Letter to the Editor

# Code Error in "Diagnostic Classification and Prognostic Prediction Using Common Genetic Variants in Autism Spectrum Disorder: Genotype-Based Deep Learning"

Catriona Miller, BSc; Theo Portlock, PhD; Denis M Nyaga, PhD; Greg D Gamble, PhD; Justin M O'Sullivan, PhD

Liggins Institute, University of Auckland, Auckland, New Zealand

**Corresponding Author:**

Justin M O'Sullivan, PhD
Liggins Institute
University of Auckland
85 Park Road, Private Bag 92019
Auckland, 1142
New Zealand
Phone: 64 099239868
Email: justin.osullivan@auckland.ac.nz

**Related Articles:**
Comment on: https://medinform.jmir.org/2021/4/e24754/
Retraction notice: https://medinform.jmir.org/2025/1/e76833

***JMIR Med Inform 2025;13:e66556;*** doi: 10.2196/66556

**Keywords:** autism prediction; machine learning; data leakage

Wang and Avillach [1] developed a convolutional neural network (CNN)–based diagnostic classifier for autism spectrum disorder (ASD). After preprocessing the genomics data from the Simons Simplex Collection (SSC) [2], common variants that may be protective or pathogenic for autism were extracted based on a $\chi^2$ test. The authors then designed a CNN-based diagnostic classifier for ASD with an accuracy and area under the receiver operating characteristic curve of 88% and 0.955, respectively. The predictor in Wang and Avillach [1] is currently considered the exemplar in the field, giving much more accurate predictions for autism than other studies [3].

However, when inspecting the code and repeating the analyses, we contend that the method used is flawed and leads to an approximately 30% overestimation of predictive ability.

Wang and Avillach [1] did not provide a GitHub link to the code that was used in their paper. However, code can be found in Dr Wang's GitHub repository [4] that matches the results and figures in the manuscript.

An error occurred in the data split for training and test sets. The methods state "...the SSC samples were partitioned into two sets based on random sampling of individuals into a training set (80%) and a hold-out test set (20%). There was no overlap of individuals across the two partitions" [1]. However, the code uses different indexing methods for the test and training sets (Multimedia Appendix 1). Because there appears to be no random seed in Wang and Avillach [1], we cannot reproduce the exact overlap that was mentioned in the manuscript. However, simulations (N=100) using the code identified an average of 80% (SD 1%) of the test dataset being represented in the training dataset.

We corrected this error in the code [5] and generated new models using the 100 features that were identified in Wang and Avillach [1] and the genomics data from the SSC [2]. Simulations (N=100) with these models result in an area under the receiver operating characteristic curve of 0.61 (SD 0.02) and an accuracy of 60% (SD 2%; Multimedia Appendix 1). This is 0.34 and 28% lower than the reported metrics, respectively, in Wang and Avillach [1].

The accuracy of the CNN-based diagnostic classifier for ASD presented in Wang and Avillach [1] is overestimated by ~28%. We contend that Wang and Avillach [1] should be retracted according to the Committee on Publication Ethics (COPE) guidelines.

## Acknowledgments

We acknowledge Simons Simplex Collection project number 15286.1.1.

## Data Availability

The code used for Wang and Avillach [1] is available on GitHub [4]. The clone of the code and corrected code are also available on GitHub [5]. Simons Simplex Collection is accessible at [6].

## Conflicts of Interest

None declared.

## Editorial Notice

The corresponding author of "Diagnostic Classification and Prognostic Prediction Using Common Genetic Variants in Autism Spectrum Disorder: Genotype-Based Deep Learning" did not submit a reply to this letter.

## Multimedia Appendix 1

Walkthrough of the steps that were taken and the results of our attempt to reproduce the work of Wang and Avillach (2021).
[PPTX File (Microsoft PowerPoint File), 418 KB-Multimedia Appendix 1]

## References

1. Wang H, Avillach P. Diagnostic classification and prognostic prediction using common genetic variants in autism spectrum disorder: genotype-based deep learning. JMIR Med Inform. Apr 7, 2021;9(4):e24754. Retracted in: JMIR Med Inform 2025;13:e76833. [doi: 10.2196/76833]
2. Fischbach GD, Lord C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron. Oct 21, 2010;68(2):192-195. [doi: 10.1016/j.neuron.2010.10.006] [Medline: 20955926]
3. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ. Sep 22, 2023;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
4. Wang H. Hms-dbmi/haishuai. GitHub. 2020. URL: https://github.com/hms-dbmi/Haishuai [Accessed 2024-09-12]
5. Miller C. Catriona-miller/sfari_paper_clone. GitHub. 2024. URL: https://github.com/Catriona-Miller/SFARI_paper_clone [Accessed 2024-09-12]
6. Simons Simplex Collection. Simons Foundation Autism Research Initiative. URL: https://www.sfari.org/resource/simons-simplex-collection/ [Accessed 2025-04-24]

## Abbreviations

**ASD:** autism spectrum disorder
**CNN:** convolutional neural network
**COPE:** Committee on Publication Ethics
**SSC:** Simons Simplex Collection

*Edited by JMIR Editorial Office ; This is a non–peer-reviewed article; submitted 21.09.2024; accepted 12.04.2025; published 06.05.2025*

*Please cite as:*
*Miller C, Portlock T, Nyaga DM, Gamble GD, O'Sullivan JM*
*Code Error in "Diagnostic Classification and Prognostic Prediction Using Common Genetic Variants in Autism Spectrum Disorder: Genotype-Based Deep Learning"*
*JMIR Med Inform 2025;13:e66556*
*URL: https://medinform.jmir.org/2025/1/e66556*
*doi: 10.2196/66556*