Regression Prediction of Parkinson’s Disease Progression Using Voice Features

Department of Chemistry
Yitian Liu

https://github.com/liulangdog1/DATA1030-Final-Project

1.Introduction

Purpose

Parkinson’s disease is a progressive neurological movement disorder that worsens over time and currently has no cure[1]. The exact cause of Parkinson’s remains unknown. Diagnosing the disease is both time-consuming and expensive, often relying on MRI scans or time-consuming clinical consultations with doctors. As a result, developing a non-invasive and cost-effective method to monitor disease progression could significantly reduce the financial burden on patients. This study aims to explore whether voice features, extracted from phone calls between researchers and patients, can be used to predict Parkinson's disease progression. Specifically, the relationship between voice features and disease severity are investigated using regression models.

Dataset

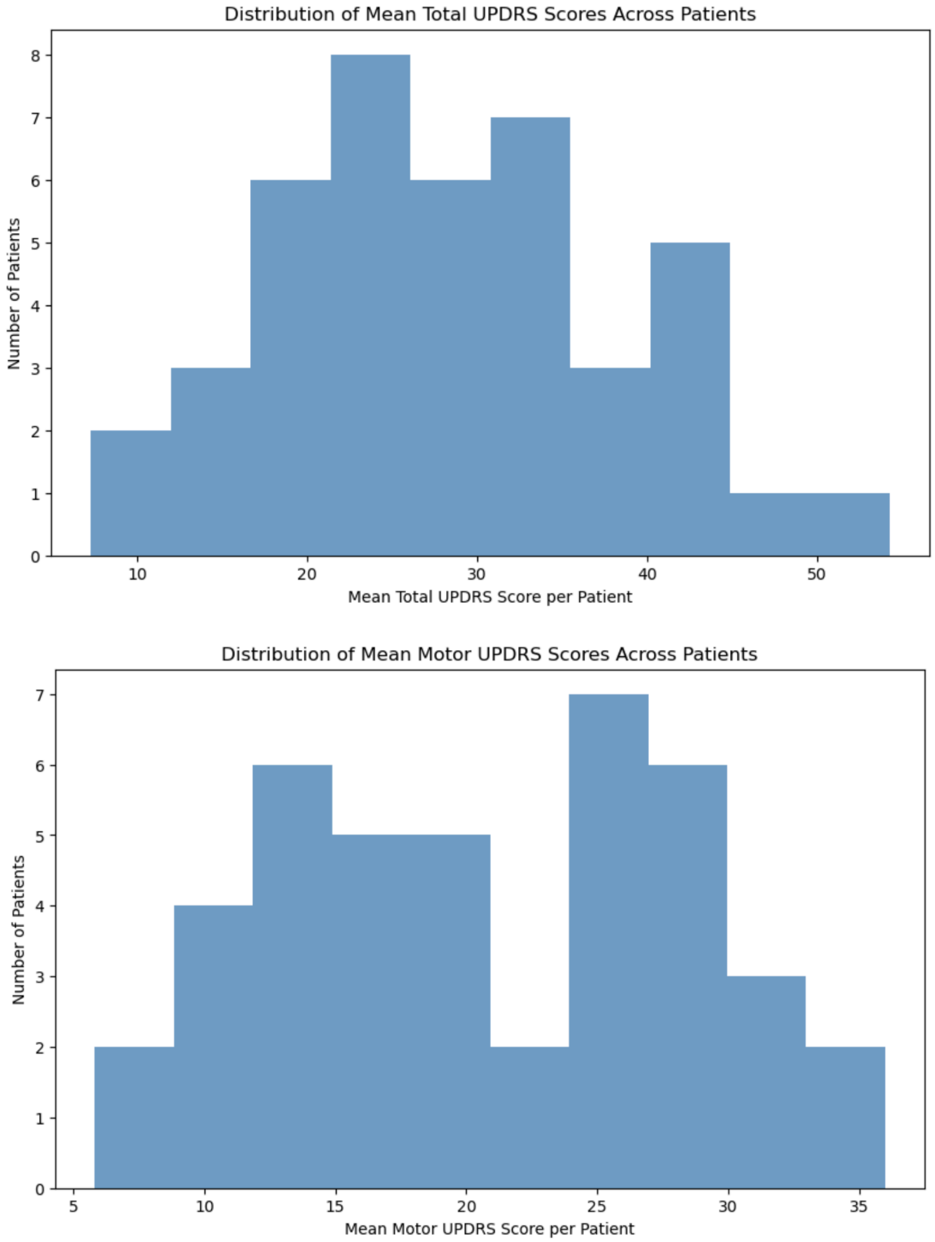
The dataset is sourced from the UCI Machine Learning Repository and includes biomedical voice features collected from 42 patients in the early stages of Parkinson’s disease[2]. These patients participated in a six-month telemonitoring study, during which their voices were recorded approximately 100 times each. The dataset contains 5,875 instances and 19 features, including patient ID, age, and various voice measurements such as jitter and shimmer. The target variables are the motor UPDRS score, which focuses solely on motor symptoms, and the total UPDRS score, which accounts for both motor and mental symptoms. All features are numerical and continuous. The dataset contains no missing value.

