Accelerating Medicines Partnership Parkinson's Disease (AMP-PD) Progression Prediction

Analysis of Protein and Peptide Levels Over Time to Predict Progression of Parkinson’s Disease

Team Members:
Ruixin Lou
Emmanuel Ruhamyankaka
Sukhpreet Sahota
Jiaxin Ying

Team Number: 10

Abstract

This should be the one paragraph that captures the significance of what you did and why you did it - this should be a summary of the work and your outcomes in brief.

Parkinson's Disease is a neurodegenerative disorder that affects millions of people worldwide. The progression of the disease is commonly measured using the Movement Disorder Society - Unified Parkinson's Disease Rating Scale (UPDRS) scores, which assess motor and non-motor symptoms. The Accelerating Medicines Partnership Parkinson's Disease (AMP-PD) Progression Prediction Kaggle competition seeks to discover unforeseen patterns and predict the progression of Parkinson's Disease using clinical UPDRS scores and peptide and protein abundance data captured from cerebrospinal fluid (CSF) samples. The AMP-PD Progression Prediction project aims to add to the emergence of data techniques, such as machine learning and artificial intelligence, in the healthcare sector, particularly Parkinson's Disease diagnosis. Using four models (Support Vector Machines, Random Forests, Gradient Boosting, and Decision Trees) as the foundation for our project, the models were iteratively optimized and trained. An analysis of the models revealed the Support Vector Machine was the best base model as it consistently performed the best across each iteration, making it a promising model for predicting Parkinson's Disease progression. Further experimentation in the study revealed that an ensemble stacking method would also be beneficial, suggesting that a combined model approach could further improve the accuracy of predicting the progression of Parkinson's Disease. Using both of these optimal-performing models on our test data revealed the ensemble stacking model garnered the best accuracy (confidence) of 57.2% in predicting the UPDRS scores. While this may seem extremely low, the result symbolically represents the complexity of the Parkinson's disease identification problem and the complexity of using data and advanced data techniques as the sole determinant (or source of discovery) within medicine. This study provides a detailed walkthrough of crucial assumptions made, the development of the predictive models, and the limitations for predicting UPDRS scores to determine Parkinson's Disease progression using peptide and protein abundance data from CSF samples. Analysis and findings from this data science competition can potentially lead to the development and breakthroughs of new medicines and medical techniques, such as specific pharmacotherapies, that aim to slow the progression or cure Parkinson's Disease. For the future expansion of this work, we recommend collaborating with medical experts on a larger clinical and protein dataset to adjust the model with expert insight while scanning more molecular data.

I. Introduction

Provide a description of the problem and the value in finding a solution, and motivate your reader as to why they should care about your problem or question.

Steve Zabielski had what one would describe as the perfect life: he was a successful lawyer who enjoyed the pleasures of life with his wife and loved ones, vacationing in exotic locations, such as the Caribbean, while also being a great, active father to his children, playing sports and mountain biking at every opportunity. On a seemingly ordinary night at the office, as Steve prepared to depart, he felt a sensation in his right arm that he noted but didn't think much of. A month later, he noticed diminished dexterity in his right hand but dismissed anything extreme again and considered it another coincidence. A few days later, while out for a run, he knew something was genuinely wrong. He felt abnormally fatigued and experienced unfamiliar pains and stiffness in his right shoulder, immediately followed by involuntary twitching in his right arm. "Steve grimaced at the thought. 'Now I know I have to see a doctor. So I go on my own, not wanting to alarm anyone. And very quickly, one thing leads to another. And then a doctor tells me, 'I think you may have Parkinson's.'" (Penn Medicine, [22])

Parkinson's Disease is a progressive neurological disorder affecting millions, like Steve and famed actor Michael J. Fox. According to the Parkinson's Foundation, more than 10 million people worldwide have been diagnosed with Parkinson's Disease, 1 million of whom currently reside in the United States (Parkinson's Foundation, [21]). As shown in the figure below, a 2016 study conducted by the Global Burden of Disease (GBD) that was funded by the Bill and Melinda Gates Foundation found that "neurological disorders were the leading source of disability globally," of which Parkinson's Disease was the fastest growing (GBD 2016 Parkinson's Disease Collaborators, [7]).

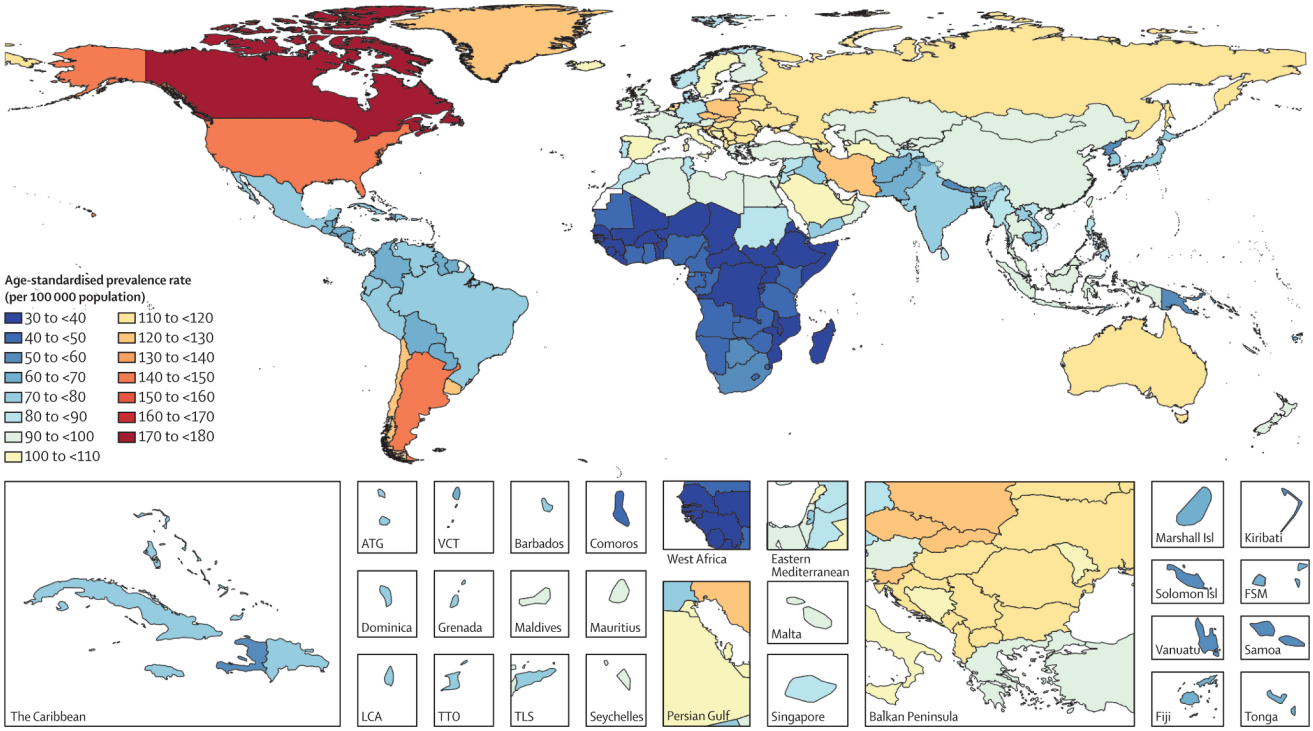


Figure 1: Age-standardized prevalence of Parkinson's Disease per 100,000 population by location for both sexes, 2016 (Source: Global Burden of Disease)

The Parkinson's Foundation extrapolated these data points and discovered that 90,000 new Parkinson's Disease diagnoses are made in the US annually. At that rate, 1.2 million individuals will be diagnosed with the disease by 2030 (Appendix Exhibit 1). Additionally, it has been found that Parkinson's Disease does affect elderly individuals, with early onset of the disease considered to be in the 50s and peak diagnoses occurring between the ages of 85-89, disproportionately as men are found to be suffering from Parkinson's more than women (Appendix Exhibit 2). As illustrated through Steve's story, several key symptoms of the disease include tremors, stiffness, and balance problems, which can significantly impact a patient's quality of life. While medications can alleviate some symptoms, there is currently no cure (Jankovic, 2008 [11]).

To address this urgent need for new treatments and to better understand the progression of the disease, the Accelerating Medicines Partnership Parkinson's Disease (AMP-PD) through the Foundation of the National Institutes of Health (FNIH) organized a competition that would use advanced machine learning algorithms to predict Parkinson's Disease progression and the molecular make-up [1]. This competition adds to the growing use of advanced data techniques like machine learning and artificial intelligence within the healthcare sector. For example, a review of machine learning applications on Parkinson's Disease by Mei, J., Desrosiers, C., & Frasnelli, J. (2021 [15]) found "previous studies have reviewed the use of machine learning in the diagnosis and assessment of PD, they were limited to the analysis of motor symptoms, kinematics, and wearable sensor data." The goal of

the competition is to go beyond this limited use to date and to train a machine learning model(s) that could accurately predict MDS-UPDRS scores by analyzing the biological make-up of patients, specifically the protein and peptide levels over time in patients with Parkinson's Disease compared to normal age-matched control subjects. By leveraging the power of machine learning, this competition has the potential to uncover new insights into Parkinson's Disease progression. The understanding of Parkinson's Disease progression from this competition can be seen as one of the preliminary studies to use machine learning to identify and validate specific protein and peptide biomarkers for Parkinson's Disease progression, which is seen by many as a significant step toward finding a cure and significantly improving the quality of life for those affected by this disease: "it is vital to find reliable molecular biomarkers that can distinguish PD from other conditions, monitor its progression, or give an indication of a positive response to a therapeutic intervention. PD biomarkers can be subdivided into four main types: clinical, imaging, biochemical, and genetic" (Emamzadeh & Surguchov, 2018 [6]). By identifying reliable biomarkers, physicians can diagnose Parkinson's Disease earlier and potentially provide more effective treatments to patients. Analysis and findings from this data science project can also potentially lead to the development and breakthroughs of new medicines and medical techniques, such as specific pharmacotherapies, that aim to slow the progression or cure Parkinson's Disease.

II. Background

This section should cite work that has been previously addressed that relates to your project, and the key takeaways of those studies/projects.

Parkinson's Disease has mystified doctors for decades, from understanding all symptoms related to the disease to the genetic components of the disease. To identify distinctive biomarkers for Parkinson's Disease, doctors and researchers have used various techniques in their medical toolbox, including imaging studies, genetics studies, and studies analyzing protein and peptide level changes. As a result of this curiosity and to better measure the progress of the disease within a victim, The Movement Disorder Society revised the well-established Unified Parkinson's Disease Rating Scale (UPDRS) in 2008, resulting in the new MDS-UPDRS, which consists of four separate parts assessing non-motor and motor aspects of daily living experiences, motor examination, and motor complications:

Components of MDS-UPDRS
Part I - Non-Motor Aspects of Experiences of Daily Living (nM - EDL)
Part II - Motor Aspects of Experiences of Daily Living (M-EDL)
Part III - Motor Examination
Part IV - Motor Complications (see Reference [8] for the MDS-UPDRS Scale)

To gauge one's motor skills and evaluate the motor symptoms of a Parkinson's Disease, a corresponding unified framework was developed, consisting of 5 unique aspects: 1) Action Selection, 2) Motor Coordination, 3) Motor Execution, 4) Motor Planning & Sequencing, and 5) Perceptual Input (as seen in the figure below) (Moustafa, 2016 [17]).

