PPSNV:A novel predictor for pathogenicity of SNV based on ensemble learning

ZHEN XUYANG

Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200030, China

Shanghai Key Laboratory of Psychotic Disorders, Shanghai 200030, China

E-mail: [zhenxuyang@sjtu.edu.cn](mailto:zhenxuyang@sjtu.edu.cn)

LIN GUANNING[[1]](#footnote-1)\*

Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200030, China

Shanghai Key Laboratory of Psychotic Disorders, Shanghai 200030, China

E-mail: [nickgnlin@sjtu.edu.cn](mailto:nickgnlin@sjtu.edu.cn)

With the next-generation sequencing(NGS) technologies developing, numerous genetic data emerges. Nonsynonymous single nucleotide variant(SNV) is a common type of genetic variant which possibly leads to diseases. However, classifying SNVs to benign of pathogenic variants with high confidence remains challenging. Inspired by ensemble learning and a machine learning algorithm, Gradient Boosting Decision Tree(GBDT), we proposed a novel prediction model named as PPSNV to identity the pathogenicity of SNVs. We integrated 14 features to train our model and tested it in two independent datasets. The results showed best performance was achieved by the proposed model compared with four commonly used predictive tools.

**CCS CONCEPTS** • Computing methodologies • Machine learning • Classification and regression trees • Ensemble methods • Boosting •

**Keywords:** Single Nucleotide Variant, Pathogenicity prediction, Machine Learning, Gradient Boosting Decision Tree, Ensemble Learning.

1. Introduction

With the next-generation sequencing(NGS) technologies developing, situations become common that genetic variant detection appers as a promising tool in clinical diagnosis. Providing useful insight into individual heredity and human diversity, genetic testing is vital for personalized medicine and targeted therapy. Among numerous output of genetic testing and sequencing, nonsynonymous single nucleotide variant(SNV) is a common type of genetic variant which probably results in disease, which means aligning SNVs with phenotypes and diseases accurately are necessary. However, identifying the pathogenicity of SNVs by experiment is costly and infeasible. Calling for computational methods to predict pathogenicity of SNVs is always urgent but real consequence of genetic variants in vivo veils behind complicated biological mechanisms through many pathways. Hench, discriminating SNVs between benign and pathogenic variants accurately remains challenging.

Inspired by different roadmaps of genetic central dogma, many tools are developed to predict the pathogenicity of genetic variants. Some tools utilize the conservation of amino acid or nucleotide to assess the damage when the reference allele is substituted, and some tools focus on the alteration of molecular mechanisms, such as the structure or functional change of protein. Trained with different algorithms and individually collected variant data, different tools may provide different results for the same genetic variant. However, most tools can still be considered as handy references for the assessment of genetic variants since extra information of SNVs that may represent deep features in machine learning are refined during the computation and prediction of developed tools in silicon. These refined features of SNVs would be valuable if utilized properly.

As a strategy of machine learning, ensemble learning is proved to acquire advantages from its base learners. Through the combination of several existing predictors, better prediction results are obtained without additional features. Inspired by it and combining several existing predictors, we propose PPSNV, a novel predictor for pathogenicity of SNV based on ensemble learning in this study. We implemented PPSNV by Gradient Boosting Decision Tree(GBDT), a machine learning algorithm that perform well in classification and regression task. With fine parameter tuning, we finished our prediction tool and tested it with two independent datasets. We compare PPSNV with four ensemble predictors and it turned out that PPSNV outperformed all compared predictors overall.

1. METHODS

We used

* 1. Gradient Boosting Decision Tree(GBDT)

Utilizing a boosting strategy, Gradient Boosting Decision Tree(GBDT) is a machine learning method which achieves state-of-the-art performances among other machine learning methods in many tasks of classification and regression. Previous research have used GBDT to perform prediction in biological problems and achieved best performance. Given the superior classification performance of GBDT, we built our pathogenicity predictors based on GBDT. Especially, we used LightGBM, a novel implementation of GBDT which succeeds in dealing with large number of data instances and sparse features, to implement our model. Friendly to categorical features and numerical features of SNVs data we collected, LightGBM is suitable for the prediction of SNV pathogenicity.

