Parkinson’s Disease Progression Prediction: Report

Department of Computer Science

Illinois Institute of Technology

April 29, 2023

Authors

Javier Echavarren A20521368

Victor Bóveda A20521372

Carlos Castillo A20520708

Abstract

# Introduction

Parkinson's disease is a brain-related health problem that leads to the slow breakdown of certain nerve cells, causing motor and non-motor symptoms. A crucial aspect of this disease is the accumulation of an abnormal protein called alpha-synuclein, which plays a significant role in its development. Identifying protein and peptide biomarkers associated with Parkinson's disease is essential for early detection and intervention.

Protein and peptide biomarker discovery has become an essential focus in Parkinson's disease research. By analyzing the levels and alterations in specific proteins and peptides linked to Parkinson's, scientists can develop more accurate diagnostic tools and improve treatment options for patients. For instance, researchers are studying the levels of alpha-synuclein in different body fluids, such as cerebrospinal fluid, blood, and saliva, hoping to discover reliable biomarkers for the disease.

Besides alpha-synuclein, other proteins and peptides are being investigated for their potential as biomarkers. For example, some researchers are examining the levels of certain enzymes, like DJ-1, which are thought to be involved in the disease's progression. By understanding how these proteins and peptides change in people with Parkinson's, scientists can develop new diagnostic tests and targeted therapies.

One of the main challenges in discovering protein and peptide biomarkers is finding specific markers that are consistently altered in Parkinson's disease patients. To achieve this, researchers often use advanced techniques, such as mass spectrometry and protein microarrays, to analyze large numbers of proteins and peptides in biological samples. These techniques allow scientists to identify patterns and changes that may be linked to the disease.

Once potential protein and peptide biomarkers are identified, they need to be validated in larger patient populations. This involves comparing the levels of these biomarkers in people with Parkinson's disease to those without the disease. If the biomarkers are consistently different between the two groups, they may be useful in diagnosing Parkinson's or predicting its progression.

In summary, focusing on protein and peptide biomarker discovery is crucial for improving early detection and treatment of Parkinson's disease. By identifying specific proteins and peptides associated with the disease, researchers can develop more accurate diagnostic tools and better understand the underlying causes of Parkinson's. This will ultimately lead to improved treatment options and a better quality of life for people with the disease and their families.

# Problem description

Parkinson's disease is a progressive neurological disorder that affects an individual's motor and non-motor functions. Accurate assessment and tracking of the disease's progression are crucial for providing timely and appropriate treatment. The Movement Disorder Society-sponsored Unified Parkinson's Disease Rating Scale (MDS-UPDRS) is a comprehensive tool used by clinicians and researchers to evaluate the severity of Parkinson's disease symptoms and monitor disease progression. In this project, our primary goal is to develop a machine learning model that can accurately predict MDS-UPDRS scores, which will ultimately assist in better understanding the disease and enhancing diagnostic and treatment strategies.

To achieve this goal, we are participating in a Kaggle competition focused on the evolution of Parkinson's disease. The competition provides a unique opportunity for our team to apply our machine learning skills to a real-world problem with significant implications for patients suffering from Parkinson's disease and the broader scientific community. By developing a robust predictive model, we hope to contribute valuable insights that can further advance research and improve patient outcomes.

The dataset for the competition consists of protein and peptide levels measured in individuals diagnosed with Parkinson's disease and age-matched healthy control subjects. These biomolecules have been identified as potential markers for the disease and could provide valuable information about its progression. By analyzing this data, we aim to identify patterns and relationships that can help predict MDS-UPDRS scores more accurately.

Data

The competition data consists of three files:

* train\_peptides.csv: Contains mass spectrometry data at the peptide level, including visit ID, visit month, patient ID, UniProt ID, peptide sequence, and peptide abundance.
* train\_proteins.csv: Provides aggregated protein expression frequencies, with details such as visit ID, visit month, patient ID, UniProt ID, and normalized protein expression (NPX).
* train\_clinical\_data.csv: Includes clinical data like visit ID, visit month, patient ID, UPDRS scores for different parts (1-4), and information on whether the patient was on medication during the UPDRS assessment.

# Competition tasks

The competition has two tasks:

1. Based on protein and peptide levels on a given visit, predict updrs levels for that time.
2. Based on protein and peptide levels on a given visit, predict updrs levels in future visits in 6,12,18 and 24 months from the given visit.

# Exploratory data analysis

The EDA was composed of the following steps:

* Check the training datasets: We firstly looked to every train dataset. This way we figured out how the data was stored and the dtypes of the features. We discovered that the data in train\_proteins and train\_peptides datasets were stored in a long format style.
* Modify the tables styles: In order to change from long to wide style, we pivot the train\_proteins and train\_peptides tables. After this step we obtained two tables with every peptide and protein as a feature.
* Merge: After obtaining the pivot tables, we merged them to the train\_clinical data frame. We used a left join so we would not miss any data from the train\_clinical dataset.
* Fill missing data: Once we had all the data in one data frame, we had to deal with the missing data. Due to we are working with time series, we thought that the best idea would be to use a linear regression model in order to fill the NA values.
* Scale the data: Finally, we had the final train data frame with all the data. Now we scaled the data so it would be easier and quicker to create a model. We use a standard scaler function from sklearn to perform this step.
* Perform linear regression of each patient evolution over time: as the updrs data we had was sparse, we performed a linear regression of each patient evolution over time, and use this regression values (intercept plus slope) to make the prediction of updrs evolution over time.

# Model selection

We propose the following model:

1. For task 1: We will use a regression model based on random forest, called XGBoost. We will fit it with the protein and peptide levels and train it to get the updrs of the visit time.
2. For task 2: We will use the updrs predicted by the task 1 model and predict its future change based on a prediction of its evolution over time. We suppose that this evolution is linear, so we will train a model to predict the slope of the updrs evolution over time based on the protein and peptide levels.

With this slope prediction, we will predict the future values of updrs with the following formula:

Model description:

For this tasks, we have implemented XGBoost (eXtreme Gradient Boosting). The XGBoost Regressor works by constructing an ensemble of decision trees, where each tree is built sequentially, focusing on the errors made by the previous trees in the ensemble. This process is achieved by minimizing a regularized objective function that consists of a loss function and a regularization term. The loss function measures the difference between the true target values and the predictions made by the model, while the regularization term penalizes the complexity of the model to prevent overfitting.

Some key features of the XGBoost Regressor model are:

* Regularized boosting: XGBoost incorporates L1 (Lasso) and L2 (Ridge) regularization techniques to control the complexity of the model and reduce overfitting.
* Column Block: XGBoost uses a column block data structure to store the dataset in-memory, which allows for efficient parallelization and cache-aware computation.
* Sparsity-aware: XGBoost can handle missing values and sparse data efficiently by using a sparsity-aware algorithm for both tree construction and prediction.
* Early stopping: XGBoost can automatically stop training if the performance on a validation set does not improve for a specified number of rounds, thus saving computation time and preventing overfitting.
* Parallelization: XGBoost supports parallelization of tree construction using multiple cores, which significantly speeds up the training process.
* Cross-validation: XGBoost provides built-in support for K-fold cross-validation to estimate the model's performance more accurately.
* Customizable loss functions: Users can define custom loss functions and evaluation metrics, which allows flexibility in solving different types of regression problems.

# Results:

# Model selection

# References