HIME: Mining and Ensembling Heterogeneous Information for Protein-Protein Interactions Prediction

Huaming Chen*, Yaochu Jin†, Lei Wang*, Chi-Hung Chi‡, Jun Shen*
*School of Computing and Information Technology, University of Wollongong, Wollongong, NSW, Australia
†Department of Computer Science, University of Surrey, United Kingdom
‡Data61, CSIRO, Australia
Email: *hc007@uowmail.edu.au, yaochu.jin@surrey.ac.uk, leiw@uow.edu.au, chihung.chi@data61.csiro.au, jshen@uow.edu.au*

***Abstract*—Research on protein-protein interactions (PPIs) data paves the way towards understanding the mechanisms of infectious diseases, however improving the prediction performance of PPIs of inter-species remains a challenge. Since one single type of sequence data such as amino acid composition may be deficient for high-quality prediction of protein interactions, we have investigated a broader range of heterogeneous information of sequences data. This paper proposes a novel framework for PPIs prediction based on Heterogeneous Information Mining and Ensembling (HIME) process to effectively learn from the interaction data. In particular, the proposed approach introduces an ensemble process together with substantial features that generate better performance of PPIs prediction task. The performance of the proposed framework is validated on real protein interaction datasets. The extensive experiments show that HIME achieves higher performance over all existing methods reported in literature so far.**

***Index Terms*—biological data, heterogeneous information, protein interaction, neural networks**

I. Introduction

Analyzing and understanding protein-protein interactions (PPIs) is of great importance and value to the study of infectious diseases, especially for inter-species interactions, such as the interactions between human and pathogens [1], [2], which is also termed as human-pathogen protein-protein interactions. It has been one of the hot topics towards the mechanism study of diseases. Since infectious diseases are still the dominant causes of death, the research of infectious diseases has solicited data from different perspectives to examine the biomedical hypothesis and propose potential therapeutics. Vast research has been conducted with a long period of biological development and examination.

As a result of decades of efforts of wet lab-based experiments in biology, the production of biological data, e.g. protein interactions, has exploded. Although there is still substantial need for further experiments, the accumulated data has benefited the research on disease mechanisms to a limited extent. One of the earliest studies was on the symptom of anthrax, which was identified as being primarily caused by the interactions between human and Bacillus anthracis. Bacillus anthracis is a type of bacterium pathogen, where people want to fully understand mechanisms with the protein interactions map between Bacillus anthracis and Homo sapiens (the host).

However, the experiment results to investigate protein-protein interactions are still very limited. There has been an incomplete picture of the protein-protein interactions relationships, where the identifications of the interactions demand a huge amount of time and resources for wet-lab experiments. Meanwhile, the nature of interaction data between different species results in a huge amount of latent interactions to be further examined and verified as positive or negative interactions by biologists. The identification of protein-protein interactions is traditionally conducted by *in vitro* and *in vivo* methods, which includes affinity purification, yeast two-hybrid assay, affinity purification-mass spectrometry (AP-MS), nu-clear magnetic resonance (NMR) and mass spectrometry methods. The processes of these methods are deemed cost-sensitive task for both time and resources.

To effectively generate high-fidelity PPIs prior to biology experiments, there has been numerous studies introducing computational methods to facilitate the process. The identified interactions data is playing an important role in the studies providing the interactions relationship. One major category is to build machine learning-based model with different protein data, such as protein sequence data [3], gene ontology data [4], and protein structure data [5], for the prediction of protein interactions. Among these, sequence information is considered as the main protein information because of its substantial accumulation in a large scale. Specifically, the proteins have been determined uniquely by the sequence information as for their physical and biochemical characteristics. By analyzing the protein sequence information hosted by the Universal Protein Resource (UniProt), the past studies had indicated that combining machine learning-based models with protein sequence data mining would benefit the prediction and analysis of protein interactions task [1], [6], [7].

More recently, Soyemi et al. have reviewed the relevant data of inter-species/host-parasite protein interaction in a comprehensive manner [8], though the quantitative evaluation is still void. Inspired from the idea in [1], [8], in this paper, we focus on constructing a computational framework towards the evaluation of machine learning-based models for prediction task of human-pathogen protein-protein interactions (HP-PPIs). One major reason is that, to the best of our knowledge, no studies conduct a comprehensive quantitative exploration from both databases perspective and computational models comparison in terms of human-pathogen protein-protein interactions [1], [6]–[8]. In this work, we have further proposed an ensemble

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

machine learning-based model through mining the heterogeneous information of protein data. The proposed framework achieves a high performance of prediction.

In summary, the novel contributions made in this work are:

- We conduct an extensive review of the existing databases for human-pathogen interactions since 2000s. By doing so, several human-pathogen protein-protein interactions datasets are carefully curated from the selected databases.
- We perform a comprehensive experiment to collect a wide scale of the prediction performances for different machine learning-based models, which also include the methods from literature focusing on the prediction of human-pathogen protein interactions. Given the void of systematic evaluation of machine learning-based HP-PPIs prediction models, the first of this kind of evaluation show that there is plenty of room for improvements to achieve a robust and efficient machine learning-based model.
- We introduce a robust and accurate framework based on Heterogeneous Information Mining and Ensembling (HIME) prediction model to harness the power of heterogeneous information, thereby greatly improving the prediction performance. The experimental results indicate that the HIME model achieves the best and most robust performance for prediction of human-pathogen protein-protein interactions compared to the state-of-the-art.

In the remainder of this paper, we firstly present a comprehensive literature review on the host-pathogen protein-protein interactions prediction, as well as the published host-pathogen interactions databases, in section II. In section III, six different datasets for HP-PPIs prediction are curated as the materials, and the HIME model is then discussed in detail together with the heterogeneous information of sequence data. The baseline models for comparison are also elaborated in section III. In Section IV, the experimental results of HIME model comparing with other baseline models are reported. Finally Section V is the conclusion of this work.

II. Literature Review

A. Review of Host-Pathogen Protein Interactions

There have been a large body of research on protein-protein interactions, aiming at developing cost-effective methods for prediction of protein interactions [9]–[12]. Since proteins present different characteristics, the methods include text mining method, network analysis method, kernel-based method, machine learning-based method and so on. However, these methods are presented as feasible and effective methods in a combination with corresponding protein characteristics, such as sequence data, gene ontology and gene expression data.

