PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	NeuroCNVscore: A tissue-specific framework to prioritizing the pathogenicity of CNVs in neurodevelopmental disorders
AUTHORS	Liu, Xuanshi Xu, Wenjian Leng, Fei Zhang, Peng Guo, Ruolan Zhang, Yue Hao, Chanjuan Ni, Xin Li, Wei

VERSION 1 - REVIEW

REVIEWER	Dr. Yu-Peng Cun Chongqing Medical University Affiliated Children's Hospital
REVIEW RETURNED	04-Apr-2023

GENERAL COMMENTS	This paper proposed a new CNVs prioritize framework, NeuroCNVscore, which give pathogenicity score of CNVs in the neurodevelopmental disorder diseases. In the paper, they employed 4 machine learning models for classification comparison, and selected XGboost for NeuroCNVscore. This paper lay an important issue in variation pathogenicity score. My option is acceptable with major reversion with following comments: Majors: 1. NeuroCNVscore was not defined clearly. A cleary definetion need gived in method part. 2. NeuroCNVscore compared with SVScore, but SVScore is SV-based pathogenicity score, not CNV. SO, a CNV-based pathogenicity score should be compared in the main text. 3. In the 4 comparison model(Page-11, line-20), the author employed 5-fold cross validation (CV) to selected best classifier. The author did described how many repeat times for 5-folds CV. Minors: 1. Page-8, line-31, refer error. 2. some typos in main text and figure legend. 3. writing need be refined.

REVIEWER	Dr. Jeremy Miles Google Inc
REVIEW RETURNED	09-May-2023

GENERAL COMMENTS	I would like to thank the authors for providing the code in an ipynb file - this makes it clear what they did and allows the readers and reviewer(s) to see all of the results. Abstract: I would refer to boosted regression, rather than XGBoost, which is one implementation of the boosting algorithm. (The specific implementations of the other algorithms is not mentioned). P6, line 25: (Error! Reference source not found.. 1). The text on some of the charts is very small, and hard to read. Table S2: Presenting p-values to such high levels of precision is (in my opinion) excessive. A p-valye of 1.1*10^-42 and a p-value of 2.51*10^-232 differ by an more orders of magnitude than there are estimated to be atoms in the universe, but they are both effectively zero. Similarly, Table S3: a p-value of 4.47*10^-3 could be written as 0.0045.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1

This paper proposed a new CNVs prioritize framework, NeuroCNVscore, which give pathogenicity score of CNVs in the neurodevelopmental disorder diseases. In the paper, they employed 4 machine learning models for classification comparison, and selected XGboost for NeuroCNVscore. This paper lay an important issue in variation pathogenicity score. My option is acceptable with major reversion with following comments:

Majors:
1. NeuroCNVscore was not defined clearly. A cleary definetion need gived in method part.
Response: We wrote the definition of NeuroCNVscore as “We developed neuroCNVscore, which utilized XGBoost and comprehensive genome-wide features to evaluate the likelihood that a given CNV contributes to the development or manifestation of NDDs” in the Methods of the revised manuscript.

2. NeuroCNVscore compared with SVScore, but SVScore is SV-based pathogenicity score, not CNV. SO, a CNV-based pathogenicity score should be compared in the main text.
Response: We totally agree with Dr. Cun that CNV based pathogenicity score is important, and we should describe it more clearly that SVScore can evaluate different types of SVs including CNVs. We revised the statements in the second paragraph of the Introduction, “Copy number variants (CNVs) are structural variants (SVs) in the genome that involve the gain or loss of large segments of DNA, which have been implicated in NDDs”. Meanwhile, we added a sentence in the Methods, “SVScore can evaluate various types of SV including CNV”.

3. In the 4 comparison model (Page-11, line-20), the author employed 5-fold cross validation (CV) to selected best classifier. The author did described how many repeat times for 5-folds CV.
Response: Thank you for raising this point. XGBoost is able to process complex features and effectively capture feature interactions. Moreover, the utilization of the XGBoost algorithm has been widely observed in successful solutions across various Kaggle competitions (https://www.kaggle.com/). We have tested our dataset with several algorithms with which XGBoost gave the best performance. We did not apply additional 5-fold cross validation with repeat times in our study.

Following your suggestion, we conducted a rigorous evaluation using a 5-fold cross validation technique with 5 repeat times. Our results indicate that XGBoost exhibited the best performance

compared to other methods (Rebuttal Table 1, Rebuttal Figure 1). Specifically, using XGBoost for 5-fold cross validation with 5 repeat times at copy number loss model, we obtained an average AUC of 0.83 with a standard deviation of 0.0016. And the average AUC was 0.82 with a standard deviation of 0.0026 at copy number gain model. We mentioned this point in the Methods of the revised manuscript.

Minors:

1. Page-8, line-31, refer error.
Response: This has been corrected.

2. some typos in main text and figure legend.
Response: The typos have been corrected.

3. writing need be refined.
Respond: We have revised the manuscript carefully to refine some statements.

Reviewer 2

1. Abstract: I would refer to boosted regression, rather than XGBoost, which is one implementation of the boosting algorithm. (The specific implementations of the other algorithms is not mentioned).
Respond: We agree with Dr. Miles. There are several implementations of boosting algorithms. As of a classification task, we have compared the performances of two popular boosting algorithms, XGBoost and AdaBoost (Rebuttal Table 2). XGBoost showed better performance in both copy number loss and copy number gain models. Therefore, we kept XGBoost for our study.

2. P6, line 25: (Error! Reference source not found.. 1).
Response: The referred Fig. 1 has been added.

3. The text on some of the charts is very small, and hard to read.
Response: We have updated the figures and changed the size of fonts.

4. Table S2: Presenting p-values to such high levels of precision is (in my opinion) excessive. A p-valye of 1.1*10^-42 and a p-value of 2.51*10^-232 differ by an more orders of magnitude than there are estimated to be atoms in the universe, but they are both effectively zero. Similarly, Table S3: a p-value of 4.47*10^-3 could be written as 0.0045.
Response: Thanks for the comments. We have revised the presentation of p-values in both main text and supplementary Table 2 and 3. We have updated all the p values (P < 1□10-4) to ~0, and changed p values (1□10-4 < P < 1) to decimal mode.

VERSION 2 – REVIEW

REVIEWER	Dr. Yu-Peng Cun Chongqing Medical University Affiliated Children's Hospital
REVIEW RETURNED	13-Jun-2023

GENERAL COMMENTS	I am satisfied with updates and answers, and think the manuscript can be acceptable.

VERSION 2 – AUTHOR RESPONSE

N/A