2019

Leveraging SMOTE in A Two-Layer Model for Prediction of Protein-Protein Interactions

Huaming Chen
University of Wollongong, hc007@uowmail.edu.au

Lei Wang
University of Wollongong, leiw@uow.edu.au

Chi-Hung Chi
Data61

Jun Shen
University of Wollongong, Massachusetts Institute of Technology, jshen@uow.edu.au

Leveraging SMOTE in A Two-Layer Model for Prediction of Protein-Protein Interactions

Abstract

The research of the mechanisms of infectious diseases between host and pathogens remains a hot topic. It takes stock of the interactions data between host and pathogens, including proteins and genomes, to facilitate the discoveries and prediction of underlying mechanisms. However, the incomplete protein-protein interactions data impediment the advances in this exploration and solicit the wet-lab experiments to examine and verify the latent interactions. Although there have been numerous studies trying to leverage the computational models, especially machine learning models, the performances of these models were not good enough to produce high-fidelity candidates of interactions data due to the nature of the proteinprotein interactions data. In this paper, we propose a two-layer model for prediction of host-pathogen protein-protein interactions tackling the challenges affiliated to the feature representation algorithms and the imbalanced data. The twolayer model consists of two essential modules, which are XGBoost to reduce the imbalanced ratio of the data and SVM to improve the performance. SMOTE technology is incorporated as a key component in our model to alleviate the bias of imbalanced ratio. In this study, we have carefully collected proteins interactions data from public databases and built a dataset following the protocol with consensus of literature. A variety of models, including traditional models, models in major literature and our model, are verified on the datasets. Results demonstrate that our model significantly improve the performance comparing with the other state-of-the-art models.

Keywords

two-layer, interactions, protein-protein, prediction, model, smote, leveraging

Disciplines

Engineering | Science and Technology Studies

Publication Details

Chen, H., Wang, L., Chi, C. & Shen, J. (2019). Leveraging SMOTE in A Two-Layer Model for Prediction of Protein-Protein Interactions. 2019 Seventh International Conference on Advanced Cloud and Big Data (CBD) (pp. 133-138). United States: IEEE.

Leveraging SMOTE in A Two-Layer Model for Prediction of Protein-Protein Interactions

Huaming Chen[1], Lei Wang[1],Chi-Hung Chi[2], Jun Shen[1,3]
1-School of Computing and Information Technology, University of Wollongong, Wollongong, NSW, Australia
2-Data61, CSIRO, Australia
3-Research Lab of Electronics, Department of EE and CS, Massachusetts Institute of Technology, Cambridge, USA
Email:hc007@uowmail.edu.au, leiw@uow.edu.au, chihung.chi@data61.csiro.au, jshen@uow.edu.au

***Abstract*—The research of the mechanisms of infectious diseases between host and pathogens remains a hot topic. It takes stock of the interactions data between host and pathogens, including proteins and genomes, to facilitate the discoveries and prediction of underlying mechanisms. However, the incomplete protein-protein interactions data impediment the advances in this exploration and solicit the wet-lab experiments to examine and verify the latent interactions. Although there have been numerous studies trying to leverage the computational models, especially machine learning models, the performances of these models were not good enough to produce high-fidelity candidates of interactions data due to the nature of the protein-protein interactions data. In this paper, we propose a two-layer model for prediction of host-pathogen protein-protein interactions tackling the challenges affiliated to the feature representation algorithms and the imbalanced data. The two-layer model consists of two essential modules, which are XGBoost to reduce the imbalanced ratio of the data and SVM to improve the performance. SMOTE technology is incorporated as a key component in our model to alleviate the bias of imbalanced ratio. In this study, we have carefully collected proteins interactions data from public databases and built a dataset following the protocol with consensus of literature. A variety of models, including traditional models, models in major literature and our model, are verified on the datasets. Results demonstrate that our model significantly improve the performance comparing with the other state-of-the-art models.**

Keywords—two-layer model; XGBoost; SVM; protein-protein interactions; imbalanced data

I. INTRODUCTION

There is a continuously broad research topic targeting on the mechanisms of infectious diseases [1, 2, 3]. These researches generally utilise the interaction data between host and pathogens, including proteins and genomes, to understand the theory of infectious diseases and anticipate to give effective solutions. One of the research issues towards this goal is the incomplete protein-protein interaction data between host and pathogens [4]. The nature of interaction data between host and pathogen introduces a huge amount of potential interaction data for biologists to examine and verify whether the relationship is positive or negative. Positive indicates there is a physical and chemical interaction between different proteins and different genomes, while negative means there is not interactions. Although the wet-lab experiments could be further facilitated by high-throughput technologies to generate the interaction data, it is still considered as a cost sensitive approach. Time and resources consumption are exponentially increased when the candidates of interaction data become a scale of millions.

