Capturing epistasis with deep learning

A comparison with polygenic risk score on phenotype prediction using genotype data

M.W. Yeung, C. Maj, P. Krawitz
Institute for Genomic Statistics and Bioinformatics, University Hospital Bonn, University of Bonn
Contact: mwyeung@uni-bonn.de

Background and Aim

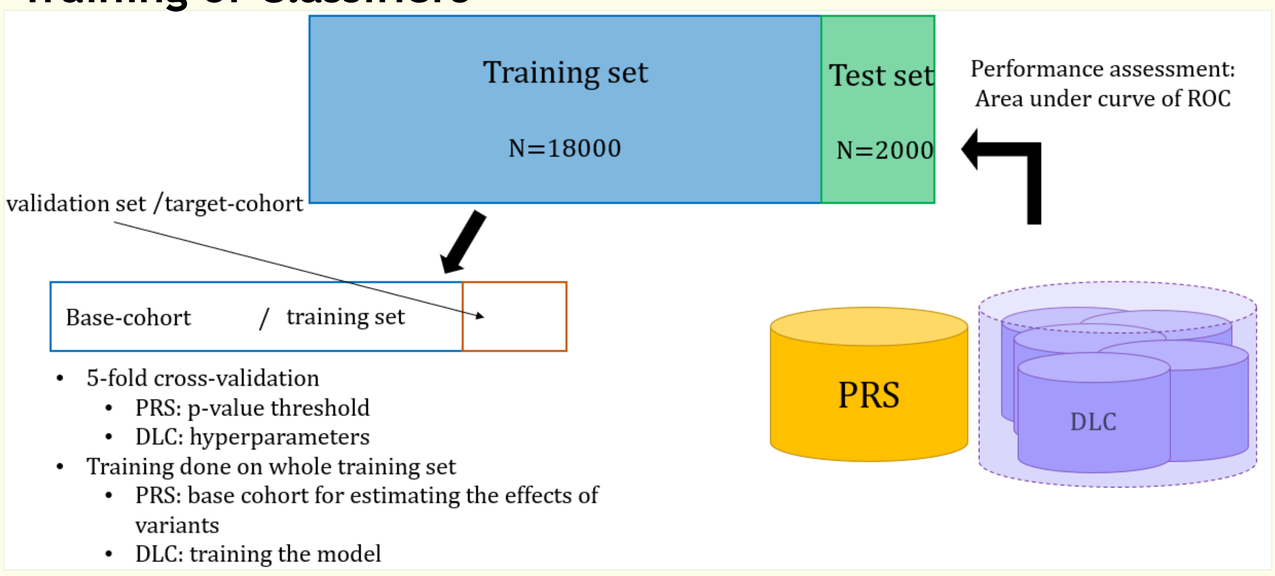
- Polygenic Risk Scores (PRS) based on variation in multiple genetic loci, typically single nucleotide polymorphisms (SNPs), is a widely used tool for investigation of genetics of complex diseases, which are influenced by a combination of multiple genes and environmental factors.
- Recent studies suggest PRS has good predictive performance on complex disease such as coronary heart disease and it has been proposed for risk (Nat Genet. 2018;50(9):1219–1224).
- However since PRS considers only the additive component of genetic variance, it is unlikely such model can fully account for the genetic architecture of all complex diseases. Substantial differences between SNP-based heritability and classical heritability have been observed for some diseases.
- Deep learning (DL) is a sub-category of machine learning capable of extracting non-linear features. In this study we evaluated the use of DL on capturing SNP-SNP interaction (epistatic) effects with simulated SNP-based genetic data.

Method

Study design

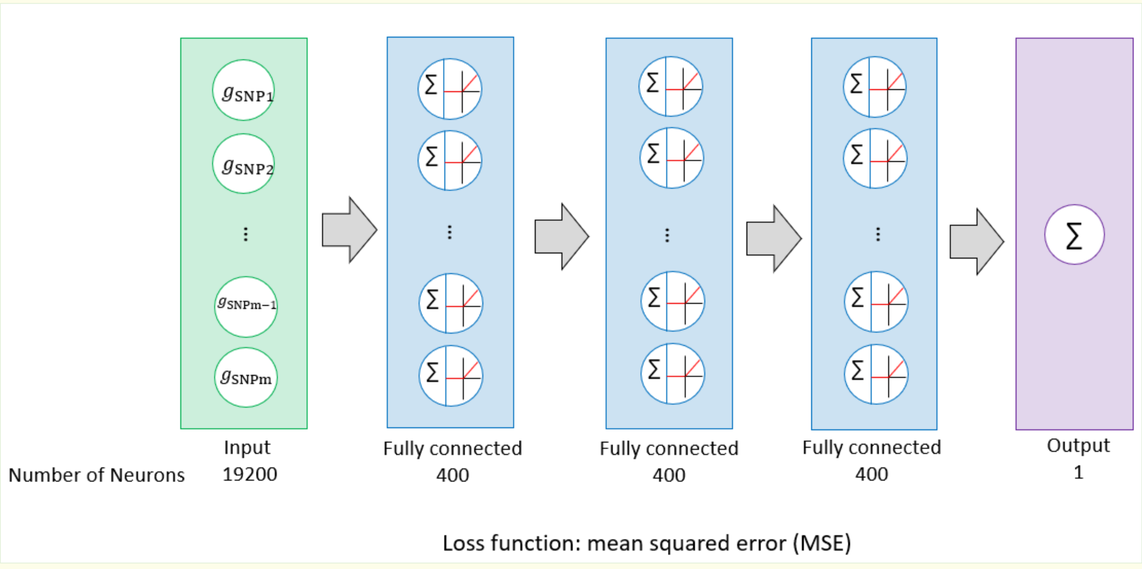
- Data size: 20000 individuals
- Number of feature: 19200 independent SNPs
- Either single or pairwise interaction effect for each SNP
- Magnitudes of effect simulated to reflect the common complex diseases
- Various genetic models considered from 0.5% (oligogenic) to 100% (omni-genic). A total of 16 datasets simulated

