Janet O. Joseph
Division of Biology and Medicine
https://github.com/JOJanet/DS1030Final

Prediction of Parkinson's Disease Based on Naturally Connected Speech

1. Introduction

Purpose

Parkinson's Disease (PD) is a progressive neurological disorder that affects the nervous system. It is a movement disorder that originates when the nerve cells in the brain do not produce enough dopamine, which could lead to tremors, slow movements, rigidity, and speech that may be slow, soft, and difficult to understand[1]. Diagnosing PD at an early stage can lead to better treatment options and slow the progression of the disease. Disruption of naturally connected speech, one of the symptoms of PD can be used as an early diagnosis of PD.

Dataset

The dataset on Early Biomarkers of Parkinson's Disease (EBPD) using naturally connected speech comprises 30 individuals diagnosed with early untreated Parkinson's Disease (PD), 50 individuals with Rapid eye movement (REM) sleep behavior disorder (RBD) who are at a heightened risk of developing PD, and 50 healthy controls (HC). Each participant engaged in reading an 80-word standardized, phonetically balanced text and delivering a 90-second monologue on topics such as their interests, jobs, family, or current activities[2].

Target Variable and Features

The EBPD dataset is a classification dataset. This is a dataset with 130 data points and 66 feature columns. There are 33 features specific to PD and or RBD participants, which include PD medication and motor examination. The other 33 features include features that were collected for all participants, which include age, gender, positive history of PD, medications, and the monologue and paragraph speech examination. Age was recorded in years, gender as male or female, medication as milligrams per day (md/day), and speech examination as a measure of time and sound in decibels. There is a classification of voiced and unvoiced speech, pause, respiration, and speed. The target variable is negative or positive for PD. With preprocessing, HC participants became 0 for negative for PD, while RBD and PD participants became 1 for positive for PD after preprocessing. In total, there are 33 features with the diagnosis of PD as the target variable.

Current Research

To date, no studies have specifically leveraged comprehensive features of naturally connected speech as a diagnostic tool for Parkinson's disease. Instead, more emphasis has been placed on developing machine learning approaches to predict voice changes assessed through vocal analysis, including phonation and articulation[2,3]. While these studies employ different methods, they provide valuable insights into the characteristics of features and their potential significance, serving as a guide for comparison. The study by Hlavnička et al. achieved accuracies ranging from 55% to 85% in models focusing on phonation and articulation, with Quan et al.'s model reaching 85%. Klempíř et al. concentrated on PD diagnosis classification based on speech signals, utilizing pattern recognition methods such as AdaBoost, Bagged trees, Quadratic SVM, and k-NN, resulting in an overall accuracy of 82.3%[4]. The strength of the EBPD dataset lies in its ability to reliably capture subliminal Parkinsonian speech deficits, even in patients with RBD, who are at a high risk of developing PD[2]. This underscores the dataset's suitability for discerning differences between individuals with and without PD. In contrast to other studies, this analysis

integrates all voice analysis aspects for prediction rather than focusing on specific vocal changes, which addresses the gap in PD diagnosis in a more thorough machine learning assessment.

2. Explanatory Data Analysis (EDA)

Feature Analysis

The analysis began with the utilization of the `.info()`, `.isnull()`, and `.sum()` methods to assess and quantify missing values within the dataset. There were 0.38 missing data in the History of Parkinson's Disease in Family feature, which was replaced with NaN.

