Using neural networks in polygenic risk score calculation

Dominique Weltevreden[1][0009-0002-9176-0509], Jalmar Teeuw[1,2][0000-0002-1637-888X], and Hilleke Hulshoff Pol[1,2][0000-0002-2038-5281]

[1] Department of Psychiatry, University Medical Center Utrecht, Utrecht, The Netherlands

[2] Department of Experimental Psychology, Utrecht University, Utrecht, The Netherlands

Abstract. Mini-review of the benefits of neural networks for calculating polygenic scores.

Keywords: Polygenic Scores · Deep Learning · Neural networks.

Human traits are influenced by genetics [18]. Variation in genetics can be caused by single nucleotide polymorphisms (SNPs): a substitution of a single allele in the DNA [20]. Genome-wide association studies (GWAS) can discover variants of these SNPs that are associated with certain traits by testing the differences in allele frequency between individuals [20]. Based on GWAS summary statistics, an individual's genetic predisposition for a trait or disease can be estimated from their genotype by calculating a polygenic (risk) score (PGS) [1]. This PGS can potentially be used clinically to classify high-risk individuals. These individuals can then be included in additional screening and preventative treatments [13].

However, so far, the predictive power of PGS for complex diseases remains limited [7]. This is partly because genome-wide association studies (GWAS) often lack the necessary statistical power and focus only on sequencing common genetic variants [7][22]. Moreover, PGS are typically calculated as the weighted sum of effect alleles [1] [13]. This assumes that (1) the effect of SNPs are linearly additive, and (2) SNPs act independent of one another [7]. These assumptions are not always met, because (1) the effects of a SNP can be non-additive [8], and (2) epistasis, or gene-gene interaction, is a ubiquitous component of disease [16]. Thus, traditional scoring methods cannot generate PGS that make the most accurate prediction in disease risk by neglecting these effects.

These limitations can be addressed by using machine learning methods, like neural networks (NN). Machine learning methods can capture non-linear relationships between SNPs and disease, assume non-additive effects of SNPs and incorporate interaction effects between genes [7]. Using activation functions, non-linearity is induced in neural networks, which allows the network to capture more complex patterns [4]. Furthermore, machine learning methods can easily include additional features, like demographics, clinical variables and measures from diagnostic imaging or lab reports [5]. A previous study has reviewed the application of traditional machine learning, (e.g. support vector machines, k-nearest neighbours, random forests) as scoring method for PGS [12]. Here, we present a mini-review of recent literature on the application of (deep) neural networks (NNs) for polygenic scoring where we address the question if these methods can benefit the predictive value of the PGS.

From literature search on keywords 'deep learning', 'neural network', and 'polygenic score', we identified 23 potential articles. From this list, 11 articles were included in this mini-review [2][3][6][11][9][10][15][17][19][23][24]. The remaining articles were excluded because they did not employ a neural network, did not calculate a PGS, or because they were a review. In 6 (55%) of the included articles, the predictive value of the NN outperformed that of traditional scoring methods [2][9][17][19][24][23]. 3 articles (27%) reported lower or equal performance of the PGS based on NNs for at least one of the phenotypes [3][6][11][10]; note that 1 article did not make a comparison between NN and traditional scoring methods [15]. It should be noted that the differences in performance between methods are often small and not always statistically validated. However, even a 1% improvement in accuracy can have strong implications for patients, as it can correspond to thousands of additional detected cases in screening [11].

Studies that improved over the traditional methods employed either a standard NN [2][24], a convolutional NN (CNN) [23], a recurrent NN [17], a NN with weight regularisation [9], or a genome local net with locally connected layers [19]. In some cases, a CNN was outperformed by traditional methods [3] or the genome local net (GLN) [19] in other studies. In the study where CNN works well [23], they also use a NN for preselection of the SNPs (as opposed to simply using the SNPs deemed significant by the GWAS), implying that the selection process for SNPs also greatly influences the results. This is supported by a the GLN model study, which is outperformed by a model based just on covariates in 70% of the 338 studied traits, and they suggest that this might because the GLN overfits [19]. Therefore, both standard NN and modern variants seem to work, but the choice of preselection of the SNPs is an important step.

The use of NN or variants thereof may improve over traditional scoring methods by incorporating non-linearity and interaction terms. These benefits only apply to traits where a non-additive effect and epistasis is expected. The studies in this review focused primarily on cancer [2][11][10][9] and Alzheimer's disease [6][17][24]. The results regarding the performance of NN over traditional methods for cancer phenotypes were mixed, but the NN performed better for all Alzheimer's studies, for which some interaction effects have been found [14][21]. In one study [19] 338 phenotypes were investigated, and it was found that their GLN worked particularly well for autoimmune diseases. This supports the idea that the trait of interest can impact whether an added benefit of the use of NN is found.