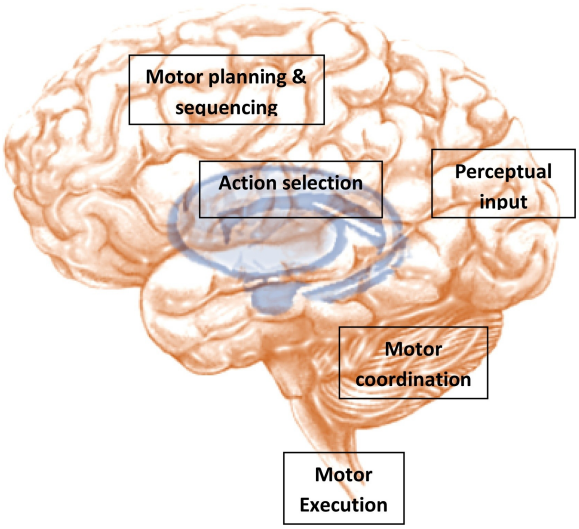


Figure 2: Motor performance relies on elemental motor processes, including action selection (basal ganglia), sequencing and planning of motor actions (cortical motor regions), and motor coordination and timing (cerebellum), among others (Moustafa, 2016 [17])

As Parkinson's Disease continues to become more prevalent in our society and doctors have been able to associate specific symptoms, there is a desire to discover the underlying factors and components to detect the disease earlier within patients. Medicine has always been a field that has been conservative in approach, and thus has become outdated and an expansive area that many have been intrigued to transform. With the evolution of technology, many assert machine learning as a possible path in combining the two areas.

As stated by Mei, J., Desrosiers, C., & Frasnelli, J. (2021 [15]), "Machine learning techniques are being increasingly applied in the healthcare sector. As its name implies, machine learning allows for a computer program to learn and extract meaningful representation from data in a semi-automatic manner. For diagnosing PD, machine learning models have been applied to a multitude of data modalities, including handwritten patterns, movement, neuroimaging, voice, cerebrospinal fluid (CSF), cardiac scintigraphy, serum, and optical coherence tomography (OCT). Machine learning also allows for combining different modalities, such as magnetic resonance imaging (MRI) and single-photon emission computed tomography (SPECT) data, in diagnosing PD. By using machine learning approaches, we may identify relevant features not traditionally used in the clinical diagnosis of PD and rely on these alternative measures to detect PD in preclinical stages or atypical forms." Additionally, several studies have analyzed changes in protein and peptide levels to identify biomarkers for Parkinson's Disease. For instance, Shi et al. (2015, [27]) used a multiplex immunoassay to measure protein levels in the cerebrospinal fluid (CSF) of Parkinson's Disease patients. The study identified a panel of biomarkers that could accurately distinguish Parkinson's Disease patients from healthy controls [27]. Another study by Mollenhauer et al. (2017, [16]) analyzed changes in CSF biomarkers over time in Parkinson's Disease patients. The study found that changes in CSF alpha-synuclein levels correlated with disease progression and could be used as a biomarker for monitoring disease progression. There have also been previous studies that have used machine learning approaches to predict disease progression in Parkinson's Disease. For example, one study by Severson et al. (2021, [26]) developed a statistical progression model that can account for intra-individual and inter-individual variability and medication effects to gain insights into Parkinson's Disease progression. Some research has also pushed the envelope in determining components to Parkinson's Disease using unprecedented means, such as using "artificial intelligence models on breathing patterns to detect Parkinson's Disease" [18]! In this regard, machine learning approaches have consistently been refined to accurately define disease states associated with particular

symptom manifestations and progression patterns, which can have clinical applications in patient counseling, prognostication, and clinical trial design.

Our proposed project aims to utilize the AMP-PD Knowledge Platform dataset provided by the Foundation on Kaggle to understand the progression of Parkinson's Disease using machine learning methods on biomarkers such as peptides and proteins. We plan to use several machine-learning techniques to predict MDS-UPDRS scores in Parkinson's patients using the data provided. The range for each UPDRS part is outlined below:

Score Ranges for each UPDRS Part (mild/moderate and moderate/severe cutoff levels respectively)
Part I - 0 - 52 *(10/11; 21/22)*
Part II - 0 - 52 *(12/13; 29/30)*
Part III - 0 - 144 *(32/33; 58/59)*
Part IV - 0 - 24 *(4/5; 12/23)*
Total - 0 - 272 (see References [2] and [28] for MDS-UPDRS Score Ranges)

Unlike previous studies that have solely focused on statistical models and peptide/protein levels, our project and this data science competition will be unique in approach as it combines this information with clinical data and analyzes it all together, potentially contributing to a better understanding of the disease and ultimately leading to improved treatments for patients using UPDRS scores. The project aligns with the need to identify early biomarkers for Parkinson's Disease and the importance of developing new therapeutic strategies to prevent or slow disease progression. We truly hope our efforts will contribute to the larger scientific community's ongoing efforts to improve the diagnosis and treatment of Parkinson's Disease.

III. Data

Describe and visualize your data in the context of the problem you are working on.

The success of a machine-learning technique and model-based approach largely depends on the data. It is essential for the data utilized to solve a problem to be representative of the population of the domain of the problem.

As part of the Kaggle competition, the AMP-PD Knowledge Platform dataset for this study has been provided. It consists of 3 distinct parts: clinical trial data, peptide data, and protein data. These three parts are each subdivided into a training set and a test set. As the Kaggle competition website notes, “The core of the dataset consists 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 multiple years while taking assessments of PD severity” [1]. Taking this notation as a background, we further examined our training data and found that the datasets related to 2,615 clinical visits made by 248 unique patients.

The unit of observation within the clinical trial data is a mass spectrometry reading made for each clinical visit made by each respective patient. It captures the following data points: the Patient ID, the month visited, UPDRS Score 1, UPDRS Score 2, UPDRS Score 3, UPDRS Score 4, and whether the patient is on medication (see Data Dictionaries within Appendix). It is equally important to note that these mass spectrometry readings were taken every three months for the patient, with patients participating in the clinical trials ranging between 6 months to 108 months (or nine years). A key assumption we are making is that patients have visited the clinic for every designated check and have not missed any scheduled appointment from their first visit to the last recorded visit (e.g., if the patient has visited the clinic for six years, the patient did not miss any appointments over those six

years). We have made this key assumption to help understand and interpret the data as we are not working with a large dataset.

As seen in the distribution below, we note a few key observations:

1) More than a third of the patients observed within the study had clinical readings taken and were part of the study between two to three years - which is a significant period to note changes within the patients' motor and non-motor skills.
2) There are two main concentrations of patients, with the first group of patients (boxed in green) being centered around 24 months (2 years) and the second group of patients (boxed in gray) being centered around 84 months (7 years). These two groups can be described as recently observed patients and established patients, respectively.

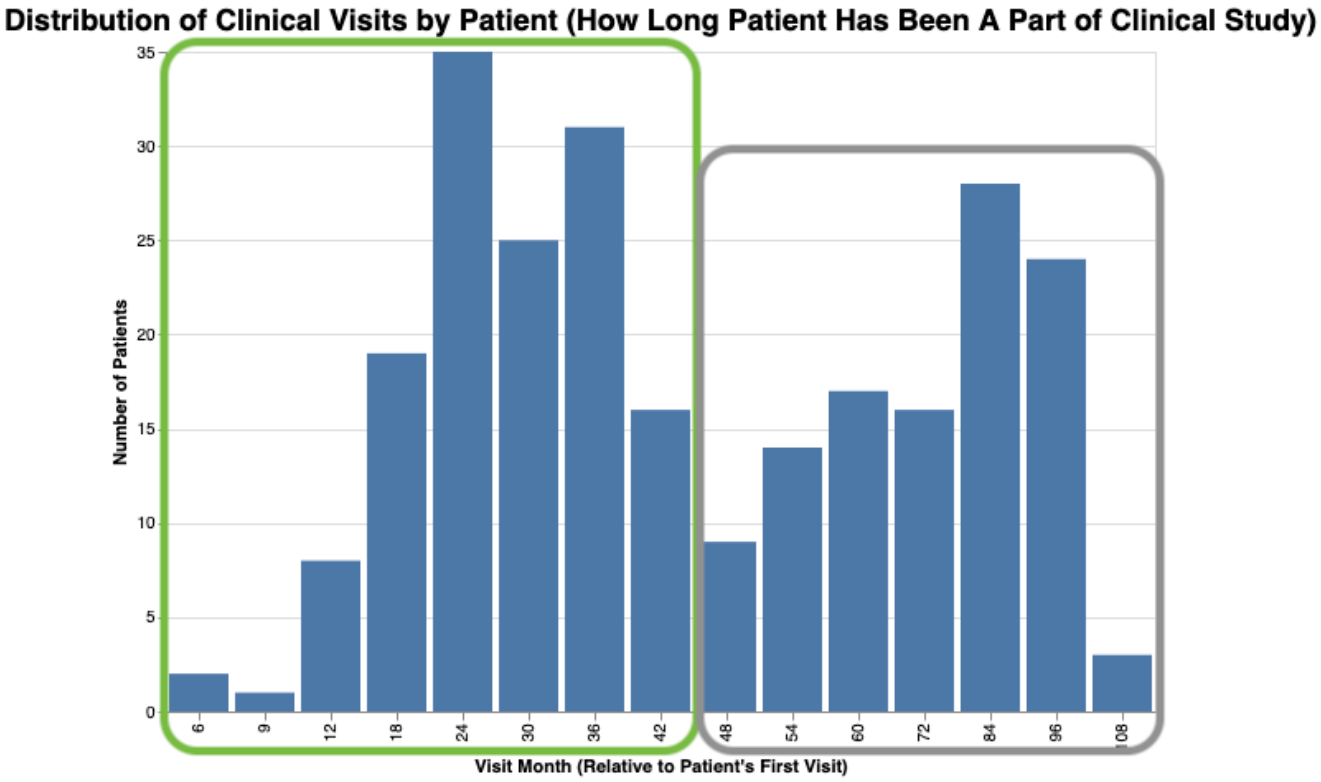
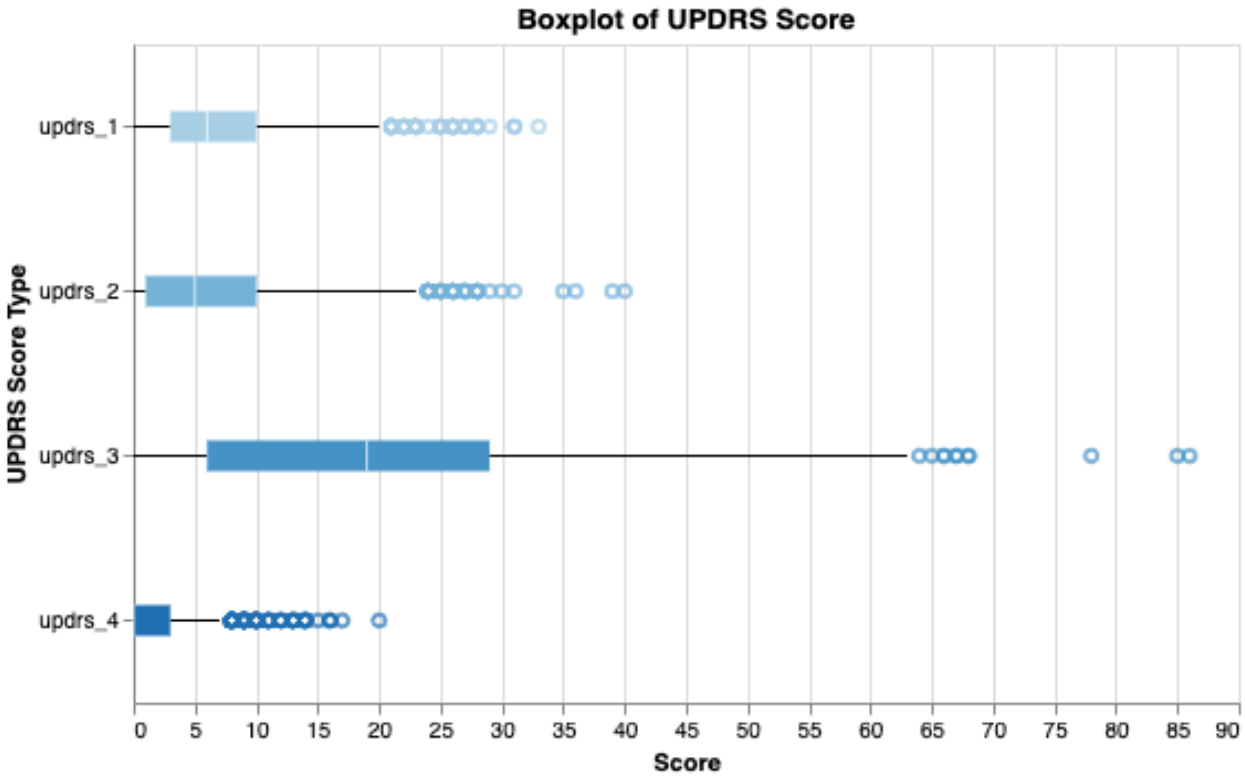