Training of Classifiers

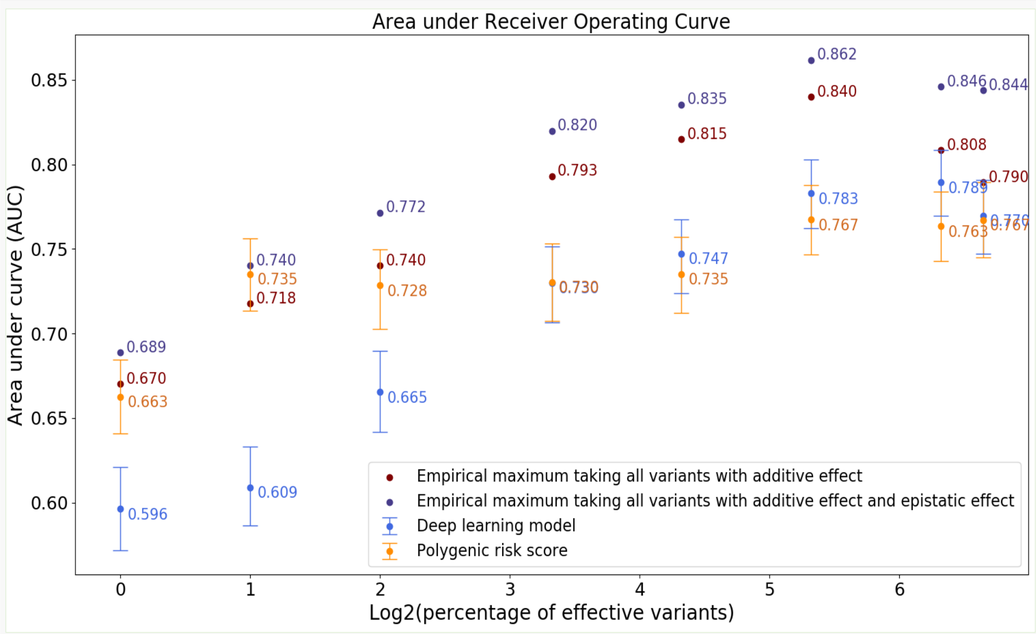


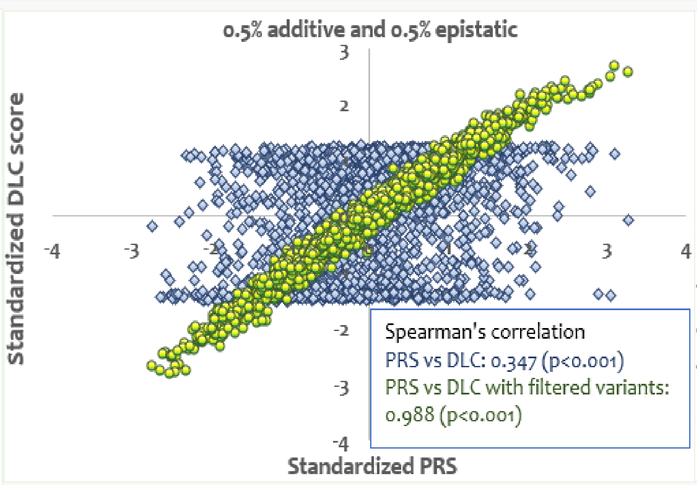
- 5-fold cross-validation
  - PRS: p-value threshold
  - DLC: hyperparameters
- Training done on whole training set
  - PRS: base cohort for estimating the effects of variants
  - DLC: training the model

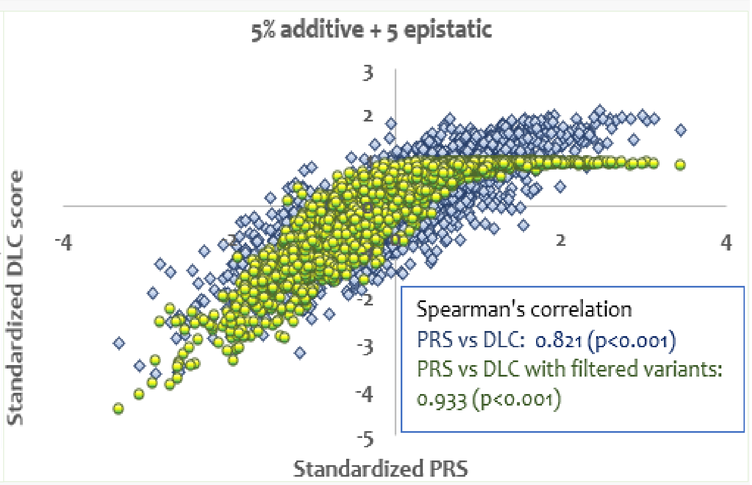
Network architecture



Results







- PRS performed better than DLC under oligogenic model and reverse was observed under omni-genic model
- Performance of DLC improved if trained on only SNPs included in PRS

Total causal variant	DLC on all variants	DLC on variants included in PRS	PRS
1%	0.596	0.664	0.663
10%	0.730	0.756	0.730

Conclusion

- We explored the utility DL for disease prediction using high dimension genetic data. In risk scenarios reflecting common complex disease, applying feature selection would improve the performance of DLC.

Conflict of Interest Disclosure: All authors have no conflicts of interest related to this study to declare.