* 1. Datasets

We used the dataset provided by a previous work, consisting of 5 subsets(*HumVar, VariBenchSelected, predictSNPSelected, ExoVar and SwissVarSelected*) and 88,332 variants in total which were collected from publicly available and commonly used benchmark datasets. For each SNV in this dataset, we extracted four primary properties(Chromosome, Location, Reference Allele and Alteration Allele) for following process. After removing overlapping variants between five subsets, *HumVar, VariBenchSelected, predictSNPSelected* were merged into a new dataset which we prepared to train our predictors. The merged training dataset, consisting of 66577 SNVs, was split into a training set and a validation set in an 7:3 ratio.

As for *left two datasets, ExoVar and SwissVarSelected,* we used them as two independent test datasets based on which we compared our proposed predictors with others. Annovar was used to annotate all variants in our datasets, by which we obtained input features for our model and output results of the compared predictors. Especially, we removed variants any comparator’s output of which was missing from the test datasets.

The details of our datasets are listed in Table 1.

Table 1. The Datasets used in this study

| Dataset | Benign | Pathogenic | Total |
| --- | --- | --- | --- |
| Training | 21,740 | 24,863 | 46,603 |
| Vaildation | 9,455 | 10,519 | 19,974 |
| ExoVar | 568 | 4,984 | 5,552 |
| SwissVarSelected | 5,004 | 4,356 | 9,360 |

* 1. Training

The classification performance of GBDT depends on how the decision trees are constructed during training. We used GridSearch provided by Scikit-learn to tune the parameters. The core parameters, such as learning rate, max depth of the decision trees and number of stacked estimators, were optimized to achieve best performance. We trained our model by maximizing the AUC score for binary pathogenicity classification in both testing dataset and validation dataset. The initial outcome of a SNV from our predictor is a probability score ranging between 0 and 1 which indicates pathogenicity of the SNV. Higher the score is, more pathogenic the SNV is. Besides, we used the threshold of 0.5 as a cut-off to classify a SNV into a benign or pathogenic variant. if the score is larger than 0.5, the SNV is identified as pathogenic otherwise benign.

* 1. Features

We included 14 features in total to train PPSNV. One part of our features is four primary features of SNV(Chromosome, Location, Reference amino acid and Alternate amino acid) and the other part is ten pathogenicity predictions scores(SIFT, Polyphen2\_HDIV, Polyphen2\_HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, RadialSVM, VEST3, CADD) obtained from nine existing tools. Missing features are not imputed because LightGBM enables the missing value handle. Actually, permitting missing value in the input features is friendly for practical application considering that the prediction tools we integrated might not return valid predictive result for SNVs in some circumstances.

Table 2. The Features used in this study

| Feature | Description |
| --- | --- |
| Chr | chromosome |
| Loc | mutational site in chromosome |
| A\_Ref | reference amino acid |
| A\_Alt | alternate amino acid |
| SIFT\_score | SIFT score |
| Polyphen2\_HDIV\_score | HDIV score of Polyphen2 |
| Polyphen2\_HVAR\_score | HVAR score of Polyphen2 |
| LRT\_score | score of LRT |
| MutationTaster\_score | score of MutationTaster |
| MutationAssessor\_score | score of MutationAssessor |
| FATHMM\_score | score of FATHMM |
| RadialSVM\_score | score of RadialSVM |
| VEST3\_score | Score of VEST3 |
| CADD\_phred | Score of CADD |

* 1. Testing

We prepared two independent test datasets to evaluate the performance of our predictor and then compared it with other predictors. In order to avoid potential basis in comparation, we assured that test datasets did not overlap with training dataset. We used trained model to predict each SNVs in test datasets and acquired pathogenic score of each SNVs. Given that several compared tools returned pathogenic score that were out of range between 0 to 1, we normalized its output to [0, 1].