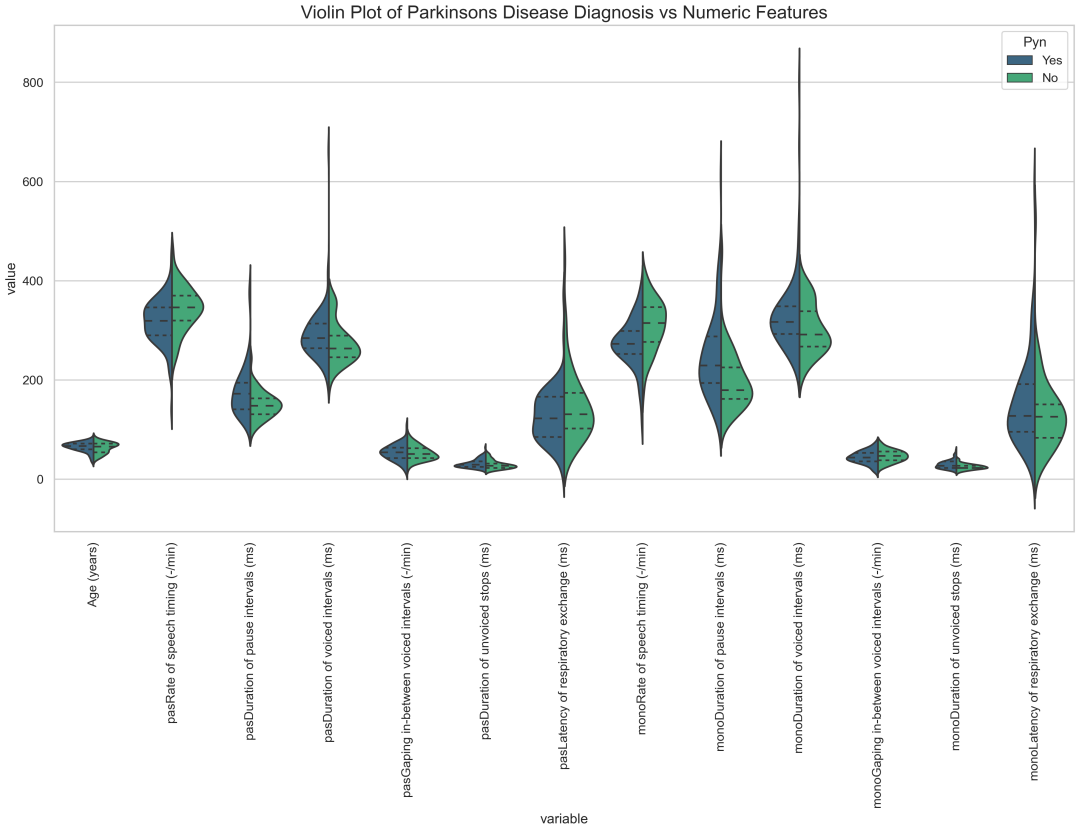


Figure 1: A violin plot generated for all numeric features, followed by filtering out features with values less than 20. This preliminary step is crucial for later comparisons with the feature importance determined by the machine learning algorithms presented in subsequent sections of the report.