Figure 3: Distribution of Clinical Visits by Patients shows that more than a third of patients had been part of the study between 2-3 years, and more than a fifth of patients had been part of the study for more than seven years

Each visit and each reading noted by the doctors provided insights into the patient's personal Parkinson's Disease journey. As shown in the boxplot and associated table below, the UPDRS Scores over the 2,615 total visits collectively demonstrated that patients scored higher for Part III (Motor Examination) with a mean score of more than 19. An analysis of the UPDRS also revealed a low average score for Part IV (Motor Complications); however, the data is heavily skewed compared to the other three parts due to a high percentage of missing observations/data points:



	count	mean	std	min	25%	50%	75%	max
Month Visited	2615	31.19	25.20	0	10.5	24	48	108
UPDRS Score 1	2614	7.11	5.53	0	3	6	10	33
UPDRS Score 2	2613	6.74	6.32	0	1	5	10	40
UPDRS Score 3	2590	19.42	15.00	0	6	19	29	86
UPDRS Score 4	1577	1.86	3.02	0	0	0	3	20

Figure 4: Boxplot of UPDRS Scores shows that average scores recorded across the 2,615 visits

There was also curiosity to investigate and better understand these UPDRS scores. Our exploratory analysis (EDA) continued with a further drill-down to understand two distinct questions: 1) is there a relationship between the UPDRS scores that can be discovered through these clinical readings, and 2) is there a difference in UPDRS score whether the patient was on medication or not? While we may not conclusively answer these questions for some scenarios, the associated EDA may help provide insights regarding these two aspects of our project.

Analyzing the correlation between the UPDRS scores will enable us to see if progression to one part measured affects the progression of another. As shown in the correlation matrix below, the highest correlation value of 0.74 is between UPDRS Score 2 and UPDRS Score 3, with the next highest correlation value being 0.66 between UPDRS Score 1 and UPDRS Score 2. Overall, there is little correlation between the individual UPDRS scores. This implies that each score/area measured offers unique information to medical experts for understanding the progression of Parkinson's Disease.

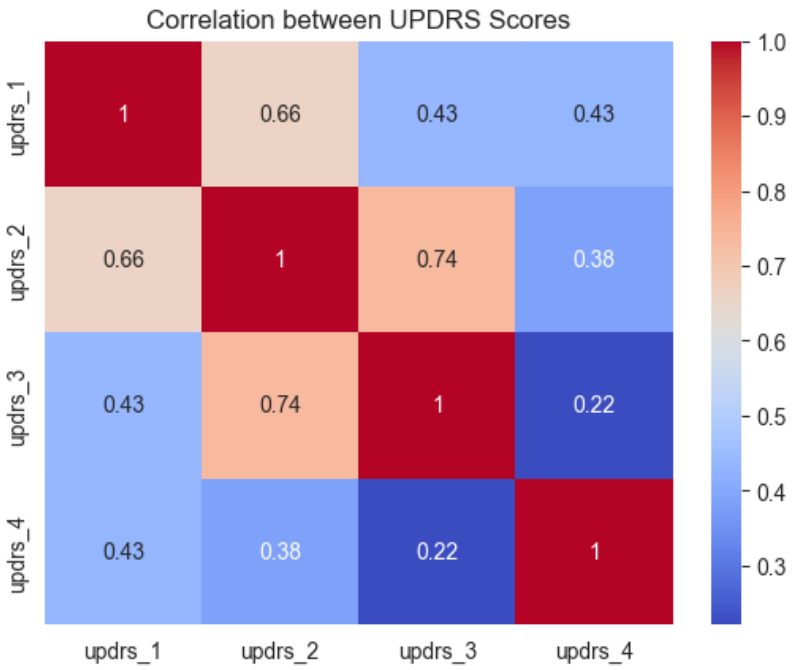


Figure 5: Correlation Matrix between UPDRS Scores shows that the areas measured by the UPDRS are not highly correlated

At a high level, an analysis of the presence of medication on the UPDRS scores shows that patients on medication score lower than patients who are not on medication (which would be the hopeful intent). As seen in the boxplot below, the median scores across all 4 UPDRS score types are lower for patients on medications (with an obvious example being for UPDRS Score 2 and UPDRS Score 3 (which had a lower score of ~8 points) - boxed in red).

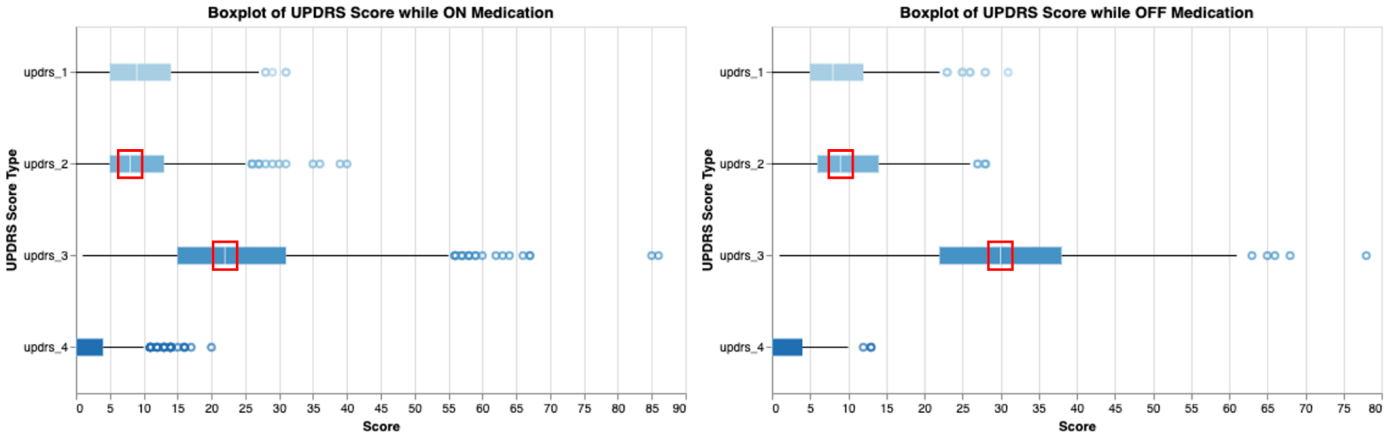


Figure 6: Side-by-side Boxplot Comparison of UPDRS Scores of whether patients are on or off medication

Additional analysis and breakdown of each UPDRS score by these factors were also investigated and have been included in the Appendix (such as the distribution of each UPDRS score, the distribution of patients being on medication, the score range for each UPDRS score by month and whether on medication, the total overall UPDRS score by month and whether on medication).

With a good foundational understanding of the clinical data, it is important to shift the viewpoint of the study and understand the molecular datasets (peptide and protein data). As we began to examine the peptide and protein

data, we found that the datasets relate to 227 unique protein structures that are associated with 968 different peptide bonds. According to Shi et al. (2015, [27]), "Finding robust biomarkers for Parkinson's disease (PD) is currently hampered by inherent technical limitations associated with imaging or antibody-based protein assays. To circumvent the challenges, we adapted a staged pipeline, starting from our previous proteomic profiling followed by high-throughput targeted mass spectrometry (MS), to identify peptides in human cerebrospinal fluid (CSF) for PD diagnosis and disease severity correlation." To this extent, capturing the frequency of protein structures (particularly the ten most and least frequent proteins present for each respective molecular dataset) is important. The frequency of proteins present within both datasets is shown in the comparison tables below:

Peptide Dataset		Protein Dataset	
Top 10 Most Frequent Protein Structures	**Top 10 Least Frequent Protein Structures**	**Top 10 Most Frequent Protein Structures**	**Top 10 Least Frequent Protein Structures**
P01009	O75326	P01011	O75326
P01024	P01780	P01023	P01780
P02647	P02655	P01024	P02655
P02649	P06310	P01042	P06310
P02751	P19827	P01834	P19827
P02768	P36980	P02790	P36980
P02774	Q562R1	P05090	Q562R1
P02787	Q6UX71	P23142	Q6UX71
P08603	Q99829	Q92520	Q99829
P10909	Q99832	Q9UHG2	Q99832

Table 1: Table outlining the top 10 most frequent and least frequent protein structures identified within the Peptide and Protein Dataset

Similar to the clinical data, we did the additional analysis on the molecular data, focusing on the NPX (Normalized Protein Expression) to understand how the frequency changed over patient visitations, which has been included in the Appendix.

IV. Experiments

Present your machine learning experiments (for supervised learning, a description of any preprocessing, feature extraction, classification/regression techniques, experimental designs, and evaluation criteria) and why you made each of the choices you did to achieve your goal. Also, include a flow chart of your methodology so the reader can easily conceptualize your solution. Describe your approach to measuring generalization performance, what metric(s) you used and why.

Now that we have an excellent foundational understanding of our datasets, our methodology to understand the progression of Parkinson's Disease involves merging of our data, feature engineering using a scalar, optimizing hyperparameters, and upon getting the optimal hyperparameters for each model, evaluating the test data for the lowest error. Below is a flowchart summarizing our methodology:

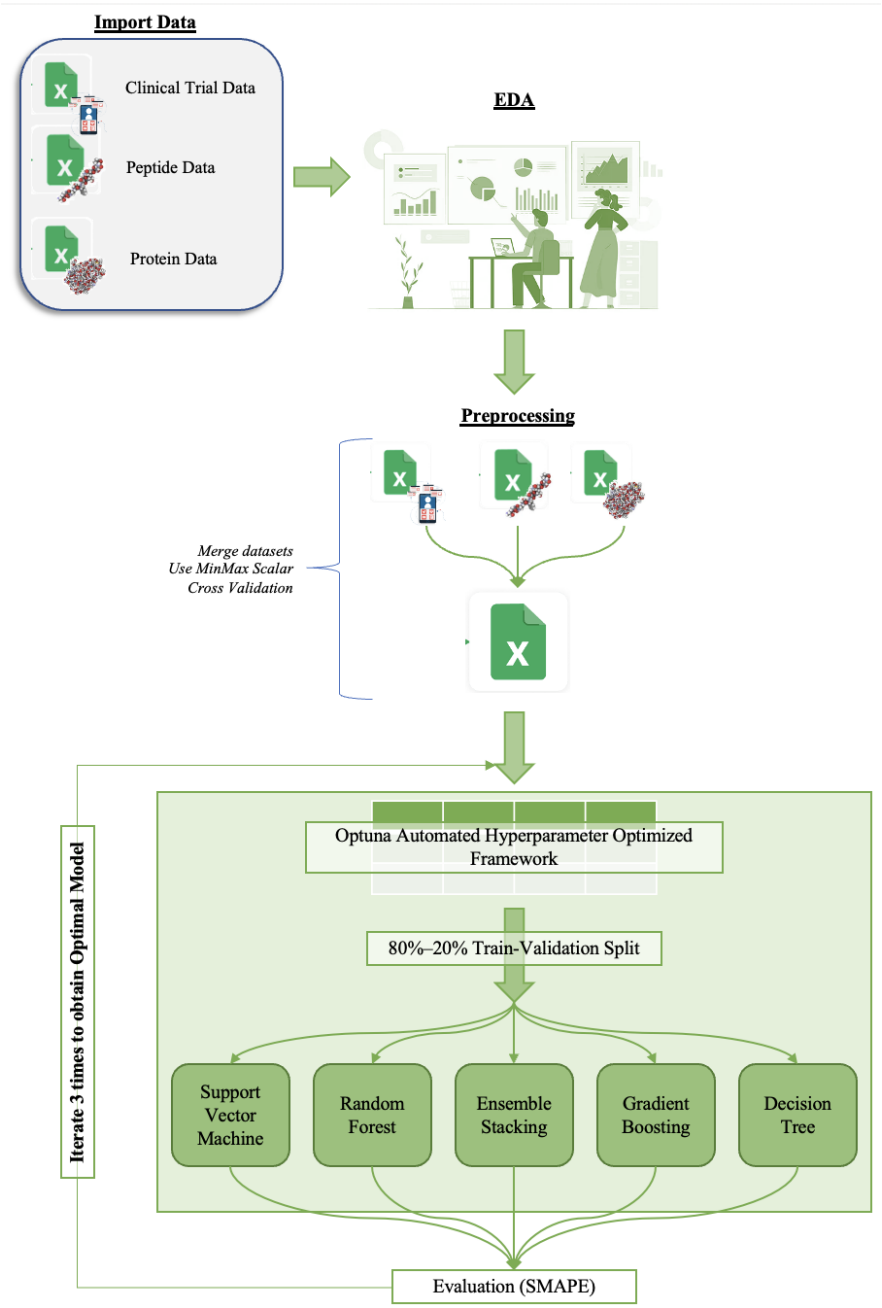


Figure 7: Flowchart of Steps Involving ML Techniques to Predict Parkinson's Disease Progression

