All Sciences Proceedings

5th International Conference on Applied Engineering and Natural Sciences

July 10-12, 2023 : Konya, Turkey

All Sciences Proceedings
http://as-proceeding.com/

https://www.icaens.com/

Improving the Prediction Performance of Protein Protein Interaction Sites Using Xgboost and Optimization

Yunus Emre GÖKTEPE

Department of Computer Engineering, Fac. Eng. Arch. Necmettin Erbakan University, Konya, Turkey

ygoktepe@erbakan.edu.tr

Abstract – Since proteins have important tasks in all processes within the cell they are vital elements for all living organisms. They are significant in regulating most of the biological processes which occur in a cell. They are widely researched to comprehend roles of them and to assist drug design studies. In these tasks, they usually work by interacting with other proteins, not alone. Thus, predicting protein-protein interactions and protein protein interaction sites is an important problem in bioinformatics. There are a number of computational methods developed for this prediction task. DeepPPISP-XGB is one of these methods and produces promising predicting results. In this study, we proposed an optimization process in order to improve prediction results of this method. This optimization process produced 1.5% better AUROC value and 3.3% better AUPRC value compared to DeepPPISP-XGB method.

Keywords – Protein-Protein Interaction Sites; Extreme Gradient Boosting; Hyperparameter Optimization

I. INTRODUCTION

Proteins are organic substances with complex structures found in all living organisms. It is known that proteins are definitely the main elements in a cell. They have crucial roles in the fulfillment of vital activities [1]. Proteins interact intensely with other proteins in order to perform these activities. Protein-protein interactions (PPIs) are fundamental for most of the biological processes. Thus, information about interactions between proteins is important for various research fields including figuring out biological operations within any cell. This information is also crucial in order to help researches which tries to produce new drugs or vaccines [2].

There are some databases growing over time that contain empirically verified or computationally predicted interactions [3]. Although these databases differ in scope and content, most commonly used databases are HPRD (Human Protein Reference Database), (BioGRID (Biological General Repository for Interaction Datasets), MINT (Molecular INTeraction database), PDB (Protein Data Bank), DIP (Database of Interacting Proteins), BIND (Biomolecular Interaction Network Database) and IntAct (IntAct molecular interaction database) which differ in scope.

It is accepted that the surfaces on which proteins communicate with each other are essential in the formation of PPIs. These surface residues are called protein-protein interaction surfaces (PPISs) and they are considered as precursors in predicting PPIs [4].

Although there are a number of wet-lab methods employed to study interactions between proteins, it is indicated that they are time-consuming and expensive [5]. It is also mentioned that these methods yield false positives and false negatives in excess of acceptable levels. [6].

Computational approaches have now become subsidiary applications to wet-lab studies, thanks to

their increased performance. Various machine learning and pattern recognition methods were performed using interaction data to reveal whether proteins interact. For this purpose, various machine learning algorithms based on support vector machines [2, 11, 18], artificial neural networks [12] and deep learning, XGBoost [13, 14] have been developed. In literature, there are a number of computational approaches which aim to predict PPIs. These approaches need different information about proteins such as genome or structure. Some of them uses information obtained from sequence of proteins in order to make protein-protein interaction predictions [7-10]. Singh et al. [12] evaluated that combining sequence and structural features raises the performance of protein-protein interaction site prediction models. They used both the local and global features of proteins. Zeng et al. [16] proposed a method called DeepPPISP which is a deep learning based model. In this model, both local contextual and global sequence features are combined in order to make predictions of PPIS.

The model proposed by [16] is enhanced by [13] using XGBoost (eXtreme Gradient Boosting) algorithm. In [13], the researchers proposed a method called DeepPPISP-XGB in order to build a classifier for PPIS prediction. In this work, based on these researches [12, 13, 16], we endeavored to increase the performance of prediction models. Parameter optimization techniques are evaluated to increase the accuracy of the model. Grid search technique was used to find optimum values of hyperparameters. Our findings may help to increase prediction capability of proposed models.

II. Materials and Method

A. *Datasets*

The amount of data available on proteins and their interactions is constantly increasing over time. Medline is an important database in which studies on this subject are kept (Medline, 2023). More than 1.3 million new citations were added to it in year 2022. The increase of number of citations over years 2021 and 2022 can be seen in Table 1.

