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

Ruixin Lou, Emmanuel Ruhamyankaka, Sukhpreet Sahota, & Jiaxin Ying

Steve's Story

Steve is a successful lawyer; on a seemingly normal 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 again, dismissed anything extreme and considered it as another coincidence. A few days later, while out for a run, he knew something was truly wrong as 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.

Parkinson's Disease by the Numbers!

A Progressive Neurological Disorder that affects millions of people, like famed actor Michael J. Fox, worldwide.

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**
90,000 new Parkinson's Disease diagnoses are made in the US every year, at which rate, 1.2 million individuals will be diagnosed with the disease by 2030

AMP-PD Progression Prediction

Kaggle Competition

Objective

Use advanced machine learning algorithms to predict Parkinson's Disease progression using clinical trial data, protein data, and peptide data

Goal

Predict the MDS-UPDRS scores for patients to determine Parkinson’s Disease

Two Types of Symptoms Measured by MDS-UPDRS Scale

Non-motor

Part I - Non-Motor Aspects of Experiences of Daily Living (nM - EDL)

Motor

Part II - Motor Aspects of Experiences of Daily Living (M-EDL)
Part III - Motor Examination
Part IV - Motor Complications

Total Score Range: 0 - 272

01

Exploratory Data Analysis

Dataset Overview

Clinical Trial Data

248 unique patients made

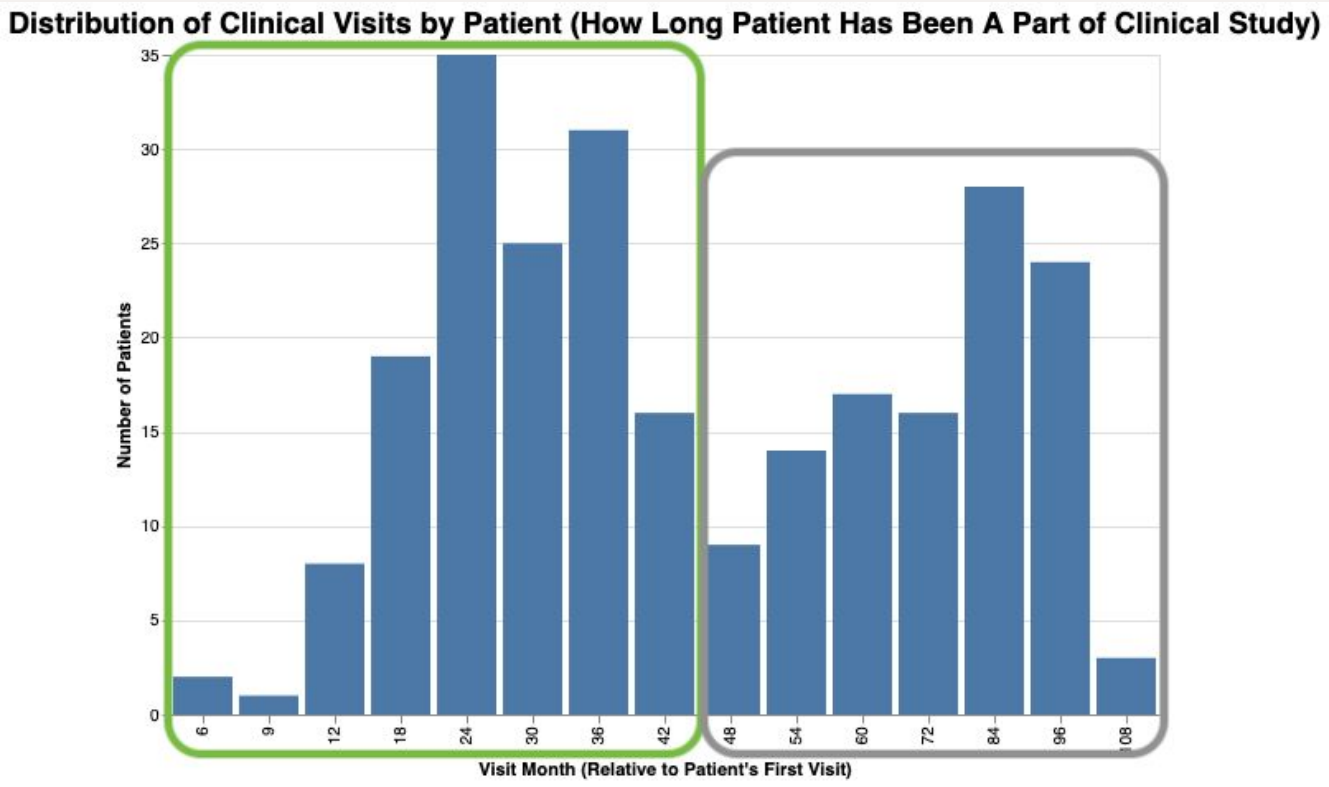
2,615 clinical visits

Peptide and Protein Data

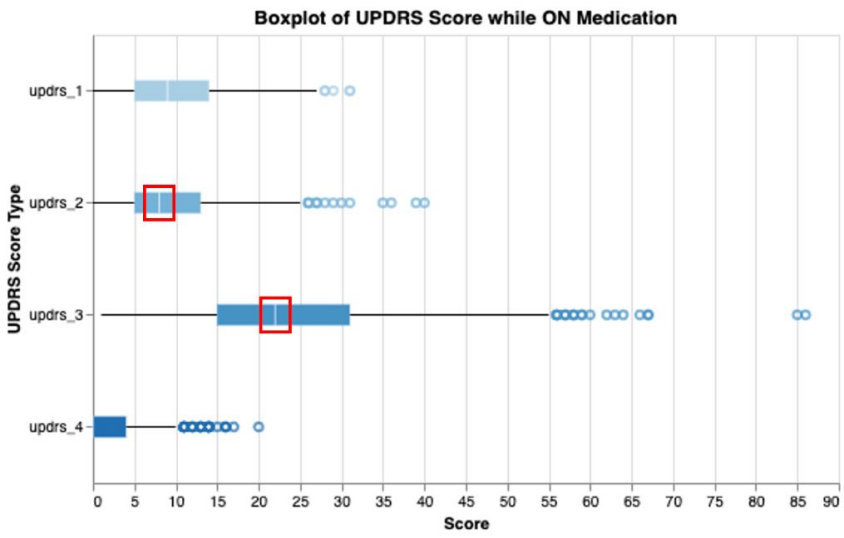
227 unique protein structures

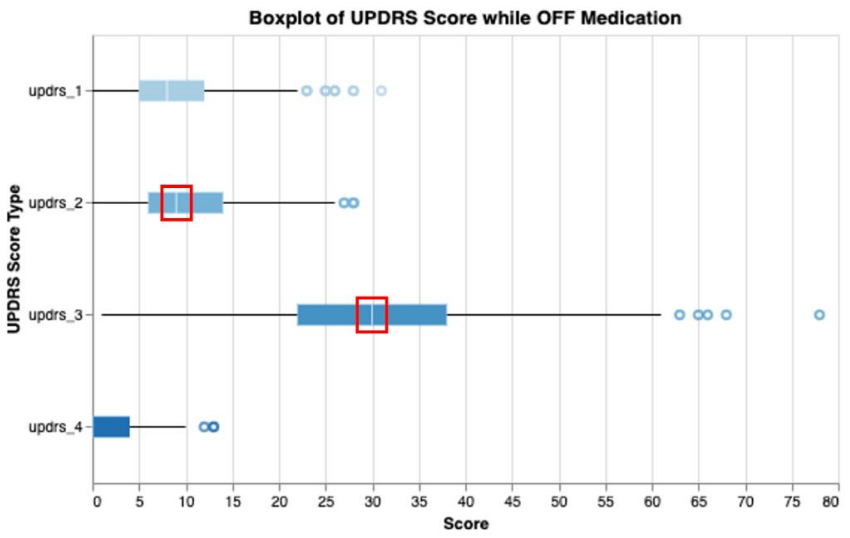
968 different peptide bonds

Duration Patient has been Part of Clinical Trials



MDS-UPDRS Score Based on Medication Status





Top 10 Most and Least Frequent Protein Structures

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

02

Experiments

The five stages of methodology

Stage 1
Preprocessing data