PPSNV are compared with four predictions tools that are commonly used and readily implemented：M-CAP, REVEL, MVP, DANN which are all prediction tools based on ensemble learning. The predictive scores for test data of these tools are annotated by Annovar. Each SNVs in our test datasets obtain all score of comparators without missing value. Several compared tools return pathogenicity scores instead of binary classification results. The optimal thresholds recommended to discriminate SNVs pathogenic or benign are not explicit, which may be misleading if we used binary metrics such as precision or recall to evaluate classification performance. Hence, we apply receiver operating characteristic(ROC) curves, the area under the ROC curve(AUC) and precision-recall (PR) curve to assess the overall performance of five predictors including the proposed.

1. RESULT
   1. Characterization of Features

PPSNV ensemble the pathogenicity predictions of ten individual tools as integrated features along with four primary features of SNV. Importance of features in PPSNP represents individual feature’s information gain during the classification. With higher importance in the algorithm, features contribute more to the final prediction result. We normalized the importance of features(Figure 1) to show the contrast between features. The top five important features in our model are VEST3\_score, FATHMM\_score, Location, MutationAssessor\_score and RadialSVM\_score. Compared with primary but basic features, features from ensemble tools earned much higher importance, which indicates that the prediction scores of existing tools represents more intrinsic information related with underlying biological mechanisms.