In recent years, protein sequence data has prevailed in numerous research areas of protein, for example protein structure prediction, protein function prediction and as in our study, PPIs prediction. Fan et al. proposed to utilise protein sequences and machine learning model for the development of Pups (pupylation site predictor), in which the pseudo-amino acid composition information was particularly employed [13]. To deal with the avalanche of newly sequenced protein data, the feature representation methods of protein sequence data were well designed as one of the important components for machine learning-based PPIs prediction models [3], [9], [14], [15]. Because sequence data was the most abundant data benefiting from high-throughput technology development, it would be beneficial to understand the performance in computational models and develop a more efficient model for HP-PPIs prediction.

Although there have been other literature reviews on the topic of inter-species PPIs prediction, such as host-pathogen interaction in [6]–[8], they barely focused on the topic discussion but failed to present the quantitative analysis. Particular models developed in [16] and [3] have demonstrated the effectiveness of encoding protein based on the sequence data to build machine learning-based models for HP-PPIs prediction.

B. Review of the Host-Pathogen Interactions Databases

Because host-pathogen interactions are the dominant interactions for infectious diseases studies and they are also mostly presented as inter-species protein interactions, we have kept our study subjects for host-pathogen interactions in this paper for brevity. Also, because it is critical to the understanding of infectious diseases, the initial development efforts of online host-pathogen interactions databases and repositories have been continuously updated by the researchers [7]. The resources cover a wide range of topics of host-pathogen interactions, including the protein-protein interactions, protein-mRNA interactions and their structural information. In our study, we have particularly filtered the online published resources by searching NCBI PubMed search engine with keywords ‘pathogen’ and ‘database’.

The preliminary results are manually examined with ‘Abstract’ from the first 400 returning items ranking by best relevance out of more than 4,000 papers. Most of the efforts and developments benefited from the strategic plan initialized by the National Institute of Allergy and Infectious Diseases (NIAID), which focus on biodefense research to define the ‘Priority Pathogens’ and to develop a subsequent watch list of genera [17], [18]. There have been several initial developments wholly or partially funded by NIAID, including the BioHealthBase [19], the pathogen interaction gateway (PIG) [20], the Virus Pathogen Database and Analysis Resource (ViPR) [21], VectorBase [22], the Pathosystems Resource Integration Center (PATRIC) [17], the Eukaryotic Pathogen Database (EuPathDB) [23].

The consolidation and facilitation of host-pathogen interactions studies have thus been promoted to elaborate the infectious and defensive mechanisms [2], [24]. The studies range from the eukaryotic pathogens, to fungi, virus, protozoa and bacteria. We herein review some of the public databases to be included in our following study. Eleven public databases are subsequently selected since their data sources mainly come from literature, domain expert manual verification and public archival databases.

TABLE I
HOST-PATHOGEN INTERACTIONS RESOURCES

Database	Data Source	Data Type	HPI Number
DIP	Literature and domain expert manual verification	Protein-protein interactions	76,882
Reactome	Literature and domain expert manual verification	Comprehensive data portal including pathway and analysis	1,016,953
APID	Public archival databases	Protein-protein interactions	133,994
IntAct	Public archival databases and literature	Molecular interaction database	857,826
MINT	Literature	Protein-protein interactions	123,892
InnateDB	Literature	Mammalian innate immunity networks, pathways and genes	24,077
PHISTO	Public archival databases	Host-pathogen and human intraspecies protein-protein interactions	90,453
PATRIC	Public archival databases	Comprehensive data portal for bacterium pathogens	618,737
Mentha	Public archival databases	Protein-protein interactions	1,272,096
HPIDB	Public archival databases and literature	Host-pathogen interactions	62,783
BioGRID	Literature	Comprehensive data portal for protein, genetic and chemical interactions	1,568,115

The Pathosystems Resource Integration Center (PATRIC) [17] targeted on all bacterial data types in its current incarnation for all NIAID priority pathogenic genera. The related data types include PPIs, genomics, transcriptomics, three-dimensional protein structures and sequence data. This relational database jointly integrates analytic and visualization tools, such as BLAST (the Basic Local Alignment Search Tool), to allow experts and computationally 'naïve' users to obtain metadata with interests. It was also built upon several other public archival databases, such as MINT [25], IntAct [26], BioGRID [27] and DIP [28]. The pathogen-host interaction search tool (PHISTO) [18] is another Web-accessible platform for HPI resources. The goal was to access a complete coverage of HPI data. The database is updated monthly.

The other databases used in our dataset curation include DIP [28], Reactome [29], APID [30], IntAct [26], MINT [25], InnateDB [31], PATRIC [17], Mentha [32], HPIDB [33] and BioGRID [27]. In TABLE. I, the databases are selected as the primary human-pathogen interactions resources.

Next, we extensively curated the downloaded data from the 11 databases and identified six different bacterium pathogens interacting with human. We specify the bacterium pathogens species by their taxonomy ID. In TABLE. II, the statistic of the collected positive human-bacterium protein-protein interactions is presented, which includes the species of *'Clostridium botulinum'*, *'Aeromonas hydrophila'*, *'Shigella paradysenteriae'*, *'Francisella tularensis subsp. tularensis (strain SCHU S4 / Schu 4)'*, *'Bacillus anthracis bacterium'* and *'Yersinia pseudotuberculosis subsp. pestis (Lehmann and Neumann 1896) Bercovier et al. 1981'*.

TABLE II
SELECTED HUMAN-PATHOGENS INTERACTIONS SYSTEMS' DATASETS

Taxonomy ID	Bacterium Pathogens	Positive Interactions	Training Dataset	Independent Dataset
1491	*Clostridium botulinum*	57	90	24
644	*Aeromonas hydrophila*	73	116	30
623	*Shigella paradysenteriae*	105	168	42
177416	*Francisella tularensis subsp. tularensis*	1207	1930	484
1392	*Bacillus anthracis bacterium*	2810	4496	1124
632	*Yersinia pseudotuberculosis subsp. pestis*	3528	5644	1412

III. MATERIALS AND MODEL

A. Datasets Curation

From the databases, the collected data are positive protein interactions data. We firstly process the data from two aspects. One is to reduce the ID information redundancy, as there may be duplicate entries when combining data from different databases. Another is related to sequence length. The proteins with less than 50 amino acids are discarded since they may be non-functional fragments. After cleansing, TABLE. II. illustrates the statistic for the selected bacterium pathogen species.

There have been discussions concerning how to select feasible negative PPIs. Currently, there is not a standard protocol defining the negative pairing strategy. In most of the literature, building a negative interaction dataset by randomly pairing proteins from the set of unknown interacting PPIs is utilized [3], [14], [16].

