MLMT-SFS model, the proposed baseline classifier yields an accuracy of 91.6 %,

Sensitivity of 63.39%, Specificity 95.72%, MCC of 0.6049, and ROC value of 0.8772 over Layer-1 respectively. Similarly, the underline model obtained an Accuracy of 96.05%, Sensitivity of 96.22%, Specificity of 95.91%, MCC of 0.9219, and ROC value of 0.9899. These empirical values shows a significant improvment classification power of the proposed model. These detail simulations are given in Table.5. Respective Model train test loss is given in Figure.2. ROC curve is calculated to signify the underline model robustness and authenticity , which is shown in Figure.3 and Figure.4

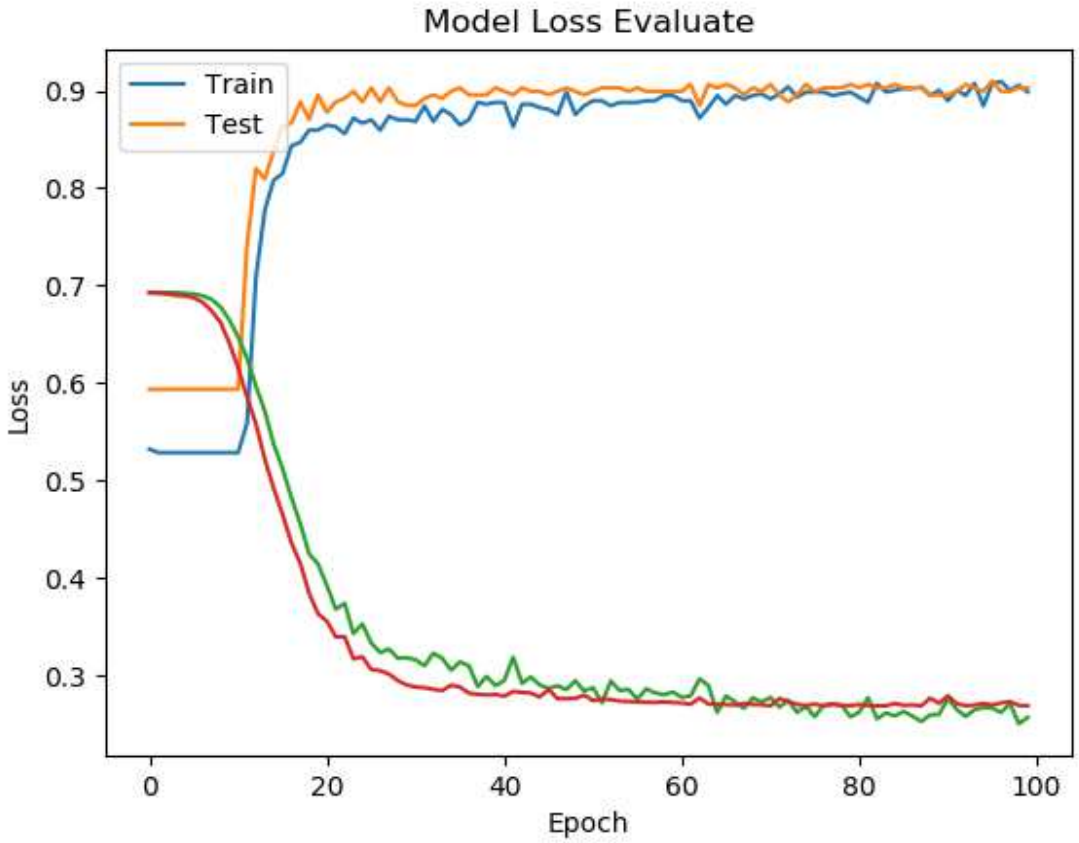

Figure 2. Model Train Test Loss Plotting

Table 5 Success rate of different classifiers over Fuse Feature with Feature Selection

Layers	Classifier	Acc%	Sn%	Sp%	F-Score	MCC	Kappa	APR	ROC	LL	GI	NGI
Layer-1	SVM	88.01	0.970	98.57	0.0195	0.6158	0.5692	0.5821	0.823	0.6099	0.2999	0.916
	MLP	81.49	27.88	89.32	0.2775	0.1714	0.1714	0.2202	0.6637	1.0696	0.1428	0.3273
	DT	78.43	25.96	86.09	0.2348	0.1113	0.1106	0.1518	0.5665	7.3701	-0.0188	-0.0431
	2D-CNN	**91.6**	**63.39**	**95.72**	**0.6481**	**0.6049**	**0.602**	**0.6235**	**0.8772**	**0.6291**	**0.3291**	**0.7543**
Layer-2	SVM	70.293	67.61	70.68	0.3688	0.27	0.2239	0.3381	0.7667	0.5522	0.2325	0.5335
	MLP	80.44	30.46	87.79	0.2857	0.1734	0.1729	0.2104	0.6591	1.4399	0.1386	0.3181

DT	76.284	28.57	83.31	0.2362	0.103	0.1008	0.1516	0.5693	8.0707	-0.0024	-0.0056
2D-CNN	**96.05**	**96.22**	**95.91**	**0.9622**	**0.9219**	**0.9212**	**0.9881**	**0.9899**	**0.1719**	**0.2445**	**0.9779**

