

# ORIE 5741 Final Report

## Prediction on UPDRS for Parkinson's Disease

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-UPDRS, which measures the progression of patients with Parkinson's Disease. 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. (Saptarshi et al., 2015).

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, `train_peptides.csv`, `train_proteins.csv`, `train_clinical_data.csv` and `supplemental_clinical_data.csv`. Each dataset follows a consistent structure, with one shared column used to distinguish individual data entries: `visit_id` (a combination of the patient id and the month of visit). The training dataset consists of 1113 distinct visit IDs, each of which corresponds to a single test for one of 248 unique patients. For each patient, there are 15 different months of visits

recorded, resulting in a total of 1113 tests. The visit ID serves as a unique identifier for each test conducted on a patient.

## 2.2 Data Visualization

### Train proteins

The dataset contains information on the presence and frequency of various proteins within the sample. A frequency plot was generated for the top 20 proteins across all patients, which revealed that certain proteins, such as P01024 and P05090, were detected in all 1113 tests. However, other proteins, such as Q99829, were only found in 489 tests. Overall, the dataset contains 227 different proteins. However, it is important to note that for a given test, only a subset of these proteins may be detected.

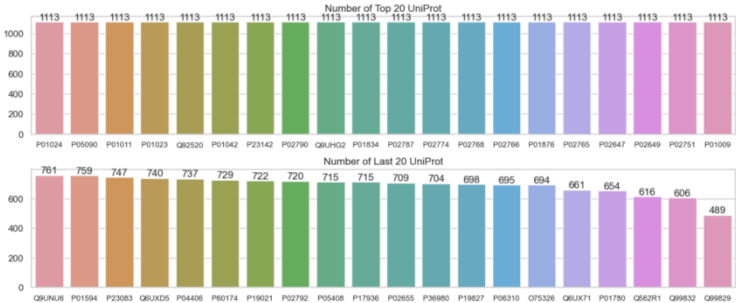


Figure 1: Frequency table for the top 20 Proteins

### Train peptide

Proteins play a critical role in many biological processes, and are composed of amino acids that are linked together through peptide bonds. Peptides are smaller units that are formed from the linking of two or more amino acids and can be further broken down into individual amino acids. A single protein can be associated with several different peptides, each with a unique sequence of amino acids.

In this dataset, the peptide abundance column records the frequency of each amino acid in the sample. This is important as it allows for the quantification of protein expression and the identification of potential biomarkers for various diseases. By examining the peptide abundance data, researchers can gain insights into the underlying biological processes and identify potential therapeutic targets. Overall, understanding the relationship between proteins and peptides is crucial for interpreting the results of this dataset and furthering our knowledge of the biological mechanisms at play for patients with Parkinson’s Disease.

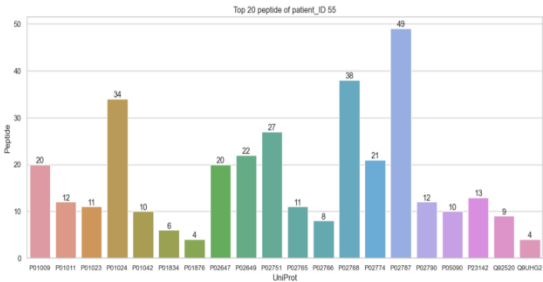


Figure 2: Frequency table for the top 20 peptides for one patient (patient\_id = 55)

Figure 2 displays the distribution of the top 20 proteins based on the number of distinct peptides detected for a single patient. For instance, patient #55 exhibits 49 distinct peptides detected under protein P02787. It is important to note that the test set may consist of peptides not found in the training set, which could impact the results of any subsequent analyses.

## Train clinical data

During the cleaning process of the `train_clinical_data.csv`, missing values were identified in multiple columns. These columns correspond to various sections of the UPDRS assessment and record the scores of each patient.

### Method 1: Mean Value

To address these missing values, we utilized mean imputation. As for the `upd23b_clinical_state_on_medication` column, which indicates whether the patient is taking medication during the UPDRS assessment, we assumed that missing values correspond to patients who are not taking medication.

### Method 2: KNN

We also implemented the KNN imputation to fill in the missing values in the clinical dataset. KNN imputation is a non-parametric method that works by finding the k-nearest neighbors of each missing value and imputing the missing value with the average of these k-nearest neighbors. By doing so, the imputed value is more likely to be similar to the actual value. To improve the performance of KNN imputation, we merge the complete records (no missing values) from `supplemental_clinical_data.csv` to `train_clinical_data.csv` to augment the clinical dataset. By doing so, the KNN algorithm has more data to work with, which can improve the imputation accuracy. Additionally, we apply the log transformation to the data to make it more normally distributed, which is explained in more detail in the later Feature Engineering section.