We merged all three datasets, including the peptide and protein data, into a single dataframe using Patient ID and the respective clinical visit month as the common identifiers. This allowed us to analyze the patient data, leveraging all available data points. Before each experiment iteration, we pre-processed the data by applying a MinMaxScaler to scale the numerical features within a range of 1s and 0s.. We chose the MinMaxScalar since the models we implemented are sensitive to the scale of input features. Their performance can be affected if the features are on different scales. MinMax scaling ensures that all features have the same scale, which can improve the performance of the chosen models and help them find patterns in the data more efficiently. Additionally, a MinMaxScaler does not assume a normal distribution of the data, making it more appropriate for this case than StandardScaler, which assumes a normal distribution. Lastly, using the MinMaxScalar ensured that all features were based on a similar scale and prevented any one feature from exerting a disproportionate influence on the others.

Once the features were scaled, we used a 10-fold split to evaluate the performance of the models, in which the models were trained on 9 portions of the data and then tested on the remaining portion. This method was repeated 10 times with different evaluations before being averaged to obtain a more reliable estimate of each model's performance. By performing feature scaling and cross-validation, we ensured that the models were trained and evaluated on the most relevant features. We evaluated the performance of four base models for predicting the progression of Parkinson's Disease: Decision Trees, Gradient Boosting, Random Forest, and Support Vector Machine. We also utilized a fifth model in ensemble stacking. This technique combines the predictions of multiple base models (the four base models identified above) to produce a more accurate prediction. Overall, ensemble stacking is an effective way to improve the accuracy of machine learning models, as it allows us to leverage the strengths of multiple individual models. To optimize the performance of the models, we utilized the Optuna Automated Hyperparameter Optimization Framework, a state-of-the-art open-source imperative hyperparameter optimization framework API that automates hyperparameter search through efficient algorithms and parallelization. As a result of these attributes, this widely used framework is faster than traditional hyperparameter optimization methods, such as random search and grid search. The hyperparameters optimized for each model are outlined in the table below (with additional insights on the importance level of each hyperparameter for each respective model highlighted in the Appendix):

Decision Tree Hyperparameters	Gradient Boosting Hyperparameters	Random Forest Hyperparameters	Support Vector Machine Hyperparameters
Maximum Number of Features (max_features)	Number of Trees (n_estimators)	Number of Trees (n_estimators)	Kernel
Maximum Depth of the Tree (max_depth)	Learning Rate (learning_rate)	Maximum Number of Features (max_features)	Regularization
Minimum Number of Samples Required for Internal Node (min_samples_split)	Maximum Depth of the Tree (max_depth)	Maximum Depth of the Tree (max_depth)	Gamma
Minimum Number of Samples Required for Leaf Node (min_samples_leaf)	Subsampling Rate (subsample)	Minimum Number of Samples Required for Internal Node (min_samples_split)	
	Maximum Number of Features (max_features)	Minimum Number of Samples Required for Leaf Node (min_samples_leaf)	
		Bootstrap Sampling Strategy (bootstrap)	

Table 2: Set of hyperparameters tuned for each selected model

After conducting the Optuna optimization process, we obtained the optimal hyperparameters for each model, which were then used to train the models on the training data. We split the training data into two subsets: 80% for training and 20% for validation. We ran 100 trials for each model, using a random sampler to choose the best hyperparameters that minimized the Symmetric Mean Absolute Percentage Error (SMAPE) score. SMAPE is a metric that calculates the symmetric mean absolute percentage error between the predicted and actual values. Since the actual UPDRS scores can vary significantly in magnitude (0 to 272) (Holden et al., 2017 [9]), SMAPE is more suitable for evaluating the accuracy of the predictions across different UPDRS scores than other absolute error metrics (i.e., mean squared error (MSE) or root-mean-squared error (RMSE)) as well as the one preferred to be used by the AMP-PD for the competition. As aforementioned before, our ultimate goal was to develop models that can accurately predict the UPDRS scores that enabled medical professionals to understand the progression of Parkinson's Disease using peptide and protein abundance data from CSF samples. Minimizing the SMAPE score ensured that the models accurately predicted the relative error in the UPDRS scores, which is essential for clinical decision-making and patient care.

As demonstrated by the distribution of the clinical patients (Figure 3), the above methodology and approach were implemented in two distinct experiments:

1) Creating a model trained on the UPDRS scores, clinical data, peptide, and protein levels across all 248 patients collectively. This would enable us to assess the two groups of patients (recently observed and established patients) together and garner input from both groups. Other than being an apparent approach,

this is ideal as the models generated would be able to analyze patients in both the short term and long term.

2) Creating two separate models trained respectively on the UPDRS scores of each group of patients (one model based solely on recently observed patients and the other based solely on established patients). As aforementioned, within the distribution analysis, it is important to analyze the time a patient has been part of the clinical trial and understand whether visiting the clinic impacts the results/progression of Parkinson's Disease. It will also let us see if the models better predict UPDRS scores in the short- or long term.

V. Results

Include a complete performance assessment that includes your validation approach (cross-validation, train/validate/test split, etc.) and the key performance metrics for the problem (ROC curves, PR curves, confusion matrices if applicable, etc.). You should also compare your outcomes to at least one baseline model to act as a point of reference for interpreting the results of your work as well as chance performance (i.e., random guessing for classification and guessing the mean/median for regression).

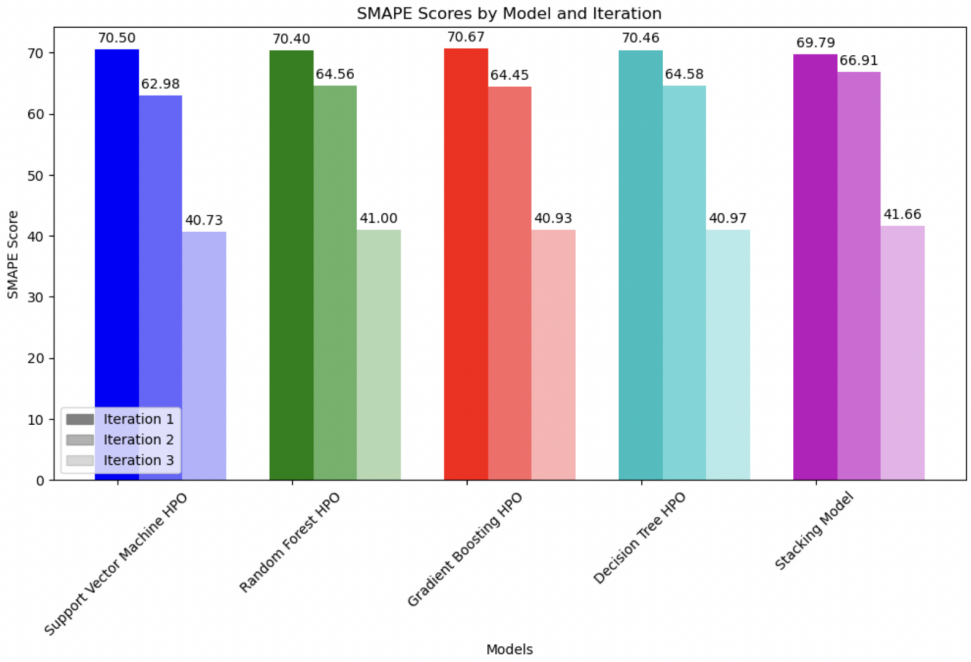
After conducting the preprocessing and validation approach described in the Experiments sections, we performed the following steps for each of the two distinct experiments highlighted above:

1) We ran three iterations of hyperparameter optimization and model training to develop our five models: Support Vector Machine, Random Forest, Gradient Boosting, Decision Tree, and Ensemble Stacking.
2) Upon obtaining each model, we evaluated the performance of the models by predicting the UPDRS score on the validation data using the SMAPE score as the evaluation metric (with the goal, again, to minimize the SMAPE score, which indicates the accuracy of the predictions of the relative error in the UPDRS scores).

Experiment #1 Results

As shown in the graph and table below, the results (SMAPE) of each iteration for the first experiment, in which we collectively used the data on all patients to train the model, highlight a few significant insights and trends:

1) Across all models, the performance improved with each iteration, which meant that the Optuna Automated Hyperparameter Framework was able to continuously get better parameters to optimize performance.
2) The Support Vector Machine performed consistently well across all iterations, making it a promising model for predicting Parkinson's Disease progression.
3) Additionally, the Ensemble Stacking method, which uses multiple models to produce an optimal model, outperformed the individual base models in the first iteration, suggesting that combining different models can further improve the accuracy of predicting.



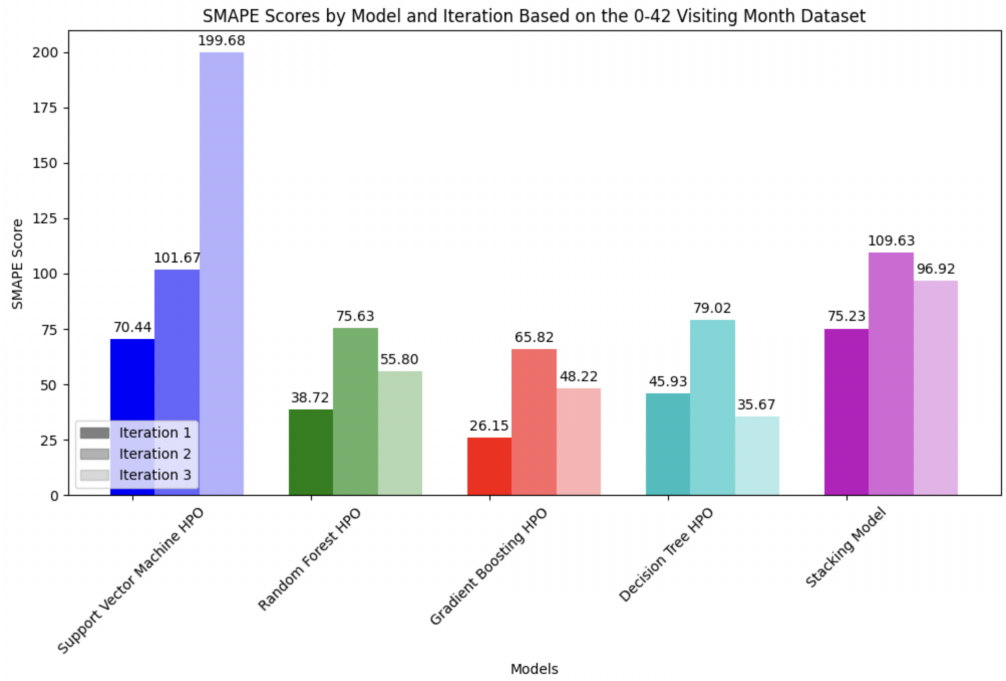
	Decision Tree	Gradient Boosting	Random Forest	Support Vector Machine	Ensemble Stacking
1st Iteration	70.4605	70.6022	70.4594	70.4969	70.0972
2nd Iteration	64.5829	64.7190	64.4017	62.9808	64.1842
3rd Iteration	40.9737	40.9467	40.9717	40.7315	41.8176