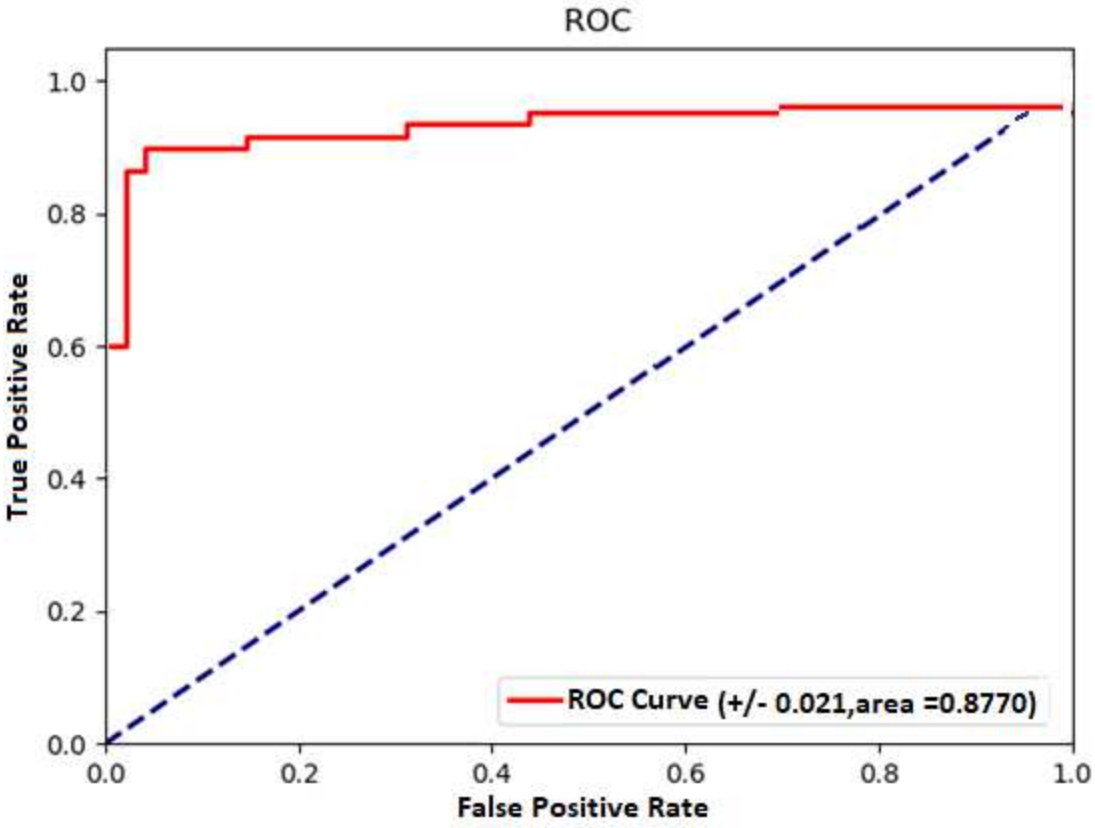

Figure 3. ROC Curve of Layer-1 of Proposed Approach

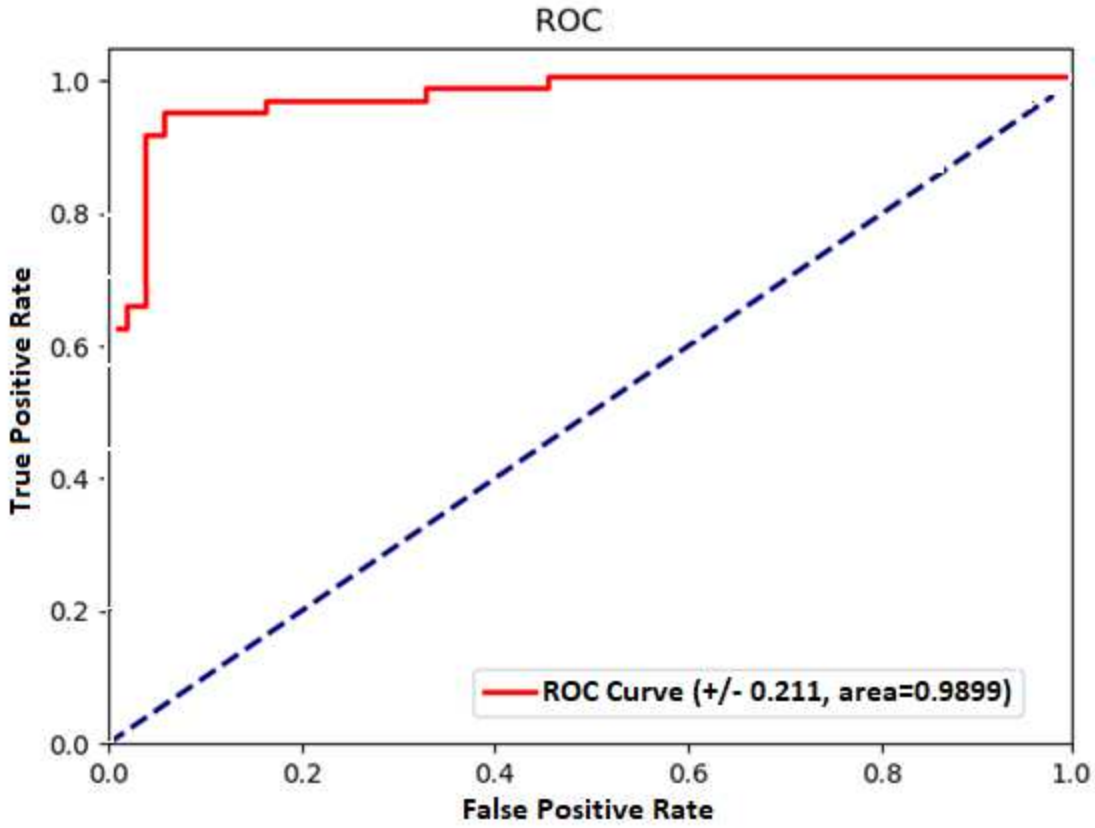

Figure 4. ROC Curve of Layer-2 of Proposed Approach

3.4. Comparative Analysis

The prime objective of building any computational predictive model is its generalization power, which should perform best on the unseen data, provided that the model is also not susceptible to overfitting or random classification. The proposed models' results are significantly encouraging compared to the existing Ding et al[20] and Hong et al [3]. It has been observed that the proposed approach outperformed in all evaluation metrics in contrast to all existing approaches. The detailed results are shown in Table 6.

Table 6 Comparison of predictive performance with exist method

		Ac%	**Sp%**	**Sn%**	**ROC**
Layer-1	*(Ding H. et al., 2016)*	84.30	81.70	87.10	-
	(Hong-Fei. et al., 2020)	85.10	88.00	83.00	-
	Proposed Study	**91.60**	**95.72**	**63.39**	**0.8772**
Layer-2	*(Ding H. et al., 2016)*	93.50	92.80	94.50	-
	(Hong-Fei. et al., 2020)	94.30	93.00	96.00	-
	Proposed Study	**96.05**	**95.91**	**96.22**	**0.9899**

4. Webserver and User Guide

Many literature studies in the field of computational biology and bioinformatics, indicates the importance and development of a user-friendly publicly accessible web server[58][22]–[25], [27], [47], [48], [59], [60]. Further, a web server simulates intuitions and signifies the importance and future direction for both academians and experimental scientists through carrying various kinds of biological (medical) computational analysis and reporting. For the ease of end-user and experimental biologist, publicly accessible web server, from where can the end-user can obtain required results without going through technical and mathematical can be accessed via http://2clphageenzyme.pythonanywhere.com/. Figure.5 shows the index page of the developed webserver.