Upon completion of the data imputation, a series of plots were produced to facilitate an analysis of the collected data. Figure 3 presents a graphical representation of the mean UPDRS scores over time, with the x-axis portraying the month of the patient's visit in relation to their initial visit, and the y-axis representing the mean score. The visualization illustrates the progression of mean scores across various time lags and highlights a notable upward trend for all four measurements. These results suggest that the severity of the condition being evaluated is gradually worsening over time. Furthermore, a comparison of the distribution of UPDRS section 1 and section 2 reveals a similarity between the two. In contrast, section 3 displays a greater level of volatility in score distribution, as evidenced by its more prominent fluctuations in mean values over time.

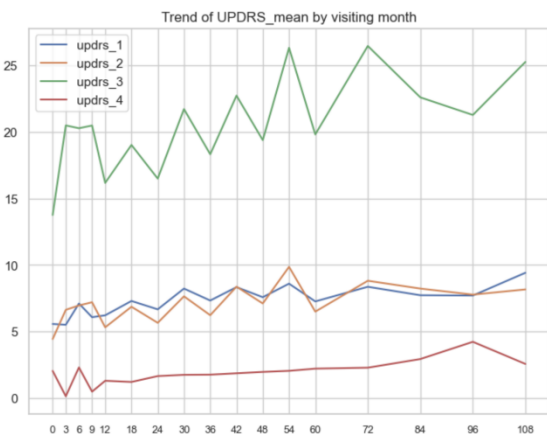


Figure 3: Frequency table of UPDRS stages

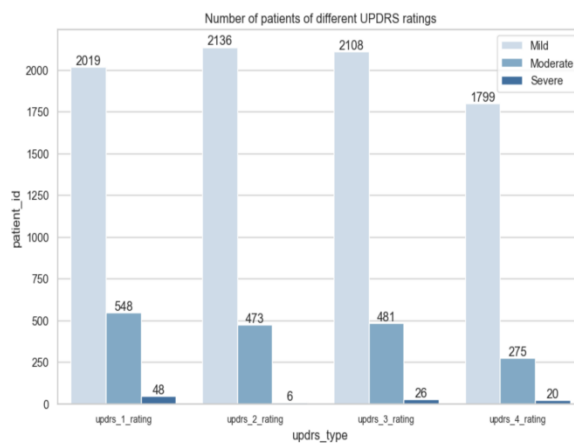


Figure 4: Frequency table of UPDRS stages


Based on our literature review, we have identified that the severity level of Parkinson's disease can be classified using scores from the UPDRS assessment tool. For example, `updrs_1`, which evaluates non-motor experiences of daily living, has a range of 0-52, with a score of 10 and below indicating mild symptoms, and 22 and above indicating severe symptoms. The bar plot shows the distribution of severity levels of four UPDRS ratings. Upon converting the UPDRS scores into severity categories, namely mild, moderate, and severe, we generated a bar chart (Figure 4) to visualize the distribution of scores across each section. The chart indicates that the majority of patients exhibit mild conditions across all four tests, while the frequency of diagnoses decreases with the increase in severity.

### 3 Feature Engineering

#### 3.1 Pivot Table

To eliminate repeated values in the protein and peptide data, we use the `groupby` function on the `UniProt` and `Peptide` columns for protein and peptide data, respectively. However, instead of computing the mean of the values as is typically done with `groupby` function, we retain all values and reset the index to create a new data frame with unique visit ID, `UniProt`, and `Peptide` combinations.

Next, we pivot the resulting data frame from a long to a wide format to create a more compact dataset with the `UniProt` and `Peptide` values as columns. We use the `visit ID` column as the index and the `NPX`, `Peptide Abundance` values as the corresponding values for each unique `UniProt`, `Peptide`, and `visit ID` combination. Finally, we merge the resulting protein and peptide Data Frames using the `visit ID` column as the key of creating a comprehensive dataset with protein and peptide abundance data for each `visit ID`. The process of pivoting the protein and peptide data frames from long to wide format provides us with additional potential features to work with. Each distinct protein and peptide are now represented as a column in the resulting merged dataset. These new features can be used in the modeling process to improve the accuracy of our analysis.