In conclusion, neural networks show marginal improvements for the calculation of the genetic risk of an individual. Improvement of the predictive power of neural networks for calculation polygenic scores may depend on the SNP selection and the trait under investigation.

References

1. Allegrini, A.G., Baldwin, J.R., Barkhuizen, W., Pingault, J.B.: Research review: A guide to computing and implementing polygenic scores in developmental research. Journal of Child Psychology and Psychiatry **63**(10), 1111–1124 (2022). https://doi.org/10.1111/jcpp.13611
2. Badré, A., Zhang, L., Muchero, W., Reynolds, J.C., Pan, C.: Deep neural network improves the estimation of polygenic risk scores for breast cancer. Journal of Human Genetics **66**(4), 359–369 (Oct 2020). https://doi.org/10.1038/s10038-020-00832-7
3. Bellot, P., de los Campos, G., Pérez-Enciso, M.: Can deep learning improve genomic prediction of complex human traits? Genetics **210**, 809–819 (Nov 2018). https://doi.org/10.1534/GENETICS.118.301298
4. Elhassani, M.E., Maisonnasse, L., Olgiati, A., Jerome, R., Rehali, M., Duroux, P., Giudicelli, V., Kossida, S.: Deep learning concepts for genomics : an overview. EMBnet.journal **27** (Jun 2022). https://doi.org/10.14806/EJ.27.0.990
5. Fritzsche, M.C., Akyüz, K., Abadía, M.C., McLennan, S., Marttinen, P., Mayrhofer, M.T., Buyx, A.M.: Ethical layering in ai-driven polygenic risk scores—new complexities, new challenges. Frontiers in Genetics **14**, 1098439 (Jan 2023). https://doi.org/10.3389/FGENE.2023.1098439
6. Hermes, S., Cady, J., Armentrout, S., O'Connor, J., Holdaway, S.C., Cruchaga, C., Wingo, T., Greytak, E.M.R.: Epistatic features and machine learning improve alzheimer's disease risk prediction over polygenic risk scores. Journal of Alzheimer's Disease **99**, 1425–1440 (Jan 2024). https://doi.org/10.3233/JAD-230236
7. Ho, D.S.W., Schierding, W., Wake, M., Saffery, R., O'Sullivan, J.: Machine learning snp based prediction for precision medicine. Frontiers in Genetics **10** (2019). https://doi.org/10.3389/fgene.2019.00267
8. Horita, N., Kaneko, T.: Genetic model selection for a case–control study and a meta-analysis. Meta Gene **5**, 1–8 (Sep 2015). https://doi.org/10.1016/j.mgene.2015.04.003
9. Kim, S.b., Kang, J.H., Cheon, M., Kim, D.J., Lee, B.C.: Penalised neural network with weight correlation descent for predicting polygenic risk score. medRxiv (2023). https://doi.org/10.1101/2023.05.23.23290438
10. Kim, S.b., Kang, J.H., Cheon, M., Kim, D.J., Lee, B.C.: Stacked neural network for predicting polygenic risk score. Scientific Reports **14**(1) (May 2024). https://doi.org/10.1038/s41598-024-62513-1
11. Klau, J.H., Maj, C., Klinkhammer, H., Krawitz, P.M., Mayr, A., Hillmer, A.M., Schumacher, J., Heider, D.: Ai-based multi-prs models outperform classical single-prs models. Frontiers in Genetics **14** (2023). https://doi.org/10.3389/FGENE.2023.1217860
12. Kruppa, J., Ziegler, A., König, I.R.: Risk estimation and risk prediction using machine-learning methods. Human Genetics **131**, 1639–1654 (Oct 2012). https://doi.org/10.1007/S00439-012-1194-Y
13. Lewis, C.M., Vassos, E.: Polygenic risk scores: from research tools to clinical instruments. Genome Medicine **12**, 44 (May 2020). https://doi.org/10.1186/s13073-020-00742-5
14. Lundberg, M., Sng, L.M.F., Szul, P., Dunne, R., Bayat, A., Burnham, S.C., Bauer, D.C., Twine, N.A.: Novel alzheimer's disease genes and epistasis identified using machine learning gwas platform. Scientific Reports **13**(1) (Oct 2023). https://doi.org/10.1038/s41598-023-44378-y