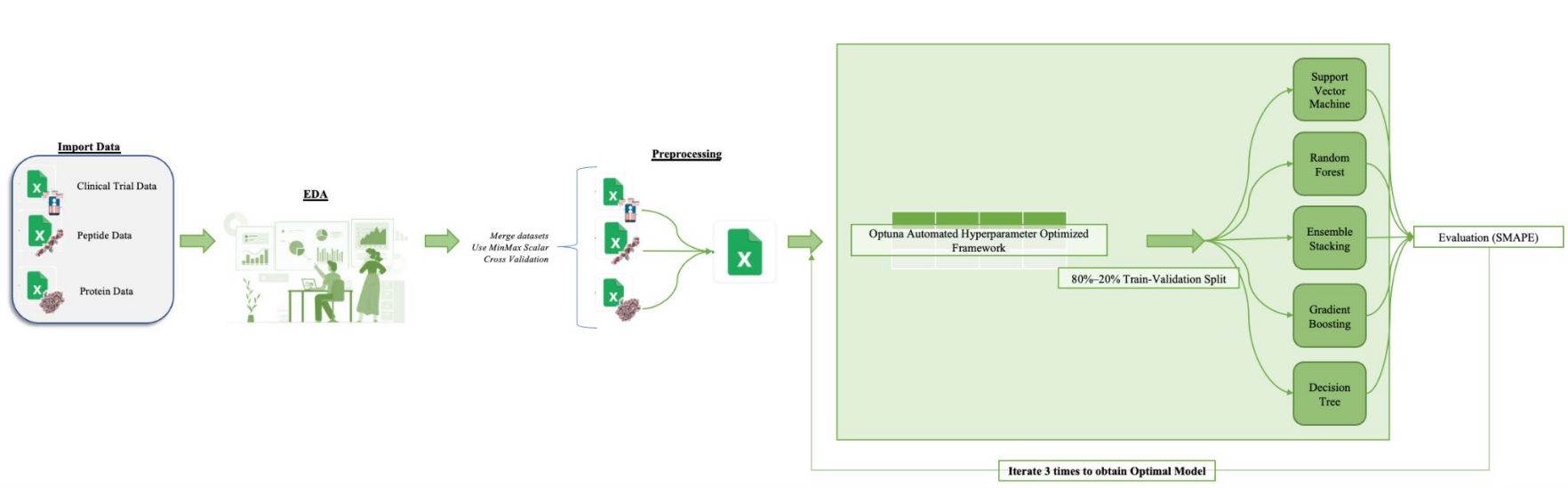
Stage 2
Feature engineering

Stage 3
Optimizing hyperparameters

Stage 4
Getting the optimal hyperparameters

Stage 5
Evaluating the test data for the lowest error.

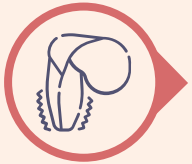
Flowchart of Steps Involving ML Techniques to Predict Parkinson's Disease Progression



Steps of Experiments

Merge and preprocess clinical, peptide, and protein datasets using MinMaxScaler.

Evaluate four base models and a fifth ensemble stacking model through 10-fold cross-validation.



Optimize hyperparameters with Optuna, train models on 80/20 split, minimize SMAPE score, and run 100 trials to predict relative error in UPDRS scores.

Set of hyperparameters tuned for each selected model

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)	

Experiments Conducted

Experiment #1

Develop a model using UPDRS scores, clinical data, peptide, and protein levels for all patients to analyze short-term and long-term Disease progression.

