NMF-based Machine Learning for Alzheimer's Disease Biomarker Identification and Diagnosis

Abraham, Aaron (School: Webber Academy)

Alzheimer’s disease (AD) diagnosis is a challenge, with misdiagnosis rates at 20-30%. Proteomics has become a field of interest in the search for biomarkers for the early diagnosis of AD. However, the prevalence of large dimensionality proteomics datasets has created a need for a methodology that can effectively reduce data, identify important biomarkers, and achieve high diagnostic accuracy. A novel method was developed that utilizes non-negative matrix factorization (NMF) and supervised machine learning to detect key biomarkers for Alzheimer’s disease using classified data (AD, mild cognitive impairment, and healthy cases) from the Alzheimer’s Disease Neuroimaging Initiative repository (n = 566 patients, 148 protein concentrations measured). NMF and a traditional principal component analysis (PCA) approach were used to reduce the data. NMF identified 11 key biomarkers that are statistically significant under ANOVA and post-hoc Tukey analysis and discriminates between different patient classes. Five different machine learning models were trained on the NMF-reduced data, PCA-reduced data and unreduced data and F1 scores were used for evaluation. Results indicated that NMF-reduced models performed 38% better than PCA-reduced models and 68% better than models trained on unreduced data. Partial dependence plots, Shapley values, and permutation importance plots were analyzed to identify how individual biomarkers affected model performances. This novel method can be applied on any omics-based dataset for biomarker selection for any disease. Through non-negative matrix factorization and supervised machine learning, a highly accurate and transparent methodology was developed for the identification of biomarkers for the early diagnosis of Alzheimer’s disease.

Awards Won:

American Statistical Association: Certificate of Honorable Mention