One of the major alternatives is to build computational models to learn from the known interactions data. There have been several studies trying to allocate computational resources to facilitate the progress and generate high-fidelity candidates for biologists to examine by subsequent wet-lab experiments. These studies indicate that machine learning model in combination with proper feature representation algorithm will benefit the success of computational models [5] [6]. However, there remains a research gap concerning the datasets and model performance. Two general questions are raised for HP-PPIs task. The first is how to build a golden dataset for HP-PPIs prediction task and the other is how to improve the model performance by incorporating different feature representation algorithms and various machine learning models. A major scheme behind this study is to build a novel model based on protein sequence information, which helps us to keep as much HP-PPIs data as possible.

In our research, we take the insight of the host-pathogen protein-protein interactions (HP-PPIs) data by considering the relevant feature representation algorithm and the imbalanced ratio between the positive and negative data, to build a machine learning model for prediction of high-fidelity HP-PPIs. A two-layer model is proposed in this paper, which consists of XGBoost [7] and support vector machine (SVM) [8, 9] as the main modules. XGBoost is the first layer to take the raw input, as it generalizes well in a large scale of datasets considering different imbalanced ratios. To further alleviate imbalanced ratio of the HP-PPIs data, SMOTE technology [10] is employed to generate a balanced data which is subsequently dealt with SVM model. Given the excellent capability of SVM in handling continuous dataset, SVM model serves as the second layer to boost our prediction result and enhance the overall performance comparing with other state-of-the-art models and traditional models.

In the remainder of this paper, the related work is introduced in section II, while the two-layer model of our work is presented in section III. We then discuss the comparison protocol and performance of metrics in section IV. The details of our curated dataset and the performance comparison discussion are reported in section V. Section VI concludes our work.

II. RELATED WORK

Considering HP-PPIs data as one of the major data sources towards the research of infectious diseases, there have been several studies proposing both statistical and machine learning based models for prediction of HP-PPIs. Being accumulated in a large volume and fast speed, the HP-PPIs data have driven the recent research taking more consideration with the machine learning model as it has proven to be successful in many real-world scenario applications, such as images, videos and language.

To build machine learning based models for HP-PPIs prediction tasks, the information of protein data is largely

involved, including the structure information, domain information, network properties and sequence information. Several studies utilized some of these information to build the computational models [11, 12, 13], however most of the original interaction data are discarded during the dataset curation process. Missing data for different protein information is one of the main causes.

For sequence information, most of the protein have been determined by the sequencing technology and the information is hosted in The Universal Protein Resource (UniProt), which has been actively updated and maintained for decade [14]. Given amino acid triplets as the feature representation algorithms for protein sequence information, [5] introduced random forests as the ensemble learning method to learn from the collected host-parasite protein interactions data. Support vector machine is employed in [6] to predict protein-protein interactions between viruses and human, especially for human papillomaviruses (HPV) and hepatitis C virus (HCV). However, these two models could not be well generalized to our HP-PPIs prediction tasks, as they have not taken consideration of the imbalanced characteristics of the HP-PPIs data.

In our work, there are two major parts, one is to build a golden dataset for HP-PPIs prediction task and the other is to build novel models to improve the prediction performance. Following the studies from [5, 6, 11, 12], a dataset for HP-PPIs is built from 11 databases. The details of the dataset will be given in section V. As the databases only contain positive interaction data, the negative interaction data are subsequently generated in three different imbalanced ratios, which are 1:25, 1:50 and 1:100. The hypothesis behind this setting is that, the number of truly interacting pairs of human-pathogen proteins is likely to be far less than the total set of protein pairs [11]. Meanwhile, we limit the study by utilizing protein sequence information as to keep the most of interactions data. Thus, local descriptor algorithm [15] is introduced in our model to map the protein sequence into vectors of same dimension.

III. TWO-LAYER MODEL

In this section, a two-layer model is presented, which includes XGBoost as the first layer to reduce the imbalanced ratio and SVM as the second layer to enhance the prediction result. An overview of the two-layer model is presented in Figure 1.