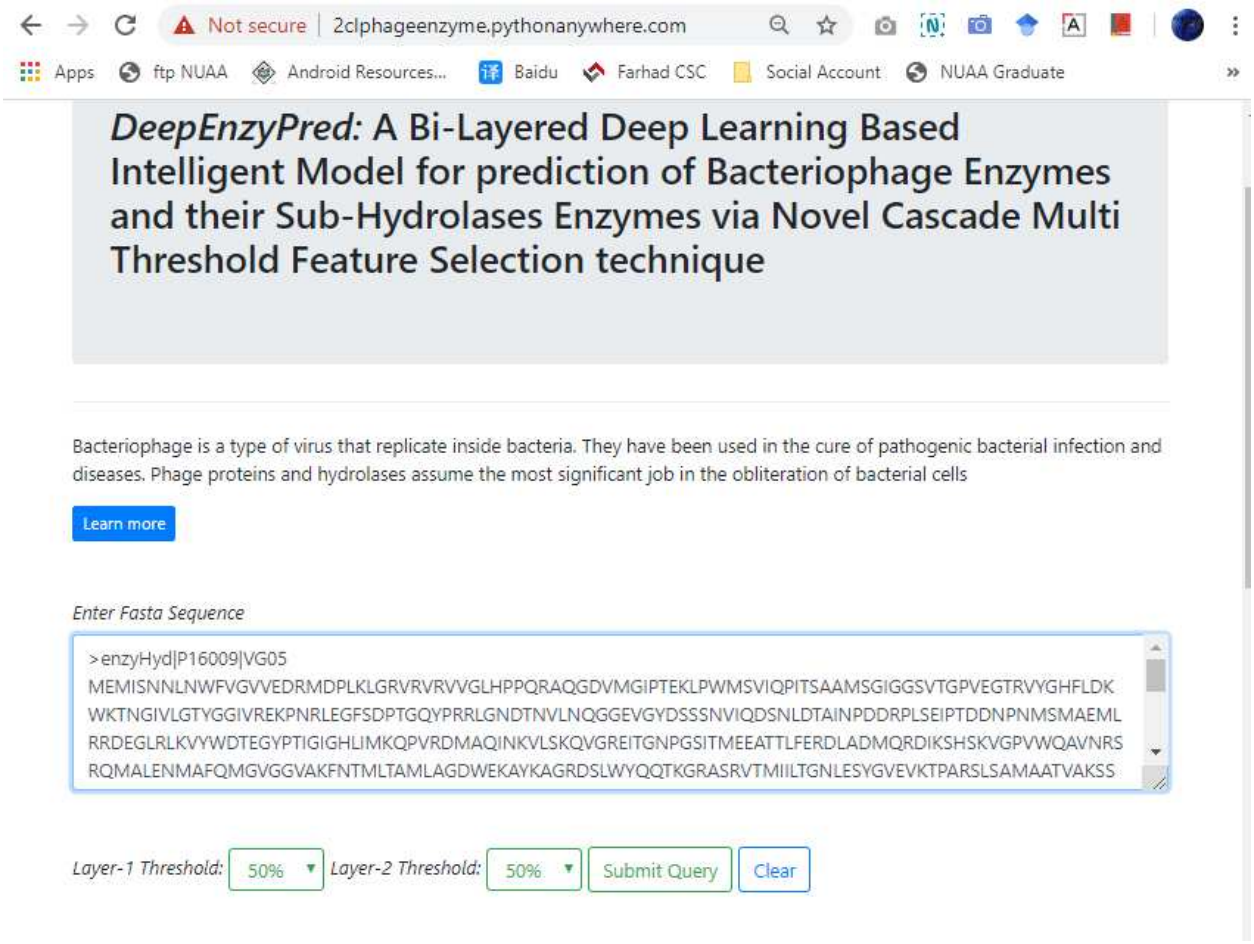

Figure 5. Index page of webserver DeepEnzyPred

5. Conclusion

In this research, we have developed a novel sequence-based automated predictor for phage enzymes and hydrolase enzymes, called DeepEnzyPred. Simulations outcomes with a training dataset and independent validation dataset have revealed the efficacy of the proposed theoretical model. The good performance of DeepEnzyPred is due to several reasons, i.e. anovative feature selection algorithm and careful construction of the prediction model through the tuned 2D-CNN classifier. We believe that the proposed research work will provide a potential insight into a further prediction of phage enzymes characteristics and functionalities. Many literature studies in the field of computational biology and bioinformatics, indicates the importance and significance of developing a user-friendly publicly accessible web server[58]. For the ease of end-user and research academia, we have made effort by establishing a robust and intelligent web server for our proposed method which can be accessed via http://deepenzypred.pythonanywhere.com. For the reproduction of the proposed methodology, all the source code and dataset can be accessed via https://github.com/zaheerkhancs/DeepPhageEnzyme.

6. Abbrivations

ORF:	phage open reading frame
ANOVA:	Analysis of Variance
SVM:	Support Vector Machine
MLP:	Multilayer Perceptron
DT:	Decision Tree
CNN:	Convolutional Neural Network
MLMT:	SFS:Multi Level Multi Threshold Subset Feature Selection
CTD:	Composition Transition Distribution
KSAACGP :	K-spaced Amino Acid Composition Group Pairs
MCC:	Mathew Correlation Coefficient
ROC:	Reciver Operating characteristics
LL:	Log-loss
GI:	Gini-index
NGI:	Normalized Gini-Index
CV:	Cross Validation
ACC:	Accuracy
Sn:	Sensitivity
Sp:	Specificity

7. Decleration

Ethics approval and consent to participate

The study does not involve participation in any human and/or animals

Consent for Publication

All the authors aware of, and there no leading consent from other party, member or person.

Availability of data and material

The data in support to the findings of this masnuscript will be furnished and provided on request.

Competing interests

The author does not declare any competing conflict of interest.

Funding

No funding were recived during this study.