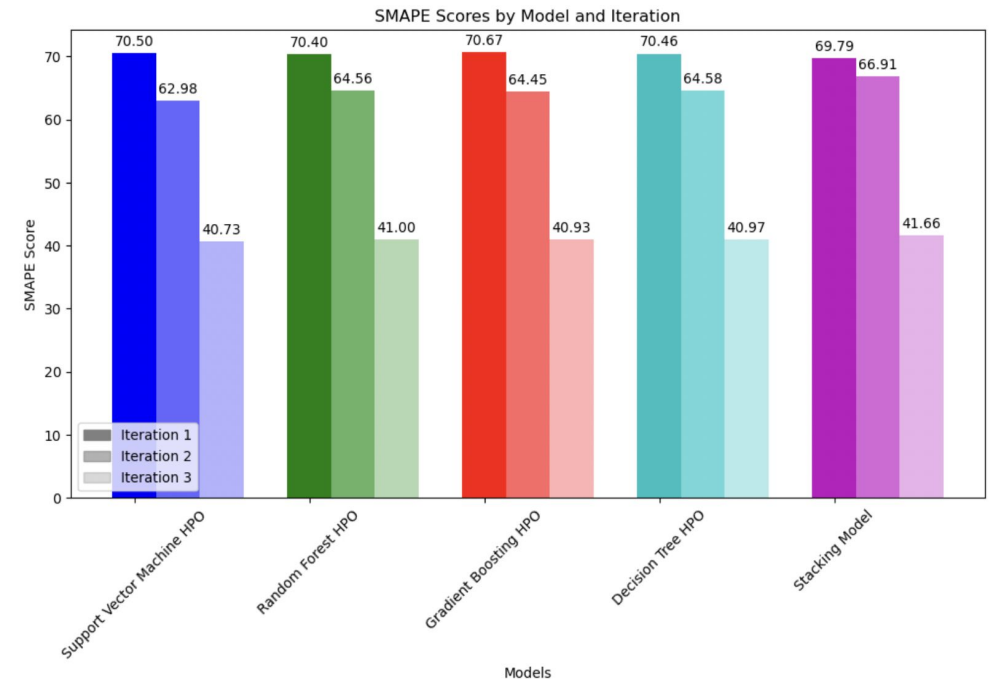
Experiment #2

Create two models for recently observed and established patients to analyze clinic visit duration's impact on progression. Compare prediction of UPDRS scores.

03

Results

Experiment #1 Results



SMAPE Scores by Model for each Iteration.