A. XGBoost

XGBoost is a scalable tree boosting system, which has proved to provide a powerful and efficient gradient boosting framework library in many applications. Benefitting from the tree boosting algorithms, XGBoost further extend the gradient boosting decision tree (GBDT) into a parallel approach to achieve a fast and accurate result.

Since XGBoost is an "extreme gradient boosting" implementation for tree ensemble models, it serves as our first layer to classify the imbalanced dataset. The predicted negative interaction data from XGBoost is considered as true negative data and we will be subsequently dealt with the rest predicted positive data. A random sampling after the first layer prediction will be conducted to generate a sampling negative interaction data. The output of the first layer will be a sampling interaction data and it will be input into the SMOTE module to generate a balanced dataset.

B. Support Vector Machine

Support vector machine is a powerful machine learning model which has demonstrated a strong generalization ability in tasks including classification, regression and distribution estimation. Given a dataset of HP-PPIs denoted as $\{x_i, y_i\}, i = 1,2, \dots, N$, SVM model outputs the prediction results according to Equa. (1):

$$y(x) = sign\left[\sum_{i=1}^{N} y_i \alpha_i * \mathrm{K}(x, x_i) + b\right] \qquad (1)$$

Here, $x_i \epsilon R^n$ and $y_i \epsilon \{+1, -1\}$. $\mathrm{K}(x_i, x_j)$ stands for the kernel function in SVM, i.e. for Radial Basis Functions (RBF) kernel, $\mathrm{K}(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$. α_i is the hyper parameters.

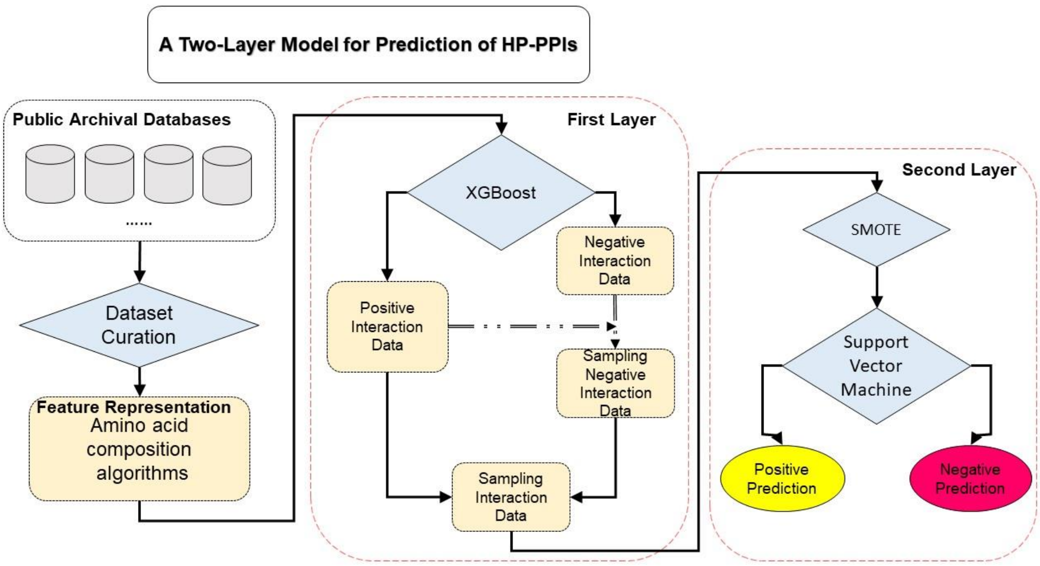

Figure 1 The Overview of Our Model

In our two-layer model, SVM serves as the second layer taking the balanced dataset as the input. Its ability to mapping data into higher dimensions space helps the two-layer model to enhance the prediction result and finally achieve a better performance.

C. SMOTE

In most cases, the real-world datasets are imbalanced with regard to the "relevant" examples and the "irrelevant" examples. The imbalanced ratio between different classes may cause machine learning model failing to yield expected prediction, especially when the ratio becomes 1:50 or even 1:100 in binary classification tasks. Thus, several algorithms have been proposed to either down-sampling the majority class [16, 17] or over-sampling the minority class [10, 18].

