

Submitter Name: Qing Lu
Submitted email: qlu@msu.edu

A Kernel-Based Neural Network for High-dimensional Risk Prediction on Massive Genetic Data

Qing Lu¹

¹Department of Epidemiology and Biostatistics, Michigan State University

Artificial intelligence (AI) is a thriving research field with many successful applications in areas such as imaging and speech recognition. Neural-network-based methods, such as deep learning, play a central role in the modern AI technology. While neural-network-based methods also hold great promise for genetic research, the high-dimensionality of genetic data, the massive amounts of study samples, and complex relationships between genetic variants and disease outcomes bring tremendous challenges to use these methods in genetic research. To address these challenges, we propose a kernel-based neural network (KNN) method. KNN inherits features from both linear mixed models (LMM) and classical neural networks, and is designed for high-dimensional genetic risk prediction analysis. Unlike the classic neural network, KNN summarizes a large number of genetic variants into kernel matrices and uses the kernel matrices as input matrices. Based on the kernel matrices, KNN builds a feedforward neural network to model the complex relationship between genetic variants and a disease outcome. Minimum norm quadratic unbiased estimation and batch training are implemented in KNN to accelerate the parameter estimation, making KNN applicable to massive datasets with millions of samples. Through simulations, we demonstrate the advantages of KNN over LMM in terms of prediction accuracy and computational efficiency. We also apply KNN to the large-scale UK Biobank dataset, investigating high-dimensional risk prediction models for substance dependence.