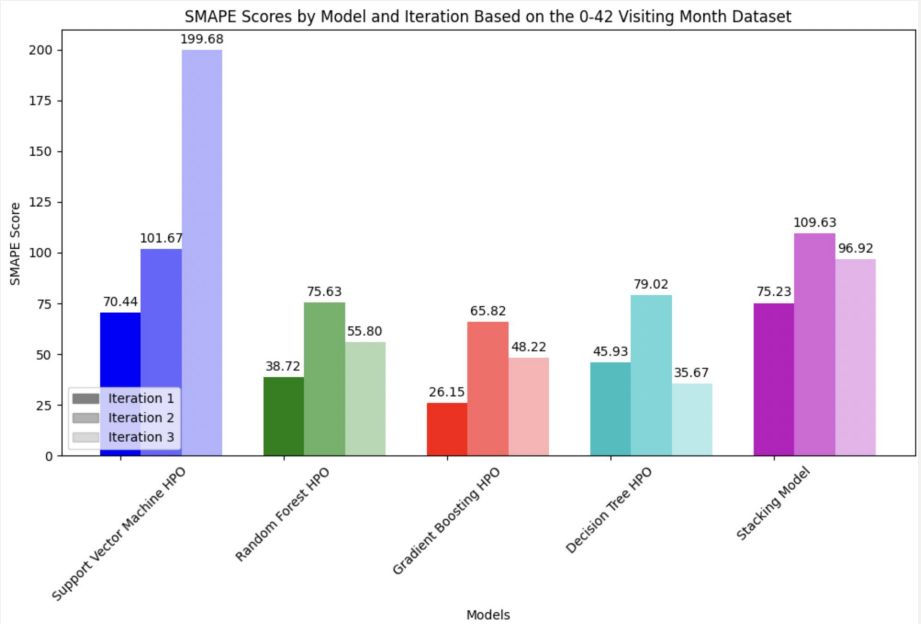
First Iteration

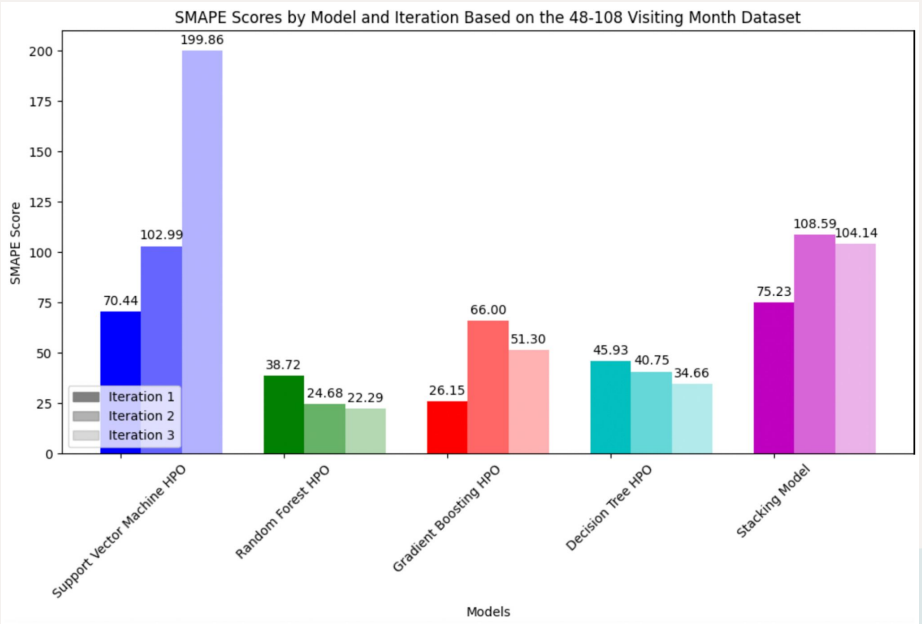
Random forest performed best

Next iterations..

Support Vector Machines performed the best from all of the models..

Experiment #2 Results





Test Results

The two optimal models that performed best on the collective dataset were Support Vector Machines and Ensemble Stacking.

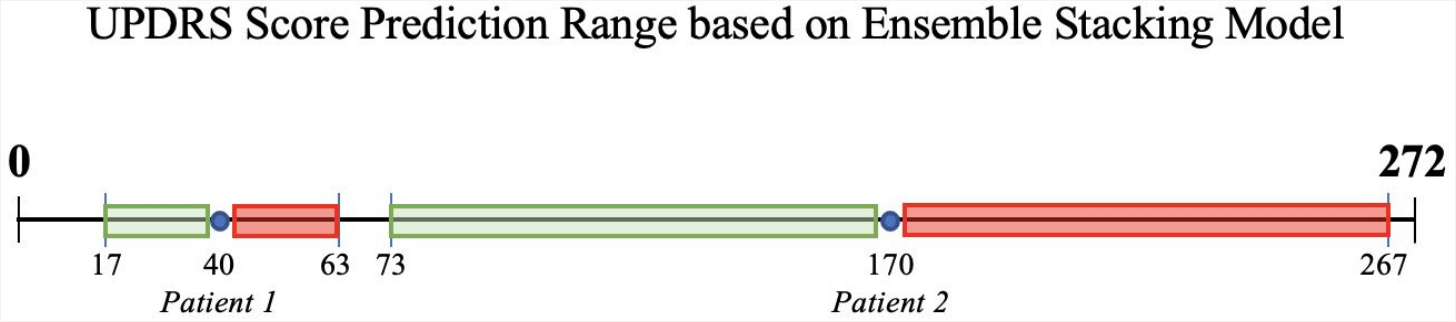
- **SVM-optimized model** garnered a **SMAPE of 57.8**.
- **Ensemble Stacking optimized model**, which amassed a **SMAPE of 57.2** and thus performed the best across all models we considered

In analyzing all patients in the dataset together, as represented by our second experiment, certain models may have advantages when analyzing the time the patient has been part of the clinical trials.

- **Short-term: Gradient Boosting and Decision Tree**
- **Long-term: Random Forest**

Interpreting SMAPE for Test Results

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 wider range and greater variance between the actual score and the predicted score



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 is preferred compared to a UPDRS score within the red range)

04

Conclusion

Conclusion

- In conclusion, this project is an important step towards using machine learning techniques to aid in understanding Parkinson’s Disease progression.
- This problem space cannot be resolved solely using one methodology, such as machine learning but requires cooperation among multiple specialities.
- Our models can complement current medical procedures, eg. be used as a pre-screening tool to determine patients who may require more medical attention in the short term than the long term.

Limitations:

- Our approach has limitations, such as limited domain expertise regarding Parkinson's Disease and a dataset with much missing data.
- Future work should focus on collecting more data and partnering with medical professionals and organizations dedicated to finding a cure for Parkinson's Disease.

Thanks!

Do you have any questions?