In our two-layer model, SMOTE is introduced to alleviate the imbalanced ratio between positive interaction data and negative interaction data. SMOTE is an over-sampling approach, which over-sample the minority class by creating "synthetic" examples [10]. The "synthetic" examples give extra training data of the minority class by operating in "feature space", which approves to be a better option than original over-sampling approach with replacement data in "data space".

D. Overall algorithms

Overall, our two-layer model combines XGBoost, SVM and SMOTE algorithm to train the model and generate better prediction results. The complete algorithm is given in Algorithm. 1.

Algorithm 1 Two-Layer Model for Prediction of HP-PPIs

1: Given the dataset $\mathrm{M} = \{x_i, y_i\}$, x_i is the vector of input, $y_i \epsilon \{+1, -1\}$ represents positive and negative interactions;
2: Training XGBoost model with M;
3: Obtain the interaction data N, which could not be classified correctly by XGBoost; the predicted negative interactions are discarded as $\mathrm{O}(Neg)$;
4: If $\mathrm{N(Neg)} < \mathrm{N(Pos)}$, randomly sample λ negative interactions from $\mathrm{O}(Neg)$:
$\lambda = \mathrm{N}(Pos)/2 - \mathrm{N(Neg)}$
5: Balance the dataset via SMOTE algorithm, obtain a subsampling interaction dataset X;
6, Training SVM model with X.

IV. Experiment

In this section, we discussed the experiment protocol and the performance metrics.

A. Experiment Protocol

In light of the imbalanced ratio for HP-PPIs dataset [11], the negative interaction data are as critical as the positive interaction data in building the final dataset. To collect the positive interaction data, a thorough investigation has been done for 11 public archival databases, including the Database of Interacting Proteins (DIP) [16], Reactome [16], the Agile Protein Interaction DataAnalyzer (APID) [18], the Molecular Interaction Database (MINT) [19], the Pathosystems Resource Integration Center (PATRIC) [20] and so on. These databases share a same important character, which is the source of the interaction data is highly trustable by verification of literature or domain experts. We carefully processed the collected data to remove the redundant interactions data and the highly homologous sequence. The goal of this step was to reduce the redundancy of the dataset, so as to reduce the bias in the training models. Once the positive interaction data is collected, we applied the ratios of 1:25, 1:50 and 1:100 on positive interactions data to build the negative interaction data, following the procedure from [11, 12].

It is required to build the training dataset as well the independent test dataset for comparison of models. Briefly, the diagram in Figure 2 illustrates our protocol. We randomly select one-fifth protein interaction data from both positive and negative data to be the independent test dataset. These data are hold till the model is trained and are unseen until the model outputs all the predicted results. The rest of the data will be the training dataset. To avoid the bias causing by random sampling method, the datasets are built five times. All the five built datasets will be used for training and testing by the models and the performance will be compared with the standard and deviation values regarding different performance metrics.

B. Performance Metrics

For an imbalanced dataset, usually accuracy is not sufficient to compare models in a full scale. Especially for an imbalanced dataset with a ratio of 1:100, the accuracy would still be very high and the difference between different models would be negligible in the worst case when giving all predictions to be the majority class. Thus, we further include other performance metrics, including precision, recall, F1-score and Matthew's correlation coefficient (MCC) score. The metrics are listed as following Equa. (2):

$$Precision = \frac{TP}{TP + FP}$$

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

TABLE I. Datasets statistics

Taxonomy ID	Bacterium Pathogens	Total number After Cleansing	Ratio 1:25		Ratio 1:50		Ratio 1:100	
			Training	Independent Testing	Training	Independent Testing	Training	Independent Testing
623	Shigella paradysenteriae	105	2184	546	4284	1071	8484	2121