In our study, the sequence data, which is dominantly published by UniProtKB database, has been used. The information helps us to build the negative inter-species PPIs as well as building the independent datasets. To obtain a sufficiently comprehensive evaluation, a dedicated preparation of independent datasets is applied, which datasets should not be used during the training and will be reported with different measurements to evaluate the model performance.

Thus, we firstly randomly select one-fifth PPIs from both positive and negative interactions as the independent dataset. The rest PPIs of positive and negative interactions are combined as the training set. Since we construct the negative interactions by a random sampling method, we apply the random sampling for the negative interactions by five times and measure the evaluation with statistic means and variations to reduce the bias caused by negative interactions. In TABLE. II, the details of the final curated datasets are shown.

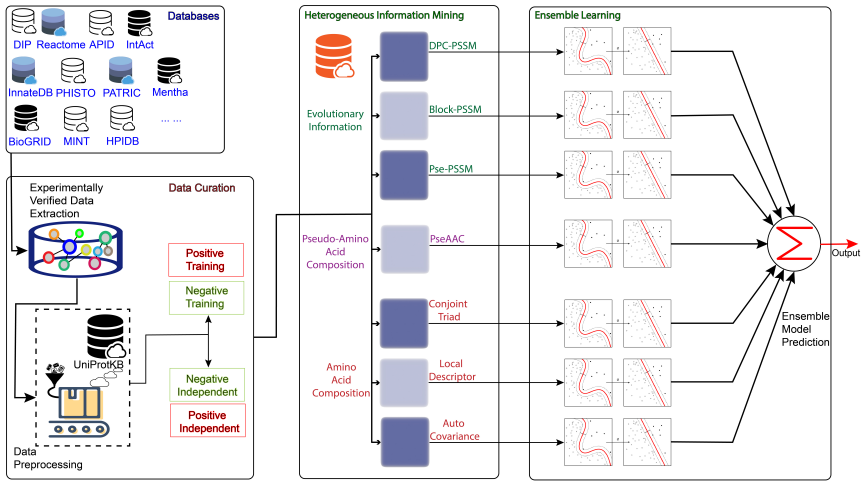

Fig. 1. The Framework of HIME Model

B. The HIME Model

In this section, we will firstly introduce the HIME model, then the details of each part of HIME model will be explained.

The proposed heterogeneous information mining and ensembling (HIME) model is shown in Fig. 1, which leverages the mining and ensembling process of heterogeneous information of sequence data, and also includes the learning process. HIME model is a sequence-based model, since the protein sequence data is considered as one of the most abundant data. The overwhelming sequence data has exclusively stimulated the ongoing research to improve the prediction performance based on novel feature representation algorithms of sequence data and machine learning models. It helps to generalize the computational models on a larger dataset and various species and genres.

HIME model tackle the heterogeneous information of sequence data in three different types, as shown in Fig. 1, which are amino acid composition information, pseudo-amino acid composition information and evolutionary information. Multiple training models are produced for different information, and HIME model subsequently utilises ensemble learning techniques to make the prediction with high performance for different human-pathogen interactions systems.

1) Heterogeneous Information of Sequence Data: Encoding sequence data as feature vectors is the first step in building computational model for prediction [3], [16]. Three different types of heterogeneous information of sequence data are explored in our proposed model, which helps to build a robust and efficient model.

a) Amino acid composition information: Amino acid composition information is dominantly inferred by the amino acids order of protein sequence data. There are several different methods converting this information into feature vectors. One was considering several adjacent amino acids as one region in the sequence, which was also called conjoint triad method feature or k-mer [34]. It considered the protein in segments to be functional between different proteins, which firstly classified the 20 different types of amino acids into seven groups according to their physiochemical characteristics. The groups were then indicated as 1-7 in numbers. When the region was limited as three adjacent amino acids, there would be a candidate of {(1,1,1), (1,2,1), ... (1,7,1), ..., (1,7,7), ..., (7,7,7)}. This encoded the sequence data into a 343-dimension vector. Also, the region can be selected as two, four, and other length adjacent amino acids.

Another approach based on amino acid composition information is to discover the auto covariance relationship among amino acids [15]. Auto covariance method considered each amino acid with its seven physicochemical properties. Thus, the 20 different amino acids were presented in a matrix of 20*7 dimension, normalized by Equa. 1:

$$\bar{P_{i,j}} = \frac{P_{i,j} - mean_j}{std_j} (i = 1, 2, 3, ..., 20; j = 1, 2, 3, ...7) \quad (1)$$

Here, $P(i, j)$ is the value of jth property for ith amino acid, $mean_j$ is the mean value of jth property of the 20 amino acids and std_j is the standard deviation of jth property over the 20 amino acids. For jth property, the auto covariance relationship was calculated for two different locations of amino acids given the maximum distance Dis in Equa. 2. The dimension of feature vector generated via auto covariance method would be $Dis * 7$, when all seven different properties are employed.

$$ACC(D, j) = \frac{1}{N-D} \sum_{i=1}^{N-D} (\bar{P_{i,j}} - \frac{1}{N} \sum_{i=1}^{N} \bar{P_{i,j}}) * (\bar{P_{i+D,j}} - \frac{1}{N} \sum_{i=1}^{N} \bar{P_{i,j}}) \quad (2)$$

The last popular method for amino acid composition information is local descriptor [35], which has divided the protein sequence information into 10 regions of six different types, including by quarter division, half division, central 50% region, first 75% region, last 75% region and central 75% region. Local descriptor specifically defined three different descriptors for each region, including composition, transition and distribution. This generated seven features for composition, 21 features for transition and 35 features for distribution. Totally with the 10 regions, local descriptor generated 630-dimension feature vector for single protein sequence.

b) Pseudo-amino acid information: Even though amino acid composition information takes consideration of sequence order to some extent, there is still some information loss when directly encoding sequence data based on composition information. Thus, pseudo-amino acid information is discovered as an important type of information of sequence data [36]. It firstly calculated the relationship between two different amino acids by Equa.3.

$$\begin{aligned} C(R_i, R_j) = & \frac{1}{3}\{[P_1(R_j) - P_1(R_i)]^2 + [P_2(R_j) - P_2(R_i)]^2 \\ & + [P_3(R_j) - P_3(R_i)]^2\} \end{aligned} \tag{3}$$

P_1, P_2 and P_3 represent different properties of amino acid R_i. Then, the pseudo-amino acid information was calculated as Equa.4. This would generate a λ-dimension feature vector.