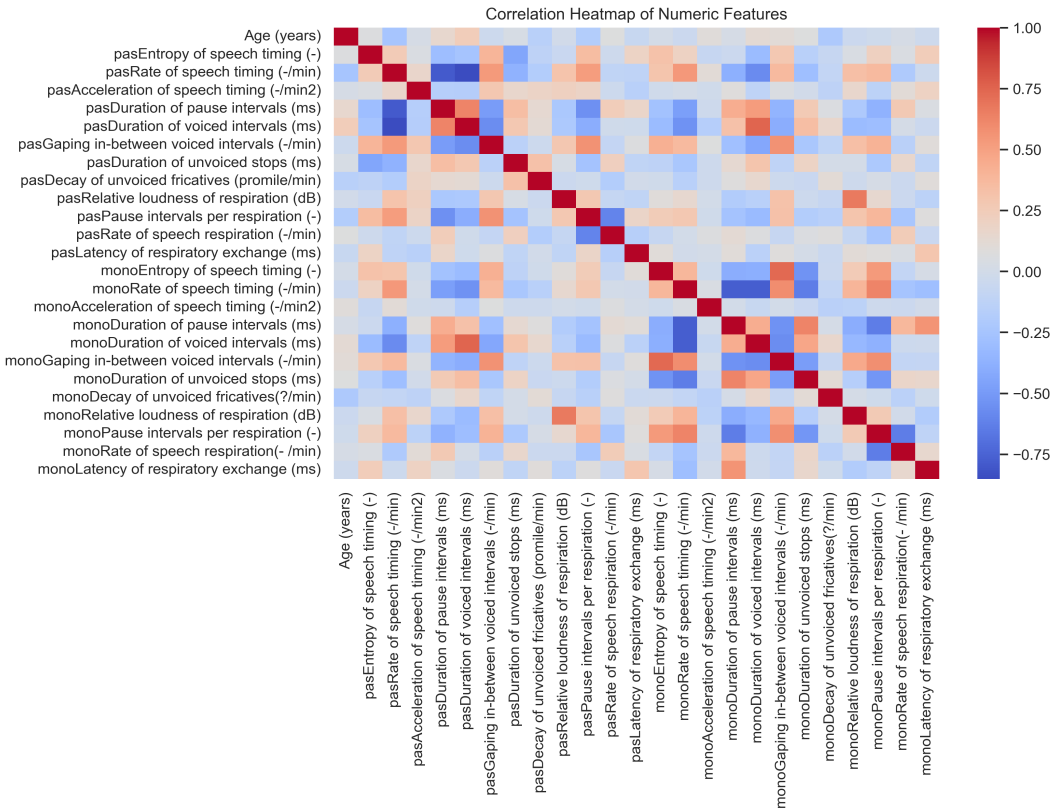


Figure 2: The heatmap incorporates all numeric features and illustrates positive correlations between certain passages and monologue features, such as the relationship between Rate of Speech Timing.

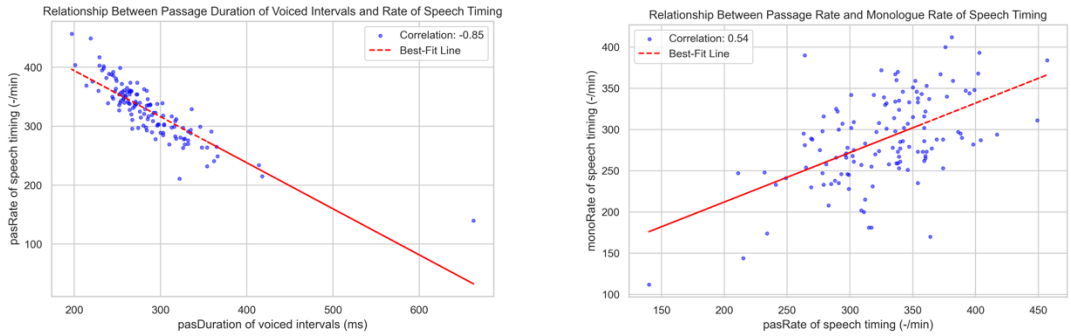


Figure 3: There are associations of the numeric features like Rate of Speech Timing and Duration of Voiced Intervals; Monologue Rate of Speech Timing and Passage Rate of Speech Timing. Identifying these correlations is valuable not only for pinpointing potentially significant features but also for aiding neurologists, scientists, and technicians in understanding relationships that could streamline data collection and assessment. Recognizing whether certain features exhibit positive or negative correlations with others can alleviate the burden of comprehensive data collection and analysis.