Figure 8: Experiment #1 SMAPE Scores by Model for each Iteration. For the first iteration, Random Forest performed the best of the 4 base models, as it had the lowest SMAPE score of 70.4594; however, the Ensemble Stacking Model performed the best overall with a SMAPE score of 70.0972. For the next two iterations, Support Vector Machines performed the best from all of the models (including the Ensemble Stacking Model) as it had the lowest SMAPE score of 62.9808 and 40.7315, respectively (highlighted in the table above)

Experiment #2 Results

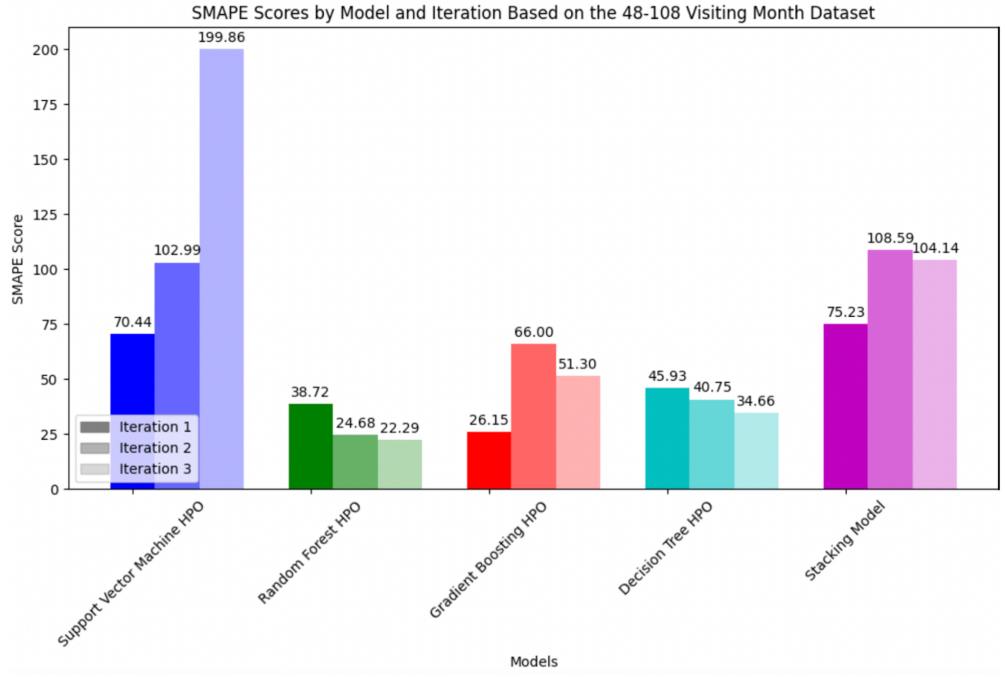
To gain a deeper understanding of the impact of patient visit duration on model performance, we conducted a second experiment after identifying the two best-performing models from the first experiment. We divided the patients into two groups based on the duration of their participation in the clinical trials, as determined by the data distribution (0-42 months and 48-108 months). Surprisingly, we found that while some optimal models performed well in both groups, some outperformed others in each group, as demonstrated in the graphs and tables below. This suggests that patient visit duration may indeed impact model performance, and it is important to consider this factor when selecting a given model for a given group of patients:

1) For the newly observed patients (group of patients between 0 - 42 months), the Gradient Boosting model significantly outperformed the other models for the first two iterations; however, the Decision Tree performed the best for the last iteration.
2) Additionally, for this group of patients, unlike Experiment #1, there was no trend of better performance during every additional iteration for any of the models.
3) For the established patients (group of patients between 48 - 108 months), while the Gradient Boosting model outperformed each of the other models for the first iteration, Random Forest was overall a better model as it had the two lowest SMAPE during all of our experimentation with a SMAPE of 24.6778 and 22.2911 respectively for the following two iterations.
4) Similar to Experiment #1, there was a trend of the performance getting better with each iteration; however, unlike that exemplary experiment where this trend was present for all models, this was only true for Random Forest and Decision Tree for the established patients.



SMAPE Scores by Model and Iteration Based on the 0-42 Visiting Month Split

	Decision Tree	Gradient Boosting	Random Forest	Support Vector Machine	Ensemble Stacking
1st Iteration	45.9318	26.1454	38.7197	70.4389	75.2309
2nd Iteration	79.0187	65.8245	75.6295	101.6725	109.6305
3rd Iteration	35.6715	48.2230	55.8009	199.6760	96.9199



SMAPE Scores by Model and Iteration Based on the 48-108 Visiting Month Split

	Decision Tree	Gradient Boosting	Random Forest	Support Vector Machine	Ensemble Stacking
1st Iteration	30.8382	28.4075	38.3151	66.0485	75.1350
2nd Iteration	40.7495	65.9985	24.6778	102.9859	108.5877
3rd Iteration	34.6593	51.2955	22.2911	199.8568	104.1366

Figure 9: Experiment #2 SMAPE Scores by Model for each Iteration based on the Time Patient has been part of the clinical trial. For both timeframes, each iteration results in a different model performing the best, respectively.

For the newly observed patients, Gradient Boosting performed the best; whereas, for established patients, Random Forest performed the best (highlighted in the tables above)

After running both experiments, we decided that using the model resultants from the first experiment were more applicable to the use case. While we had key takeaways from performing the second experiment, the first experiment was more generalizable as it analyzed and took data from both groups of patients at the same time and performed better over each iteration of the experiment.

Test Results

As outlined in the findings after performing the first experiment, the two optimal models that performed the best on the collective dataset were Support Vector Machines and Ensemble Stacking. We first used Support Vector Machines to predict the UPDRS scores for the provided test data. The test data initially consists of 2 unique patients, and the goal is to predict the UPDRS scores for the first visit and then the visits 6 months, 12 months, and 24 months later (and this accounts for 13% of the test data that is revealed for initial competition positioning). Using the SVM-optimized model, the model garnered a SMAPE of 57.8. Next, we ran the Ensemble Stacking optimized model on the same test data, which amassed a SMAPE of 57.2 and thus performed the best across all models we considered (output is provided in the Appendix Exhibit 8). As the experiment section highlights, this outcome results from the ensemble stacking model combining all base models. Even though we are analyzing all patients in the dataset together, as represented by our second experiment, specific models may have advantages when analyzing the time the patient has been part of the clinical trials (such as Gradient Boosting and Decision Tree for short-term and Random Forest for long-term).

Taking a step back, while an independent observer may look at this SMAPE measure and contemplate whether this is satisfactory, it is worth noting that the top-performing SMAPE for the competition is 54.1. Hence "in the grand scheme of things," having a SMAPE that is 3.1 away from the best-performing model is an accomplishment our team is proud of. In addition, all three of these SMAPE measurements further prove the complexity of using just the provided datasets in determining the UPDRS scores and hence, Parkinson's Disease progression.

In indicating that the Ensemble Stacking model performed the best, it is also crucial to truly understand what the SMAPE of 57.2 represents. As shown in the illustrative example below, this SMAPE accounts for the percentage of error within our predicted UPDRS score and the actual observed UPDRS score. This, in turn, means that the total UPDRS score for a given patient will have a broader range and greater variance between the actual score and the predicted score:

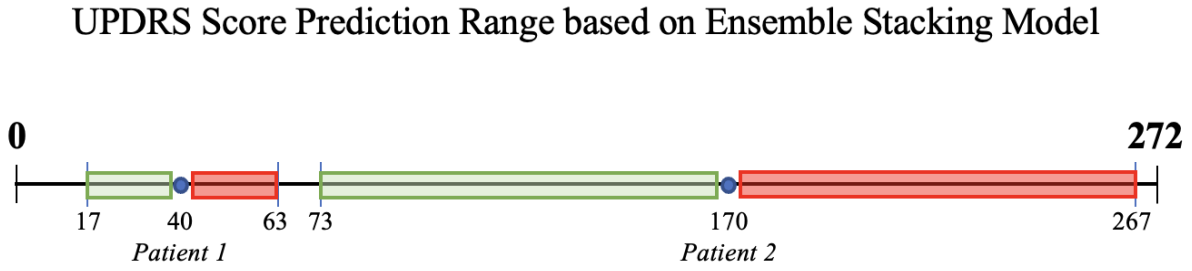


Figure 10: Illustrative example of two patients - Patient 1 has a predicted UPDRS score total of 40 while Patient 2 has a predicted UPDRS score total of 170. The respective ranges are calculated based on the SMAPE of 57.2. Hence, Patient 1's actual UPDRS score could be between 17 to 63, while Patient 2's actual UPDRS score could be between 73 to 267 (with a UPDRS score within the green range preferred compared to a UPDRS score within the red spectrum)

While we highlight the substantial problem of having a high SMAPE in the illustration above, as the range for the UPDRS score is substantial, our findings still underscore the potential of machine learning models in predicting the progression of Parkinson's Disease using peptide and protein abundance data from CSF samples. The optimized SMAPE score ensured that the models accurately predicted the relative error in the UPDRS scores, which is essential for clinical decision-making and patient care.

VI. Conclusions

Very succinctly recap the problem you were studying and what was your approach to the solution. Focus on explaining the key takeaways from your work - as you're writing your conclusions think about if the reader took nothing else away from reading your report, what would you want them to know most? Did you identify one particular approach that worked well? Was there a challenge that you faced that opens the door to working on solving a new problem? What avenues of research would you pursue next?

As represented through this research and data science application, Parkinson's Disease is an extremely complex disease that doctors haven't fully understood. To summarize our approach in using machine learning techniques to understand and possibly reveal hidden characteristics and features of Parkinson's Disease, we optimized our chosen models (Decision Tree, Gradient Boosting, Random Forest, Support Vector Machine, and Ensemble Stacking). To identify the best machine learning model for this problem, we created two distinct experiments, in which we analyzed the data in different manners - the first in which we analyzed all of the patients with no exclusions and the second in which we divided the patients by the amount of time that they had been a part of the clinical trials. Through this experimentation, we confidently selected and applied the ensemble stacking model on all patients as the best and preferred model to predict UPDRS scores for patients. The ensemble stacking model achieved our lowest SMAPE of 57.2. While this initial outcome may seem less ideal, we recommend that machine learning tools, such as our ensemble stacking model, be considered complementary to current medical procedures. For example, medical professionals can use this application as a pre-screening tool, in conjunction with their expertise and the patient's medical history, to determine patients who may require more medical attention in the short-term compared to the long term. Our recommendation highlights that while the end results were not exemplary, we were able to garner insights in regards to not only the solution in general but the complexity of using machine learning techniques to predict the progression of Parkinson's Disease.

Through this process, we can also assert that this project as a whole and our approach have several drawbacks. The first is apparent: while we, as data scientists, conducted extensive research, our domain expertise regarding Parkinson's Disease is still very limited. Refining this process by seeking insights from medical professionals (through interviewing, guidance, etc.) would be critical. The second limitation was that the dataset provided contained a large quantity of missing data, as depicted in Figure 4. Since we were working with only 2,615 clinical visits, we found it more insightful to include data observations with one of the score types missing rather than excluding it to avoid working with an even smaller dataset. Similar to the first limitation aforementioned, if we were to resolve these missing values with imputation, it would be best to get medical expertise and guidance. For example, there has been some research that has shown that the UPDRS scores can be considered stable in the short-term (for the first 6 - 12 months) [16].

Lastly, two areas could be explored for future work. First, building on the insight gained from our second experiment, it becomes evident that having more data available on a patient is beneficial. It is important to note that it is not only the quantity of data but also the duration of data captured, as the longer the patient has visited a clinic, the better. As mentioned within our limitations, it would also be crucial for the success of a real-world project, with practical applications, to not only work with medical professionals but also partner with foundations

and organizations dedicated to the sole purpose of finding not only a cure to Parkinson's Disease one day but additional insights. For example, as recently as April 13th, 2023, the Michael J. Fox Foundation announced a groundbreaking discovery in finding and linking the biomarker pathologically responsible for causing Parkinson's Disease; however, this was discovered using medical and scientific methodologies (Chute, 2023 [4]). This discovery reiterates the passion and urge to find answers and that this problem space cannot solely be resolved using one methodology (such as machine learning) but requires cooperation amongst multiple specialties.