$$\begin{aligned} \theta_1 &= \frac{1}{N-1}\sum_{i=1}^{N-1} C(S_i, S_(i+1)) \\ \theta_2 &= \frac{1}{N-2}\sum_{i=1}^{N-2} C(S_i, S_(i+2)) \quad \lambda < T \\ &... \\ \theta_\lambda &= \frac{1}{N-\lambda}\sum_{i=1}^{N-\lambda} C(S_i, S_{i+\lambda}) \end{aligned} \tag{4}$$

c) Evolutionary information: Another important information of sequence data is the evolutionary information, which represents the continuous change and evolution trends in a given reference protein database. The information is referred as a scoring matrix to indicates the probability of related amino acid types in corresponding position. It is commonly derived by aligning a set of sequence, which is considered to be functionally related. One important matrix firstly derived is called the position-specific scoring matrix (PSSM), which is a T*20 matrix for a given protein sequence. T represents the length of its corresponding protein sequence. Several algorithms have been developed to generate feature vector for single protein sequence. The first one is pseudo position-specific score matrix (Pse-PSSM), which combines the idea of pseudo-amino acid composition [37]. Pse-PSSM represented the original PSSM by compressing the matrix values vertically into their corresponding mean value. This means, after transformation, PSSM becomes a 20-dimension Pse-PSSM vector. Another one is called Block-PSSM by dividing sequence data into 20 equal blocks [38]. Each block represents five percent of a sequence. For each block, a 20-dimension vector is extracted. This generates a 20*20=400-dimension vector totally with 20 blocks. The last one is the traditional dipeptide composition PSSM (DPC-PSSM) [39]. It calculated the covariance of two adjacent amino acid and represented the information in a 400-dimension feature vector.

The heterogeneous information of sequence data have been categorized in three different types, as shown in Fig. 1. Different algorithms including conjoint triad method (CTM) [34], auto covariance (ACC) [15], local descriptor (LD) [35], PseAAC [36], pseudo position-specific score matrix (Pse-PSSM) [37], transition dipeptide composition PSSM (DPCPSSM) [39] and block PSSM (BlockPSSM) [38] algorithms, are subsequently incorporated in HIME model.

2) Ensemble Learning: Machine learning-based models have been widely applied for prediction of bioinformatics tasks recently. Mostly, the models are compared and the best of the models is selected as the applied computational model.

Ensemble learning model is designed with multiple machine learning models, which are called 'base learner' for same task [40]. Typically, ensemble learning model benefits from the integration of individual base learners to achieve a robust and superior performance. Even though there are different categories of ensemble learning model, various applications have shown that none of them could be outstanding consistently [41]–[43].

Algorithm 1: Heterogeneous Information Ensembling Process

Input: Dataset $D = (x_1, y_1), (x_2, y_2), ..., (x_m, y_m)$;
Heterogeneous information feature representation algorithms $\Re_1, \Re_2, ..., \Re_T$;
Base learner algorithms $\mathcal{L}_1, \mathcal{L}_2, ..., \mathcal{L}_T$;
Ensemble learner $\mathcal{L}$.

Process:

for $t = 1$ *to* T **do** *Heterogeneous information mining*
  $\mathcal{D}_t = \Re_t()$ %Mining heterogeneous information
    %and applying the different feature
    %representation algorithms;
end
for $t = 1$ *to* T **do**
  h_t = $\mathcal{L}_t(\mathcal{D}_t)$ %Training a base learner algorithm h_t
    %by applying the base leanrner
    %algorithm $\mathcal{L}_t$ to the dataset $\mathcal{D}_t$;
end
$\mathcal{D}'$ = $\varnothing$ %Collect the base learners;
for $i = 1$ *to* m **do**
  for $t = 1$ *to* T **do**
    z_{it} = $h_t(x_i)$ %Use h_t to classify the Dataset D;
  end
  $\mathcal{D}'$ = $\mathcal{D}' \cup \{((z_{i1},z_{i2},...,z_{iT}),y_i)\}$;
end
h' = $\mathcal{L}(\mathcal{D}')$;

Output: $H(x) = h'(h_1(x), ..., h_T(x))$.

Generally, the ensemble learning model can be deployed either vertically or horizontally [43]. To avoid building a single strong machine learning model in the task, HIME model leverages the heterogeneous information and plenarily exerts the various base learners in a horizontal way. lightGBM [44], one of the recently popular tree-based models, is selected as the base learner in the model to build HIME for prediction of human-pathogen protein-protein interactions.

Algorithms 1 illustrates the procedure of HIME model. Our model not only leverages the precision and diversity from base learner, but also emphasises the diversity from the heterogeneous information mining process. As a result, HIME model is capable to enhance the performance fueled by the

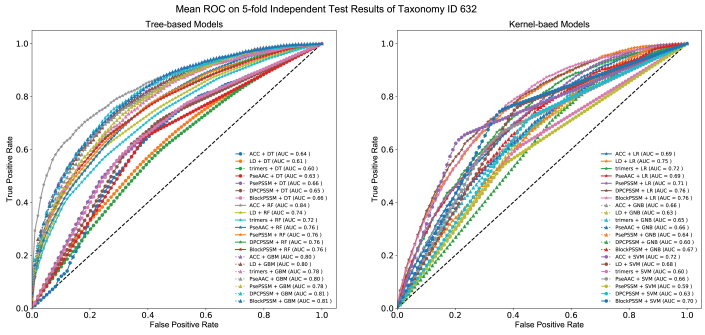

Fig. 2. The ROC Curves for 'HB_6' of Traditional Models

designed information mining and ensembling procedure.

C. Baseline Models

In this study, different methods, such as [16] and [3] from literature, and traditional machine learning models including random forest, support vector machine, logistic regression model, Gaussian naïve Bayes, decision tree and gradient boosting machine, are used in the prediction task of HP-PPIs. These models explicitly demonstrate different capabilities on different tasks, such as classification task and time series regression task.

Since these models are traditionally used in different tasks, our evaluation tasks explore a comprehensive experiments to generate the performance for the prediction task. In our study, we take advantage of mining all the available information of sequence data and further present a comprehensive evaluation, which compares the performance with corresponding machine learning models for HP-PPIs prediction. Particularly, we have included the performance of different groups of feature representation algorithms and machine learning models. This results in 42 different combinations as the first group baseline models. The hyperparameters are subsequently obtained by 5-fold cross validation for classifier according to the dataset.

$$
\begin{aligned}
Accuracy &= \frac{TP+TN}{TP+FP+TN+FN} \\
Precision &= \frac{TP}{TP+FP} \\
Recall &= \frac{TP}{TP+FN} \\
Specificity &= \frac{TN}{TN+FP} \\
F1 &= \frac{2*Precision}{Precision+Recall} \\
MCC &= \\
&\frac{(TP*TN)-(FN*FP)}{\sqrt{(TP+FN)*(TN+FP)*(TP+FP)*(TN+FN)}}
\end{aligned}
\tag{5}
$$

Secondly, we compared with two methods from literature, which are [3] and [16], with the same methods settings. Wuchty et al. presented a random forests model as the ensemble model to learn from the host-parasite protein-protein interactions [16]. A variant version of amino acid triplets algorithm was used as the feature representation algorithm. Cui et al. applied SVM as the computational model with the proposed protein sequence representation algorithm to predict the human-pathogen protein-protein interactions [3].