15. Montañez, C.A., Fergus, P., Montañez, A.C., Hussain, A., Al-Jumeily, D., Chalmers, C.: Deep learning classification of polygenic obesity using genome wide association study snps. Proceedings of the International Joint Conference on Neural Networks **2018-July** (Jul 2018). https://doi.org/10.1109/IJCNN.2018.8489048
16. Moore, J.H.: The Ubiquitous Nature of Epistasis in Determining Susceptibility to Common Human Diseases. Human Heredity **56**(1-3), 73–82 (Nov 2003). https://doi.org/10.1159/000073735
17. Peng, J., Bao, Z., Li, J., Han, R., Wang, Y., Han, L., Peng, J., Wang, T., Hao, J., Wei, Z., Shang, X.: Deeprisk: A deep learning approach for genome-wide assessment of common disease risk. Fundamental Research **4**, 752–760 (Jul 2024). https://doi.org/10.1016/J.FMRE.2024.02.015
18. Polderman, T.J.C., Benyamin, B., de Leeuw, C.A., Sullivan, P.F., van Bochoven, A., Visscher, P.M., Posthuma, D.: Meta-analysis of the heritability of human traits based on fifty years of twin studies. Nature Genetics **47**(7), 702–709 (May 2015). https://doi.org/10.1038/ng.3285
19. Sigurdsson, A.I., Louloudis, I., Banasik, K., Westergaard, D., Winther, O., Lund, O., Ostrowski, S., Erikstrup, C., Pedersen, O., Nyegaard, M., Consortium, D., Banasik, K., Bay, J., Boldsen, J.K., Brodersen, T., Brunak, S., Burgdorf, K., Chalmer, M.A., Didriksen, M., Dinh, K.M., Dowsett, J., Erikstrup, C., Feenstra, B., Geller, F., Gudbjartsson, D., Hansen, T.F., Hindhede, L., Hjalgrim, H., Jacobsen, R.L., Jemec, G., Kaspersen, K., Kjerulff, B.D., Kogelman, L., Larsen, M.A.H., Louloudis, I., Lundgaard, A., Mikkelsen, S., Mikkelsen, C., Nielsen, K.R., Nissen, I., Nyegaard, M., Ostrowski, S.R., Pedersen, O.B., Henriksen, A.P., Rohde, P.D., Rostgaard, K., Schwinn, M., Stefansson, K., Stefónsson, H., Sørensen, E., Thorsteinsdóttir, U., Thørner, L.W., Bruun, M.T., Ullum, H., Werge, T., Westergaard, D., Brunak, S., Vilhjálmsson, B., Rasmussen, S.: Deep integrative models for large-scale human genomics. Nucleic Acids Research **51**, e67–e67 (Jul 2023). https://doi.org/10.1093/NAR/GKAD373
20. Uffelmann, E., Huang, Q.Q., Munung, N.S., de Vries, J., Okada, Y., Martin, A.R., Martin, H.C., Lappalainen, T., Posthuma, D.: Genome-wide association studies. Nature Reviews Methods Primers 2021 1:1 **1**, 1–21 (Aug 2021). https://doi.org/10.1038/S43586-021-00056-9
21. Wang, H., Bennett, D.A., De Jager, P.L., Zhang, Q.Y., Zhang, H.Y.: Genome-wide epistasis analysis for alzheimer's disease and implications for genetic risk prediction. Alzheimer's Research & Therapy **13**(1) (Mar 2021). https://doi.org/10.1186/s13195-021-00794-8
22. Wray, N.R., Lee, S.H., Mehta, D., Vinkhuyzen, A.A., Dudbridge, F., Middeldorp, C.M.: Research review: Polygenic methods and their application to psychiatric traits. Journal of Child Psychology and Psychiatry **55**(10), 1068–1087 (2014). https://doi.org/10.1111/jcpp.12295
23. Yin, B., Balvert, M., Spek, R.A.V.D., Dutilh, B.E., Bohté, S., Veldink, J., Schönhuth, A.: Using the structure of genome data in the design of deep neural networks for predicting amyotrophic lateral sclerosis from genotype. Bioinformatics **35**, i538–i547 (Jul 2019). https://doi.org/10.1093/BIOINFORMATICS/BTZ369
24. Zhou, X., Chen, Y., Ip, F.C.F., Jiang, Y., Cao, H., Lv, G., Zhong, H., Chen, J., Ye, T., Chen, Y., Zhang, Y., Ma, S., Lo, R.M.N., Tong, E.P.S., Alzheimer's Disease Neuroimaging Initiative, Mok, V.C.T., Kwok, T.C.Y., Guo, Q., Mok, K.Y., Shoai, M., Hardy, J., Chen, L., Fu, A.K.Y., Ip, N.Y.: Deep learning-based polygenic risk analysis for alzheimer's disease prediction. Communications Medicine 2023 3:1 **3**, 1–20 (Apr 2023). https://doi.org/10.1038/S43856-023-00269-X