VII. Roles

Provide detail on your individual role. Each team member should clearly articulate an individual role.

Below are the roles and responsibilities of our team:

- Ruixin Lou: Responsible for data analysis and exploratory data analysis (EDA). She was also involved in developing and optimizing machine learning models and evaluating their performance. As an undergraduate psychology major, she provided input and background gained from academia.

- Emmanuel Ruhamyankaka: Responsible for developing and optimizing machine learning models, evaluating their performance, original draft report outline, and reviewing the report. He also developed and maintained the project's codebase and implemented the models in a test environment.

- Sukhpreet Sahota: Responsible for data analysis and exploratory data analysis (EDA) and developing and optimizing machine learning models and evaluating their performance. He led research efforts, produced presentation slides, and was responsible for the overall story communicated within the project report. He was also involved in project management and coordinating the work of other team members.

- Jiaxin Ying: Responsible for data cleaning, preprocessing, and feature engineering. She was also involved in developing and optimizing machine learning models and evaluating their performance. She generated plots that outlined models, their respective performances, and other key findings.

Overall, each team member had a crucial role in the project's success, but we worked collaboratively to achieve the project's objectives.

References

Include an alphabetical list of references cited in this work. A minimum of 15 are required (a minimum of 10 must be technical papers/reports or conference papers, rather than blogs or websites).

[1] *AMP®-Parkinson's Disease Progression Prediction*. Kaggle. (n.d.). Retrieved March 3, 2023, from https://www.kaggle.com/competitions/amp-parkinsons-disease-progression-prediction/data

[2] Balestrino, R., Hurtado-Gonzalez, C. A., Stocchi, F., Radicati, F. G., Chaudhuri, K. R., Rodriguez-Blazquez, C., Martinez-Martin, P., & The PDCS European Study Group. (2019, November 27). *Applications of the European Parkinson's Disease Association sponsored Parkinson's Disease Composite Scale (PDCS)*. Nature News. Retrieved April 13, 2023, from https://www.nature.com/articles/s41531-019-0097-1

[3] Blink, J. C. (2023, March 14). *AMP - EDA + Models*. Kaggle. Retrieved April 5, 2023, from https://www.kaggle.com/code/jcblink/amp-eda-models#2.1.2---Visit-Month-vs-Protein-Data

[4] Chute, B. (2023, April 13). *Breaking News: Parkinson's Disease Biomarker Found*. The Michael J. Fox Foundation for Parkinson's Research | Parkinson's Disease. Retrieved April 13, 2023, from https://www.michaeljfox.org/news/breaking-news-parkinsons-disease-biomarker-found?pn_cid=pn-a1b1R00000AlQ1B

[5] Duncan, R. P., & Earhart, G. M. (2012). *Randomized Controlled Trial of Community-Based Dancing to Modify Disease Progression in Parkinson's Disease*. Sage Publication Journal. Retrieved March 8, 2023, from https://journals.sagepub.com/doi/10.1177/1545968311421614

[6] Emamzadeh, F. N., & Surguchov, A. (2018, August 30). *Parkinson's Disease: Biomarkers, Treatment, and Risk Factors*. Frontiers in Neuroscience. Retrieved April 13, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6125353/

[7] GBD 2016 Parkinson's Disease Collaborators. (2018, October 1). *Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016*. ScienceDirect. Retrieved March 9, 2023, from https://www.sciencedirect.com/science/article/pii/S1474442218302953

[8] Goetz, C. G., Tilley, B., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M. B., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A. E., Lees, A., Leurgans, S., LeWitt, P. A., Nyenhuis, D., … LaPelle, N. (n.d.). MDS-Unified Parkinson's Disease Rating Scale (MDS-UPDRS). Retrieved April 2, 2023, from https://www.movementdisorders.org/MDS/MDS-Rating-Scales/MDS-Unified-Parkinsons-Disease-Rating-Scale-MDS-UPDRS.htm

[9] Holden, S. K., Finseth, T., Sillau, S. H., & Berman, B. D. (2017, September 22). *Progression of MDS-UPDRS Scores Over Five Years in De Novo Parkinson Disease from the Parkinson's Progression Markers Initiative Cohort*. Movement Disorders Clinical Practice. Retrieved April 2, 2023, from https://pubmed.ncbi.nlm.nih.gov/29662921/

[10] Inácio, P. (2022, November 30). *Changes in Protein Structure May Act as Parkinson's Biomarkers*. Parkinson's News Today. Retrieved March 10, 2023, from https://parkinsonsnewstoday.com/news/changes-protein-structure-may-act-parkinsons-biomarkers/

[11] Jankovic, J. (2008, April). *Parkinson's disease: Clinical features and diagnosis*. Journal of neurology, neurosurgery, and psychiatry. Retrieved March 8, 2023, from https://pubmed.ncbi.nlm.nih.gov/18344392/

[12] Kalia, L. V., & Lang, A. E. (2015, April 19). *Parkinson's disease*. Lancet (London, England). Retrieved March 8, 2023, from https://pubmed.ncbi.nlm.nih.gov/25904081/

[13] Mandal, I., & Sairam, N. (2012, September 24). *New machine-learning algorithms for prediction of Parkinson's disease*. Taylor & Francis. Retrieved March 10, 2023, from https://www.tandfonline.com/doi/full/10.1080/00207721.2012.724114

[14] Masliah, E., Rockenstein, E., Veinbergs, I., Sagara, Y., Mallory, M., Hashimoto, M., & Mucke, L. (2001, September 25). *β-Amyloid peptides enhance α-synuclein accumulation and neuronal deficits in a transgenic mouse model linking Alzheimer's disease and Parkinson's disease*. PNAS. Retrieved March 11, 2023, from https://sci-hub.se/10.1073/pnas.211412398

[15] Mei, J., Desrosiers, C., & Frasnelli, J. (2021, May 6). *Machine Learning for the Diagnosis of Parkinson's Disease: A Review of Literature*. Frontiers in aging neuroscience. Retrieved March 10, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8134676/

[16] Mollenhauer, B., Caspell-Garcia, C. J., Coffey, C. S., Taylor, P., Shaw, L. M., Trojanowski, J. Q., Singleton, A., Frasier, M., Marek, K., Galasko, D., & Parkinson's Progression Marker Initiative. (2017, November 7). *Longitudinal CSF biomarkers in Patients with Early Parkinson Disease and Healthy Controls*. Neurology. Retrieved March 6, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5679418/

[17] Moustafa, A. A., Chakravarthy, S., Phillips, J. R., Gupta, A., Keri, S., Polner, B., Frank, M. J., & Jahanshahi, M. (2016, July 12). *Motor symptoms in parkinson's disease: A unified framework*. Neuroscience & Biobehavioral Reviews. Retrieved March 9, 2023, from https://www.sciencedirect.com/science/article/pii/S0149763415300919#fig0005

[18] Ouyang, A., & Abdul Latif Jameel Clinic for Machine Learning in Health. (2022, August 22). *Artificial intelligence model can detect Parkinson's from breathing patterns*. MIT News | Massachusetts Institute of Technology. Retrieved March 10, 2023, from https://news.mit.edu/2022/artificial-intelligence-can-detect-parkinsons-from-breathing-patterns-0822

[19] Padilla-Godínez, F. J., Ramos-Acevedo, R., Martínez-Becerril, H. A., Bernal-Conde, L. D., Garrido-Figueroa, J. F., Hiriart, M., Hernández-López, A., Argüero-Sánchez, R., Callea, F., & Guerra-Crespo, M. (2021, November 18). *Protein Misfolding and Aggregation: The Relatedness between Parkinson's Disease and Hepatic Endoplasmic Reticulum Storage Disorders*. MDPI. Retrieved March 5, 2023, from https://www.mdpi.com/1422-0067/22/22/12467

[20] *Parkinson's 101*. The Michael J. Fox Foundation for Parkinson's Research | Parkinson's Disease. (n.d.). Retrieved March 10, 2023, from https://www.michaeljfox.org/parkinsons-101

[21] Parkinson's Foundation. (n.d.). *Statistics - Get informed about Parkinson's disease with these key numbers*. Parkinson's Foundation. Retrieved April 3, 2023, from https://www.parkinson.org/understanding-parkinsons/statistics

[22] Penn Medicine. (n.d.). *Steve's Story*. Penn Medicine. Retrieved April 4, 2023, from https://www.pennmedicine.org/for-patients-and-visitors/find-a-program-or-service/neurology/movement-disorders/patient-stories/steve-story

[23] Prashanth, R., Roy, S. D., Mandal, P. K., & Ghosh, S. (n.d.). *High-Accuracy Detection of Early Parkinson's Disease through Multimodal Features and Machine Learning*. ResearchGate. Retrieved March 8, 2023, from https://www.researchgate.net/profile/Prashanth-R/publication/297626811_High-Accuracy_Detection_of_Early_Parkinson's_Disease_through_Multimodal_Features_and_Machine_Learning/links/620537067b05f82592decc05/High-Accuracy-Detection-of-Early-Parkinsons-Disease-through-Multimodal-Features-and-Machine-Learning.pdf

[24] Pierson, J., Andrén, P. E., Caprioli, R. M., Norris, J. L., Aerni, H.-R., & Svenningsson, P. (2004, March). *Molecular profiling of experimental parkinson's disease: Direct analysis of peptides and proteins on brain tissue sections by MALDI mass spectrometry*. Journal of proteome research. Retrieved March 9, 2023, from https://pubmed.ncbi.nlm.nih.gov/15113106/

[25] Saini, S. (2023, March 31). *AMP-PDPP- Random Forest+ Lasso+ LGBM + others*. Kaggle. Retrieved April 4, 2023, from https://www.kaggle.com/code/suneetsaini/amp-pdpp-random-forest-lasso-lgbm-others

[26] Severson, K. A., Chahine, L. M., Smolensky, L. A., Dhuliawala, M., Frasier, M., Ng, K., Ghosh, S., & Hu, J. (2021, July 29). *Discovery of Parkinson's disease states and disease progression modelling: a longitudinal data study using machine learning*. The Lancet Digital Health. Retrieved March 7, 2023, from https://www.thelancet.com/journals/landig/article/PIIS2589-7500(21)00101-1/fulltext#%20

[27] Shi, M., Movius, J., Dator, R., Aro, P., Zhao, Y., Pan, C., Lin, X., Bammler, T. K., Stewart, T., Zabetian, C. P., Peskind, E. R., Hu, S.-C., Quinn, J. F., Galasko, D. R., & Zhang, J. (2015, January 2). *Cerebrospinal fluid peptides as potential parkinson disease biomarkers: A staged pipeline for discovery and validation*. Molecular & cellular proteomics : MCP. Retrieved March 6, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4349976/

[28] Skorvanek, M., Martinez-Martin, P., Kovacs, N., Rodriguez-Violante, M., Corvol, J.-C., Taba, P., Seppi, K., Levin, O., Schrag, A., Foltynie, T., Alvarez-Sanchez, M., Arakaki, T., Aschermann, Z., Aviles-Olmos, I., Benchetrit, E., Benoit, C., Bergareche-Yarza, A., Cervantes-Arriaga, A., Chade, A., … Mendoza-Rodriguez, A. (2017, March 11). *Differences in MDS-UPDRS Scores Based on Hoehn and Yahr Stage and Disease Duration*. Movement Disorders Clinical Practice. Retrieved April 16, 2023, from https://pubmed.ncbi.nlm.nih.gov/30363418/

Appendix

A1. Data Dictionary

Clinical Trial Data

- visit_id - ID code for the visit.
- visit_month - The month of the visit, relative to the first visit by the patient.
- patient_id - An ID code for the patient.
- updrs_[1-4] - The patient's score for part N of the Unified Parkinson's Disease Rating Scale. Higher numbers indicate more severe symptoms. Each sub-section covers a distinct category of symptoms, such as mood and behavior for Part 1 and motor functions for Part 3.
- upd23b_clinical_state_on_medication - Whether or not the patient was taking medication such as Levodopa during the UPDRS assessment. Expected to mainly affect the scores for Part 3 (motor function). These medications wear off fairly quickly (on the order of one day) so it's common for patients to take the motor function exam twice in a single month, both with and without medication.