D. Performance Measurements

To evaluate the performance of our model, numerous metrics are compared, including the accuracy, precision, recall, specificity, F1-score, the area under curve (AUC) value and Matthew's correlation coefficient (MCC) score. We have also collected the receiver operating characteristic curves (ROC). The Equa. 5 show the definitions of these metrics.

IV. Results and Discussion

We collected the results of a 5-fold independent test of the six different taxonomy IDs datasets. Herein, the performances with both the mean values and the deviations are presented.

A. Baseline models

Firstly, we discuss the evaluations on traditional machine learning models, including decision tree (DT), random forest (RF), gradient boosting machine (GBM), logistic regression (LR), Naïve Bayesian and support vector machine (SVM). Seven different feature representation algorithms of sequence data are included and the corresponding models are built upon six traditional machine learning models, which result in 42 different models. TABLE. III includes the accuracy and F1 score for all the evaluated models, including HIME model. The performances of traditional models, '$Model_1$' and '$Model_2$', share a same fluctuation trend concerning different datasets, which worst performances are all observed with 'HB_6'.

B. HIME Model Performance and Comparison

In TABLE. III, the best models are indicated in bold fonts. We can clearly observe that for five prediction tasks, which are 'HB_1', 'HB_3', 'HB_4', 'HB_5' and 'HB_6', the best performances are all achieved by our proposed HIME model. This indicates that mining and ensembling heterogeneous information of sequence data indeed help boosting the model performance.

In Fig. 2 and Fig. 3, we have shown partial results of the ROC curves for discussion due to the limited space. The ROC curves show that, different types of protein sequence information generate diverse learners, which generate different performance. One particularly selected information may