Authors' contributions

‘YW’ wrote and initiate the idea and formulate the overall algorithm, ‘ZUK’ was the corresponding author of the manuscript, made analysis, athuntication and formulate of the whole roadmap of the current study. He also devised and framed the algorithm of the technique used. ‘SA’ proofread the manuscript and performed unit tests of the study. ‘MH’ also proofread the manuscript and removed extensive grammatical and typographical erros. He also devised model validation and statical metrics validations. All authors read and approved the final manuscript.

Acknowledgements

The author is gratefully acknowledging the support of Jiangsu's key laboratory for the donation of Titan XP GPU used for this research study.

Informed Consent

All authors are aware of and approve the manuscript

References

[1] S. McGrath and D. van Sinderen, *Bacteriophage: genetics and molecular biology.* Caister Academic Press, 2007.

[2] K. M. Parmar, N. A. Dafale, H. Tikariha, and H. J. Purohit, "Genomic characterization of key bacteriophages to formulate the potential biocontrol agent to combat enteric pathogenic bacteria," *Arch. Microbiol.*, vol. 200, no. 4, pp. 611–622, 2018.

[3] H.-F. Li, X.-F. Wang, and H. Tang, "Predicting Bacteriophage Enzymes and Hydrolases by Using Combined Features," *Front. Bioeng. Biotechnol.*, vol. 8, p. 183, 2020.

[4] E. C. Keen, "Phage therapy: concept to cure," *Front. Microbiol.*, vol. 3, p. 238, 2012.

[5] C. O. Wilke, "Bringing molecules back into molecular evolution," *PLoS Comput. Biol.*, vol. 8, no. 6, 2012.

[6] T. Parfitt, "Georgia: an unlikely stronghold for bacteriophage therapy," *Lancet*, vol. 365, no. 9478, pp. 2166–2167, 2005.

[7] K. Thiel, "Old dogma, new tricks—21st century phage therapy," *Nat. Biotechnol.*, vol. 22, no. 1, pp. 31–36, 2004.

[8] A. Pirisi, "Phage therapy—advantages over antibiotics?," *Lancet*, vol. 356, no. 9239, p. 1418, 2000.

[9] Z. Atamer, M. Samtlebe, H. Neve, K. J. Heller, and J. Hinrichs, "elimination of bacteriophages in whey and whey products," *Front. Microbiol.*, vol. 4, p. 191, 2013.

[10] K. Kimura and Y. Itoh, "Characterization of poly-γ-glutamate hydrolase encoded by a bacteriophage genome: possible role in phage infection of Bacillus subtilis encapsulated with poly-γ-glutamate," *Appl. Environ. Microbiol.*, vol. 69, no. 5, pp. 2491–2497, 2003.

[11] L. Rodríguez-Rubio, N. Quiles-Puchalt, B. Martínez, A. Rodríguez, J. R. Penadés, and P. García, "The peptidoglycan hydrolase of Staphylococcus aureus bacteriophage ϕ11 plays a structural role in the viral particle," *Appl. Environ. Microbiol.*, vol. 79, no. 19, pp. 6187–6190, 2013.

[12] A. O. Kovalenko *et al.*, "Investigation of the calcium-induced activation of the bacteriophage T5 peptidoglycan hydrolase promoting host cell lysis," *Metallomics*, vol. 11, no. 4, pp. 799–809, 2019.

[13] D. Liu, G. Li, and Y. Zuo, "Function determinants of TET proteins: the arrangements of sequence motifs with specific codes," *Brief. Bioinform.*, vol. 20, no. 5, pp. 1826–1835, 2019.

[14] H. Lin and Q.-Z. Li, "Eukaryotic and prokaryotic promoter prediction using hybrid approach," *Theory Biosci.*, vol. 130, no. 2, pp. 91–100, 2011.

[15] H. Ding, P.-M. Feng, W. Chen, and H. Lin, "Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis.," *Mol. Biosyst.*, vol. 10, no. 8, pp. 2229–2235, Aug. 2014, doi: 10.1039/c4mb00316k.

[16] P.-M. Feng, H. Ding, W. Chen, and H. Lin, "Naive Bayes classifier with feature selection to identify phage virion proteins," *Comput. Math. Methods Med.*, vol. 2013, 2013.

[17] I. Riede, H. Schwarz, and F. Jähnig, "Predicted structure of tail- fiber proteins of T- even type phages," *FEBS Lett.*, vol. 215, no. 1, pp. 145–150, 1987.

[18] S. W. White, "Prediction of DNA-binding regulatory proteins in bacteriophage T7," *Protein Eng. Des. Sel.*, vol. 1, no. 5, pp. 373–376, 1987.

[19] L. Song, D. Li, X. Zeng, Y. Wu, L. Guo, and Q. Zou, "nDNA-Prot: identification of DNA-binding proteins based on unbalanced classification.," *BMC Bioinformatics*, vol. 15, p. 298, Sep. 2014, doi: 10.1186/1471-2105-15-298.

[20] H. Ding *et al.*, "PHYPred: a tool for identifying bacteriophage enzymes and hydrolases," *Virol. Sin.*, vol. 31, no. 4, pp. 350–352, 2016.

[21] K. Qu, L. Wei, and Q. Zou, "A Review of DNA-binding Proteins Prediction Methods," *Curr. Bioinform.*, vol. 14, no. 3, pp. 246–254, 2019.

[22] I. A. Khan, D. Pi, Z. U. Khan, Y. Hussain, and A. Nawaz, "HML-IDS: A Hybrid-Multilevel Anomaly Prediction Approach for Intrusion Detection in SCADA Systems," *IEEE Access*, vol. 7, pp. 89507–89521, 2019.

[23] D. Pi, P. Yue, B. Li, Z. U. Khan, Y. Hussain, and A. Nawaz, "An efficient behaviour specification and bidirectional Gated Recurrent Units based intrusion detection method for industrial control systems," *Electron. Lett.*, 2019.

[24] Z. U. Khan, F. Ali, I. Ahmad, M. Hayat, and D. Pi, "iPredCNC: Computational prediction model for cancerlectins and non-cancerlectins using novel cascade features subset selection," *Chemom. Intell. Lab. Syst.*, vol. 195, p. 103876, 2019, doi: https://doi.org/10.1016/j.chemolab.2019.103876.

[25] Z. U. Khan, M. Hayat, and M. A. Khan, "Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model," *J. Theor. Biol.*, vol. 365, pp. 197–203, 2015, doi: https://doi.org/10.1016/j.jtbi.2014.10.014.