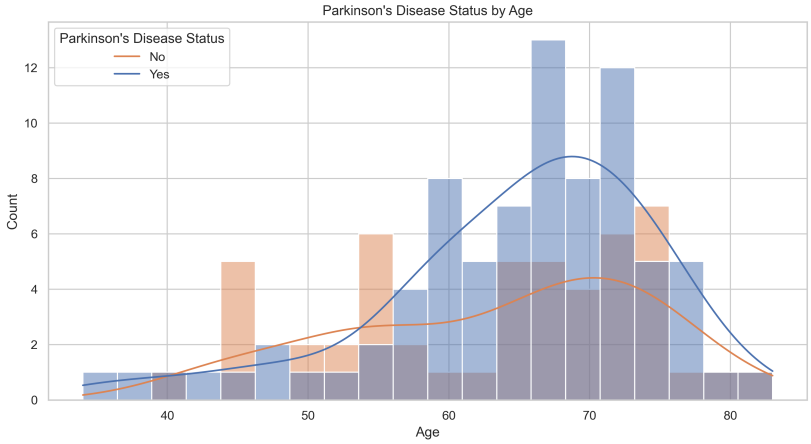


Figure 4: This figure illustrates the distribution of PD across different age groups, providing a key validation indicator for the dataset. The expected concentration of PD cases in later age stages aligns with the recognized age-related risk for PD. This visual assessment reinforces the dataset's credibility, aligning with the established epidemiological correlation between age and Parkinson's Disease.

Target variable

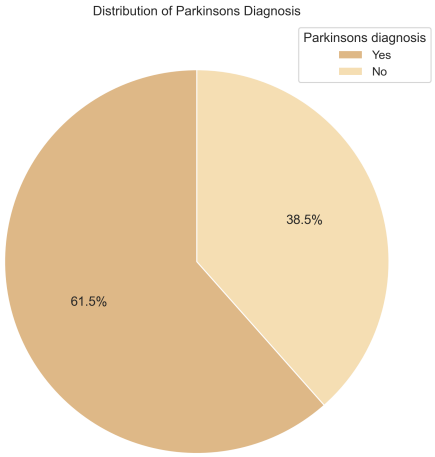


Figure 5: Distribution of Parkinson's Diagnosis. The target variable in this dataset is 'Parkinson's Diagnosis'. Participants in the study are categorized into two groups: 'Yes' or 'No' for PD. 80 participants are positive for PD and 50 participants are negative. A total of 130 data points and 33 features will be utilized in this analysis.

3. Methods

Dataset Split - Data preprocessing

To accommodate the small EBPD dataset, a 6:2:2 (Train:Validation:Test) split ratio was employed across machine learning models (Logistic Regression – LR, Decision Trees – DT, K-Nearest

Neighbors – KNN, Support Vector Machines – SVM, and XGBoost). This distribution ensures sufficient training data while maintaining distinct subsets for validation and testing, enhancing the models' robust evaluation. Additionally, 10-fold cross-validation with shuffling was applied to address potential bias or variance issues, providing a balanced approach to effective training, robust validation, and reliable testing. This strategy aims to comprehensively assess model performance, mitigating risks of overfitting or underfitting in the context of a limited dataset like EBPD.

Preprocessor

In the EBPD dataset, both continuous and categorical values are present. The preprocessing of features involved using OneHotEncoder for categorical features such as Gender, Positive history of Parkinson's disease in the family, and Antidepressant therapy. Additionally, MinMaxScaler was applied to features like Age (years), Levodopa equivalent (mg/day), Clonazepam (mg/day), pasEntropy of speech timing (-), monoEntropy of speech timing (-), and pasRate of speech timing (-/min). MinMaxScaler was specifically chosen for Age to ensure it shares a comparable scale with the speech features. All other measures were normalized using MinMaxScaler.

Features

The features were reduced from 66 to 33 to account for only features that are generalizable to PD, RBD, and HC participants. An extra column was created for the target variable, which is negative or positive for PD based on participants' status in the study as PD positive, RBD positive, or HC.

ML Pipeline

Logistic Regression, Decision Tree, K-Nearest Neighbors, Support Vector Machine, Random Forest, and Extreme Gradient Boosting (XGBoost) machine-learning pipelines were employed in this analysis and several parameters were tuned for each algorithm (Table).

ML Algorithms	Parameters
Logistic Regression	C: [0.001, 0.01, 0.1, 1.0, 10, 100]
Decision Trees	DT__max_depth: [3, 5, 7, 10, 13]
K-Nearest Neighbors	n_neighbors : [1, 3, 5, 7, 10, 20, 50]
Support Vector Machines	C: [1e-1, 1e0, 1e1, 1e2]; Gamma: [1e-3, 1e-1, 1e3]
Random Forest	n_estimators: [50, 100, 200]; max_depth: [3, 5, 7, None] min_samples_split: [2, 5, 10]; min_samples_leaf: [1, 2, 4] max_features: ['sqrt', 'log2', None]
XGBoost	XGB__reg_alpha: [1e-2, 1e-1, 1e0, 1e1, 1e2] XGB__max_depth: [1, 3, 5, 10, 50]

Table 1: Tuned Parameters for Each Machine Learning Algorithm

The metric employed to assess the performance of these models is accuracy, aligning with the approach advocated by other peer-reviewed studies for Parkinson's disease (PD) diagnosis using machine learning. The dataset used is relatively balanced, with 61.5% of participants testing positive for PD and 38.5% negative. Opting for accuracy over F1 score is considered more informative for gaining insights into the dataset. In this machine learning pipeline, diverse classification models, namely Logistic Regression, Decision Tree, K-Nearest Neighbors, Support Vector Machine, Random Forest, and XGBoost, underwent evaluation on a dataset featuring features relevant to Parkinson's disease. Each model underwent a 10-fold cross-validation, incorporating hyperparameter tuning via GridSearchCV. At the conclusion of the loop, the best model, along with its validation and test scores, was saved to a list. Evaluation involved multiple

random splits of the data, and the best hyperparameters were chosen based on their performance on the validation set. Subsequently, the test set was utilized to gauge the generalization performance of each model. The results exhibited variability within random states of models and within the models themselves.

4. Results

ML Algorithms	Mean Test Score	Standard Deviation	(Mean test score – baseline) / standard deviation	Baseline Model
Logistic Regression	0.661538	0.037684	17.512110	0.3846 to 0.6923
Decision Trees	0.638462	0.113053	5.651122	0.4231 to 0.7308
K-Nearest Neighbors	0.676923	0.071336	9.500151	0.3462 to 0.8077
Support Vector Machines	0.692308	0.048650	14.231319	0.3846 to 0.7308
Random Forest	0.700000	0.078446	8.945890	0.4615 to 0.8077
XGBoost	0.676923	0.071336	9.500151	0.4231 to 0.7692

Table 2: Mean values and Standard Deviation of Machine Learning Algorithms with their Range of Baseline Model Across Random States (21, 42, 84, 168, & 336). The baseline accuracy score is 0.6154, calculated by dividing the positive Parkinson's disease participants by all participants. In terms of mean test scores, Random Forest outperformed with an average score of 0.70, closely followed by K-Nearest Neighbors at 0.6769. Logistic Regression stood out with a higher ratio of mean test score to standard deviation (17.5), indicating significant improvement over the baseline with low variability.

Figure 6: Comparison of mean test scores and standard deviations of all models compiled from five random states. They all range from 0.64 – 0.70, with variabilities in their standard deviation from 0.03-0.11.