Previous Study

Previous research[3] using the same dataset utilized various statistical regression methods to predict UPDRS scores, including classical least squares regression, iterative reweighted least squares, least absolute shrinkage and selection operator (LASSO), and classification and regression trees (CART). The study used mean absolute error (MAE) as the evaluation metric and conducted 1,000 runs of 10-fold cross-validation. Results showed that non-linear models achieved lower MAE scores compared to linear models, indicating that nonlinear machine learning methods better capture the relationship between voice features and UPDRS scores.

2.EDA

Two histograms were first used to examine the distribution of the average motor UPDRS and total UPDRS scores.



Distribution of Mean Total UPDRS Scores Across Patients

Distribution of Mean Motor UPDRS Scores Across Patients

Figure 1. The distribution of average motor UPDRS and total UPDRS

These distributions are important because they show that most scores lie between 10 and 35, which helps set expectations for the models' predictions.

To investigate multicollinearity among features, a heatmap was generated to visualize pairwise correlations. The heatmap indicates that several features within the same category have strong correlations, such as *Shimmer* and *Shimmer:APQ3*. Features with a correlation greater than 0.95 were removed to reduce redundancy. From the heatmap, three features—Age, HNR (Harmonics-to-Noise Ratio), and PPE (Pitch Period Entropy)—show strong to moderate correlations with the motor and total UPDRS scores.