Peptide Data

- visit_id - ID code for the visit.
- visit_month - The month of the visit, relative to the first visit by the patient.
- patient_id - An ID code for the patient.
- UniProt - The UniProt ID code for the associated protein. There are often several peptides per protein.
- Peptide - The sequence of amino acids included in the peptide. See this table for the relevant codes. Some rare annotations may not be included in the table. The test set may include peptides not found in the train set.
- PeptideAbundance - The frequency of the amino acid in the sample.

Protein Data

- visit_id - ID code for the visit.
- visit_month - The month of the visit, relative to the first visit by the patient.
- patient_id - An ID code for the patient.
- UniProt - The UniProt ID code for the associated protein. There are often several peptides per protein. The test set may include proteins not found in the train set.
- NPX - Normalized protein expression. The frequency of the protein's occurrence in the sample. May not have a 1:1 relationship with the component peptides as some proteins contain repeated copies of a given peptide.

Exhibit 1: Prevalence of Parkinson's Disease within the United States

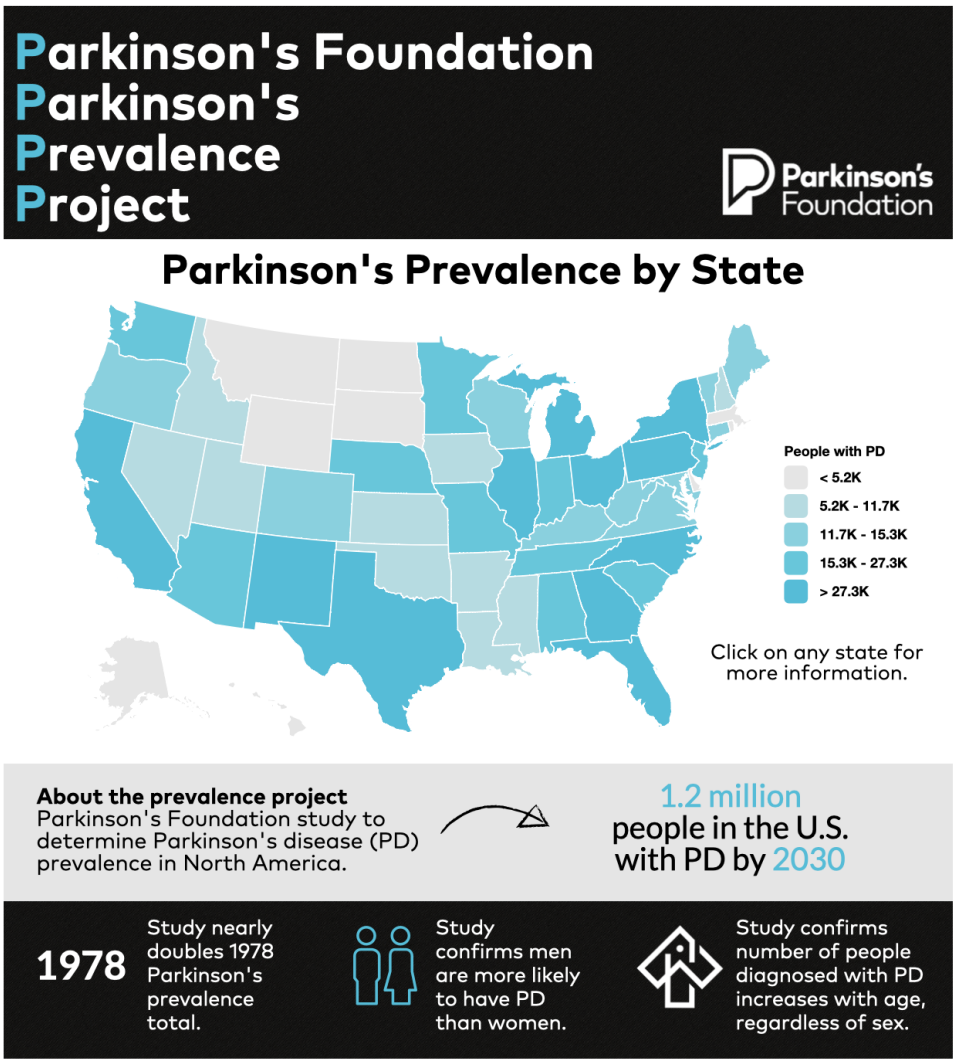


Exhibit 2: Global Prevalence of Parkinson's Disease by age and sex, 2016 (Source: Global Burden of Disease)