[26] Z. U. Khan, F. Ali, I. A. Khan, Y. Hussain, and D. Pi, "iRSpot-SPI: Deep learning-based recombination spots prediction by incorporating secondary sequence information coupled with physio-chemical properties via Chou's 5-step rule and pseudo components," *Chemom. Intell. Lab. Syst.*, 2019.

[27] F. Ali *et al.*, "DBPPred-PDSD: Machine learning approach for prediction of DNA-binding proteins using Discrete Wavelet Transform and optimized integrated features space," *Chemom. Intell. Lab. Syst.*, vol. 182, pp. 21–30, 2018, doi: https://doi.org/10.1016/j.chemolab.2018.08.013.

[28] K.-C. Chou and Y.-D. Cai, "Prediction of Membrane Protein Types by Incorporating Amphipathic Effects," *J. Chem. Inf. Model.*, vol. 45, no. 2, pp. 407–413, Mar. 2005, doi: 10.1021/ci049686v.

[29] Z. Chen *et al.*, "iLearn: an integrated platform and meta-learner for feature engineering,

machine-learning analysis and modeling of DNA, RNA and protein sequence data," *Brief. Bioinform.*, Apr. 2019, doi: 10.1093/bib/bbz041.

[30] X. Zhao, W. Zhang, X. Xu, Z. Ma, and M. Yin, "Prediction of protein phosphorylation sites by using the composition of k-spaced amino acid pairs," *PLoS One*, vol. 7, no. 10, pp. e46302–e46302, 2012, doi: 10.1371/journal.pone.0046302.

[31] T.-Y. Lee, Z.-Q. Lin, S.-J. Hsieh, N. A. Bretaña, and C.-T. Lu, "Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences," *Bioinformatics*, vol. 27, no. 13, pp. 1780–1787, May 2011, doi: 10.1093/bioinformatics/btr291.

[32] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, no. Oct, pp. 1205–1224, 2004.

[33] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, 2018.

[34] Y. Sun, "Iterative RELIEF for feature weighting: algorithms, theories, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1035–1051, 2007.

[35] F. Ali and M. Hayat, "Classification of membrane protein types using Voting Feature Interval in combination with Chou′s Pseudo Amino Acid Composition," *J. Theor. Biol.*, vol. 384, pp. 78–83, 2015, doi: https://doi.org/10.1016/j.jtbi.2015.07.034.

[36] M. Hayat and A. Khan, "Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition," *J. Theor. Biol.*, vol. 271, no. 1, pp. 10–17, 2011.

[37] K.-C. Chou and H.-B. Shen, "Recent progress in protein subcellular location prediction.," *Anal. Biochem.*, vol. 370, no. 1, pp. 1–16, Nov. 2007, doi: 10.1016/j.ab.2007.07.006.

[38] I. A. Gheyas and L. S. Smith, "Feature subset selection in large dimensionality domains," *Pattern Recognit.*, vol. 43, no. 1, pp. 5–13, 2010.

[39] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1–2, pp. 273–324, 1997.

[40] A. Chokka and K. S. Rani, "AdaBoost with Feature Selection Using IoT to Bring the Paths for Somatic Mutations Evaluation in Cancer," in *Internet of Things and Personalized Healthcare Systems*, Springer, 2019, pp. 51–63.

[41] S. Maldonado and R. Weber, "A wrapper method for feature selection using Support Vector Machines," *Inf. Sci. (Ny).*, vol. 179, no. 13, pp. 2208–2217, 2009, doi: https://doi.org/10.1016/j.ins.2009.02.014.

[42] S. Das, "Filters, wrappers and a boosting-based hybrid for feature selection," in *Icml*, 2001, vol. 1, pp. 74–81.

[43] H.-H. Hsu, C.-W. Hsieh, and M.-D. Lu, "Hybrid feature selection by combining filters and wrappers," *Expert Syst. Appl.*, vol. 38, no. 7, pp. 8144–8150, 2011.

[44] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr.*

Eng., vol. 40, no. 1, pp. 16–28, 2014, doi: https://doi.org/10.1016/j.compeleceng.2013.11.024.

[45] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005, doi: 10.1109/TPAMI.2005.159.

[46] R. Yang, C. Zhang, L. Zhang, and R. Gao, "A Two-Step Feature Selection Method to Predict Cancerlectins by Multiview Features and Synthetic Minority Oversampling Technique," *Biomed Res. Int.*, 2018, doi: 10.1155/2018/9364182.

[47] Z. U. Khan and M. Hayat, "Hourly based climate prediction using data mining techniques by comprising entity demean algorithm," *Middle-East J. Sci. Res*, vol. 21, no. 8, pp. 1295–1300, 2014.

[48] Z. U. Khan, M. Sohail, M. N. Hayat, and H. Khan, "Face Recognition using Principle Component Analysis Based feature selection Feature Vector."

[49] M. R. Jani, M. T. Khan Mozlish, S. Ahmed, N. S. Tahniat, D. M. Farid, and S. Shatabda, "iRecSpot-EF: Effective sequence based features for recombination hotspot prediction.," *Comput. Biol. Med.*, vol. 103, pp. 17–23, Dec. 2018, doi: 10.1016/j.compbiomed.2018.10.005.

[50] D. Cohn, O. Zuk, and T. Kaplan, "Enhancer Identification using Transfer and Adversarial Deep Learning of DNA Sequences," *bioRxiv*, p. 264200, 2018.

[51] A. Telenti, C. Lippert, P.-C. Chang, and M. DePristo, "Deep learning of genomic variation and regulatory network data," *Hum. Mol. Genet.*, vol. 27, no. R1, pp. R63–R71, 2018.

[52] Y. Li, C. Huang, L. Ding, Z. Li, Y. Pan, and X. Gao, "Deep learning in bioinformatics: Introduction, application, and perspective in the big data era," *Methods*, 2019, doi: https://doi.org/10.1016/j.ymeth.2019.04.008.

[53] M. Tahir, H. Tayara, and K. T. Chong, "iDNA6mA (5-step rule): Identification of DNA N6-methyladenine sites in the rice genome by intelligent computational model via Chou's 5-step rule," *Chemom. Intell. Lab. Syst.*, vol. 189, pp. 96–101, 2019, doi: https://doi.org/10.1016/j.chemolab.2019.04.007.