Table 1. Number of PUBMED production statistics (Medline, 2023)

	FY2022	FY2021
MEDLINE Citations Indexed (Annual)	1,369,611	1,291,807
MEDLINE Citations Cumulative Total	29,807,639	28,444,654
MEDLINE Journal Titles	5,282	5,282
PubMed Citations (Annual)	1,714,780	1,733,089
PubMed Citations Cumulative Total	34,693,538	33,136,289
PubMed Searches	2.58 Billion	2.57 Billion
Web/Interactive	1.283 Billion	1.186 Billion
Script/E-Utilities	1.303 Billion	1.391 Billion

Similar to referred studies, we used the same three dataset Dset_72, Dset_186 and Dset_164 which are exploited widespread in the literature [12]. The number of protein sequences exist in these datasets are 72, 186 and 164, respectively. These protein sequences are obtained from PDB (Protein Data Bank) with selecting sequences with less than %25 homology identity and less than 3.0 Å resolution. Interaction sites involved in these datasets are 5517, 1923 and 6069, respectively [14].

B. *DeepPPISP*

DeepPPISP is a machine learning approach which is recently proposed by researchers [16]. This approach uses both local and global features for protein-protein interaction site prediction using a deep learning based model. A sliding window was used to get local features of the sequences. Raw protein sequences, PSSM (Position Specific Scoring Matrix) feature extractor and a convolutional neural network model was run in order to reveal global features. Then the local and global features are concatenated before the classification step.

C. *XGBoost*

XGBoost (eXtreme Gradient Boosting) is a decision tree-based algorithm. It is proposed by Chen & Guestrin [15]. It is a high-performance version of the Gradient Boosting algorithm optimized by performing various arrangements. XGBoost algorithm offers high prediction accuracy and can manage the overfitting problem very well [15].

Wang et al. [13], have proposed a new model called DeepPPISP-XGB to predict sites of proteins in which interaction between them occurs. They implemented eXtreme Gradient Boosting method upon the model called DeepPPISP proposed by Zeng et al. [16]. With applying XGBoost, the proposed classifier [13] was acquired promising results. As the researchers pointed out, the results of 0.68 and 0.34 were achieved for the AUROC and AUPRC evaluation terms respectively.

XGBoost parameters are considered by researchers in three basic groups. These groups are general, booster and learning task parameters. Learning rate (eta), gamma, alpha and lambda are among the most important boosting parameters of the XGBoost method.

D. Selecting Optimum Parameters

In machine learning, parameter optimization method is applied to ensure that the learning algorithms give the best possible result. For this purpose, the best values are sought for a set of hyperparameters. A hyperparameter can be considered as a value that affects the success of the learning model.

In this study, we used "Grid Search" technique which is one of the most used parameter tuning method. With this method, parameter scanning is performed on a manually determined subset of the hyperparameter space, guided by a performance criterion such as cross-validation.

In order to perform parameter optimization using the grid search method, we used the following hyperparameter set, which includes several important parameters of the XGBoost model.

```
params={"reg_lambda": [0.35, 0.42, 0.49, 0.56, 0.63],
"min_split_loss": [0, 0.15, 0.25, 0.35, 0.45],
"reg_alpha": [0, 0.1, 0.2, 0.3, 0.4],
"learning_rate": [0.05, 0.06, 0.07, 0.8]
}
```

With the grid search optimization process, values of eta, gamma, alpha and lambda parameters were determined as 0.07, 0.15, 0 and 0.49, respectively.

E. Evaluation

Since we aimed to increase the performance values of a previously proposed method, the evaluation criteria used in that method were preferred. Thus, AUROC (the area under the receiver operating characteristic curve) and AUPRC (area under the precision-recall curve) were used for the evaluation process. AUROC and AUPRC curves are widely preferred performance metrics for classification problems.

ROC is a graph where the x-axis shows the false positive (FPR) rate and the y-axis shows the true positive (TPR) rate (recall or sensitivity), with the area under the curve forming the AUROC value. PRC is a graph where the x-axis shows the recall rate and the y-axis shows the precision ratio, the area under the curve creates the AUPRC value. The following formulas (1) are used to calculate above metrics:

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

$$FPR = \frac{FP}{TN + FP} \qquad (1)$$

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

III. RESULTS

Figure 1 depicts the ROC curves of XGBoost model four other methods (Random forest, ExtraTrees, DecisionTreeClassifier, and SVM) obtained by Wang et al. [13]. Similarly, Figure 2 shows the precision-recall curves of XGBoost model four other methods (Random forest, ExtraTrees, DecisionTreeClassifier, and SVM) obtained by Wang et al. [13]. The researchers showed in these graphs that the XGBoost model achieved higher values than other compared methods.