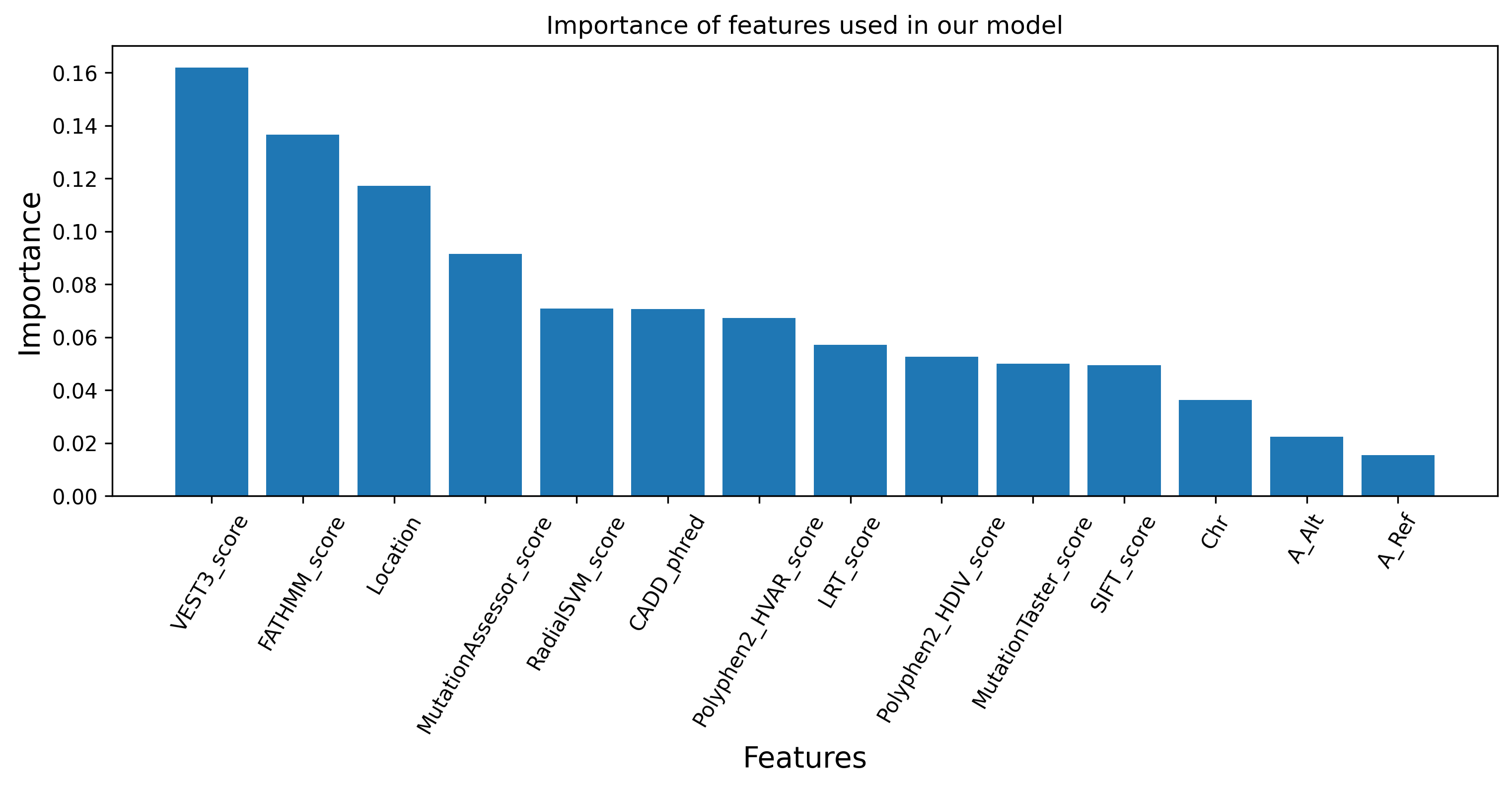


Figure 1. Importance of features used in the proposed prediction model, the top five features are VEST3\_score, FATHMM\_score, Location, MutationAssessor\_score and RadialSVM\_score

* 1. Performance of PPSNV compared with other methods

PPSNV performed best in the ExoVar dataset with a highest AUC score of 0.979 among comparators, which means PPSNV could identify the pathogenicity of SNVs in ExoVar accurately. Three comparators, M-CAP, REVEL, MVP achieved slightly worse performance of which the best one is REVEL with an AUC of 0.963. In the SwissVarSelected dataset, PPSNV outperformed M-CAP, REVEL and DANN with an AUC of 0.744 but failed to surpass REVEL with an AUC of 0.757. Considering that REVEL ensembled 18 individual scores of other tools as features, PPSNV with less ensembled features, may gain improvement of performance through more optimization in future work.

All included tools performed significantly worse in SwissVarSelected in contrast to ExoVar. The gap between different performance in ExoVar and SiwssVarSelected may be due to the fact that SwissVarSelected contains the newest variants compared with other datasets. Furthermore，deterioration of tool’s predictive function when dealing with novel variants reveals lack of generalization of existing tools.

