
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

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

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@sdaau.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 consists 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 innovative bi-layered model named 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 respectively.

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.
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

<table>
<thead>
<tr>
<th>Physicochemical properties</th>
<th>Group I</th>
<th>Group II</th>
<th>Group III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polarity</td>
<td>DEKNQR</td>
<td>AGHPSTY</td>
<td>CFILMVW</td>
</tr>
<tr>
<td>Solvent accessibility</td>
<td>ACDGPST</td>
<td>EILNQV</td>
<td>FHKMRWY</td>
</tr>
<tr>
<td>Charge</td>
<td>CFILMVYW</td>
<td>AGPST</td>
<td>DEHKNQR</td>
</tr>
<tr>
<td>Polarizability</td>
<td>ADGST</td>
<td>CEILNPQV</td>
<td>FHKMRWY</td>
</tr>
<tr>
<td>Vander waals volume</td>
<td>KR</td>
<td>DE</td>
<td>ADFGHILMNPQSTVWY</td>
</tr>
<tr>
<td>Secondary structure</td>
<td>AEHKLMQR</td>
<td>CFITVWY</td>
<td>DGNPS</td>
</tr>
<tr>
<td>Hydrophobicity</td>
<td>ACFGILVW</td>
<td>DEKNQR</td>
<td>HMPSTY</td>
</tr>
</tbody>
</table>

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

\[
\left( \frac{G_1}{L}, \frac{G_2}{L}, \frac{G_3}{L} \right)
\]

(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

\[
\left( \frac{G_{a_1a_2} + G_{a_2a_1}}{L}, \frac{G_{a_1a_3} + G_{a_3a_1}}{L}, \frac{G_{a_2a_3} + G_{a_3a_2}}{L} \right)
\]

(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:

\[
\left( \frac{G_{11}}{L}, \frac{G_{12}}{L}, ..., \frac{G_{15}}{L}, \frac{G_{21}}{L}, \frac{G_{22}}{L}, ..., \frac{G_{25}}{L}, \frac{G_{31}}{L}, \frac{G_{32}}{L}, ..., \frac{G_{35}}{L} \right)
\]

(3)

here \( G_{11}, G_{12}, G_{13}, G_{14} \) and \( G_{15} \) 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 prevent 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. Coheens-kappa statistics, measure the inter-observer agreement, the reliability of the system. Average precision is a singular value metric to evaluate the model prediction.
\[
\begin{align*}
Sn &= 1 - \frac{N_+}{N} \\
Sp &= 1 - \frac{N_-}{N} \\
Acc &= 1 - \frac{N_+ + N_-}{N^+ + N^-} \\
MCC &= \frac{1-(N_+ + N_-)}{\sqrt{(1+N_+ - N_-)(1+N_- - N_+)}} \\
Fscore &= (2/recall^{-1} + precision^{-1}) \\
kappa &= 1 - \frac{1-p_0}{1-p_e}
\end{align*}
\]

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

<table>
<thead>
<tr>
<th>Feature</th>
<th>Classifier</th>
<th>ACC</th>
<th>Sn</th>
<th>Sp</th>
<th>F-Score</th>
<th>MCC</th>
<th>Kappa</th>
<th>APR</th>
<th>ROC</th>
<th>LL</th>
<th>GI</th>
<th>NGI</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTDC</td>
<td>SVM</td>
<td>83.08</td>
<td>30.76</td>
<td>90.73</td>
<td>0.3168</td>
<td>0.2206</td>
<td>0.2204</td>
<td>0.2788</td>
<td>0.6979</td>
<td>0.6183</td>
<td>0.1727</td>
<td>0.3958</td>
</tr>
<tr>
<td></td>
<td>MLP</td>
<td>76.77</td>
<td>24.76</td>
<td>84.43</td>
<td>0.2149</td>
<td>0.0824</td>
<td>0.0814</td>
<td>0.1436</td>
<td>0.5454</td>
<td>0.8024</td>
<td>0.1103</td>
<td>0.2531</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>76.03</td>
<td>48.57</td>
<td>80.08</td>
<td>0.3423</td>
<td>0.2258</td>
<td>0.2111</td>
<td>0.2256</td>
<td>0.6607</td>
<td>0.7321</td>
<td>0.14</td>
<td>0.3213</td>
</tr>
<tr>
<td></td>
<td>2DCNN</td>
<td>91.32</td>
<td>60.65</td>
<td>95.82</td>
<td>0.6241</td>
<td>0.5795</td>
<td>0.5762</td>
<td>0.6052</td>
<td>0.8686</td>
<td>0.5009</td>
<td>0.3215</td>
<td>0.7372</td>
</tr>
<tr>
<td>KSGPAAC</td>
<td>SVM</td>
<td>64.54</td>
<td>67.61</td>
<td>64.09</td>
<td>0.3287</td>
<td>0.2166</td>
<td>0.1668</td>
<td>0.2854</td>
<td>0.6936</td>
<td>0.6289</td>
<td>0.1687</td>
<td>0.3871</td>
</tr>
<tr>
<td></td>
<td>MLP</td>
<td>79.77</td>
<td>32.69</td>
<td>86.65</td>
<td>0.2918</td>
<td>0.1769</td>
<td>0.1755</td>
<td>0.2161</td>
<td>0.6473</td>
<td>1.2213</td>
<td>0.1285</td>
<td>0.2946</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>78.48</td>
<td>30.47</td>
<td>85.55</td>
<td>0.2667</td>
<td>0.1444</td>
<td>0.1429</td>
<td>0.1615</td>
<td>0.5787</td>
<td>7.434</td>
<td>0.13</td>
<td>0.2982</td>
</tr>
<tr>
<td></td>
<td>2DCNN</td>
<td>93.82</td>
<td>73.0</td>
<td>96.92</td>
<td>0.7513</td>
<td>0.718</td>
<td>0.7165</td>
<td>0.8136</td>
<td>0.9379</td>
<td>0.2729</td>
<td>0.3819</td>
<td>0.8757</td>
</tr>
</tbody>
</table>

