Enhancing Prediction Models for Genetic Neurodegenerative Diseases Using TabDDPM-Based Data Augmentation

Wang, Angela (School: University School of Milwaukee)

Genetic neurodegenerative diseases affect more than 57 million people worldwide. With aging populations, their prevalence is rising, making early prediction critical for enabling timely interventions that can slow disease progression and improve quality of life. Single Nucleotide Polymorphisms (SNPs) provide valuable insights into disease susceptibility. However, while prediction models can detect patterns and drive earlier diagnosis, there is limited availability of labeled SNP datasets to train these models due to privacy regulations like HIPAA, which restrict genetic data sharing to prevent re-identification risks. This study investigates the use of TabDDPM, a diffusion-based generative AI model that learns the distribution of tabular genomic data and generates high-quality synthetic datasets. These datasets are then used to train a convolutional neural network (CNN) model, which predicts disease susceptibility risk. Utilizing Alzheimer's disease as a model, TabDDPM analyzes the distribution of SNP arrays from the APOE gene on chromosome 19, a well-established genetic risk factor for Alzheimer's disease, and then generates synthetic data. The CNN model was then trained using real data augmented with synthetic datasets. Model performance was evaluated using Mean Squared Error (MSE) and R^2. The results showed that in each trial, adding synthetic data to the real dataset significantly improved the model's prediction accuracy. This study demonstrates the feasibility of using synthetic data to advance research on genetic neurodegenerative diseases, enhancing prediction accuracy while reducing barriers to accessing patient data.

Awards Won: