Predicting the Age of Disease Onset for Parkinson's Disease

Kate Gilbert
DATA1030 Fall 2023
Github Repo: https://github.com/kateegilbert/Data1030Project

Introduction

Parkinson's Disease (PD) is a highly prevalent neurodegenerative disease that affects thousands yearly. The Parkinson Foundation estimates that 90,000 new diagnoses will have been made by the end of 2023, amounting to 30 million total diagnoses by 2030.[1] PD involves the impairment or death of nerve cells in the body, which are unable to be regenerated.[2] Healthy nerve cells normally produce the neurotransmitter dopamine, which is responsible for instilling feelings of pleasure and motivation, as well as controlling memory, learning, concentration and movement.[3] Nerve cell death leads to lower levels of dopamine, severely impacting the main functions of control and movement in the body, and leading to depression.[2] Characteristic motor skill symptoms of PD include tremors, gait difficulties, postural instability, and bradykinesia. Nonmotor skill symptoms of PD include behavioral issues, gastrointestinal issues, joint pain, and loss of sense of smell.[2]

This project constructs a model to predict the age at which a patient will develop PD based on speech and motor pattern measurements. The data was sourced from Kaggle, from a study that was using analysis of speech to determine early biomarkers of PD in patients with rapid eye movement sleep behavior disorder (RBD).[4,5] RBD patients are at high risk for developing PD, and this study sought to record the differences in speech from patients with RBD, patients with PD, and healthy patients. Measurements from this automated vocal analysis hope to improve screening and diagnostic procedures that are currently used for PD.

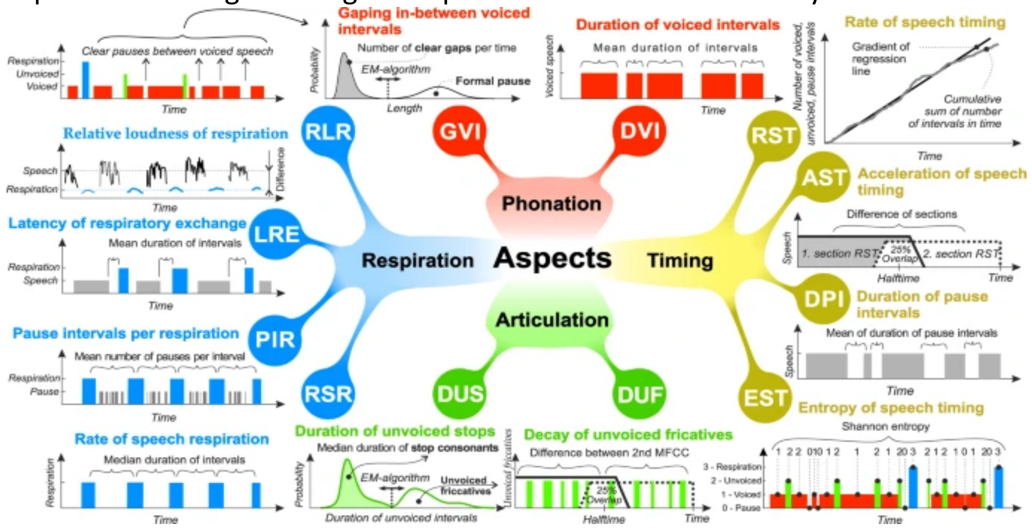
Aspects
Respiration
Phonation
Articulation
Timing
Clear pauses between voiced speech
Respiration
Unvoiced
Voiced
Time
Gaping in-between voiced intervals
Number of clear gaps per time
EM-algorithm
Formal pause
Probability
Length
GVI
Duration of voiced intervals
Mean duration of intervals
Voiced speech
Time
DVI
Rate of speech timing
Gradient of regression line
Cumulative sum of number of intervals in time
Number of voiced, unvoiced, pause intervals
Time
RST
RLR
Relative loudness of respiration
Speech
Respiration
Difference
Time
LRE
Latency of respiratory exchange
Mean duration of intervals
Respiration
Speech
Time
PIR
Pause intervals per respiration
Mean number of pauses per interval
Respiration
Pause
Time
RSR
Rate of speech respiration
Median duration of intervals
Respiration
Time
DUS
Duration of unvoiced stops
Median duration of stop consonants
EM-algorithm
Unvoiced fricatives
Probability
Duration of unvoiced intervals
DUF
Decay of unvoiced fricatives
Difference between 2nd MFCC
Unvoiced fricatives
25% Overlap
Halftime
Time
AST
Acceleration of speech timing
Difference of sections
Speech
1. section RST
25% Overlap
2. section RST
Halftime
Time
DPI
Duration of pause intervals
Mean of duration of pause intervals
Speech
Time
EST
Entropy of speech timing
Shannon entropy
1 2 2 0 1 0 1 2 1 2 1 2 0 3 2 1 2 1 0 1 2 0 3
3 - Respiration
2 - Unvoiced
1 - Voiced
0 - Pause
Time

Figure 1. The biomarker measurements taken as data from the automated speech pattern analysis. Figure adapted from [5].

The study collected data such as family history, current medications, and age, and also performed speech analysis for two tests: the patient reciting a pre-written passage, and the patient reciting a personal monologue. The measurements taken were scores for the level of tremor (movement) each extremity was experiencing, the pauses and durations of the speech, the rate of speech, and more (Figure 1). The target variable for this project was the reported age at which the patient developed PD. For healthy patients with no age of disease onset, this result was reported as 117 years, the age of the oldest living person. This result essentially means that this person will not develop PD within their lifetime. The purpose of this project was to develop a machine learning model that predicts the age of disease onset based on speech pattern measurements and medications prescribed to patients. This project is therefore a regression problem, as the model is predicting a specific age, and not an age-class.

Exploratory Data Analysis

In this dataset, there were 30 patients with RBD, 50 patients that were newly diagnosed and untreated for PD, and 50 healthy patients. For a majority of the speech pattern and tremor measurements, there were missing values for the healthy patients. The automated system scores these movements from 1-4, and missing values were assigned the values of 0 to mean no movement recorded. Next, knowing that PD leads to symptoms of depression, I looked at the fraction of patients (by gender) that were on antidepressant medications (Figure 2). As shown, many of the patients were not currently taking antidepressant medication, and there was not a significant difference by gender.

Fraction of Patients (By Gender) Currently Prescribed Antidepressants

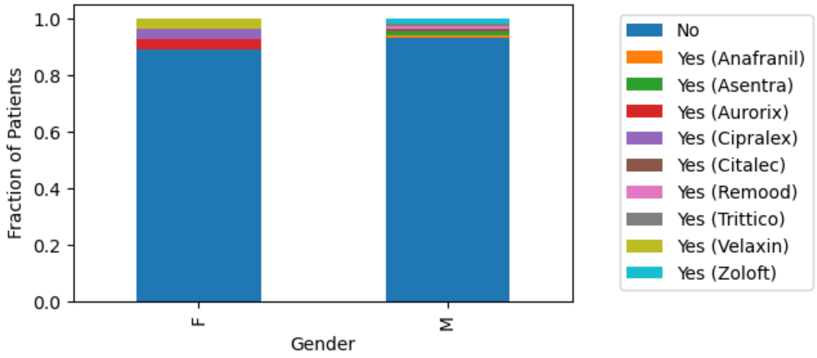

Figure 2. Fraction of patients, sorted by gender, that are currently prescribed antidepressant medication.

I then looked at the differences of a select few speech measurements between the patient's reciting a pre-written passage versus their personal monologue. The violin plots below show the correlation between a patient having a positive family history of PD and the duration of speech pauses in both their monologue and passage (Figure 3). There was no significant difference found in the length of pause between the monologue and the passage, positive family history or not.

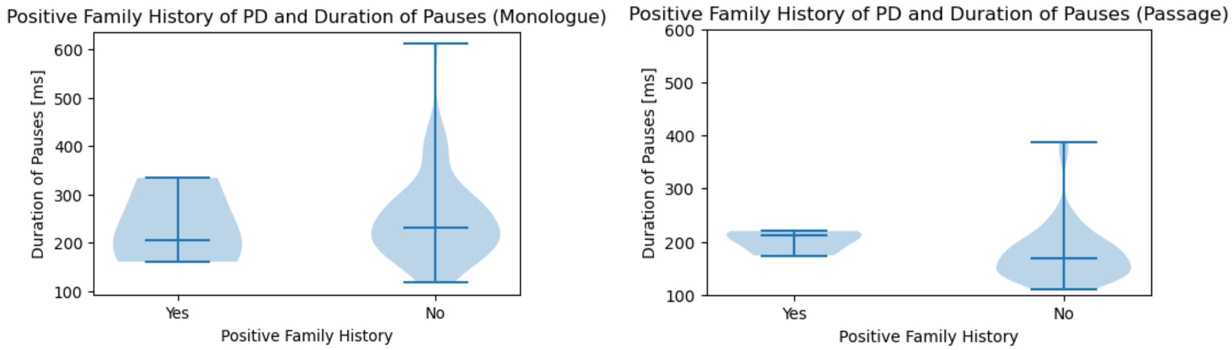

Figure 3. Violin plots of duration of speech pauses vs. positive family history of PD. This is shown for the monologue (left) and the passage (right). The speech pauses were measured is milliseconds.

Lastly for EDA, I explored the rate of speech timing versus age, as well as the rate of the monologue against the rate of the passage. The heatmaps below in Figure 4 show the. Rate of speech in words per minute versus the age of the patient in years. There was no strong correlation found between these two features, and no significant difference found between the monologue and the passage. Old age did not seem to slow speech significantly in either the monologue or the passage. Figure 5 shows the correlation between the two rates, not including age.

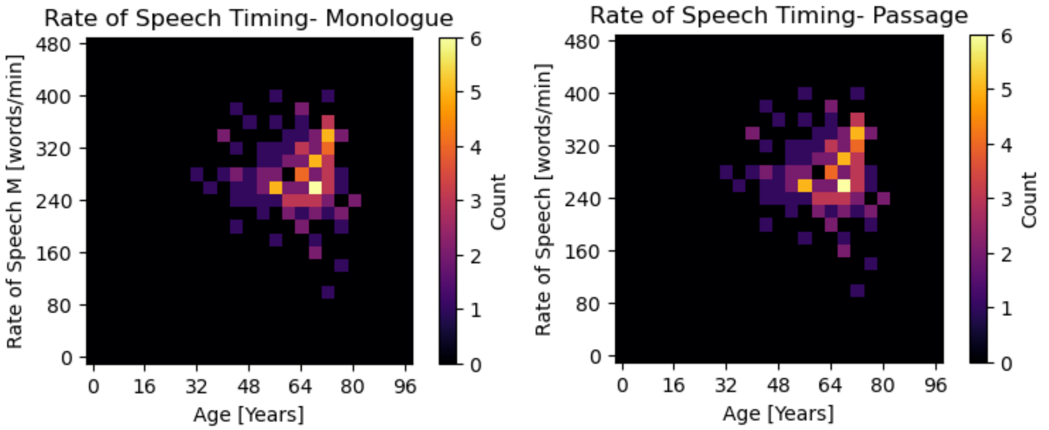

Figure 4. Rate of speech in words per minute versus age in years for the monologue (left) and the passage (right).

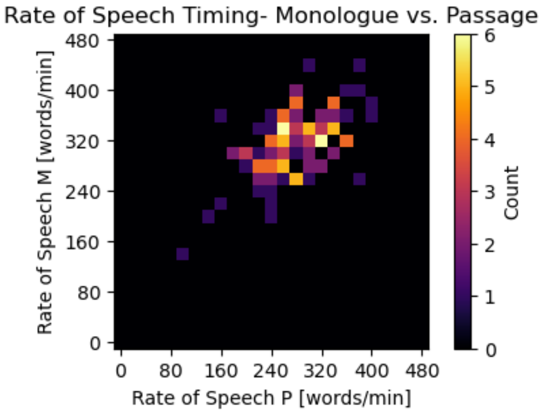

Figure 5. Rate of speech in words per minute of monologue (y-axis) and passage (x-axis).

The ages of patients ranged from about 30 to 80 years old, and the range of the target variable was from 30 to 117 years. 117 years meaning that the patient will not develop disease within their lifetime.

Methods

Data Splitting

The data were split using the train_test_split function. With 130 total datapoints and balanced (30, 50, 50) types of patients, the train set used 60% of the datapoints, with 20% each for the test and validation sets. The sets were confirmed to have a balanced portion of the three types of patients after splitting.

Preprocessing

The target variable 'age_of_disease_onset' was separated into y, with the rest of the data used in X. The feature 'participant_code' was dropped, as this is the equivalent to an 'ID' column and serves no purpose to the training of the model. This feature was used to ensure that the types of patients were balanced after splitting but was no longer needed from this point forward. There were three different types of features in this dataset: categorical, numerical, and ranked scores from 1-4. The categorical features included gender, history (family history of PD), and four others for whether the patient was on certain medications, and the names of those medications. Missing values for history were changed to being reported as "No", and missing values for medication were reported as "None". These features were preprocessed using the OneHotEncoder. The numerical features included age, amount of medication (in mg), and all duration and rate measurements. These were preprocessed using the StandardScaler. Lastly, the ordinal features included the tremor and rigidity scores. As stated, each extremity and the head and neck area were scored from 1-4 for each patient depending on the amount of movement present in the video of the speech. Missing values in these features were reported as "0", meaning no movement was recorded or analyzed from the patient. These features were all preprocessed using the OrdinalEncoder with categories of ['0','1','2','3','4']. These preprocessing pipelines were combined into one preprocessor using ColumnTransformer. This was used to fit_transform the training set and fit the test and validation sets.

Features

Do the use of the OneHotEncoder, the number of features increased from 63 to 75. Some features had missing value percentages around 25%, but this was also fixed in preprocessing. An absence of a value in this situation meant an absence of a measurement, indicating a healthy biomarker.

ML Pipeline

Six machine learning models were trained and evaluated for this project. The evaluation metric I chose was mean squared error, with the goal of minimizing this metric to produce a more accurate model. This metric was chosen as the datapoints can be arranged in order of age onto a scatter plot, with the predicted values superimposed onto the true values. The MSE would provide insight into how far off and inaccurate the model's predictions were from the true values (Figure 6). The baseline score was calculated to be an MSE of 823.96. The models trained were Linear Regressor, Ridge Regressor, Lasso Regressor, Support Vector Regressor, Random Forest Regressor, and XGBoost Regressor. As XGBoost showed the most promise, the hyperparameters that were tuned using GridSearchCV were the number of estimators (n_estimators) from 50-400, the max depth from 3-9, and the learning rate from 0.01 to 0.3. Hyperparameter tuning of this model is summarized below in Table 1.

Table 1. Parameters tuned for XGBoost Regressor model.

Parameter	Tuning Range
n_estimators	50, 100, 200, 300, 350, 400
max_depth	3, 5, 7, 8, 9
learning_rate	0.01, 0.02, 0.03, 0.04, 0.05, 0.075, 0.1, 0.2, 0.3

Results

The results for best MSE for each model trained can be seen in Table 2. The results for mean MSE and standard deviation can be seen in Figure 6.

Table 2. Best MSE calculated for each model trained.

Model	Best MSE
Baseline	823.96
Linear Regressor	1.39e24
Ridge Regressor	898.22
Lasso Regressor	459.47
Support Vector Regressor	659.68
Random Forest Regressor	18.01
XGBoost Regressor	5.28

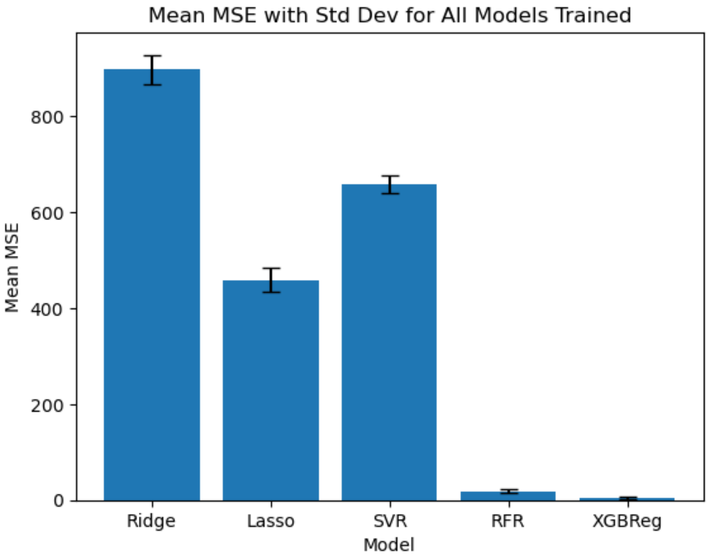

Figure 6. Mean MSE for each model trained and standard deviation bars. Linear Regressor was not included due to the very high MSE of 1.29e24.

The worst model was found to be linear regression, with the best MSE being 1.29e24 (Figure 7). The best model was found to be XGBoost Regressor with n_estimators of 200, a max_depth of 7, and a learning rate of 0.05 (Figures 8 and 9). This model had an MSE of 5.28.

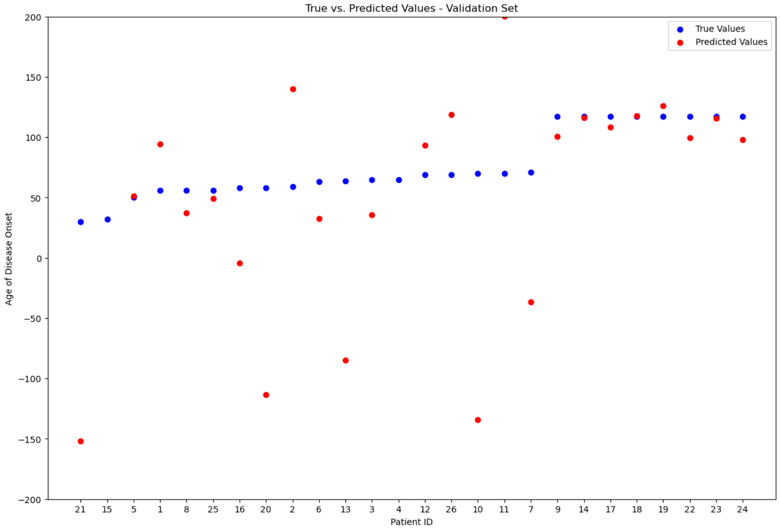

Figure 7. Scatter plot for the linear regression model trained of age of disease onset for each patient in the validation set. True values of the validation set are in blue and the model's predicted values are in red. The 'Patient ID' on the x-axis is an arbitrary label given to each member of the dataset, as the 'participant code' was dropped as feature before preprocessing.

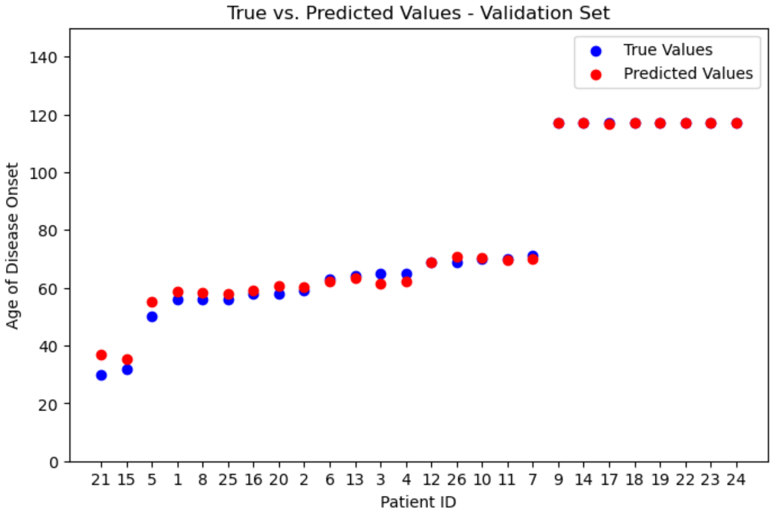

Figure 8. Scatter plot for the XGBoost model trained of age of disease onset for each patient in the validation set. This model was found to have the best MSE (lowest MSE) of all models trained in this project.

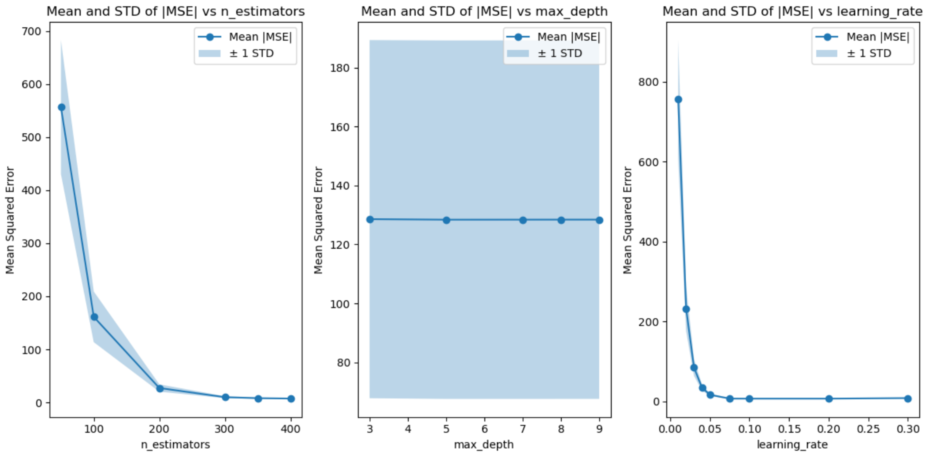

Figure 9. Plots for hyperparameter tuning of the best model, XGBoost Regressor. As shown, the best parameters were found to be n_estimators = 200, max_depth = 7, and learning_rate = 0.05. These values were found using multiple iterations of hyperparameter tuning and adding additional values to see convergence.

Lastly, the top 10 most important features for the XGBoost Regressor model were calculated (Figure 10). Interestingly, the highest of which are current medications that are prescribed to the patients. Higher doses or more medications currently prescribed led to a lower prediction for age of disease onset, meaning they will develop the disease sooner in life. Clonazepam is a seizure medication given to minimize early symptoms of PD. Levodopa is a tremor medication, and benzodiazepines are depressants used as muscle relaxers. Also among the top 10 features were antidepressant medications, duration of disease (if they had just been diagnosed but not treated), and family history. Parkinson's biomarkers are hereditary, explaining the prevalence of the family history feature. Gender was also ranked in the top 10 most important features, as the model predicted lower ages of disease onset for men. This supports the overall observation in PD diagnoses that men are twice as likely to develop PD over women.

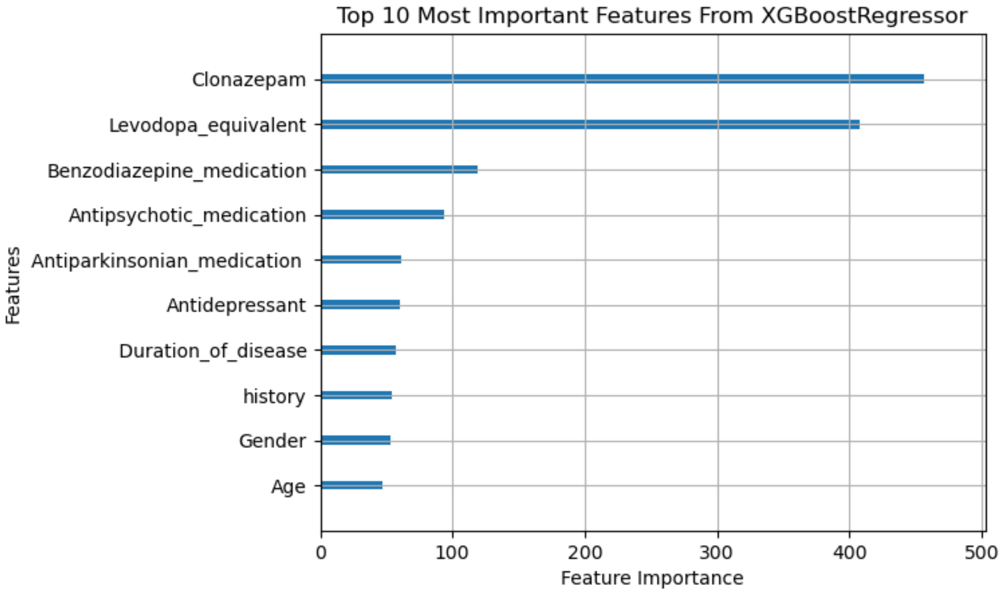

Figure 10. Top 10 most important features found for the best model trained, XGBoost Regressor.

Outlook

To improve the predictive power of this model, more data would be needed. This dataset was quite small, and a model trained on few data points is limited in the potential for accuracy. I did not use fold splitting for this set, as I wanted each of the train, test, and validation sets to have a balanced number of RBD, PD, and healthy patients. A training set with only healthy patients would predict a high age of disease onset for every patient, which would be inaccurate. Interpreting this model is easy, as the input is from an automated tool analyzing speech patterns, and the output is simply an age. Improving this model could include training on only features for speech pattern measurements, and not including current medications. Prescribed medications were found to be in the top 10 most important features, and predictions could change if only based on the speech patterns themselves. In conclusion, the model trained was able to predict the age of disease onset somewhat accurately but would need improvements on the dataset size and features included to increase in its predictive power.

References

1. Parkinson’s Foundation. https://www.parkinson.org/?utm_source=google&utm_medium=cpc&utm_campaign=Google_AlwaysOn_Brand&utm_term=parkinson%27s%20foundation&gad_source=1&gclid=CjwKCAiApuCrBhAuEiwA8VJ6JjrQI0_3PpbSbJxTw_pTLYC7ld9y7hgRjl3vesESAqd5JOQpM53DqRoCXGUQAvD_BwE
2. Singapore Brain Parkinson’s Disease. https://singaporebrain.org/brain/parkinsons-disease/
3. Health Direct. https://www.healthdirect.gov.au/dopamine#:~:text=Dopamine%20acts%20on%20areas%20of,movement%20and%20other%20body%20functions.
4. Early Biomarkers of Parkinson's Disease Dataset. https://www.kaggle.com/datasets/ruslankl/early-biomarkers-of-parkinsons-disease
5. Hlavnička, J.; Čmejla, R.; Tykalovà, T.; Šonka, K. Ružička, E. Rusz, J. Automated analysis of connected speech reveals early biomarkers of Parkinson’s disease in patients with rapid eye movement sleep behaviour disorder. *Nature Scientific Reports* **2017**, *7*(12).