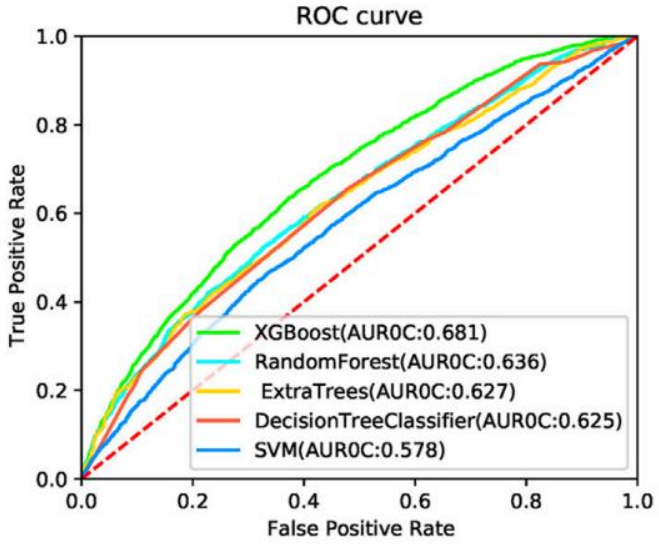

Figure 1. The ROC curve of XGBoost model compared to four methods (Random forest, Extra Trees, Decision Tree Classifier, and SVM) [13].

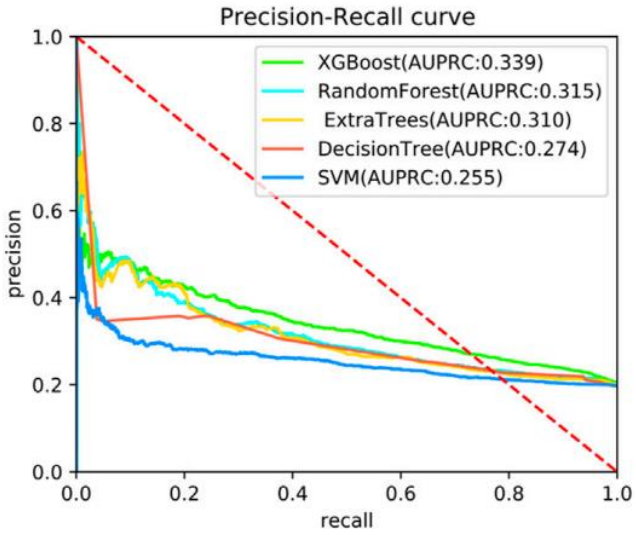

Figure 2. The precision-recall curve of XGBoost model compared to four methods (Random forest, Extra Trees, Decision Tree Classifier, and SVM) [13].

As shown in Figure 3, XGBoost model has obtained the best performance with a result of 0.681 for AUROC evaluation metric. Based on the AUPRC metric, as seen in Figure 4, the XGBoost model gave the best result with a value of 0.339. These values have been slightly improved with the optimization process we have made. As given in Figure 3, the AUROC value was increased from 0.681 to 0.691. As seen in Figure 4, the 0.339 AUPRC value produced by the XGBoost model has been increased to 0.35.

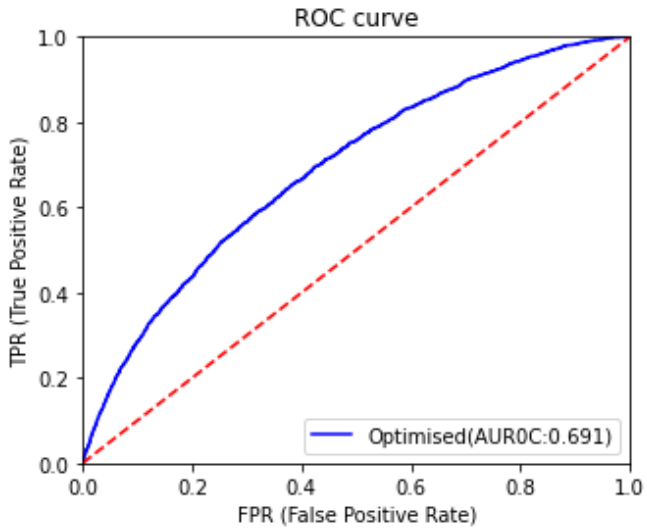

Figure 3. The ROC curve of XGBoost model with optimized parameters.