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

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

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 improvement 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.

![Model Loss Evaluate](image)

**Figure** 2. Model Train Test Loss Plotting

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

**Table 5** Success rate of different classifiers over Fuse Feature with Feature Selection
<table>
<thead>
<tr>
<th>Method</th>
<th>AUC</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT</td>
<td>76.284</td>
<td>28.57</td>
<td>83.31</td>
<td>0.2362</td>
<td>0.103</td>
<td>0.1008</td>
<td>0.5693</td>
<td>8.0707</td>
<td>-0.0024</td>
</tr>
<tr>
<td>2D-CNN</td>
<td>96.05</td>
<td>96.22</td>
<td>95.91</td>
<td>0.9622</td>
<td>0.9219</td>
<td>0.9212</td>
<td>0.9881</td>
<td>0.9899</td>
<td>0.1719</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Layer</th>
<th>Ac%</th>
<th>Sp%</th>
<th>Sn%</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Ding H. et al., 2016)</td>
<td>84.30</td>
<td>81.70</td>
<td>87.10</td>
</tr>
<tr>
<td></td>
<td>(Hong-Fei. et al., 2020)</td>
<td>85.10</td>
<td>88.00</td>
<td>83.00</td>
</tr>
<tr>
<td>Proposed Study</td>
<td><strong>91.60</strong></td>
<td><strong>95.72</strong></td>
<td><strong>63.39</strong></td>
<td><strong>0.8772</strong></td>
</tr>
<tr>
<td></td>
<td>(Ding H. et al., 2016)</td>
<td>93.50</td>
<td>92.80</td>
<td>94.50</td>
</tr>
<tr>
<td>Layer-2</td>
<td>(Hong-Fei. et al., 2020)</td>
<td>94.30</td>
<td>93.00</td>
<td>96.00</td>
</tr>
<tr>
<td>Proposed Study</td>
<td><strong>96.05</strong></td>
<td><strong>95.91</strong></td>
<td><strong>96.22</strong></td>
<td><strong>0.9899</strong></td>
</tr>
</tbody>
</table>
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](image)

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. Abbreviations

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: Receiver Operating characteristics
LL: Log-loss
GI: Gini-index
NGI: Normalized Gini-Index
CV: Cross Validation
ACC: Accuracy
Sn: Sensitivity
Sp: Specificity

7. Declaration

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 manuscript will be furnished and provided on request.

**Competing interests**

The author does not declare any competing conflict of interest.

**Funding**

No funding were received 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, authentication 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 errors. 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


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


Figure 1

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
DeepEnzyPred: A Bi-Layered Deep Learning Based Intelligent Model for prediction of Bacteriophage Enzymes and their Sub-Hydrolases Enzymes via Novel Cascade Multi Threshold Feature Selection technique

Bacteriophage is a type of virus that replicate inside bacteria. They have been used in the cure of pathogenic bacterial infection and diseases. Phage proteins and hydrolases assume the most significant job in the obliteration of bacterial cells.