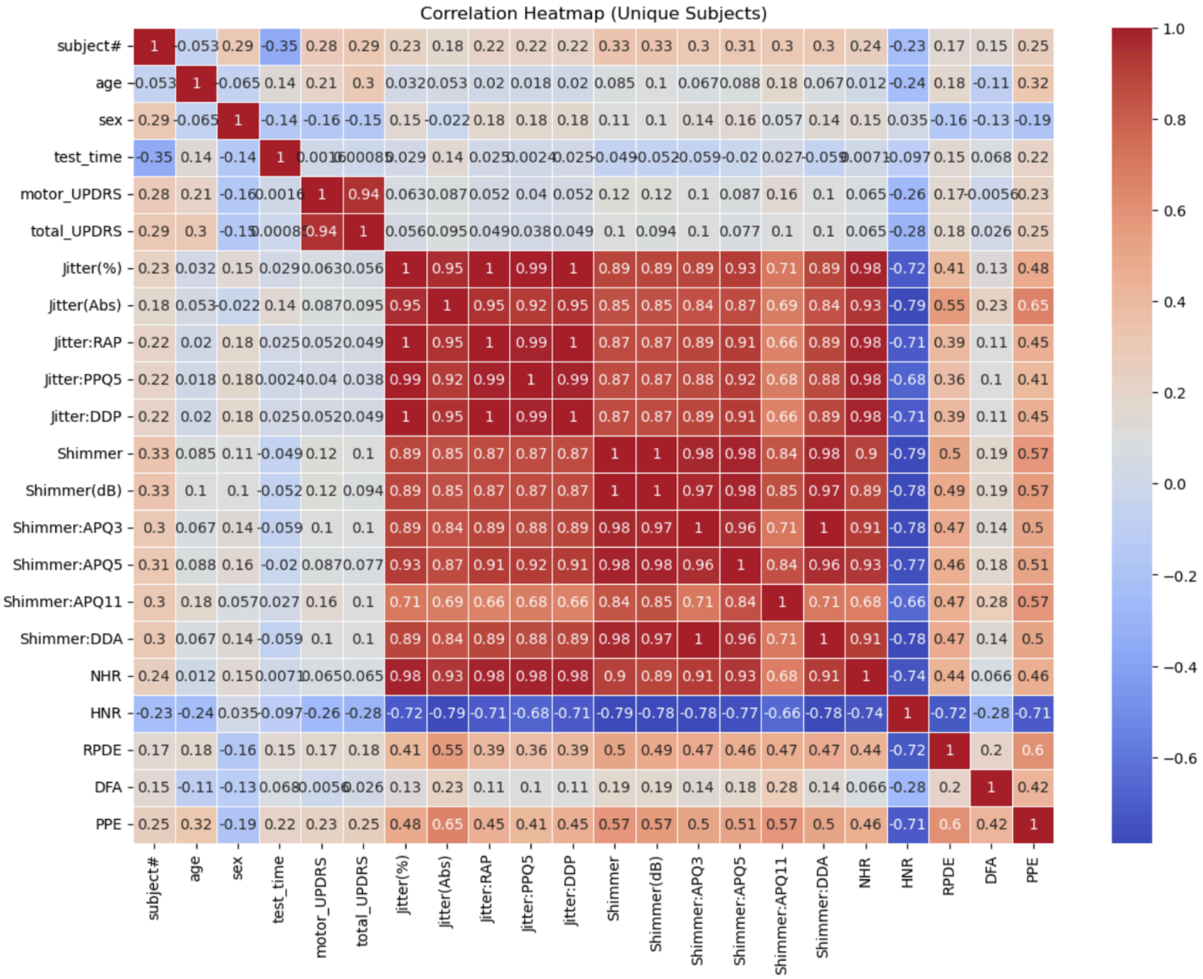


Figure 2. The heatmap of all variables. The color difference indicate whether the mutual relationship is positive or negative

While Age is a well-known factor associated with Parkinson's disease progression, HNR and PPE are less commonly discussed in this context. To better understand the relationship between these features and the target variables, linear regression analyses were conducted.

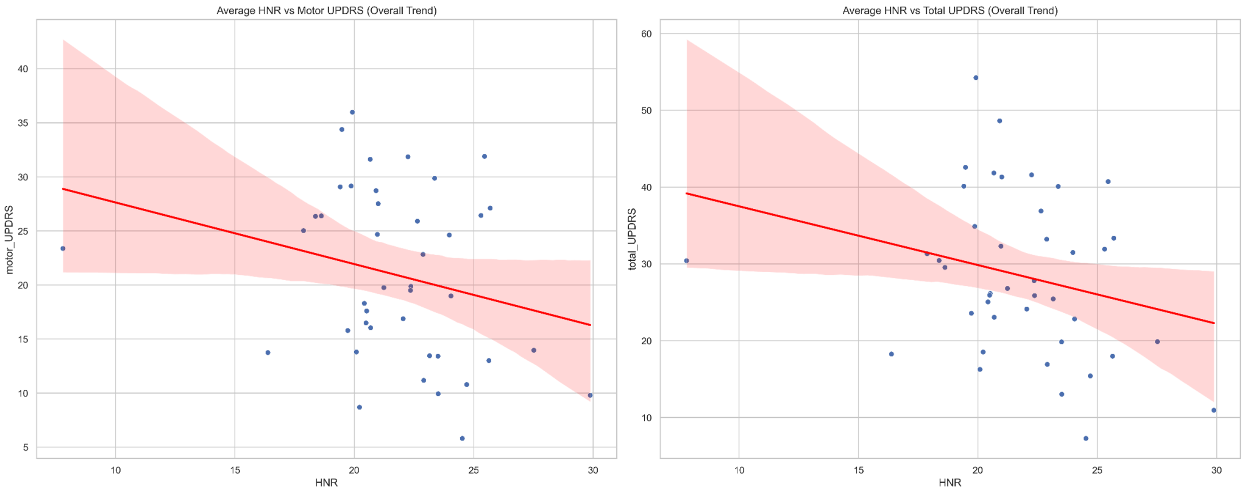


Figure 3. The relationship between average HNR and Motor/Total UPDRS score. Best fit lines were plotted