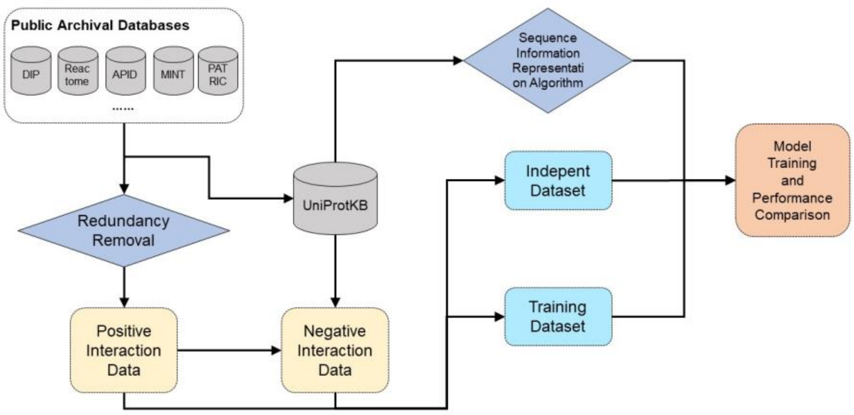

Figure 2 Experiment Dataset Curation

$$F1 = \frac{2 * Precision}{Precision + Recall}$$

$$MCC = \frac{(TP * TN) - (FN * FP)}{\sqrt{(TP + FN) * (TN + FP) * (TP + FP) * (TN + FN)}} \quad (2)$$

V. RESULT AND DISCUSSION

A. Datasets

The experimental HP-PPIs dataset consists of the protein interactions between homo sapiens (taxonomy ID 9606) as host species and Shigella paradysenteriae as the bacterium pathogen (taxonomy ID 623). TABLE I shows the statistics after the data cleansing and negative interaction data building, which results in a total number of 118 for positive interaction data, and a total number of 2184, 4284, 8484 for different ratios of 1:25, 1:50 and 1:100 for negative interaction data.

We applied local descriptor algorithm [15] as sequence information representation algorithm. Local descriptor algorithm considers the protein sequence in regions, which has a capability of keeping regional sequence order information. Ten regions are calculated, including dividing the sequence into four equal regions, diving the sequence into two equal regions, taking the central 50% region, taking the first 75% region, taking the final 75% region and the central 75% region of the sequence. Within these regions, three types of descriptors are calculated, which are Composition (C), Transition (T) and Distribution (D). In details, the local descriptor algorithm applied the diploe and volume classification method to group the 20 basic amino acids into seven groups. This results in 7 composition values, 21 transition values and 35 distribution values for each protein sequence. In a HP-PPI pair, the local descriptor algorithm generates a vector of 1260 features [4] for each HP-PPIs pair.

B. Discussion

In the experiments, the results were collected against five different HP-PPIs datasets randomly sampling for taxonomy ID ‘623’. Both standard and deviation values are recorded in terms of accuracy, precision, recall, F1 and MCC. We firstly briefly describe the methods from the literature, as well as the methods of traditional machine learning models.

Since the study limits the protein information as sequence information to retain a most portion of protein interaction data, [5] and [6] are selected as our most comparable methods from the literature.

[5] applied random forest as their ensemble learning method to train the computational model for host-parasite protein interactions. The protein sequence information was mapped as vectors of amino acid triplets, which also groups amino acids in 7 classes and obtain a total 7*7*7=343 possible amino acid triplets for a protein. These classes were further transferred as the frequency $f_i, i = \{1,2,3, \ldots, 343\}$ by Equa. (3). n_i is the occurrence of amino acid triplets combination in protein and i is a combination over all 343 amino acid triples combinations.

$$f_i = \frac{n_i}{\sum_{l=1}^{343} n_l} \quad (3)$$

In [6], both machine learning model and sequence representation algorithm were different. [6] considered amino acids types based on the biochemical similarity, which turns out to be six classes: {IVLM}, {FYW}, {HKR}, {DE}, {QNTP} and {ACGS}. Totally, there will be 6*6*6=216 possible amino acid triplets. Given each amino acid triplets a frequency $f_i, i = \{1,2, \ldots, 216\}$, the corresponding feature d_i is calculated as below Equa. (4):

$$d_i = \left\{ e^{\frac{f_i - \min\{f_1, f_2, \ldots, f_{216}\}}{\max\{f_1, f_2, \ldots, f_{216}\} - \min\{f_1, f_2, \ldots, f_{216}\}}} \right\} - 1 \quad (4)$$

Here, d_i ranges from 0 to 1.714. The machine learning model selected in [6] is support vector machine with radial basis function (RBF) kernel.

Since a different feature representation algorithm is introduced in this paper, which is local descriptor algorithm, we also test the traditional machine learning model, including support vector machine, random forest, logistic regression, naïve Bayes, gradient boosting machine and decision tree. The hyper parameters are all selected via 5-fold grid searching and the optimal settings are used in the models.