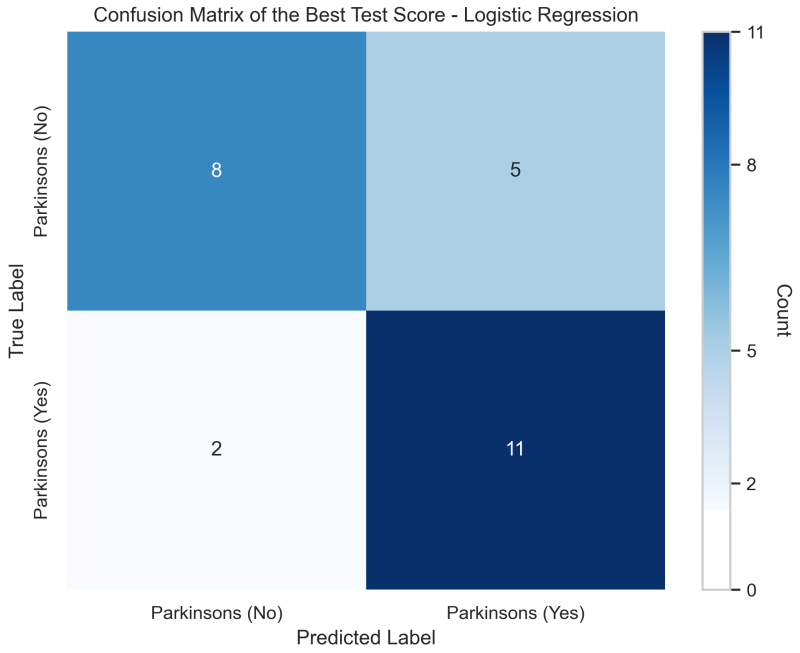
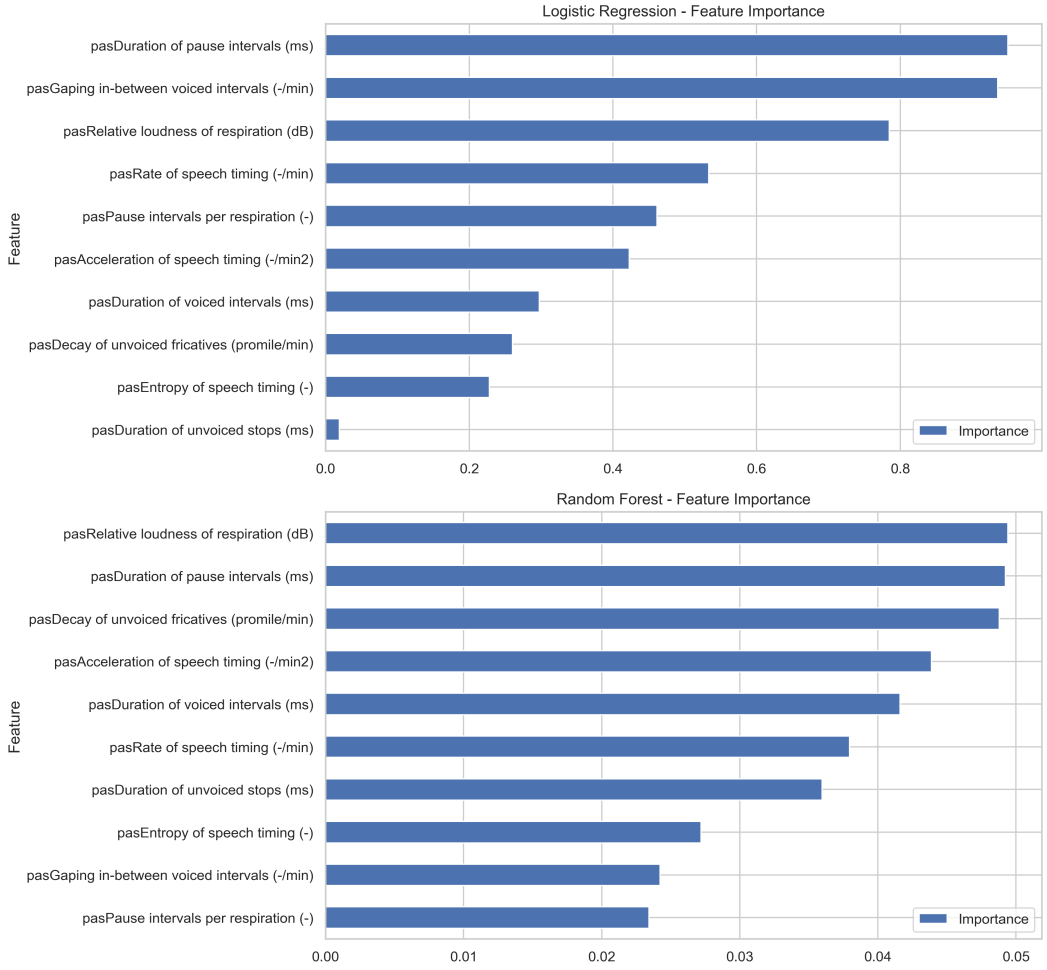


Figure 7: The Best Logistic Regression model out of the 5 random states predicted 8 out of 13 true negative and 11 out of 13 true positive cases for Parkinson's disease.