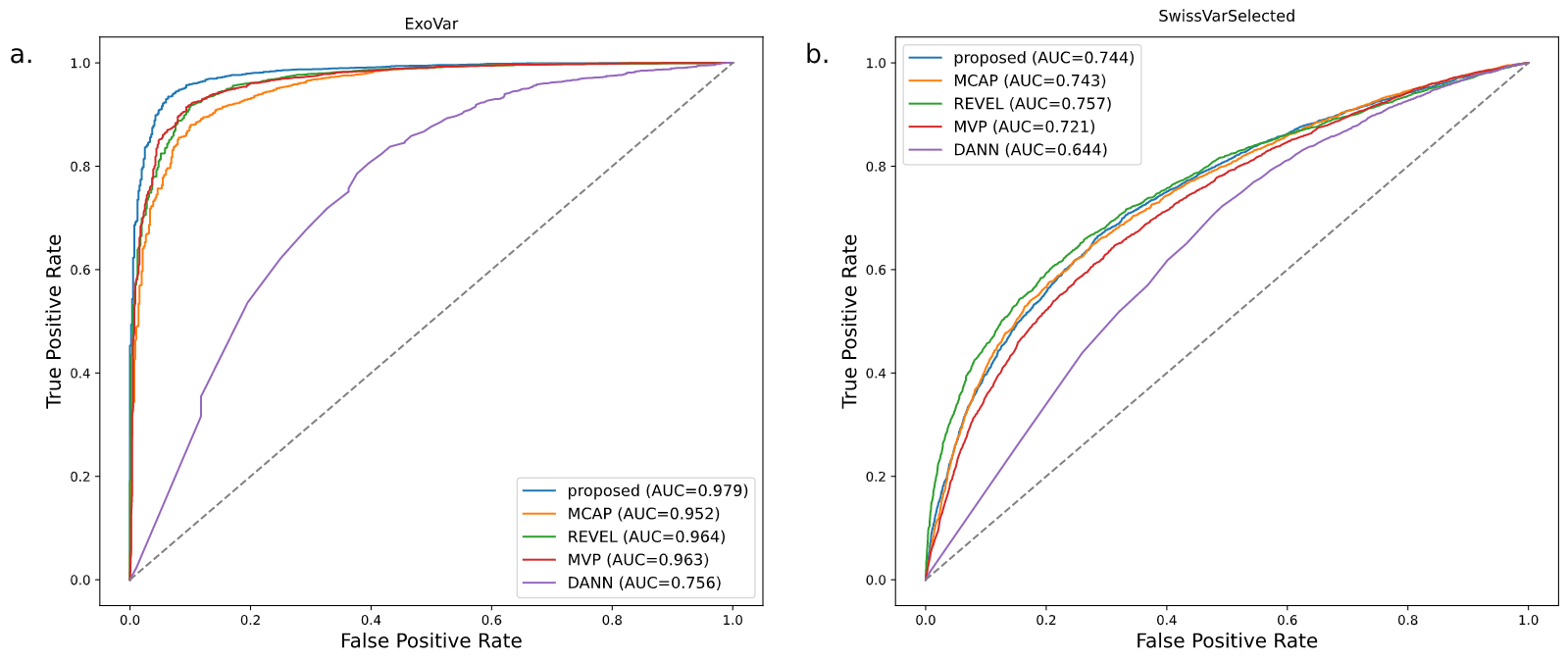


Figure 2. Performance of PPSNV and compared predictors. (a) Performance of predictors tested by ExoVar, PPSNP performs best. (b) Performance of predictors tested by SwissSelected, PPSNP performs slightly worse that REVEL

1. DISCUSSION

With numerous genome sequencing data, how to map genetic variants with phenotypes and disease is concerning. Concurrently, the development of personalized medicine demands accurate interpretation for genetic testing. Aiming to predict the pathogenicity of genetic variants in silico, there are many computational methods proposed based on which ensemble learning may achieve better performance. Combining existing prediction result as input features, many ensembled pathogenicity prediction tools are developed and widely applied.

In this work, we proposed PPSNV, a novel predictor for pathogenicity of SNV based on ensemble learning. We have tested the performance of our predictor and demonstrated that the proposed model outperform M-CAP, MVP and DANN in both test datasets and REVEL in one test dataset. Even though REVEL perform slightly better than PPSNV in SwissvarSelected dataset, it integrated more features which indicates there is space for PPSNV, the relatively light-weight one to improve.

Although our proposed tool achieves promising performance in pathogenicity prediction of SNVs, there are several obvious limitations of PPSNV. PPSNV focus on nonsynonymous single nucleotide variants which are small proportion of genetic variants, which means PPSNV may not be suitable for synonymous SNV prediction. Even source code and trained model of PPSNV are provided publicly, it is worthy to optimize and release PPSNV as a online tool which remains future work.

ACKNOWLEDGMENTS

This work was supported by grants from the National Natural Science Foundation of China (nos. 81971292), the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning (no. 1610000043), and the Innovation Research Plan supported by Shanghai Municipal Education Commission (ZXWF082101).