TABLE II. RESULTS OF ACCURACY, PRECISION AND RECALL

Model	**Accuracy**			**Precision**			**Recall**		
	1:25	**1:50**	**1:100**	**1:25**	**1:50**	**1:100**	**1:25**	**1:50**	**1:100**
[5]	0.975092 (0.0897)	*0.981326 (0.0)*	0.991985 (0.0)	1.000000 (0.0)	1.000000 (0.0)	1.000000 (0.0)	0.352381 (0.023328)	0.047619 (0.0)	0.190476 (0.0)
[6]	0.971795 (0.000897)	*0.981326 (0.0)*	*0.992362 (0.000462)*	0.942857 (0.069985)	1.000000 (0.0)	0.847619 (0.106053)	0.285714 (0.0)	0.047619 (0.0)	0.285714 (0.0)
RF	0.970330 (0.001371)	0.981139 (0.000373)	0.991985 (0.000298)	0.695922 (0.045611)	0.900000 (0.200000)	0.960000 (0.080000)	0.419048 (0.087287)	0.047619 (0.000000)	0.200000 (0.019048)
SVM	*0.979121 (0.001465)*	0.980952 (0.000457)	0.992268 (0.000971)	0.907143 (0.068760)	0.800000 (0.244949)	0.877778 (0.173561)	0.514286 (0.019048)	0.047619 (0.000000)	0.266667 (0.038095)
LR	0.971795 (0.002741)	0.980766 (0.000747)	0.991702 (0.000377)	0.805000 (0.074833)	0.766667 (0.290593)	0.860000 (0.127192)	0.352381 (0.048562)	0.047619 (0.000000)	0.200000 (0.019048)
Naïve Bayes	0.677289 (0.015341)	0.694304 (0.017995)	0.680717 (0.009091)	0.094328 (0.004231)	0.044737 (0.002568)	0.026024 (0.000730)	0.857143 (0.000000)	0.714286 (0.000000)	0.857143 (0.000000)
GBM	0.971429 (0.004719)	0.978711 (0.001811)	0.988213 (0.002109)	0.750999 (0.153070)	0.406926 (0.090796)	0.406117 (0.208510)	0.409524 (0.023328)	0.152381 (0.019048)	0.200000 (0.035635)
DT	0.952381 (0.007141)	0.971242 (0.001712)	0.988685 (0.001606)	0.397069 (0.083398)	0.213095 (0.030152)	0.383333 (0.178263)	0.409524 (0.038095)	0.171429 (0.023328)	0.133333 (0.035635)
Ours	**0.980586 (0.002484)**	**0.981513 (0.001089)**	**0.992834 (0.000693)**	0.905505 (0.090918)	0.643333 (0.124544)	0.908333 (0.130171)	0.561905 (0.035635)	0.133333 (0.035635)	0.314286 (0.023328)

TABLE II includes the results of accuracy, precision and recall values. In TABLE II, the accuracy result between different models are very small due to the high imbalanced ratio of the HP-PPIs dataset. The best results from other models for accuracy is 00.979121±0.001465 of ratio 1:25 from SVM, 0.981326±0.0 of ratio 1:50 from [5, 6] and 0.992362±0.000462 of ratio 1:100 from [6]. However, our proposed two-layer model outperforms all of them by 0.980586±0.002484 for ratio 1:25, 0.981513±0.001089 for ratio 1:50, and 0.992834±0.000693 for ratio 1:100. Since the precision and recall values indicate a different ability for the models, and in TABLE II the results of precision and recall values give different trends for the model, we further combine precision and recall as F1-score to validate their performance. Furthermore, the F1-score and MCC results are listed in TABLE III. Both values in italic style are the second best results for each metric in TABLE II and III. All the results are given by the mean values with deviation values in brackets for the five independent tests experiments.

For F1-score and MCC value, the closer the value is to 1.0 indicates the better the trained model is. In TABLE III, the results show that for ratio 1:25 and ratio 1:100, our model achieve F1-score as 0.690496±0.032247 and 0.465441±0.038502 respectively. When the ratio is 1:50, the gradient boosting machine presents a better capability of F1-score as 0.219974±0.030379. For our model, the F1-score of the ratio of 1:50 is 0.219316±0.05392, which is closer. Both these two results are better than the other models. Concerning MCC values, our proposed two-layer model delivers the best results for all three different imbalanced ratios.

