ORIE 5741 Final Report

                 Jiacheng Cao (jc2992), Fuhan Cong (fc334)

May 10, 2023

**1 Introduction**

Parkinson's disease is a severe neurological condition that impacts multiple bodily functions and lacks a cure. By 2037, an estimated 1.6 million individuals in the US will be impacted, resulting in substantial economic consequences. With the assistance of data science, comprehending these anomalies may lead to significant breakthroughs in the creation of novel pharmacotherapies that can either mitigate or eradicate Parkinson's disease.

In our project, we will build a predictive model for MDS-UPDR scores, which measures the progression of patients with Parkinson’s Disease. The MDS-UPDRS, sponsored by the Movement Disorder Society, is a thorough evaluation that covers both the motor and non-motor symptoms linked to Parkinson's disease. A higher MDS-UPDRS score typically indicates more severe Parkinson's Disease symptoms. In the study published in the Journal of Proteome Research, researchers analyzed the levels of proteins and peptides in the cerebrospinal fluid (CSF) of Parkinson's Disease patients over time. They identified several proteins and peptides that were significantly associated with disease progression, including alpha-synuclein and DJ-1. The authors concluded that monitoring protein and peptide levels in CSF may be useful in predicting disease progression in Parkinson's patients.1

The project is developed based on data on protein and peptide levels over time of subjects with Parkinson’s Disease. The goal of the project is to develop a predictive model that utilizes this data to forecast the course of Parkinson's Disease in patients. By analyzing changes in protein and peptide levels over time, we aim to identify potential biomarkers of Parkinson's Disease and their association with disease progression (represented by MDS-UPDR scores). This will provide valuable insights into the underlying mechanisms of the disease and may lead to the development of novel pharmacotherapies that target these biomarkers. Ultimately, the goal of this project is to improve the diagnosis and management of Parkinson's Disease and to help pave the way toward finding a cure for this debilitating condition.

**2 Data**

**2.1 Exploring the data**

Our project is established on the dataset consisting of protein abundance values derived from mass spectrometry readings of cerebrospinal fluid (CSF) samples gathered from several hundred patients. Each patient contributed several samples over the course of multiple years while they also took assessments of Parkinson’s Disease severity. We worked on four related datasets. First, train\_peptides.csv includes mass spectrometry data at the peptide level (peptides are the component subunits of proteins). Second, train\_proteins.csv, which includes protein expression frequencies aggregated from the peptide level data. Third, train\_clinical\_data.csv includes the patient's score for different parts of the Unified Parkinson's Disease Rating Scale. Parkinson's Disease patients’ scores for mood and behavior are recorded in Part 1, and their motor function scores are in Part 3. Also, whether the patient was taking medication such as Levodopa during the UPDRS assessment is also included in this table. Eventually, the **supplemental\_clinical\_data.csv** provides additional context about the typical progression of Parkinson's Disease.

**2.2 Data Cleaning**

First, we focus on analyzing the first three datasets. After cleaning them, the resulting dataset train\_peptides.csv contains 981834 rows and 6 columns, train\_proteins.csv contains 232741 rows and 5 columns, and train\_clinical\_data.csv contains 2615 rows and 8 columns. Among the three datasets above, we have only found the missing values from clinical train data.

**2.3 Data Visualization**

**5 Reference**

1. Mollenhauer, B., El-Agnaf, O. M., Marcus, K., Trenkwalder, C., Schlossmacher, M. G., & The Systemic Synuclein Sampling Study (S4) Group. (2019). The utility of cerebrospinal fluid alpha-synuclein levels in Parkinson's disease: a systematic review and meta-analysis. Journal of proteome research, 18(11), 4373-4385.