[54] B. Liu, S. Wang, R. Long, and K. C. Chou, "IRSpot-EL: Identify recombination spots with an ensemble learning approach," *Bioinformatics*, 2017, doi: 10.1093/bioinformatics/btw539.

[55] B. Liu, L. Fang, R. Long, X. Lan, and K.-C. Chou, "iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition," *Bioinformatics*, vol. 32, no. 3, pp. 362–369, 2015.

[56] K.C. Chou, "Some remarks on predicting multi-label attributes in molecular biosystems.," *Mol. Biosyst.*, vol. 9, pp. 1092–1100, 2013.

[57] K. C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of Theoretical Biology*. 2011, doi: 10.1016/j.jtbi.2010.12.024.

[58] K.-C. Chou and H.-B. Shen, “Recent advances in developing web-servers for predicting protein attributes,” *Nat. Sci.*, vol. 1, no. 02, p. 63, 2009.

[59] Z. U. Khan, F. Ali, I. A. Khan, Y. Hussain, and D. Pi, “iRSpot-SPI: Deep learning-based recombination spots prediction by incorporating secondary sequence information coupled with physio-chemical properties via Chou’s 5-step rule and pseudo components,” *Chemom. Intell. Lab. Syst.*, vol. 189, 2019, doi: 10.1016/j.chemolab.2019.05.003.

[60] I. A. Khan *et al.*, “Efficient behaviour specification and bidirectional gated recurrent units-based intrusion detection method for industrial control systems,” *Electron. Lett.*, vol. 56, no. 1, pp. 27–30, Jan. 2020, doi: 10.1049/el.2019.3008.

Figures

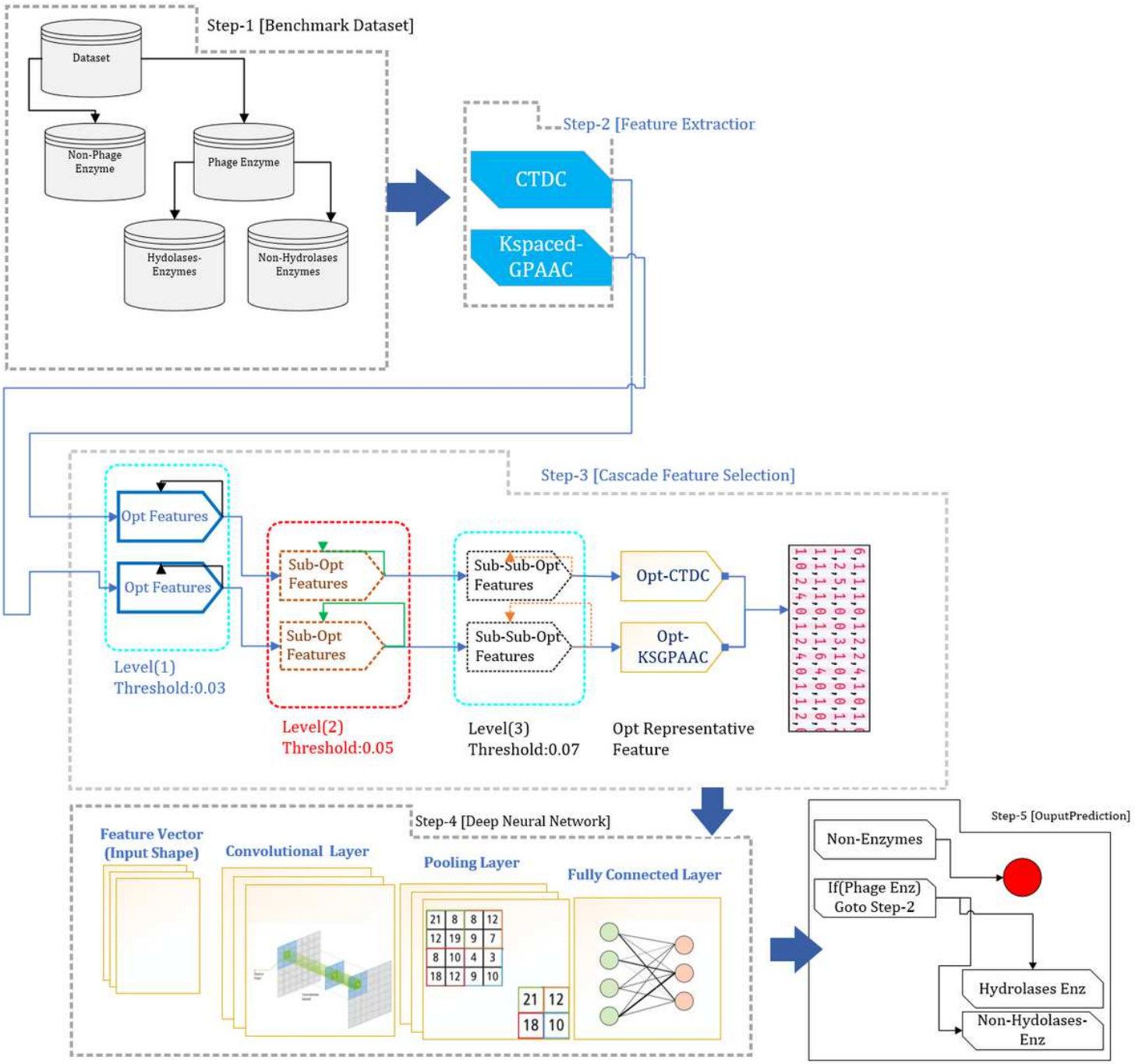

Figure 1

Schematic Workflow Model of the proposed model. Step (1): Training and Testing dataset, Step (2): Feature Extraction, Step (3): MLMT-SFS and Step (4): Classifier Evaluation and Prediction of Phage Enzymes and their sub-types)

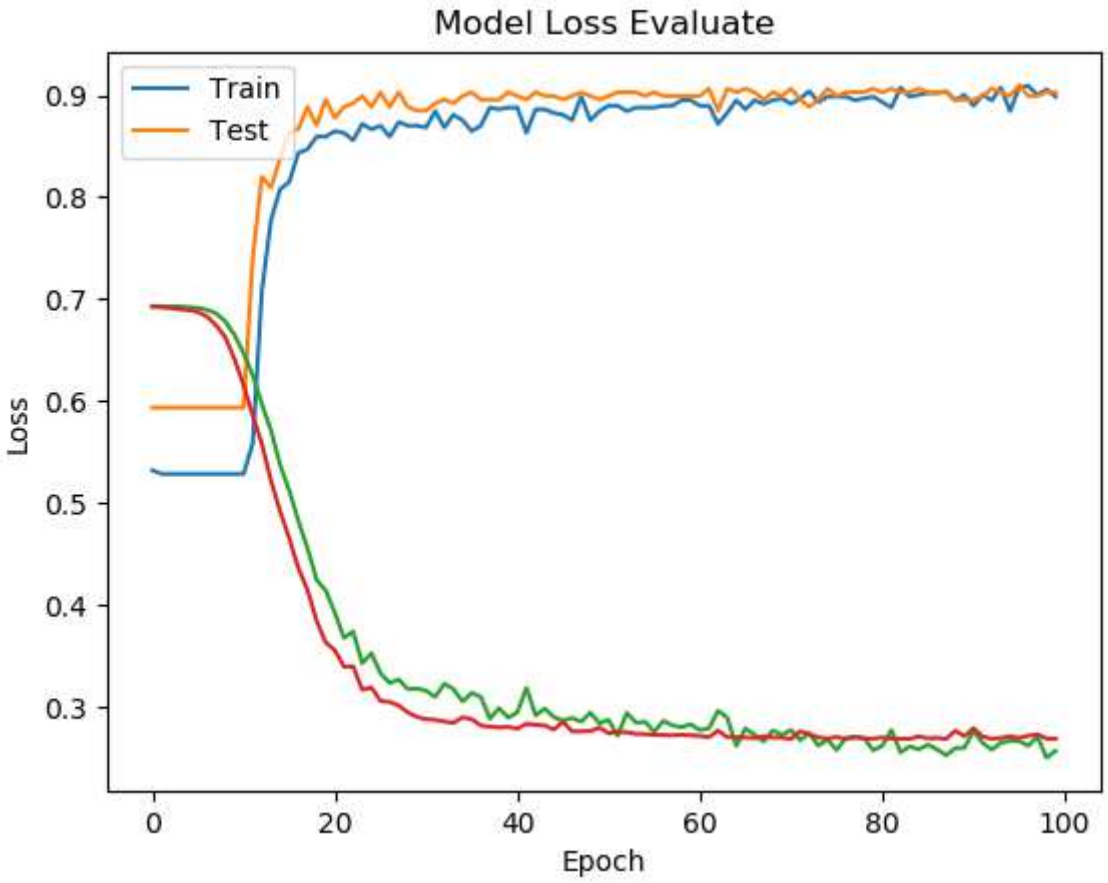

Figure 2

Model Train Test Loss Plotting

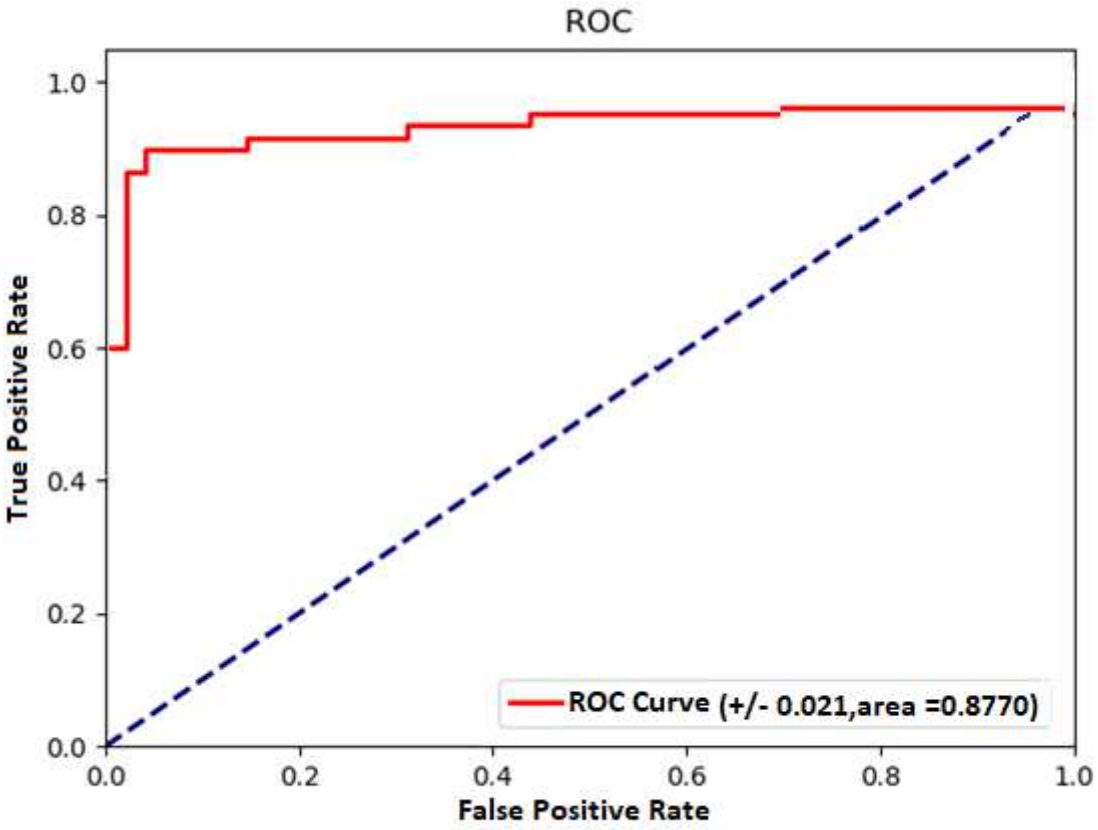

Figure 3

ROC Curve of Layer-1 of Proposed Approach