REFERENCES

1. Atul Adya, Paramvir Bahl, Jitendra Padhye, Alec Wolman, and Lidong Zhou. 2004. A multi-radio unification protocol for IEEE 802.11 wireless networks. In Proceedings of the IEEE 1st International Conference on Broadnets Networks (BroadNets’04) . IEEE, Los Alamitos, CA, 210–217. https://doi.org/10.1109/BROADNETS.2004.8
2. Sam Anzaroot and Andrew McCallum. 2013. UMass Citation Field Extraction Dataset. Retrieved May 27, 2019 from <http://www.iesl.cs.umass.edu/data/data-umasscitationfield>
3. Martin A. Fischler and Robert C. Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24, 6 (June 1981), 381–395. https://doi.org/10.1145/358669.358692
4. Chelsea Finn. 2018. Learning to Learn with Gradients. PhD Thesis, EECS Department, University of Berkeley.
5. Jon M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. J. ACM 46, 5 (September 1999), 604–632. https://doi.org/10.1145/324133.324140
6. Matthew Van Gundy, Davide Balzarotti, and Giovanni Vigna. 2007. Catch me, if you can: Evading network signatures with web-based polymorphic worms. In Proceedings of the first USENIX workshop on Offensive Technologies (WOOT ’07) . USENIX Association, Berkley, CA, Article 7, 9 pages.
7. James W. Demmel, Yozo Hida, William Kahan, Xiaoye S. Li, Soni Mukherjee, and Jason Riedy. 2005. Error Bounds from Extra Precise Iterative Refinement. Technical Report No. UCB/CSD-04-1344. University of California, Berkeley.
8. David Harel. 1979. First-Order Dynamic Logic. Lecture Notes in Computer Science, Vol. 68. Springer-Verlag, New York, NY. <https://doi.org/10.1007/3-540-09237-4>
9. Jason Jerald. 2015. The VR Book: Human-Centered Design for Virtual Reality. Association for Computing Machinery and Morgan & Claypool.
10. Prokop, Emily. 2018. The Story Behind. Mango Publishing Group. Florida, USA.
11. R Core Team. 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
12. Brian K. Reid. 1980. A high-level approach to computer document formatting. In Proceedings of the 7th Annual Symposium on Principles of Programming Languages. ACM, New York, 24–31. <https://doi.org/10.1145/567446.567449>
13. John R. Smith and Shih-Fu Chang. 1997. VisualSEEk: a fully automated content-based image query system. In Proceedings of the fourth ACM international conference on Multimedia (MULTIMEDIA ’96). Association for Computing Machinery, New York, NY, USA, 87–98. https://doi.org/10.1145/244130.244151
14. TUG 2017. Institutional members of the LaTeX Users Group. Retrieved May 27, 2017 from <http://wwtug.org/instmem.html>
15. Alper Yilmaz, Omar Javed, and Mubarak Shah. 2006. Object tracking: A survey. ACM Comput. Surv. 38, 4 (December 2006), 13–es. https://doi.org/10.1145/1177352.1177355

Abdelghaffar, H., Kamel, S., and Duquenoy, P. (2010). “Studying E-Government Trust in Developing Nations: Case of University and Colleges Admissions and Services in Egypt”. Proceedings of the International Information Management Association Conference, Utrecht, The Netherlands.

Abdelsalam, H., Reddick, C., Gamal, S., and Al-shaar, A. (2013). “Social Media in Egyptian Government Websites: Presence, Usage, and Effectiveness”. Government Information Quarterly, 30(4): 406-416.

Baldassare, M. (2000). “California in the New Millennium: The Changing Social and Political Landscape”. Berkeley: University of California Press.

Bertot, J. and Grimes, J. (2012). “Promoting Transparency and Accountability through ICTs, Social Media, and Collaborative E-Government”. Transforming Government: People, Process and

Authors’ Information

|  |  |  |  |
| --- | --- | --- | --- |
| Your Name | Title\* | Research Field | Personal website |
| First Author |  |  |  |
| Second Author |  |  |  |
| Third Author |  |  |  |

\*This form helps us to understand your paper better, and the form itself will not be published.

\*Title can be chosen from: Master Student, PhD Candidate, Doctor, Assistant Professor, Lecturer, Senior Lecturer, Associate Professor, Full Professor, Researcher, Senior Researcher

1. \* Corresponding author [↑](#footnote-ref-1)