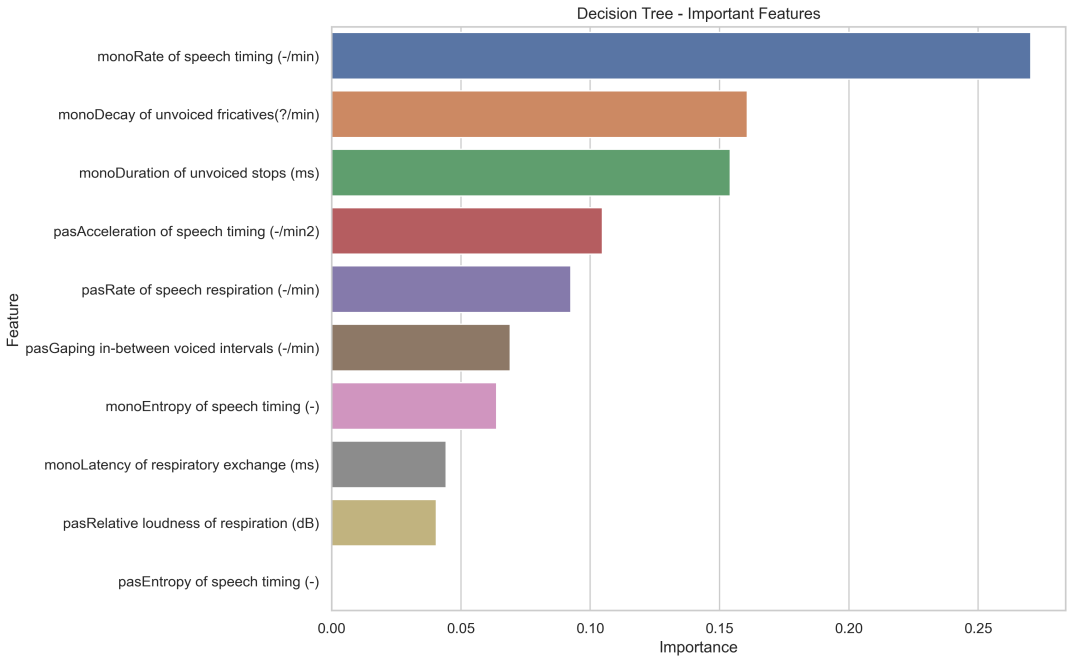


Figure 8 A, B, & C shows the top 10 features for Logistic Regression, Random Forest, and Decision Tree. All the top features are passage-based, except for Decision Tree importance which incorporates monologue speech in its top 3. It is interesting to note that a feature of least importance, such as Passage Relative Loudness of Respiration is the top feature in both Random Forest and Logistic Regression. The 3 most important in the models are Passage Duration of Pause Intervals, Passage Relative Loudness of Respiration, and Acceleration of Speech Timing. The 3 least important are Passage Duration of Unvoiced Stops, Passage Entropy of Speech Timing, and Monologue Latency of Respiratory Exchange.