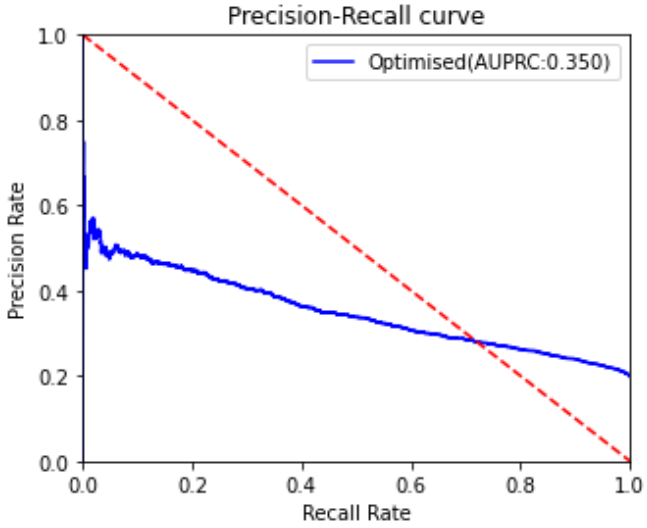

Figure 4. The precision-recall curve of XGBoost model with optimized parameters.

IV. CONCLUSION

XGBoost (eXtreme Gradient Boosting) is proposed by Chen & Guestrin [15]. It is a high-performance version of the Gradient Boosting algorithm and offers high prediction accuracy. A new model called DeepPPISP-XGB was proposed by Wang et al. [13].

The optimization studies we have carried out furthered the results of this model. AUROC and AUPRC values of the model were improved as shown in the results section. These results showed that the AUROC value was improved by 1.5%. The improvement rate in the AUPRC value was determined as 3.3%. These results enable a model in

which interaction sites in protein-protein interactions can be predicted more successfully.

REFERENCES

[1] Li, M., Gao, H., Wang, J., & Wu, F. X. (2019). Control principles for complex biological networks. Briefings in bioinformatics, 20(6), 2253-2266.

[2] Göktepe, Y. E., & Kodaz, H. (2018). Prediction of protein-protein interactions using an effective sequence based combined method. Neurocomputing, 303, 68-74.

[3] Lehne B, Schlitt T. Protein-protein interaction databases: keeping up with growing interactomes. Hum Genomics. 2009 Apr;3(3):291-7. doi: 10.1186/1479-7364-3-3-291. PMID: 19403463; PMCID: PMC3500230.

[4] Wang, X., Yu, B., Ma, A., Chen, C., Liu, B., & Ma, Q. (2019). Protein–protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique. Bioinformatics, 35(14), 2395-2402.

[5] Aumentado-Armstrong, T. T., Istrate, B., & Murgita, R. A. (2015). Algorithmic approaches to protein-protein interaction site prediction. Algorithms for Molecular Biology, 10(1), 1-21.

[6] Ho, Y. et al, 2002. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry, Nature, 415, pp 180-183.

[7] Pan, X.Y. et al, 2010. Large scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features, Journal of Proteome Research.

[8] Nanni, L., 2006. Comparison among feature extraction methods for HIV-1 Protease Cleavage Site Prediction, Pattern Recognition. 39, pp 711–713.

[9] Zhang J, Kurgan L. SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences. Bioinformatics. 2019 Jul 15;35(14):i343-i353. doi: 10.1093/bioinformatics/btz324. PMID: 31510679; PMCID: PMC6612887.

[10] Wang, Y., Mei, C., Zhou, Y., Wang, Y., Zheng, C., Zhen, X., ... & Wang, B. (2019). Semi-supervised prediction of protein interaction sites from unlabeled sample information. BMC bioinformatics, 20(25), 1-10.

[11] Nanni, L. et al, 2010. High performance set of PseAAC and sequence based descriptors for protein classification, Journal of Theoretical Biology. 266, pp 1-10.

[12] Singh, G., Dhole, K., Pai, P. P., & Mondal, S. (2014). SPRINGS: prediction of protein-protein interaction sites using artificial neural networks (No. e266v2). PeerJ PrePrints.

[13] Wang, P., Zhang, G., Yu, Z. G., & Huang, G. (2021). A Deep Learning and XGBoost-Based Method for Predicting Protein-Protein Interaction Sites. Frontiers in genetics, 12.

[14] Deng, A., Zhang, H., Wang, W., Zhang, J., Fan, D., Chen, P., & Wang, B. (2020). Developing computational model to predict protein-protein interaction sites based on the XGBoost algorithm. International Journal of Molecular Sciences, 21(7), 2274.

[15] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[16] Zeng, M., Zhang, F., Wu, F. X., Li, Y., Wang, J., & Li, M. (2020). Protein–protein interaction site prediction through combining local and global features with deep neural networks. Bioinformatics, 36(4), 1114-1120.

[17] MEDLINE, 2023. MEDLINE PubMed Production Statistics https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html, 22.06.2023..

[18] Göktepe, Y. E., İlhan, İ., & Kahramanlı, Ş. (2016). Predicting protein-protein interactions by weighted pseudo amino acid composition. International Journal of Data Mining and Bioinformatics, 15(3), 272-290.