TABLE III
RESULTS OF ACCURACY AND F1 SCORE

Model		Accuracy						F1 Score					
		HB_1[a]	HB_2	HB_3	HB_4	HB_5	HB_6	HB_1	HB_2	HB_3	HB_4	HB_5	HB_6
RF	$\Re_1$[b]	1.000±0.000	0.967±0.000	0.824±0.036	0.725±0.008	0.773±0.011	0.757±0.008	1.000±0.000	0.966±0.000	0.811±0.040	0.730±0.005	0.770±0.011	0.752±0.007
	$\Re_2$	1.000±0.000	0.967±0.000	0.757±0.038	0.696±0.007	0.689±0.010	0.661±0.015	1.000±0.000	0.966±0.000	0.771±0.034	0.715±0.011	0.710±0.009	0.691±0.012
	$\Re_3$	0.975±0.033	0.967±0.000	0.752±0.053	0.686±0.011	0.670±0.015	0.651±0.012	0.974±0.036	0.966±0.000	0.768±0.049	0.707±0.009	0.696±0.014	0.680±0.011
	$\Re_4$	1.000±0.000	0.967±0.000	0.795±0.039	0.682±0.017	0.701±0.005	0.684±0.016	1.000±0.000	0.966±0.000	0.807±0.033	0.705±0.016	0.723±0.006	0.711±0.012
	$\Re_5$	1.000±0.000	0.973±0.013	0.876±0.035	0.671±0.014	0.680±0.004	0.683±0.009	1.000±0.000	0.972±0.014	0.878±0.038	0.696±0.008	0.704±0.004	0.711±0.006
	$\Re_6$	1.000±0.000	0.973±0.013	0.814±0.063	0.679±0.010	0.690±0.015	0.678±0.009	1.000±0.000	0.974±0.013	0.831±0.049	0.709±0.012	0.712±0.011	0.707±0.008
	$\Re_7$	1.000±0.000	**0.993±0.013**	0.838±0.066	0.678±0.014	0.687±0.011	0.676±0.006	1.000±0.000	**0.993±0.014**	0.852±0.057	0.707±0.010	0.709±0.010	0.709±0.004
SVM	$\Re_1$	1.000±0.000	0.867±0.000	0.800±0.024	0.700±0.013	0.653±0.015	0.719±0.012	1.000±0.000	0.846±0.000	0.775±0.016	0.674±0.014	0.656±0.013	0.697±0.012
	$\Re_2$	0.975±0.033	0.960±0.013	0.676±0.053	0.696±0.012	0.708±0.016	0.676±0.007	0.977±0.031	0.959±0.013	0.705±0.041	0.722±0.012	0.701±0.017	0.703±0.005
	$\Re_3$	1.000±0.000	0.860±0.033	0.790±0.046	0.651±0.008	0.696±0.007	0.597±0.009	1.000±0.000	0.835±0.046	0.792±0.036	0.678±0.009	0.702±0.007	0.599±0.007
	$\Re_4$	1.000±0.000	0.700±0.101	0.752±0.049	0.666±0.019	0.604±0.025	0.661±0.018	1.000±0.000	0.762±0.061	0.741±0.040	0.670±0.020	0.539±0.125	0.657±0.014
	$\Re_5$	1.000±0.000	0.767±0.060	0.729±0.058	0.583±0.008	0.665±0.007	0.588±0.005	1.000±0.000	0.734±0.062	0.722±0.028	0.531±0.010	0.682±0.005	0.567±0.002
	$\Re_6$	0.992±0.017	0.853±0.086	0.648±0.035	0.601±0.010	0.642±0.016	0.635±0.007	0.992±0.016	0.877±0.064	0.701±0.031	0.615±0.008	0.644±0.026	0.663±0.005
	$\Re_7$	1.000±0.000	0.947±0.027	0.900±0.035	0.673±0.011	0.635±0.047	0.699±0.009	1.000±0.000	0.948±0.026	0.908±0.032	0.683±0.006	0.665±0.090	0.713±0.008
LR	$\Re_1$	0.942±0.033	0.967±0.000	0.819±0.032	0.635±0.018	0.656±0.012	0.645±0.006	0.946±0.031	0.966±0.000	0.818±0.030	0.642±0.020	0.661±0.011	0.654±0.004
	$\Re_2$	0.983±0.033	0.953±0.016	0.695±0.051	0.709±0.017	0.709±0.013	0.686±0.010	0.985±0.031	0.953±0.016	0.691±0.039	0.719±0.018	0.720±0.014	0.696±0.006
	$\Re_3$	0.983±0.020	0.960±0.013	0.829±0.028	0.681±0.013	0.692±0.010	0.659±0.010	0.984±0.020	0.959±0.013	0.828±0.029	0.684±0.015	0.698±0.007	0.659±0.011
	$\Re_4$	0.975±0.033	0.887±0.045	0.843±0.061	0.673±0.017	0.676±0.014	0.643±0.015	0.974±0.036	0.892±0.041	0.840±0.065	0.673±0.011	0.664±0.007	0.624±0.024
	$\Re_5$	1.000±0.000	0.953±0.016	0.857±0.052	0.678±0.008	0.662±0.008	0.648±0.008	1.000±0.000	0.953±0.016	0.856±0.055	0.692±0.006	0.670±0.005	0.659±0.006
	$\Re_6$	0.992±0.017	0.960±0.025	0.876±0.038	0.712±0.018	0.706±0.016	0.694±0.013	0.992±0.016	0.961±0.024	0.881±0.035	0.723±0.016	0.713±0.012	0.702±0.011
	$\Re_7$	1.000±0.000	0.973±0.025	0.895±0.029	0.667±0.015	0.702±0.017	0.695±0.014	1.000±0.000	0.973±0.025	0.899±0.028	0.675±0.015	0.711±0.016	0.707±0.013
Naïve Bayes	$\Re_1$	1.000±0.000	0.967±0.000	0.767±0.061	0.661±0.007	0.625±0.002	0.608±0.019	1.000±0.000	0.966±0.000	0.794±0.043	0.714±0.005	0.672±0.003	0.653±0.013
	$\Re_2$	1.000±0.000	0.967±0.000	0.667±0.050	0.658±0.007	0.634±0.008	0.597±0.015	1.000±0.000	0.966±0.000	0.697±0.030	0.708±0.006	0.687±0.006	0.646±0.009
	$\Re_3$	1.000±0.000	0.967±0.000	0.733±0.049	0.621±0.016	0.639±0.011	0.609±0.013	1.000±0.000	0.966±0.000	0.724±0.034	0.685±0.006	0.667±0.003	0.614±0.026
	$\Re_4$	1.000±0.000	0.947±0.040	0.619±0.054	0.638±0.022	0.580±0.009	0.574±0.006	1.000±0.000	0.947±0.036	0.524±0.088	0.622±0.039	0.341±0.032	0.380±0.011
	$\Re_5$	1.000±0.000	0.967±0.000	0.833±0.040	0.609±0.008	0.592±0.014	0.589±0.009	1.000±0.000	0.966±0.000	0.828±0.039	0.574±0.005	0.539±0.016	0.545±0.007
	$\Re_6$	1.000±0.000	0.960±0.013	0.619±0.045	0.571±0.015	0.565±0.005	0.586±0.007	1.000±0.000	0.958±0.015	0.673±0.033	0.618±0.010	0.602±0.013	0.654±0.003
	$\Re_7$	1.000±0.000	0.967±0.000	0.752±0.061	0.627±0.014	0.610±0.003	0.632±0.007	1.000±0.000	0.966±0.000	0.771±0.052	0.647±0.010	0.616±0.001	0.665±0.005
GBM	$\Re_1$	0.892±0.090	0.933±0.021	0.771±0.089	0.719±0.009	0.744±0.013	0.724±0.011	0.900±0.079	0.934±0.020	0.758±0.097	0.727±0.010	0.747±0.013	0.733±0.009
	$\Re_2$	0.975±0.020	0.953±0.016	0.814±0.038	0.728±0.013	0.753±0.017	0.721±0.017	0.976±0.020	0.953±0.016	0.817±0.037	0.731±0.013	0.754±0.015	0.727±0.013
	$\Re_3$	0.867±0.085	0.960±0.013	0.795±0.039	0.714±0.016	0.735±0.006	0.700±0.010	0.860±0.098	0.959±0.013	0.794±0.040	0.719±0.017	0.737±0.006	0.700±0.006
	$\Re_4$	1.000±0.000	0.953±0.034	0.833±0.058	0.720±0.020	0.737±0.013	0.719±0.014	1.000±0.000	0.954±0.032	0.842±0.054	0.724±0.022	0.741±0.011	0.725±0.010
	$\Re_5$	1.000±0.000	0.933±0.021	0.886±0.059	0.728±0.015	0.720±0.010	0.710±0.006	1.000±0.000	0.935±0.020	0.893±0.052	0.737±0.014	0.722±0.011	0.720±0.005
	$\Re_6$	0.992±0.017	0.987±0.016	0.824±0.072	0.725±0.011	0.738±0.007	0.734±0.010	0.992±0.016	0.987±0.016	0.830±0.068	0.735±0.010	0.738±0.009	0.743±0.008
	$\Re_7$	1.000±0.000	0.967±0.030	0.910±0.038	0.719±0.011	0.743±0.007	0.729±0.008	1.000±0.000	0.967±0.029	0.915±0.035	0.729±0.009	0.748±0.005	0.739±0.008
DT	$\Re_1$	0.900±0.077	0.927±0.033	0.676±0.061	0.628±0.015	0.604±0.017	0.624±0.013	0.905±0.077	0.929±0.030	0.681±0.072	0.655±0.009	0.640±0.012	0.656±0.011
	$\Re_2$	0.942±0.057	0.953±0.016	0.729±0.081	0.650±0.028	0.593±0.003	0.590±0.021	0.936±0.070	0.953±0.016	0.722±0.092	0.653±0.028	0.587±0.055	0.591±0.038
	$\Re_3$	0.883±0.085	0.980±0.016	0.695±0.087	0.640±0.018	0.609±0.015	0.576±0.009	0.875±0.099	0.979±0.017	0.712±0.079	0.634±0.029	0.611±0.047	0.571±0.023
	$\Re_4$	1.000±0.000	0.953±0.034	0.705±0.094	0.633±0.010	0.644±0.019	0.629±0.008	1.000±0.000	0.954±0.032	0.685±0.105	0.640±0.010	0.651±0.017	0.636±0.010
	$\Re_5$	0.983±0.020	0.933±0.021	0.829±0.065	0.632±0.011	0.644±0.006	0.632±0.012	0.984±0.020	0.934±0.020	0.834±0.061	0.639±0.006	0.650±0.009	0.641±0.014
	$\Re_6$	0.992±0.017	0.947±0.054	0.710±0.070	0.640±0.023	0.640±0.013	0.634±0.015	0.992±0.016	0.949±0.051	0.735±0.049	0.648±0.028	0.648±0.011	0.648±0.015
	$\Re_7$	1.000±0.000	0.960±0.033	0.771±0.072	0.632±0.018	0.631±0.008	0.630±0.012	1.000±0.000	0.960±0.031	0.764±0.096	0.639±0.022	0.639±0.005	0.639±0.014
$Model_1$[c]		1.000±0.000	0.900±0.000	0.800±0.029	0.710±0.012	0.742±0.004	0.719±0.019	1.000±0.000	0.889±0.000	0.779±0.028	0.705±0.011	0.735±0.004	0.716±0.016
$Model_2$[d]		0.992±0.017	0.967±0.000	0.810±0.030	0.689±0.021	0.731±0.015	0.706±0.008	0.992±0.016	0.966±0.000	0.810±0.026	0.682±0.018	0.728±0.013	0.705±0.007
HIME (proposed)		**1.000±0.000**	0.967±0.000	**0.929±0.037**	**0.757±0.009**	**0.801±0.009**	**0.783±0.010**	**1.000±0.000**	0.966±0.000	**0.931±0.032**	**0.763±0.009**	**0.798±0.007**	**0.783±0.008**