Additionally, we collected the time cost for training models and Figure 3 shows the result. Undoubtedly, naïve Bayes model obtains the fastest training speed, while random forests becomes less efficient when the ratio becomes higher. The time costs by our proposed model are gradually increased by the imbalanced ratios. Although our two-layer model is not fastest, we excel in trading off time cost and accuracy considering the accuracy is better than the other models.

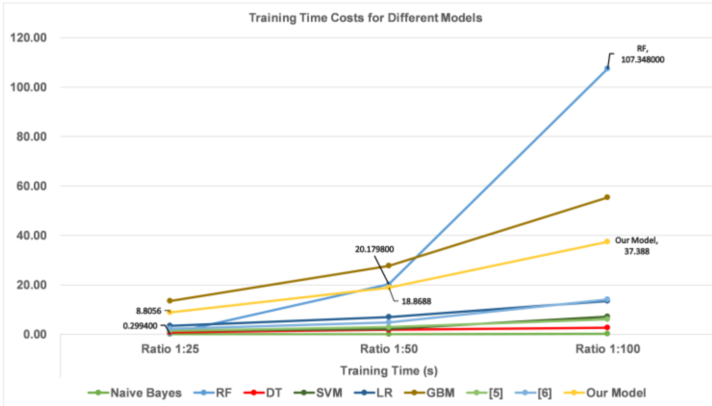

Figure 3 Time Costs Comparison of Different Models

VI. CONCLUSION

In this paper, we studied the ever-challenging HP-PPIs prediction problem, especially we targeted on the imbalanced dataset issue and proposed a two-layer model. A detailed two layered structure leveraging XGBoost model and SMOTE technology to ease the burden of imbalanced dataset and enhancing the model performance by SVM is presented. Results indicated a better performance comparing with other models reported in similar literature and most traditional models. However, the F1-score in TABLE II is still not considered as high enough to generate high-fidelity candidates of HP-PPIs. The future work will be to address the imbalanced datasets by focusing on not only the model aspect but also the feature aspect.

TABLE III. RESULTS OF F1-SCORE AND MCC

Model	**F1-score**			**MCC**		
	1:25	**1:50**	**1:100**	**1:25**	**1:50**	**1:100**
[5]	0.520690 (0.025340)	0.090909 (0.0)	0.32 (0.0)	0.585766 (0.019551)	0.216169 (0.0)	0.434680 (0.0)
[6]	0.438095 (0.007776)	0.090909 (0.0)	0.426032 (0.014395)	0.510286 (0.020527)	0.216169 (0.0)	0.488573 (0.031846)
RF	0.515472 (0.057686)	0.090119 (0.001581)	0.330462 (0.027493)	0.522309 0.044992	0.202909 0.026521	0.435610 0.031072
SVM	*0.654747 (0.015513)*	0.089328 (0.001936)	0.406553 (0.058855)	*0.673402 0.021537*	0.189648 0.032481	0.479377 0.074487
LR	0.488988 (0.056970)	0.088603 (0.003047)	0.322872 (0.025335)	0.520864 0.056991	0.183660 0.040926	0.410708 0.034654
Naïve Bayes	0.169922 (0.006850)	0.084184 (0.004546)	0.050514 (0.001375)	0.212762 0.008699	0.122152 0.007389	0.113248 0.002763
GBM	0.527137 (0.054758)	**0.219974 (0.030379)**	0.255816 (0.052325)	0.540288 0.073108	0.238834 0.039974	0.271733 0.075399
DT	0.401192 (0.057632)	0.189447 (0.024179)	0.190145 (0.046456)	0.377599 0.061403	0.176362 0.024848	0.215932 0.066885
Ours	**0.690496 (0.032247)**	*0.219316 (0.05392)*	**0.465441 (0.038502)**	**0.703311 0.038889**	**0.285516 0.061405**	**0.530828 0.052146**

ACKNOWLEDGMENT

This work is supported by the scholarship from the China Scholarship Council (CSC), Faculty Strategic Investments Grant for DP 2019 development and The University Global partnership Network (UGPN) Research Collaboration Fund, while the first author pursues his PhD degree in the University of Wollongong.

References

[1] S. Blasche, S. Arens, A. Ceol, G. Siszler, M. A. Schmidt, R. Häuser, F. Schwarz and e. al, "The EHEC-host interactome reveals novel targets for the translocated intimin receptor," Scientific Reports, vol. 4, no. 1, pp. 22-26, 2014.