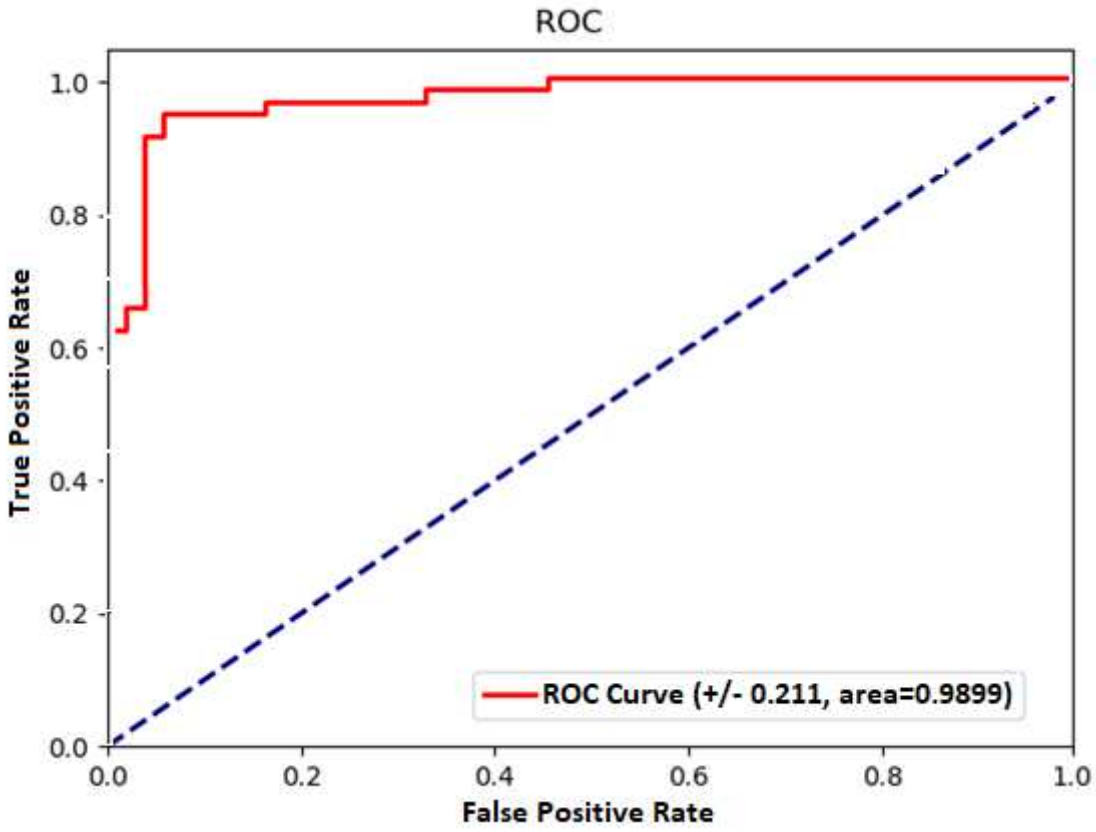

Figure 4

ROC Curve of Layer-2 of Proposed Approach

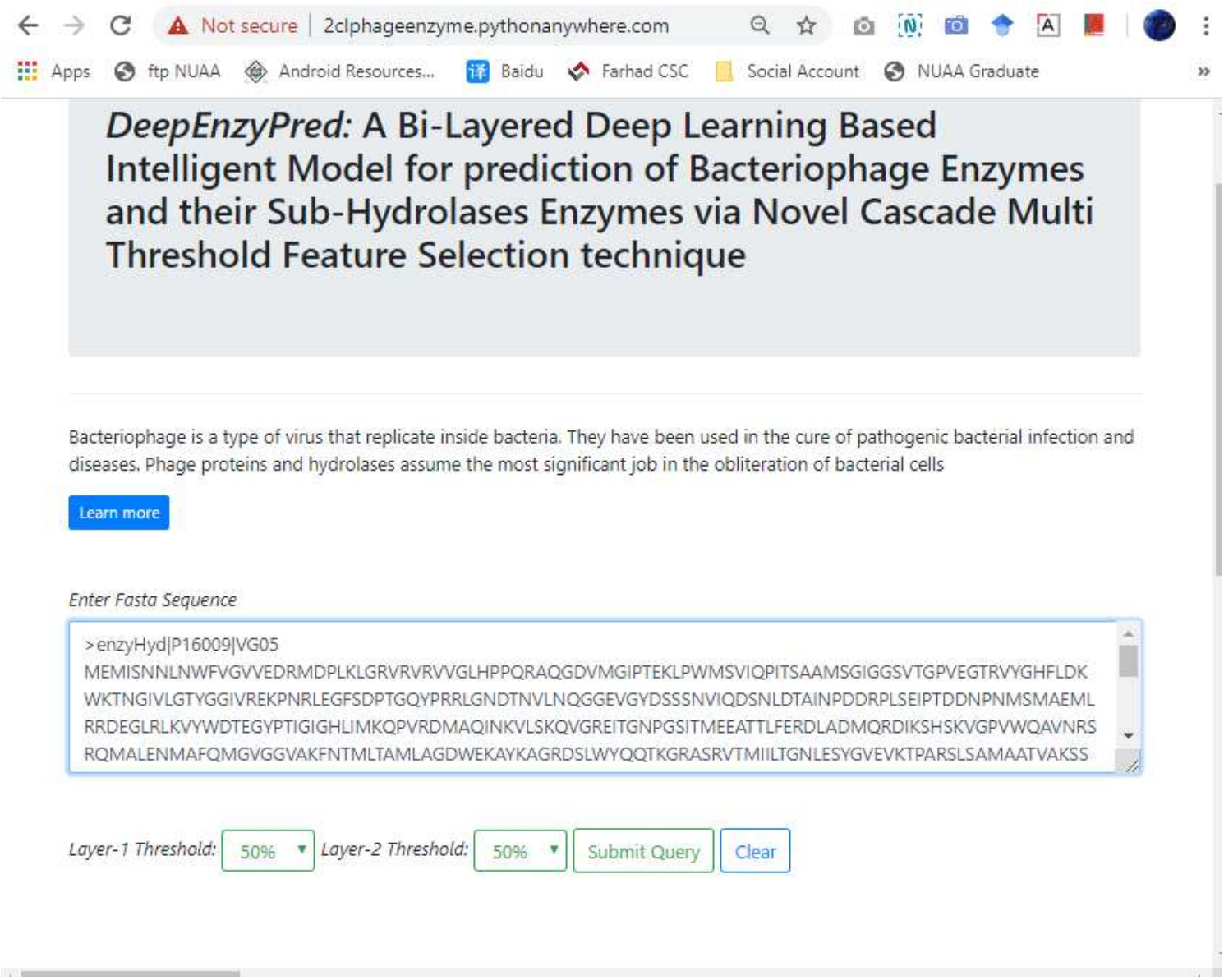

Figure 5

Index page of webserver DeepEnzyPred