	visit_id	UniProt	NPX
1	55_0	O00391	11254.3
2	55_0	O00533	732430.0
3	55_0	O00584	39585.8

visit_id	O00391	O00533	O00584
55_0	11254.3	732430.0	39585.8

Table 1: Pivot Transformation

#### 3.2 Log Transformation

To improve the accuracy of our KNN imputation, we first concatenated the two data frames `clinical_train_data` and `supplemental_clinical_data` vertically to help KNN find patterns in the data. Next, we identified a subset of columns that we wanted to normalize and applied the  $x = \log(1 + x)$  transformation to them. In this case, we selected columns UPDRS scores for normalization. The log transformation is a commonly used technique to normalize skewed data. By taking the logarithm of the data, we can reduce the impact of outliers and make the data more normally distributed. Additionally, by

adding 1 to the data before taking the logarithm, we can avoid taking the logarithm of 0 or negative values.

After normalizing the data, we performed KNN imputation on the normalized data. We then imputed the missing values in the original data frame using the imputed values from the normalized data frame. Finally, we applied the inverse transformation to obtain the final imputed values and return the data to its original scale. This step improves the reliability of the imputation.

## **4 Data Modeling and Analysis**

### **4.1 Linear Regression**

The first model applied was linear regression. In this model, the protein and peptide features created earlier, along with the visit month, were selected as the independent variables, and the UPDRS scores from the dataset were taken as the dependent variable. The resulting mean squared error (MSE) for the linear regression model was calculated for each of the four UPDRS sections separately. The MSE values were 300.2198, 353.3514, 1740.6120, and 27.7125 respectively. These values indicate that the model was not able to predict the UPDRS scores accurately, particularly for the third section where the MSE was very high.

A high MSE suggests that the predictions made by the model are far from the actual values. There could be several reasons for this, such as the presence of outliers, non-linearity in the relationship between the independent and dependent variables, underfitting or overfitting of the model, or insufficient data. Further investigation and adjustments to the model may be necessary to improve its accuracy and reduce the MSE values.

### **4.2 Random Forest**

The exploration of alternative models to minimize the MSE and improve prediction accuracy ensued after the implementation of the Linear regression model. One such popular algorithm in machine learning is Random Forest, which leverages decision trees to make predictions based on their combined outputs. Thus, we opted to employ the Random Forest method and began by examining the complete model utilizing all potential features.

Subsequently, we generated four datasets, one for each updrs\_1 to updrs\_4 score, by merging the protein and peptide data with their respective UPDRS scores. These resulting datasets were then divided into training and validation sets. Thereafter, we initialized a random forest model using the TensorFlow Decision Forests library, compiled it using the MSE metric, and fitted it to the training data using the fit function. Finally, the evaluation function was utilized to assess model performance and calculate the MSE for the validation set. However, it is noteworthy to mention that the explanatory variable matrix is not ideal due to the slightly higher number of columns than rows. This divergence from the common scenario might impact the prediction outcome from the model.

The output generated from the Random Forest indicates the Mean Squared Error (MSE) for each of the updrs\_1 to updrs\_4 scores. The value for updrs\_1 is 22.9682, while updrs\_2 has a higher value of 29.7070. However, updrs\_3 records the highest MSE among all scorers with a value of 172.7797. Conversely, updrs\_4 has the lowest MSE of 6.0370. These findings provide insight into the performance of the Random Forest model. In addition to generating the MSE results, the Random Forest model can also furnish us with further insights into feature importance. The following Figure 5 has listed the top 20 most important feature generated from the random forest model.