[a] HB_1~HB_6 represent the six built dataset in the order from TABLE II, which are '1491', '644', '623', '177416', '1392', '632', respectively;
[b] $\Re_1$~$\Re_7$ are the different feature representations algorithms, representing ACC, LD, CTM, PseAAC, PsePSSM, DPCPSSM, BlockPSSM;
[c] $Model_1$ is the method from [3];[d] $Model_2$ is the method from [16]

not be sufficient to produce a robust model. Moreover, the performance will become worse when the dataset is larger.

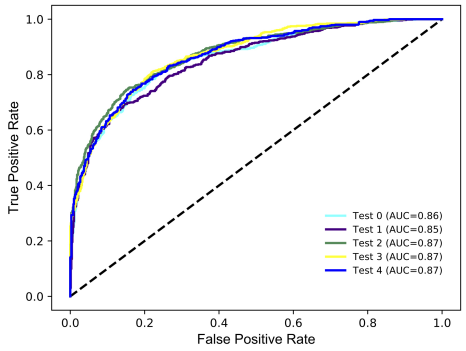

Fig. 3. The ROC Curves for 'HB_6' of HIME Model

In comparison with Fig. 2, the ROC curves for five-times independent test of 'HB_6' with HIME model is diagrammed in Fig. 3. As for our proposed HIME model utilizing heterogeneous information, the model obtains a more robust and accurate performance than the other baseline models. Regarding the performance metrics including Specificity, MCC and AUC values, we have also observed the same performance comparison results, in which HIME model outperforms the others. However, they are not presented in this paper due to the space limit as well. The performance comparison demonstrates that, the proposed HIME model outperforms most of the predictor compared in this study for different human-pathogen PPIs prediction tasks. Hence, the heterogeneous information mining and ensembling strategy benefits the performance improvement in this work.

V. CONCLUSION

In this paper, we have firstly presented an extensive study covering the pathogens database since 2000s and conducted an evaluation for human-pathogen protein-protein interaction prediction task, given protein sequence data. To the best of our knowledge, our work is by far the first comprehensive quantitative review focusing on this area. The prediction of HP-PPIs could help the study of infectious disease mechanisms. A robust performance of the prediction model is desired to achieve for different pathogen species. Through mining the heterogeneous information of sequence data, we have proposed HIME model leveraging the abundant information. Furthermore, the horizontal ensemble procedure with heterogeneous information has greatly exerted the base learners to boost the performance in the prediction task. The performances are evaluated on six different datasets and indicate HIME model outperforms the others. However, in this study, the performance declines when dataset size becomes larger. The future work will be targeted to boost the performance by incor-

porating more dedicated feature representation algorithms and novel machine learning models. Meanwhile, we will enlarge the dataset by including more comprehensive experiments settings to response to the biology meanings accordingly.

ACKNOWLEDGMENT

This work is supported by the China Scholarship Council (CSC) scholarship, Faculty Strategic Investments Grant for DP 2019 development and The University Global Partnership Network (UGPN) Research Collaboration Fund.

REFERENCES

[1] A. K. Halder, P. Dutta *et al.*, "Review of computational methods for virus–host protein interaction prediction: a case study on novel ebola–human interactions," *Briefings in functional genomics*, vol. 17, no. 6, pp. 381–391, 2017.

[2] K. Asehnoune, J. Villadangos, and R. Hotchkiss, "Understanding host–pathogen interaction," *Intensive care medicine*, vol. 42, no. 12, pp. 2084–2086, 2016.

[3] G. Cui, C. Fang, and K. Han, "Prediction of protein-protein interactions between viruses and human by an svm model," in *BMC bioinformatics*, vol. 13, no. 7. BioMed Central, 2012, p. S5.

[4] H. Chen, W. Guo *et al.*, "Structural principles analysis of host-pathogen protein-protein interactions: A structural bioinformatics survey," *IEEE Access*, vol. 6, pp. 11 760–11 771, 2018.

[5] E. A. Franzosa and Y. Xia, "Structural models for host-pathogen protein-protein interactions: assessing coverage and bias," in *Biocomputing 2012*. World Scientific, 2012, pp. 287–298.

[6] E. Nourani, F. Khunjush, and S. Durmuş, "Computational approaches for prediction of pathogen-host protein-protein interactions," *Frontiers in microbiology*, vol. 6, p. 94, 2015.

[7] H. Zhou, J. Jin, and L. Wong, "Progress in computational studies of host–pathogen interactions," *Journal of bioinformatics and computational biology*, vol. 11, no. 02, p. 1230001, 2013.

[8] J. Soyemi, I. Isewon *et al.*, "Inter-species/host-parasite protein interaction predictions reviewed," *Current Bioinformatics*, vol. 13, no. 4, pp. 396–406, 2018.

[9] A. Ben-Hur and W. S. Noble, "Kernel methods for predicting protein–protein interactions," *Bioinformatics*, vol. 21, no. suppl_1, pp. i38–i46, 2005.

[10] N. Papanikolaou, G. A. Pavlopoulos *et al.*, "Protein–protein interaction predictions using text mining methods," *Methods*, vol. 74, pp. 47–53, 2015.