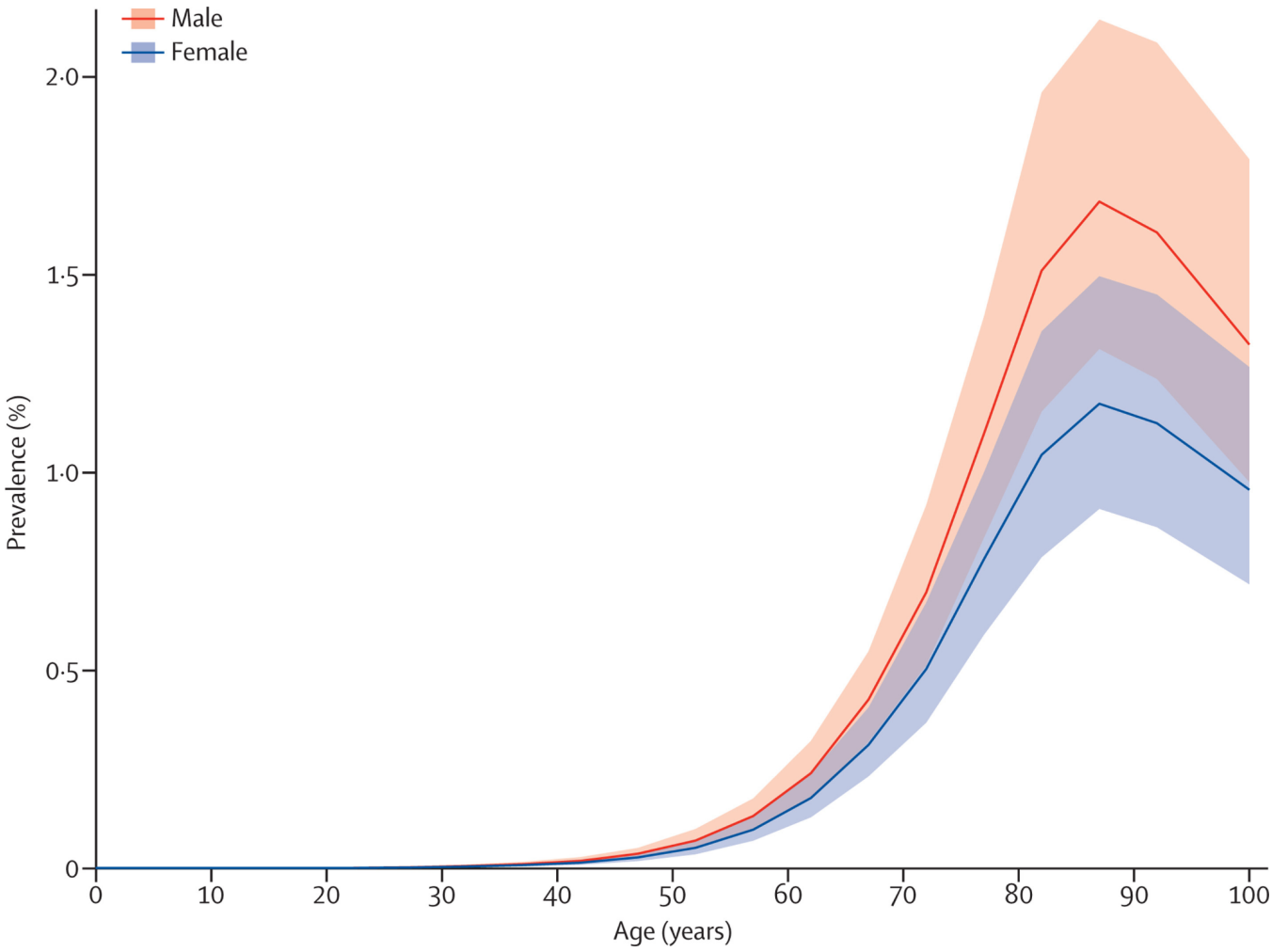


Exhibit 3: Distribution of each UPDRS Score

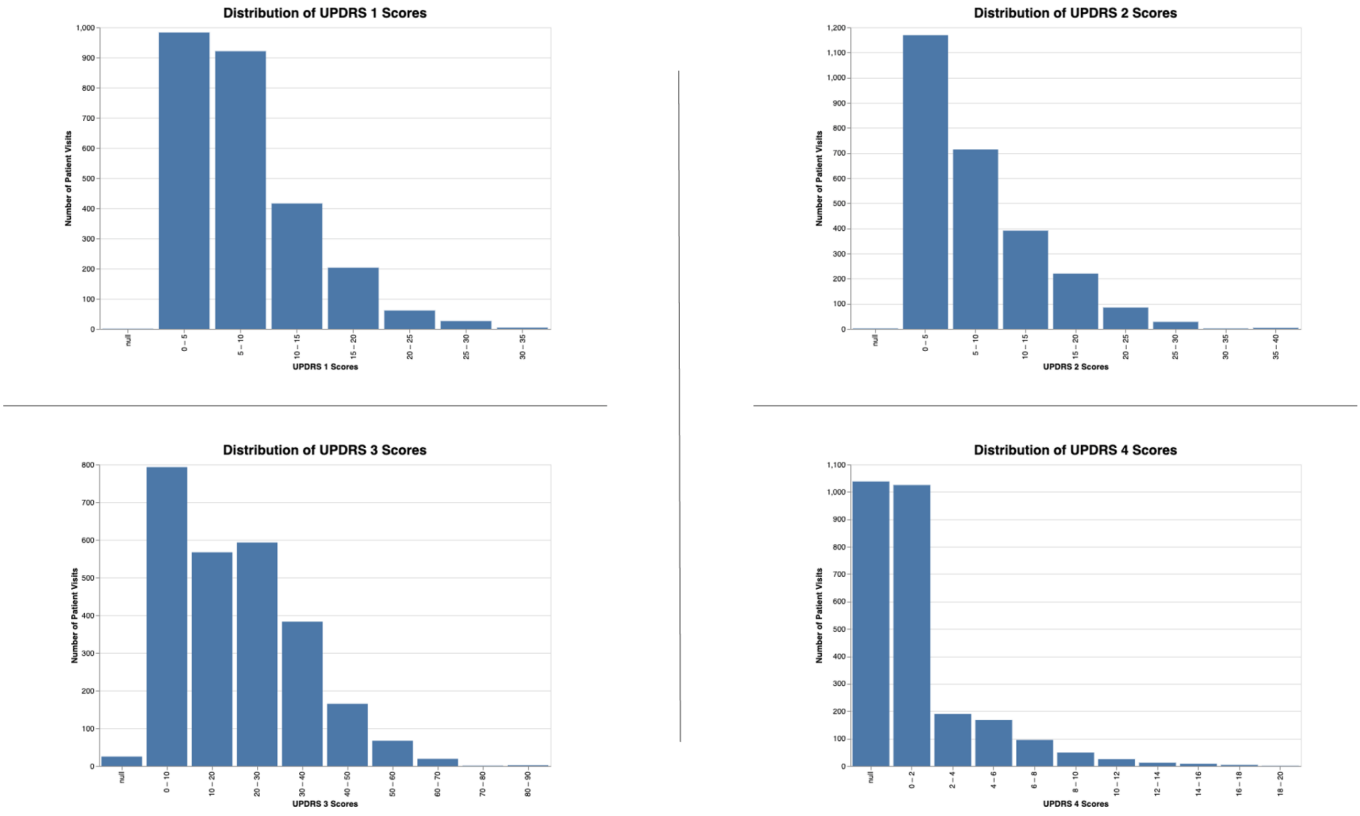


Exhibit 4: Distribution of Patient Medication Status

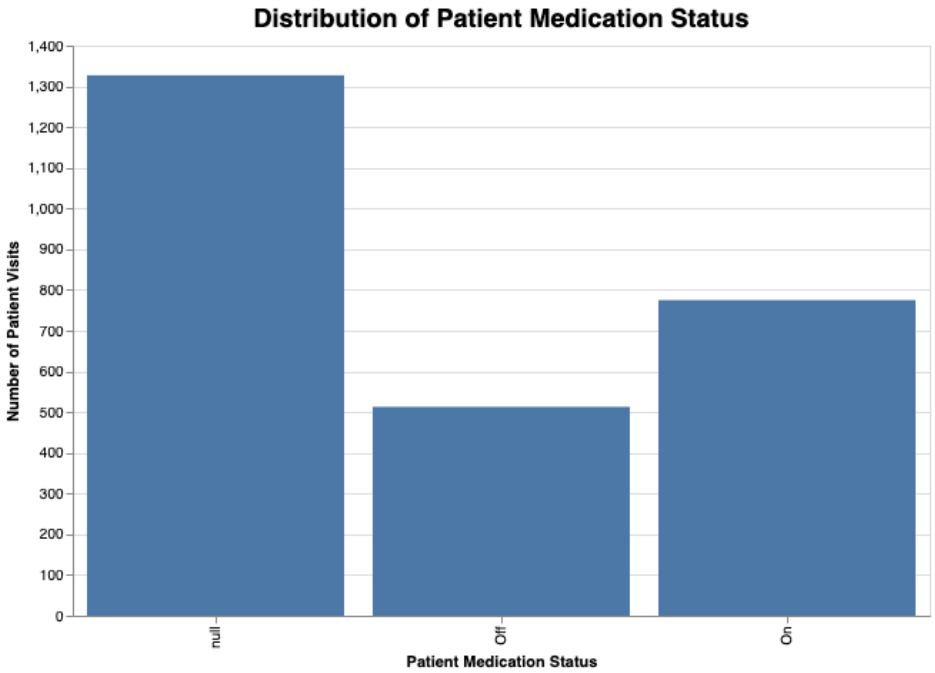


Exhibit 5: Boxplot of UPDRS Scores by Clinical Visit Month and Medication Status

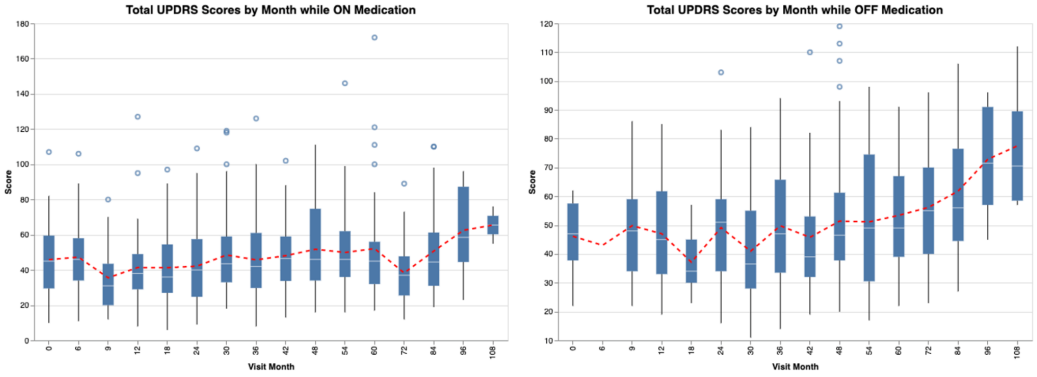


Exhibit 6: Side-by-side Boxplot Comparison of each UPDRS Scores by Clinical Visit Month and Medication Status

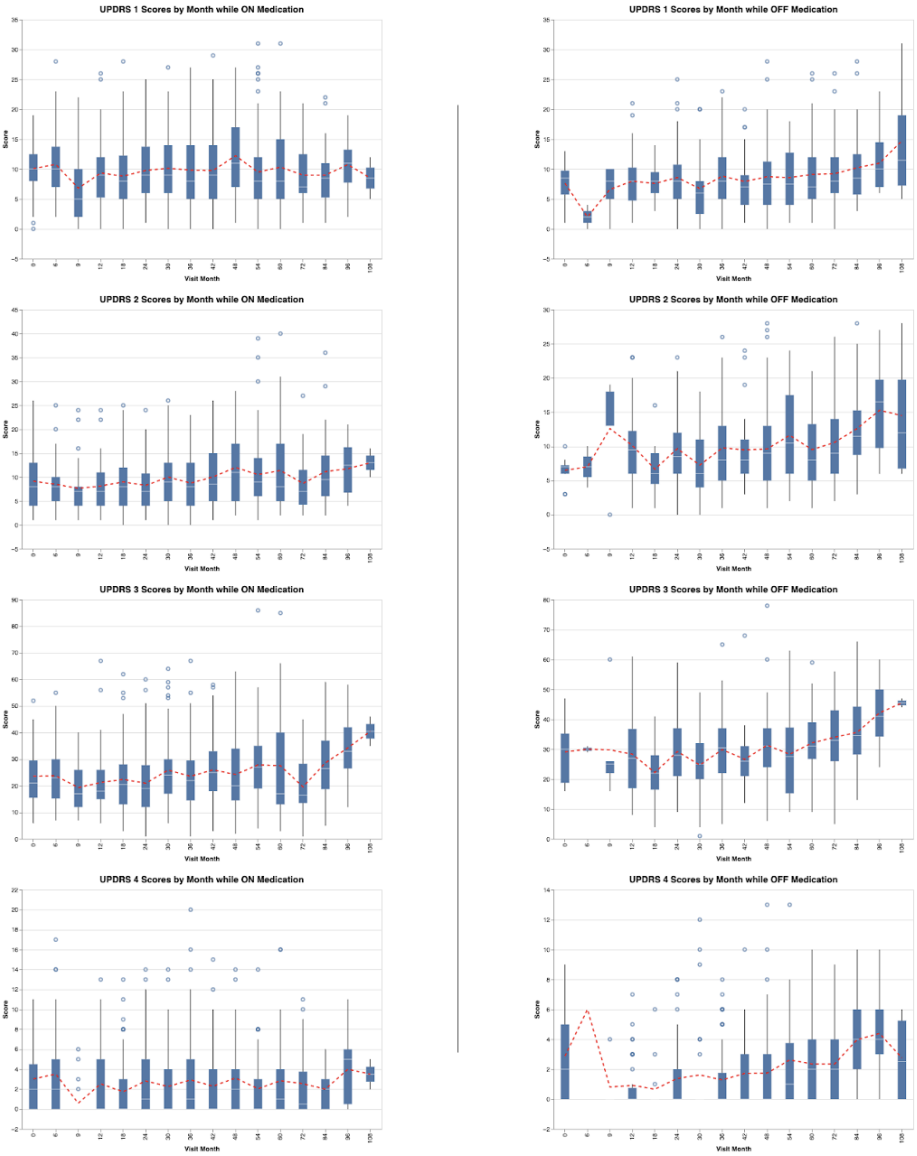


Exhibit 7: Boxplot of Frequency of Normalized Protein Expression by Clinical Visit Month

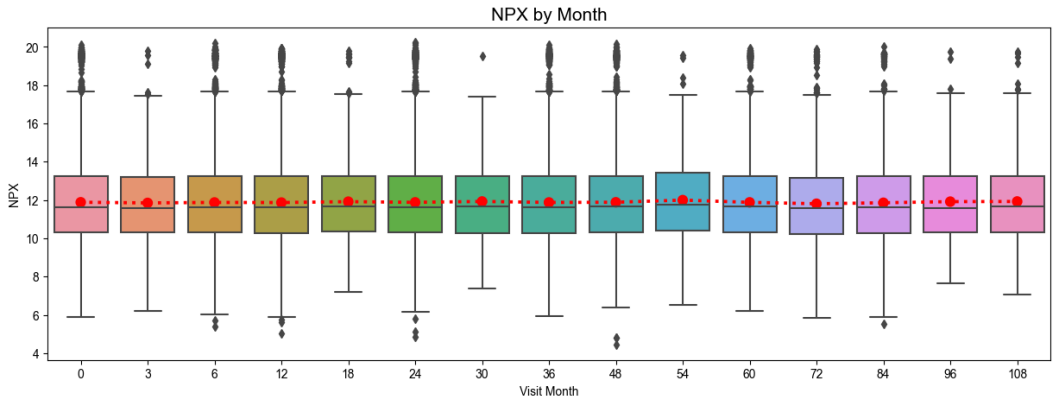


Exhibit 8: Prediction Output of Optimized Ensemble Stacking Model on Test Data

Prediction	Predicted UPDRS Score
3342_0_updrs_1_plus_0_months	6
3342_0_updrs_1_plus_6_months	6
3342_0_updrs_1_plus_12_months	6
3342_0_updrs_1_plus_24_months	6
3342_0_updrs_2_plus_0_months	9
3342_0_updrs_2_plus_6_months	9
3342_0_updrs_2_plus_12_months	9
3342_0_updrs_2_plus_24_months	9
3342_0_updrs_3_plus_0_months	24
3342_0_updrs_3_plus_6_months	24
3342_0_updrs_3_plus_12_months	24
3342_0_updrs_3_plus_24_months	24
3342_0_updrs_4_plus_0_months	0
3342_0_updrs_4_plus_6_months	0
3342_0_updrs_4_plus_12_months	0
3342_0_updrs_4_plus_24_months	0
50423_0_updrs_1_plus_0_months	6
50423_0_updrs_1_plus_6_months	6
50423_0_updrs_1_plus_12_months	6
50423_0_updrs_1_plus_24_months	6
50423_0_updrs_2_plus_0_months	9
50423_0_updrs_2_plus_6_months	9
50423_0_updrs_2_plus_12_months	9
50423_0_updrs_2_plus_24_months	9
50423_0_updrs_3_plus_0_months	24
50423_0_updrs_3_plus_6_months	24
50423_0_updrs_3_plus_12_months	24
50423_0_updrs_3_plus_24_months	24
50423_0_updrs_4_plus_0_months	0
50423_0_updrs_4_plus_6_months	0
50423_0_updrs_4_plus_12_months	0
50423_0_updrs_4_plus_24_months	0
3342_6_updrs_1_plus_0_months	6
3342_6_updrs_1_plus_6_months	6
3342_6_updrs_1_plus_12_months	6
3342_6_updrs_1_plus_24_months	6
3342_6_updrs_2_plus_0_months	9
3342_6_updrs_2_plus_6_months	9
3342_6_updrs_2_plus_12_months	9
3342_6_updrs_2_plus_24_months	9
3342_6_updrs_3_plus_0_months	24
3342_6_updrs_3_plus_6_months	24
3342_6_updrs_3_plus_12_months	24
3342_6_updrs_3_plus_24_months	24
3342_6_updrs_4_plus_0_months	0
3342_6_updrs_4_plus_6_months	0
3342_6_updrs_4_plus_12_months	0
3342_6_updrs_4_plus_24_months	0
50423_6_updrs_1_plus_0_months	6
50423_6_updrs_1_plus_6_months	6
50423_6_updrs_1_plus_12_months	6
50423_6_updrs_1_plus_24_months	6
50423_6_updrs_2_plus_0_months	9
50423_6_updrs_2_plus_6_months	9
50423_6_updrs_2_plus_12_months	9
50423_6_updrs_2_plus_24_months	9
50423_6_updrs_3_plus_0_months	24
50423_6_updrs_3_plus_6_months	24
50423_6_updrs_3_plus_12_months	24
50423_6_updrs_3_plus_24_months	24
50423_6_updrs_4_plus_0_months	0
50423_6_updrs_4_plus_6_months	0
50423_6_updrs_4_plus_12_months	0
50423_6_updrs_4_plus_24_months	0