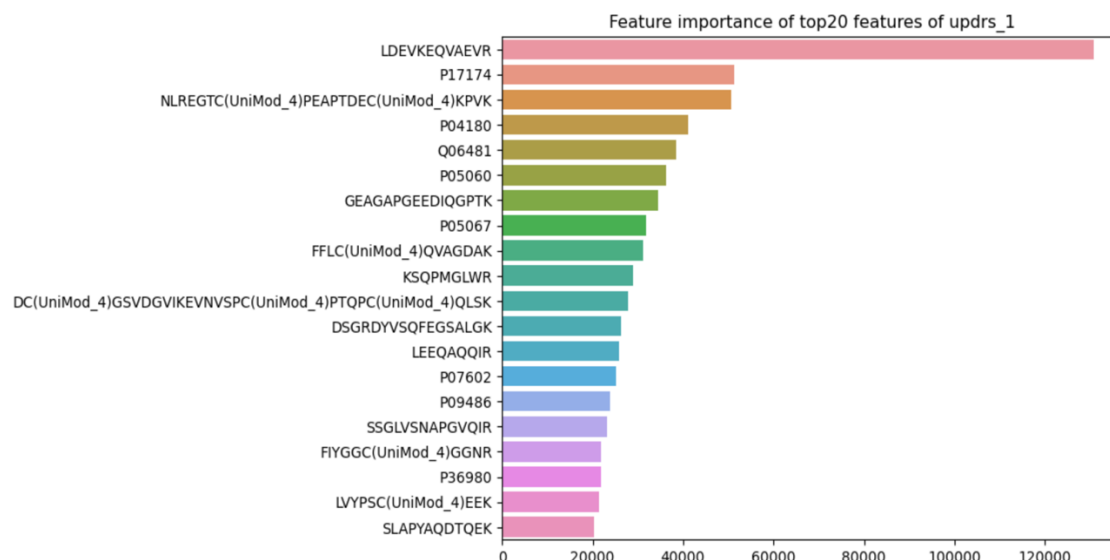


Figure 5: Random Forest – top 20 important features

To avoid the overfitting of the model and control the depth of the trees within a reasonable range, we need to test for the optimal number of trees. The below graph illustrates the performance of the Random Forest model in predicting the updrs\_1 score with varying numbers of trees. The x-axis represents the number of trees used in the model, while the y-axis denotes the corresponding Mean Squared Error (MSE) value.

As the number of trees increases, the MSE generally decreases, indicating better model performance. However, after a certain number of trees (around 32 in this case), the MSE begins to stabilize, and further increases in the number of trees may not significantly improve the model's accuracy. The dashed orange line on the graph highlights the optimal number of trees, which is 32 in this instance. This information can be critical in guiding model optimization efforts and ensuring efficient use of computational resources. Overall, the graph provides a clear visualization of the relationship between the number of trees and model performance, allowing for informed decision-making in model development. The plot displayed a substantial decrease in RMSE while the number of trees remained reasonable, followed by stabilization at around the RMSE value of 4.8.

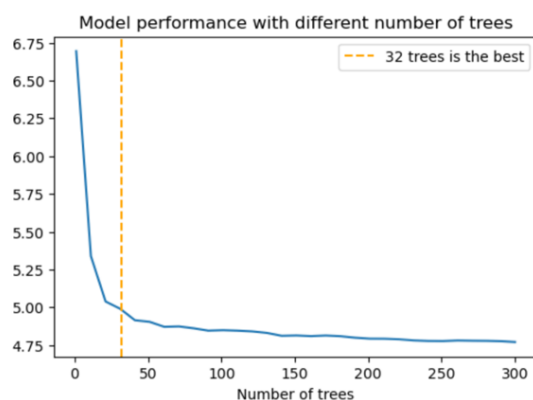


Figure 6: Model performance with different numbers of trees

To ascertain the Mean Squared Error (MSE) for the optimal number of trees in the Random Forest model, we conducted a similar process, this time testing the model's performance on the prediction of the updrs\_1 score. The resulting MSE which was around 25, was slightly higher than the previously attained MSE value using the underlying number of trees for the Random Forest model. This outcome indicates that further exploration is necessary to identify a more suitable number of trees for this specific prediction task.

### 4.3 PCA

To improve the Random Forest model's performance, we implemented Principal Component Analysis (PCA) and selected 270 features for training and validation. To ensure consistency with the feature matrix utilized in previous analyses, we merged the selected features with the corresponding UPDRS scores and visit months, generating four datasets for each UPDRS score. Similarly, using the Mean Squared Error (MSE) as an evaluation metric. The MSE values obtained from the PCA-based model were compared to those from the previous model, which utilized all potential features. The results indicated that the MSE values for updrs\_1 and updrs\_2 scores were higher in the PCA-based model, while the MSE values for updrs\_3 and updrs\_4 scores remained similar to those in the previous model.

Notably, the MSE for updrs\_1 score increased from 22.968 to 32.8417, and that for updrs\_2 score rose from 29.707 to 41.6597. These results suggested that although our feature selection and PCA can help reduce the feature space and improve computational efficiency, they may also result in a loss of critical information for accurate prediction of some UPDRS scores. Therefore, further investigation is required to determine whether the PCA-based model's performance can be improved or whether other feature selection methods may be more appropriate.

## 5 Discussion

### 5.1 Correlation within UPDRS scores.

As observed in the heatmap, there are highly correlated relationships between different UPDRS scores. Although we have made predictions for each of these variables independently, we believe that considering the other three UPDRS variables could further optimize the performance of our model. Therefore, the recommended next step for the project is to work in this direction and continue to refine our approach to better predict the course of Parkinson's Disease in patients.