[11] U. Kuzmanov and A. Emili, "Protein-protein interaction networks: probing disease mechanisms using model systems," *Genome medicine*, vol. 5, no. 4, p. 37, 2013.

[12] R. Jansen, H. Yu *et al.*, "A bayesian networks approach for predicting protein-protein interactions from genomic data," *science*, vol. 302, no. 5644, pp. 449–453, 2003.

[13] Y.-X. Fan and H.-B. Shen, "Predicting pupylation sites in prokaryotic proteins using pseudo-amino acid composition and extreme learning machine," *Neurocomputing*, vol. 128, pp. 267–272, 2014.

[14] M. D. Dyer, T. Murali, and B. W. Sobral, "Supervised learning and prediction of physical interactions between human and hiv proteins," *Infection, Genetics and Evolution*, vol. 11, no. 5, pp. 917–923, 2011.

[15] Y. Guo, L. Yu *et al.*, "Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences," *Nucleic acids research*, vol. 36, no. 9, pp. 3025–3030, 2008.

[16] S. Wuchty, "Computational prediction of host-parasite protein interactions between p. falciparum and h. sapiens," *PLoS One*, vol. 6, no. 11, p. e26960, 2011.

[17] A. R. Wattam, D. Abraham *et al.*, "Patric, the bacterial bioinformatics database and analysis resource," *Nucleic acids research*, vol. 42, no. D1, pp. D581–D591, 2013.

[18] S. Durmuş Tekir, T. Çakır *et al.*, "Phisto: pathogen–host interaction search tool," *Bioinformatics*, vol. 29, no. 10, pp. 1357–1358, 2013.

[19] B. Squires, C. Macken *et al.*, "Biohealthbase: informatics support in the elucidation of influenza virus host–pathogen interactions and virulence," *Nucleic acids research*, vol. 36, no. suppl_1, pp. D497–D503, 2007.

[20] T. Driscoll, M. D. Dyer *et al.*, "Pig—the pathogen interaction gateway," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D647–D650, 2008.

[21] B. E. Pickett, E. L. Sadat *et al.*, "Vipr: an open bioinformatics database and analysis resource for virology research," *Nucleic acids research*, vol. 40, no. D1, pp. D593–D598, 2011.

[22] K. Megy, S. J. Emrich *et al.*, "Vectorbase: improvements to a bioinformatics resource for invertebrate vector genomics," *Nucleic acids research*, vol. 40, no. D1, pp. D729–D734, 2011.

[23] C. Aurrecoechea, A. Barreto *et al.*, "Eupathdb: the eukaryotic pathogen genomics database resource," *Nucleic acids research*, vol. 45, no. D1, pp. D581–D591, 2016.

[24] S. Braxton, D. Onstad *et al.*, "Description and analysis of two internet-based databases of insect pathogens: Edwip and vidil," *Journal of invertebrate pathology*, vol. 83, no. 3, pp. 185–195, 2003.

[25] L. Licata, L. Briganti *et al.*, "Mint, the molecular interaction database: 2012 update," *Nucleic acids research*, vol. 40, no. D1, pp. D857–D861, 2011.

[26] S. Kerrien, B. Aranda *et al.*, "The intact molecular interaction database in 2012," *Nucleic acids research*, vol. 40, no. D1, pp. D841–D846, 2011.

[27] A. Chatr-Aryamontri, R. Oughtred *et al.*, "The biogrid interaction database: 2017 update," *Nucleic acids research*, vol. 45, no. D1, pp. D369–D379, 2017.

[28] I. Xenarios, L. Salwinski *et al.*, "Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions," *Nucleic acids research*, vol. 30, no. 1, pp. 303–305, 2002.

[29] A. Fabregat, S. Jupe *et al.*, "The reactome pathway knowledgebase," *Nucleic acids research*, vol. 46, no. D1, pp. D649–D655, 2017.

[30] C. Prieto and J. De Las Rivas, "Apid: agile protein interaction dataanalyzer," *Nucleic acids research*, vol. 34, no. suppl_2, pp. W298–W302, 2006.

[31] K. Breuer, A. K. Foroushani *et al.*, "Innatedb: systems biology of innate immunity and beyond—recent updates and continuing curation," *Nucleic acids research*, vol. 41, no. D1, pp. D1228–D1233, 2012.

[32] A. Calderone, L. Castagnoli, and G. Cesareni, "Mentha: a resource for browsing integrated protein-interaction networks," *Nature methods*, vol. 10, no. 8, p. 690, 2013.

[33] M. G. Ammari, C. R. Gresham *et al.*, "Hpidb 2.0: a curated database for host–pathogen interactions," *Database*, vol. 2016, 2016.

[34] J. Shen, J. Zhang *et al.*, "Predicting protein–protein interactions based only on sequences information," *Proceedings of the National Academy of Sciences*, vol. 104, no. 11, pp. 4337–4341, 2007.

[35] M. N. Davies, A. Secker *et al.*, "Optimizing amino acid groupings for gpcr classification," *Bioinformatics*, vol. 24, no. 18, pp. 1980–1986, 2008.

[36] K.-C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins: Structure, Function, and Bioinformatics*, vol. 43, no. 3, pp. 246–255, 2001.

[37] K.-C. Chou and H.-B. Shen, "Memtype-2l: a web server for predicting membrane proteins and their types by incorporating evolution information through pse-pssm," *Biochemical and biophysical research communications*, vol. 360, no. 2, pp. 339–345, 2007.

[38] J. cheol Jeong, X. Lin, and X.-w. Chen, "On position-specific scoring matrix for protein function prediction," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 8, no. 2, pp. 308–315, 2010.

[39] S. Zhang, F. Ye, and X. Yuan, "Using principal component analysis and support vector machine to predict protein structural class for low-similarity sequences via pssm," *Journal of Biomolecular Structure and Dynamics*, vol. 29, no. 6, pp. 1138–1146, 2012.

[40] Z.-H. Zhou, "Ensemble learning," *Encyclopedia of biometrics*, pp. 411–416, 2015.

[41] Á. Győrfi, L. Kovács, and L. Szilágyi, "Brain tumor detection and segmentation from magnetic resonance image data using ensemble learning methods," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2019, pp. 909–914.

[42] S. Akodad, S. Vilfroy *et al.*, "An ensemble learning approach for the classification of remote sensing scenes based on covariance pooling of cnn features," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

[43] F. Fahiman, S. M. Erfani, and C. Leckie, "Robust and accurate short-term load forecasting: A cluster oriented ensemble learning approach," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[44] G. Ke, Q. Meng *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, 2017, pp. 3146–3154.