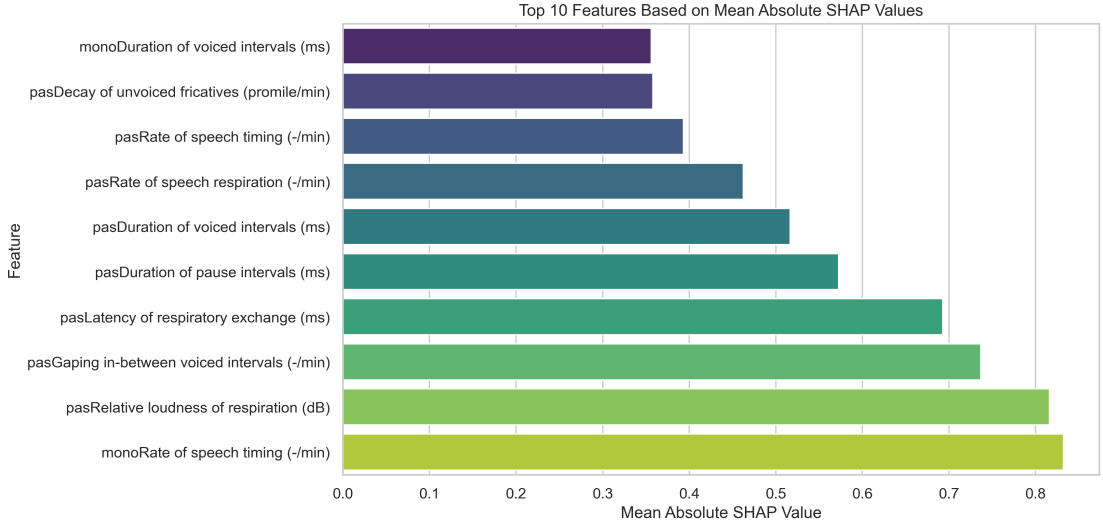


Figure 9: The top 10 SHAP values include features from Logistic Regression, Random Forest, and Decision Tree models, with the most influential being Monologue Rate of Speech Timing, Passage Relative Loudness of Respiration, and Passage Gap between voiced intervals. Unlike other models, SHAP considers how these features work in conjunction, addressing nonlinearities and multicollinearity issues, and offering a consistent framework for interpretation across different model types. This makes SHAP valuable for comprehending features and aiding decision-making in diverse scenarios.

It is important to note that the rate of speech timing, loudness of respiration, pause intervals, and voiced intervals all contribute significantly to all models. This shows that the models rightly incorporate different arrays of connected speech into the prediction. This importance is also reflected in the violin plot (Figure 1).

5. Outlook

Out of all the models, Logistic Regression maintains a reliable and competitive performance with an average score of 0.6615, demonstrating consistency and effectiveness, as reflected in its higher ratio of mean test score to standard deviation (17.5). Logistic Regression is ultimately not an ideal model for PD diagnosis using this data, but it can be improved. To enhance the performance, feature engineering techniques could be incorporated to extract more relevant information from the existing variables. Additionally, optimizing hyperparameters through techniques like grid search or random search can fine-tune the model for better results. Lastly, expanding the dataset or employing data augmentation methods may provide the model with a more diverse set of examples, potentially improving its generalization and robustness.

6. Reference

1. Bastiaan R Bloem, Michael S Okun, Christine Klein, “Parkinson's disease,” The Lancet, Volume 397, Issue 10291, Pages 2284-2303, ISSN 0140-6736, 2021. https://doi.org/10.1016/S0140-6736(21)00218-X
2. Hlavnička, J., Čmejla, R., Tykalová, T. et al. Automated analysis of connected speech reveals early biomarkers of Parkinson’s disease in patients with rapid eye movement sleep behaviour disorder. Sci Rep 7, 12 (2017). https://doi.org/10.1038/s41598-017-00047-5
3. C. Quan, K. Ren and Z. Luo, "A Deep Learning Based Method for Parkinson’s Disease Detection Using Dynamic Features of Speech," in IEEE Access, vol. 9, pp. 10239-10252, 2021. https://doi.org/10.1109/ACCESS.2021.3051432
4. Ondřej Klempíř, Radim Krupička, “Machine Learning Using Speech Utterances For Parkinson Disease Detection,” Lekar a technika – Clinician and Technology, vol. 48(2), pp. 66–71, ISSN 0301-5491 (Print), ISSN 2336-5552 (Online), 2018. https://ojs.cvut.cz/ojs/index.php/CTJ/article/view/4881