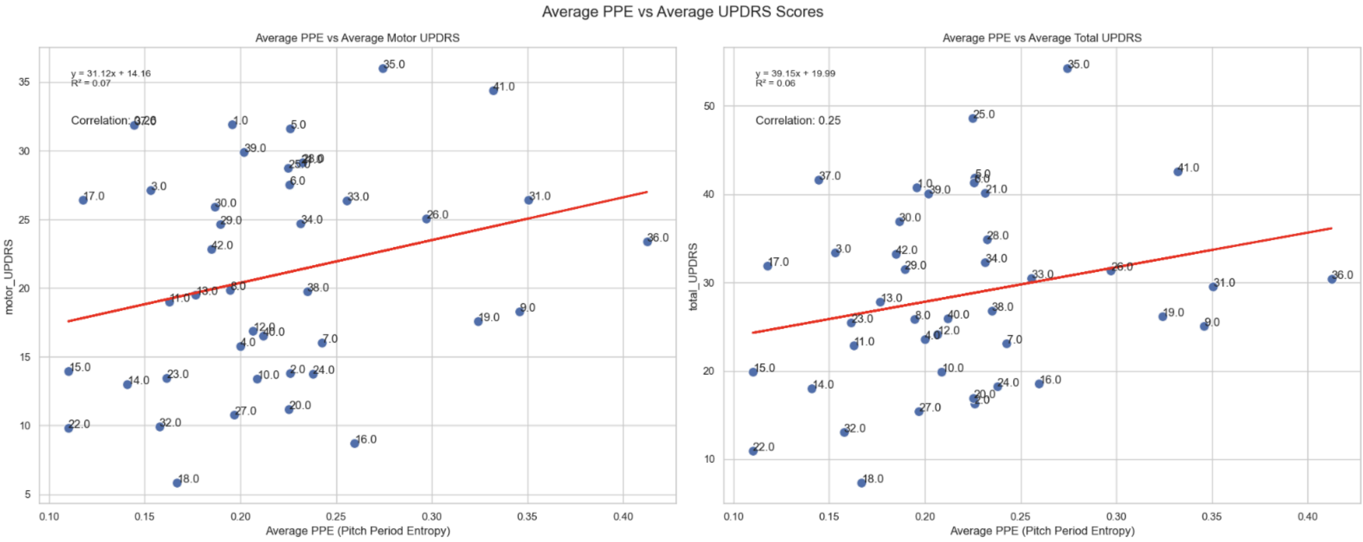


Figure 4. The relationship between average PPE and Motor/Total UPDRS score. Best fit lines were plotted, the number around each dot refers to patient ID

From the figures, it can be concluded that HNR has a positive relationship with UPDRS scores, while PPE shows a negative relationship with UPDRS scores.

3. Methods

Splitting

The dataset is non-iid since each patient was measured approximately 100 times over a six-month period. Although the features and target variables were averaged, intrinsic temporal dependencies still exist. To address this, the data was split using the train_test_split method, with 33 patients (80%) in the training set, 4 patients (10%) in the validation set, and 5 patients (10%) in the test set. Splitting was performed at the patient level using the patient ID as a unique identifier to prevent data leakage between sets. A fixed random_state was used to ensure reproducibility of the split.

Preprocess

Two preprocessing steps were applied to prepare the data for machine learning:

1. Standardization: The features were normalized using StandardScaler to have a mean of 0 and a variance of 1, which is essential for models like KNN and SVR.
2. Feature Selection: Highly correlated features (correlation > 0.95) were removed to reduce multicollinearity. Before feature selection, there were 19 features, the final set of features included: Age, Jitter (%), Shimmer, HNR, RPDE, DFA, and PPE.

Evaluation metric

The primary evaluation metric used was Root Mean Square Error (RMSE), which is suitable for continuous target variables like UPDRS scores. Also, RMSE squares the residuals before averaging them. This means larger errors are penalized more heavily, which is particularly important in this study, as large prediction errors can lead to incorrect predictions of disease severity. Also, RMSE has the same unit as the target variable, make the result easy to interpret.

ML pipeline

Five machine learning algorithms were implemented: ElasticNet (Linear Model), K-Nearest Neighbors (KNN), Random Forest, XGBoost, and Support Vector Regression (SVR). GridSearchCV was used to find the best hyperparameter. The table below summarizes the hyperparameters tuned and their respective ranges.

Model	Hyperparameters	Tuning Range	CV method
Linear Regression	Alpha, L1 ratio	[0.0001,0.001,0.01,0.1, 1] [0.1,0.3,0.5,0.7]	3-fold CV
KNN	n_neighbors	[1, 3, 5, 7, 10]	3-fold CV
Random Forest	Max_depth,, min_samples_split	[1, 3, 10], [0.01, 0.1]	3-fold CV
XGBoost	Max_depth, n_estimators	[3, 10, 30], [100, 300]	3-fold CV
SVR	c,gamma,kernel	[0.1, 1, 10], [0.01, 0.1, 1], ['rbf']	3-fold CV

Table 1. The ML models, hyperparameters, tuning ranges and CV methods

Evaluation metric uncertainty

To calculate model uncertainty, 10 iterations were performed, with each iteration generating a new random train-test split. For deterministic models, variability arises solely from the data split. In contrast, for non-deterministic models, variation results from both the data split and the model's internal randomness. The results of this analysis are summarized in the following figure. From the figure, it can be seen that Total UPDRS has a larger uncertainty compared to Motor UPDRS. Also, XGBoost shows a larger uncertainty compared to other models.

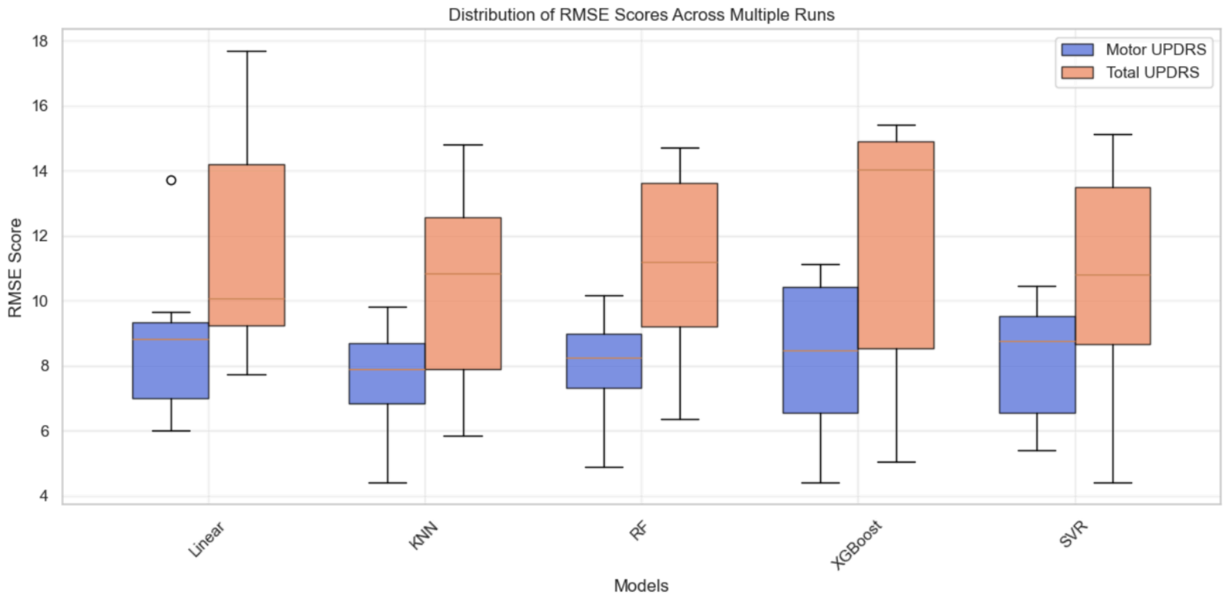


Figure 5. The uncertainty across each ML model

Baseline

The baseline used is the mean of the target variables (motor UPDRS score and total UPDRS score) from the training set. This baseline serves as a computationally efficient benchmark to evaluate and compare model performance. Since this benchmark predicts the average target value for all samples, any model that cannot outperform this baseline indicates poor predictive power.

4.Result

The table below summarizes the best hyperparameters identified during the cross-validation (CV) process.

	Motor UPDRS	Total UPDRS
Linear	Alpha:1.0 l1_ratio:0.1	Alpha:1.0 l1_ratio:0.7
KNN	n_neighbors=10	n_neighbors=10

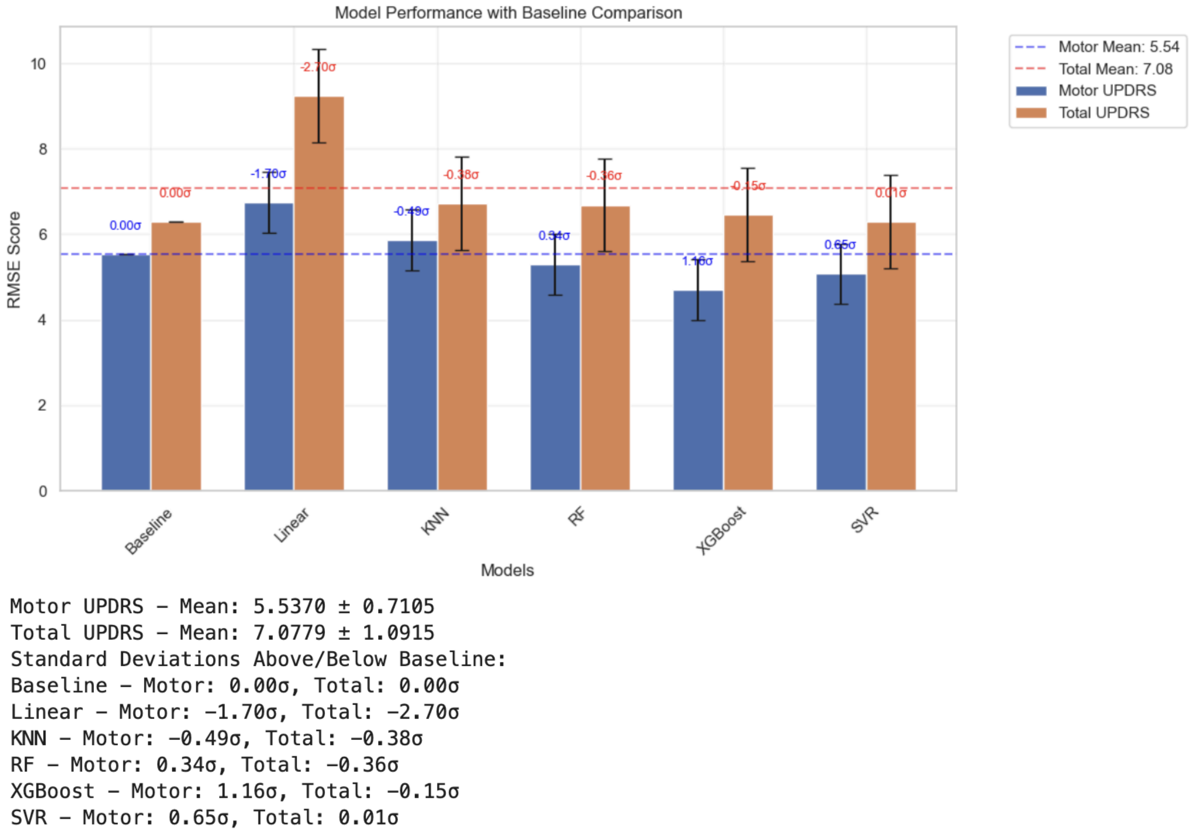
RF	max_depth=1
	min_samples_split=0.01
		max_depth=1
		min_samples_split=0.01

XGBoost	max_depth=3
	n_estimattors=100
		max_depth=3
		n_estimattors=100

SVR	C=10
	gamma=1
	kernel=rbf
		C=0.1
		gamma=0.01
		kernel=rbf

Table 2. The best hyperparameter found through CV

The following figure compares the performance of each model against the baseline with standard deviation. From the figure, it is evident that most models underperformed relative to the baseline, with the linear regression model showing particularly poor results. However, SVR and XGBoost outperformed the baseline, showing superior predictive performance.



```
Motor UPDRS - Mean: 5.5370 ± 0.7105
Total UPDRS - Mean: 7.0779 ± 1.0915
Standard Deviations Above/Below Baseline:
Baseline - Motor: 0.00σ, Total: 0.00σ
Linear - Motor: -1.70σ, Total: -2.70σ
KNN - Motor: -0.49σ, Total: -0.38σ
RF - Motor: 0.34σ, Total: -0.36σ
XGBoost - Motor: 1.16σ, Total: -0.15σ
SVR - Motor: 0.65σ, Total: 0.01σ
```

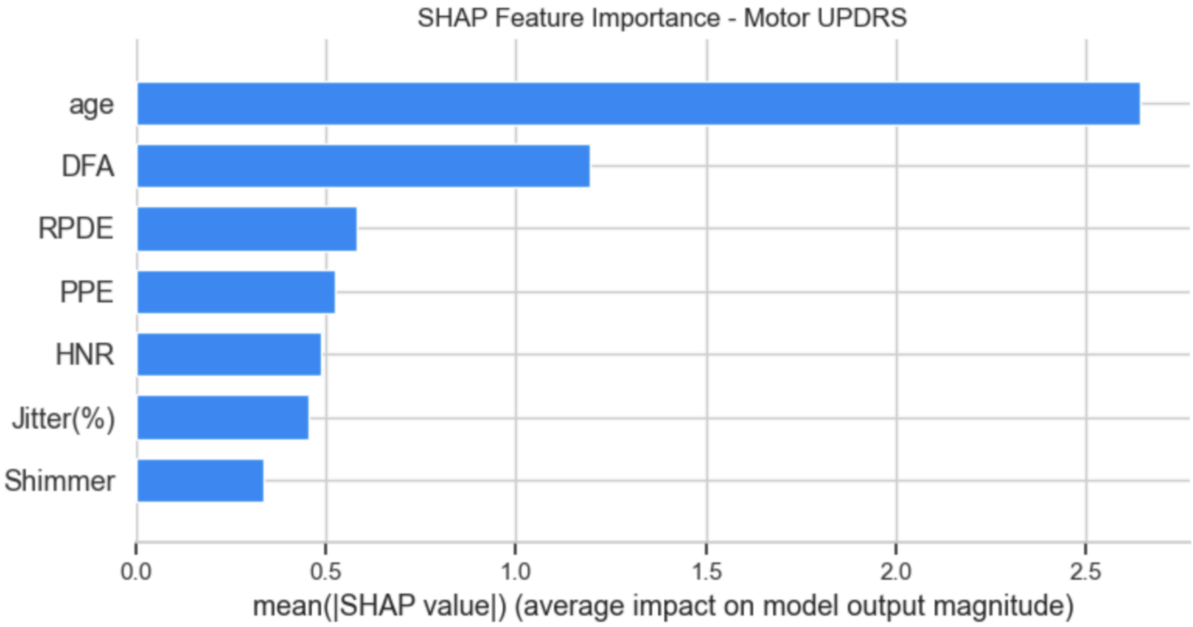
Figure 6. The Models’ Performances VS Baseline with standard deviation and mean value

The following table shows the percentage change in each model's performance relative to the baseline. A positive percentage indicates that the model outperformed the baseline. From the table, it is evident that SVR consistently outperformed the baseline for both target variables. In contrast, XGBoost achieved the best performance for the motor UPDRS score but did not perform as well for the total UPDRS score. Therefore, SVR can be identified as the best-performing model, as it surpassed the baseline for both target variables.

```
           Model  Motor RMSE  Total RMSE Motor Improvement Total Improvement
0       Baseline      5.5319      6.2965             0.00%             0.00%
1         Linear      6.7386      9.2387           -21.81%           -46.73%
2            KNN      5.8767      6.7133            -6.23%            -6.62%
3  Random Forest      5.2895      6.6866             4.38%            -6.20%
4        XGBoost      4.7075      6.4647            14.90%            -2.67%
5            SVR      5.0729      6.2864             8.30%             0.16%
```

Table 3. The model performance comparison

The following figures illustrate the SHAP feature importance analysis. From the results, age is the most influential feature for both motor UPDRS and total UPDRS scores, followed by DFA. For motor UPDRS, the third most important feature is RPDE, while for total UPDRS, it is PPE.



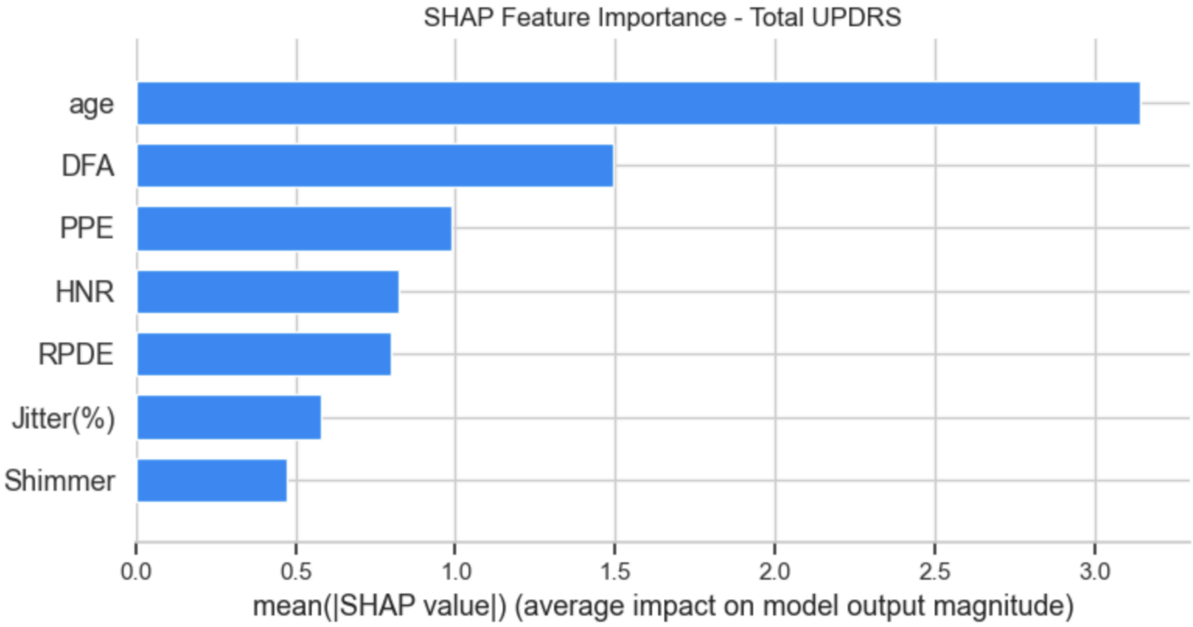


Figure 7. The SHAP Feature Importance Analysis

In the final step, feature importance was compared using different methods, specifically Random Forest (RF), XGBoost, and SVR. The results indicate that age remains the most important feature across all models, followed by DFA as the second most significant feature.

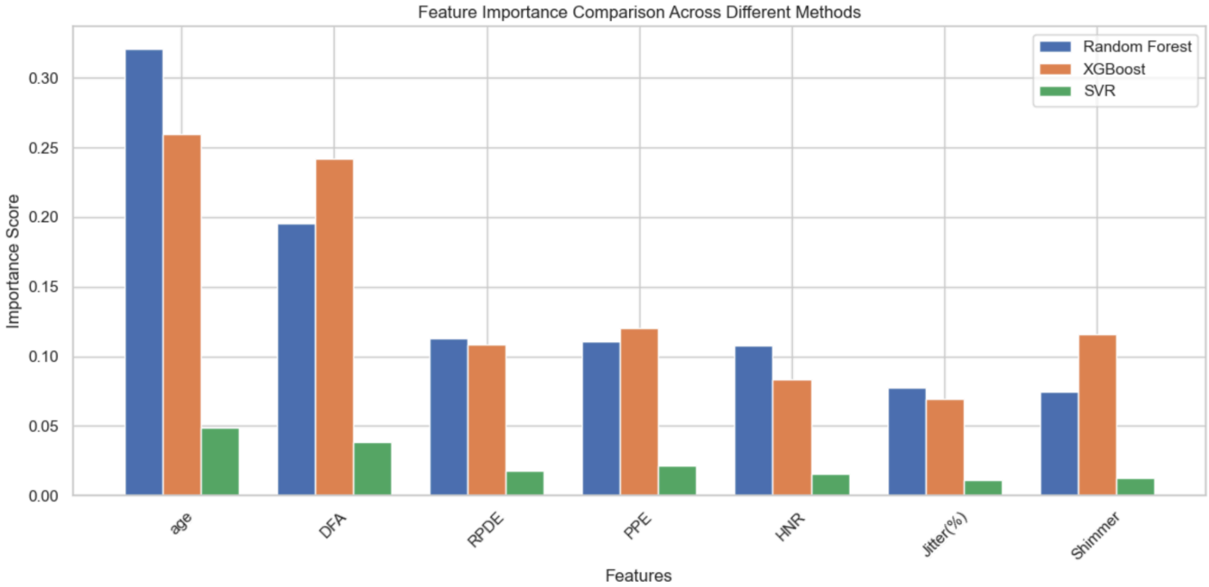


Figure 8. The Feature Importance Comparison Using Different Methods

In summary, The SVR model performed best across all models, outperforming the baseline for both motor UPDRS and total UPDRS scores. This is because the SVR model is able to capture complex non-linear relationships between voice features and UPDRS scores. XGBoost was the second-best model, showing significant improvement over the baseline for motor UPDRS but underperforming for total UPDRS. Other models struggled to surpass the baseline. These results align with previous findings, indicating that non-linear models outperform linear models in capturing the relationship between voice features and UPDRS scores.

Age consistently emerged as the dominant predictor across all models and analyses, showing a strong correlation with both motor and total UPDRS scores. This finding aligns with common understanding: as individuals age, the likelihood of developing Parkinson’s disease increases. Interestingly, HNR and PPE, which quantify voice quality and pitch variability, also show as significant predictors of disease progression. These findings suggest that voice features could serve as valuable biomarkers for Parkinson's disease monitoring, making telemonitoring of Parkinson’s disease progression possible.

5.Outlook

There are several potential improvements for this analysis. The most significant limitation is the small sample size, as only 42 patients were included. Increasing the number of patients would likely enhance model accuracy and improve performance relative to the baseline. Additionally, the majority of patients in this study have UPDRS scores ranging from 10 to 30, with very few high UPDRS scores represented. Including more patients with higher UPDRS scores would improve the generalizability and applicability of the findings.

One weakness of the current approach is the utilization of basic feature preprocessing and selection methods. There are other techniques that can reveal more non-linear relationships between features and target variables, such as polynomial transformation.

Furthermore, this research relied on telemonitoring data collected over a decade ago. With advancements in technology, modern smartphones can now record voice data at higher resolutions and with better accuracy. By using modern cell phone technology for voice recording, the accuracy of this research could be improved.

6.Reference

(1)
National Institute on Aging. *Parkinson's Disease: Causes, Symptoms, and Treatments*. National Institute of Aging. https://www.nia.nih.gov/health/parkinsons-disease/parkinsons-disease-causes-symptoms-and-treatments.

(2)
Tsanas, A. *UCI Machine Learning Repository:Parkinsons Telemonitoring*. archive.ics.uci.edu. https://archive.ics.uci.edu/dataset/189/parkinsons+telemonitoring.

(3)
Tsanas, A.; Little, M. A.; McSharry, P. E.; Ramig, L. O. Accurate Telemonitoring of Parkinson's Disease Progression by Noninvasive Speech Tests. *IEEE Transactions on Biomedical Engineering* **2010**, *57* (4), 884–893. https://doi.org/10.1109/tbme.2009.2036000.

7. Github Link

https://github.com/liulangdog1/DATA1030-Final-Project