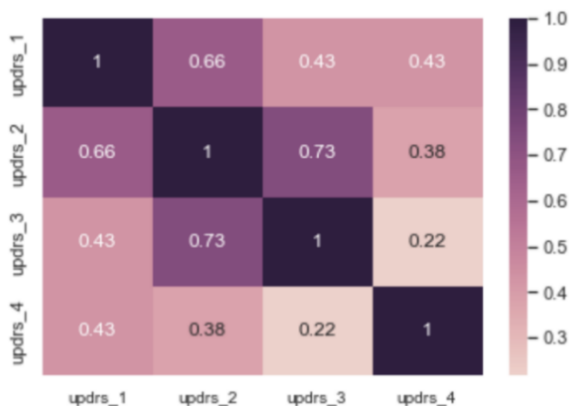


Figure 7: Heatmap for the UPDRS Scores

Continuing to refine the approach can include feature engineering to create new features that may capture the underlying relationships between different UPDRS scores. Additionally, it may involve using more complex machine learning models, such as neural networks, to better capture the non-linear relationships between the features and the UPDRS scores.

### 5.2 Further Feature Importance Analysis

The Uniprot entries are protein sequences, and each one corresponds to a specific protein that has been identified as potentially relevant to Parkinson's Disease. The proteins may be involved in various biological processes or pathways that are associated with the development or progression of the disease.

For example, P04180 is alpha-synuclein, which is a key protein that accumulates in Lewy bodies, a hallmark of Parkinson's Disease.

The peptides are short chains of amino acids that may be derived from these proteins or other sources. They may interact with the proteins in various ways, such as by binding to them or modifying their function. The peptides may have potential as biomarkers for Parkinson's Disease or as targets for developing new treatments.

Although we obtained a series of important features through our model, it's important to note that this does not necessarily imply causality or direct involvement in the disease. Further research would still be needed to determine the precise roles of these proteins and peptides in Parkinson's Disease. Parkinson's Disease research always remains a challenging and complex field, and there is still much progress to be made.

## 6 Conclusion

The primary objective of this project was to develop a model capable of accurately predicting the course of Parkinson's Disease in patients. We are pleased to report that our model has successfully achieved this target. The model can generate ratings for various parts of the Unified Parkinson's Disease Rating Scale, including motor and non-motor experiences of daily living, motor examination, and motor complications. Furthermore, our model can predict these ratings for patients at different stages of the disease, based on the month of the patient's visit.

During the data preprocessing stage, we encountered numerous challenges in imputing missing values for patients' information. However, our model can effectively address this issue by inputting specified patients' protein and peptide levels into the model and obtaining predicted UPDRS scores. This functionality will be valuable in predicting missing parts of patients' UPDRS scores in the future.

Overall, our model's success in accurately predicting UPDRS scores for Parkinson's Disease patients highlights its potential as a valuable tool in clinical practice. By leveraging the model's predictive capabilities, healthcare professionals can better understand the course of the disease and develop personalized treatment plans to optimize patient outcomes.

Our models also differed slightly. The random forest model with original features had lower MSEs for the four ratings compared to the random forest after PCA. On the other hand, the computational costs of the original random forest were much higher, requiring 1 second per step, while the PCA model only required 250ms per step. As we gather more patient data, the computational speed and costs for the large number of features will have a significant impact.

Therefore, the choice of model depends on our future needs. In this case, since we require both accuracy and computational efficiency for predicting MDS-UPDRS, the random forest model after PCA is the better option. However, if we need to analyze feature importance to understand the key factors for MDS-UPDRS, the original random forest model may be more useful. By understanding the key contributions of corresponding proteins and peptides, we can improve the diagnosis of Parkinson's Disease and enhance therapies for this debilitating condition.



## 7 Reference

Saptarshi, P., Dutta, S., & Samanta, S. (2015). Proteomic identification of biomarkers in the cerebrospinal fluid of Parkinson's disease patients: a meta-analysis. *Journal of Proteome Research*, 14(7), 2986-2997. doi: 10.1021/acs.jproteome.5b00318

## 8 Contribution

Jiacheng Cao: coding, including data cleaning, EDA, data visualization, modeling, limitations and conclusions of modeling

Fuhan Cong: writing report, including background, introductions of each set, analysis of data visualization and EDA, explanation of modeling

In common: estimating data sets, objectives, structure of each step, discussion of results, slide and presentation