[2] H. Chen, W. Guo, J. Shen, L. Wang and J. Song, "Structural Principles Analysis of Host-Pathogen Protein-Protein Interactions : A Structural Bioinformatics Survey," IEEE Access, vol. 6, pp. 11760-11771, 2018.

[3] E. Guven-Maiorov, C. J. Tsai, B. Ma and R. Nussinov, "Prediction of Host–Pathogen Interactions for Helicobacter pylori by Interface Mimicry and Implications to Gastric Cancer," Journal of Molecular Biology, vol. 429, no. 24, pp. 3925-3941, 2017.

[4] H. Chen, J. Shen, L. Wang and J. Song, "Towards data analytics of pathogen-host protein-protein interaction: a survey," in 2016 IEEE International Congress on Big Data (BigData Congress), San Francisco, USA, 2016.

[5] S. Wuchty, "Computational prediction of Host-Parasite protein interactions between P. falciparum and H. sapiens," PLoS ONE, vol. 6, no. 11, pp. e26960: 1-8, 2011.

[6] G. Cui, C. Fang and K. Han, "Prediction of protein-protein interactions between viruses and human by an SVM model," BMC Bioinformatics, vol. 13, no. Suppl 7, p. S5, 2012.

[7] T. Chen and G. Carlos, "Xgboost: A scalable tree boosting system," in the 22nd acm sigkdd international conference on knowledge discovery and data mining, San Francisco, USA, pp. 785-794. ACM, 2016.

[8] C.-C. Chang and L. Chih-Jen, "LIBSVM: A library for support vector machines," ACM transactions on intelligent systems and technology (TIST), vol. 2, no. 3, p. 27, 2011.

[9] H. Drucker, D. Wu and V. N. Vapnik, "Support vector machines for spam categorization," IEEE Transactions on Neural networks, vol. 10, no. 5, pp. 1048-1054., 1999.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hal and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of artificial intelligence research, vol. 16, pp. 321-357, 2002.

[11] M. D. Dyer, T. M. 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.

[12] M. Kshirsagar, J. Carbonell and J. Klein-Seetharaman, "Multitask learning for host-pathogen protein interactions," Bioinformatics, vol. 29, no. 13, pp. 217-226, 2013.

[13] T. Huo, W. Liu, Y. Guo and e. al., "Prediction of host-pathogen protein interactions between Mycobacterium tuberculosis and Homo sapiens using sequence motifs," BMC Bioinformatics, vol. 16, no. 1, p. 100, 2015.

[14] T. U. Consortium, "UniProt: a worldwide hub of protein knowledge," Nucleic Acids Res., vol. 47, no. D1, pp. D506-D515, 2019.

[15] M. N. Davies, A. Secker, A. A. Freitas and e. al, "Optimizing amino acid groupings for GPCR classification," Bioinformatics, vol. 24, no. 18, pp. 1980-1986, 2008.

[16] J. Laurikkala, "Improving identification of difficult small classes by balancing class distribution," in Proceedings of the 8th Conference on Artificial Intelligence in Medicine in Europe, Cascais, Portugal, 2001.

[17] I. a. Z. I. Mani, "kNN approach to unbalanced data distributions: a case study involving information extraction," in Proceedings of workshop on learning from imbalanced datasets, Washington DC, USA, 2003.

[18] H. a. B. Y. a. G. E. A. a. L. S. He, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 2008.

[19] I. Xenarios, L. Salwínski, X. J. Duan, P. Higney, S.-M. Kim and D. Eisenberg, "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.

[20] G. Joshi-Tope, M. Gillespie, I. Vastrik, D&apos, P. Eustachio, E. Schmidt and e. al., "Reactome: A knowledgebase of biological pathways," Nucleic Acids Research, vol. 33, no. Database issue, pp. 428-432, 2005.

[21] C. Prieto and J. De Las Rivas, "APID: Agile protein interaction DataAnalyzer," Nucleic Acids Research, vol. 34, no. Web Server issue, pp. 298-302, 2006.

[22] L. Licata, L. Briganti, D. Peluso and e. al., "MINT, the molecular interaction database: 2012 Update," Nucleic Acids Research, vol. 40, no. Database issue, pp. D857-D861, 2012.

[23] A. R. Wattam, D. Abraham, O. Dalay and e. al, "PATRIC, the bacterial bioinformatics database and analysis resource," Nucleic Acids Research, vol. 42, no. Database issue, pp. D581-D591, 2014.