Journal of Computer Science
https://dergipark.org.tr/en/pub/bbd
Anatolian Science

ISSN,e-ISSN: 2548-1304
Volume:8, Issue:2, pp:66-82, 2023
https://doi.org/10.53070/bbd.1350356
Research Paper

Predicting Parkinson's Disease Progression: A Non-Invasive Method Leveraging Voice Inputs

Ahmad Hassan*[1], Arslan Ahmed[2]

[1]Department of Computer Science, COMSATS University Islamabad, Wah Campus, Wah, Pakistan

[2]Department of Medicine & Allied, West Medical Ward, Mayo Hospital, King Edward Medical University, Lahore, Pakistan

(ahmad.hassan.hattar@gmail, arslanahmed37201@kemu.edu.pk)

Received:Aug.22,2023 Accepted:Sep.22, 2023 Published:Dec.20, 2023

Abstract— Parkinson's Disease (PD) is a complex neurodegenerative condition with a global impact, demanding precise disease progression prediction to facilitate effective treatment strategies. To assess PD symptoms, the Unified Parkinson's Disease Rating Scale (UPDRS) is widely adopted, encompassing both motor and non-motor assessments. This research delves into voice inputs as a non-intrusive method to predict total UPDRS and motor UPDRS scores, offering new possibilities for Parkinson's assessment. Feature engineering and data augmentation techniques address challenges related to class imbalance and diverse demographics, including an original imbalanced dataset with more females than males. Additionally, three new datasets are created: oversampled balanced, only-female, and only-male datasets. An ensemble based stacking model, including random forest and extreme gradient boosting as base models and the gradient boosting regressor as the meta-regressor is used for predicting UPDRS scores. The study employs two distinct validation methods to evaluate model performance: a 70/30 dataset split and 5-fold cross-validation. Notably, the results obtained from the 70/30 split method exhibit impressive performance, with R2 values ranging from 0.991 to 0.995 across different datasets. This method also showcases the model's effectiveness in minimizing prediction errors, as indicated by low MSE, RMSE, and MAE. However, when subjected to 5-fold cross-validation, the model's performance experiences a decline, with R2 value levels ranging from 0.781 to 0.873 across the same datasets. Similarly, the other performance metrics values show variations, suggesting that the model's performance may be sensitive to the choice of validation method. Nevertheless, these findings underscore the potential of voice-based methods for non-invasive PD assessment, offering the prospect of remote and continuous monitoring. This study could significantly enhance the quality of life for PD patients and can facilitate in effective treatment plans.

Keywords : *parkinson's assessment, UPDRS score estimation, ensemble-based stacking, extreme gradient boosting, random forests, gradient boosting regressor.*

1. Introduction

Parkinson's Disease (PD) is a long-term neurodegenerative disorder that mainly affects the motor system (De Miranda & Greenamyre, 2017). It is a prevalent condition worldwide, particularly in individuals over 60, approximately 1% of this age group is diagnosed with PD. The exact cause of PD is still unknown, and currently, no cure is available (Rizek et al., 2016). The parkinson's disease arises due to the deterioration of essential neurons residing in the basal ganglia, a crucial brain structure situated in the center of the brain. Common symptoms for indication of PD include shakiness, slowness of movement, muscle rigidity, and balance issues, with symptom severity often increasing as the time passes, and the disease progresses (McGregor & Nelson, 2019).

To estimate the severity of Parkinsonian symptoms, the Unified Parkinson's Disease Rating Scale (UPDRS) was developed (Disease, 2003). The UPDRS comprises five segments that test PD's motor and cognitive symptoms. During the motor portion of the test, patients perform various muscle movements while a qualified movement disorder clinician rates the impairment of movement for each task (Stamate et al., 2018). The total UPDRS score is obtained by summing each task's values (Trudelle, 2006).

Researchers have recently explored voice recordings as a non-invasive and practical approach for diagnosing and monitoring PD (Van Den Bergh et al., 2021). Voice tests can provide valuable insights into vocal impairments, which are common symptoms of PD and early indicators of the disease (Pastor-Sanz et al., 2011). With technological advancements, at-home recording devices like those developed by Intel for PD telemonitoring enable convenient and remote health monitoring for PD patients (Tsanas et al., 2021). The integration of telemonitoring capabilities offers the potential for continuous real-time monitoring of PD symptoms for early detection of changes and personalized treatment strategies (Polverino et al., 2022).

This study endeavours to explore the potential of voice-based methods in predicting motor and total UPDRS scores, showcasing a promising avenue for non-invasive PD assessment. By leveraging voice inputs and telemonitoring, the research contributes to improved disease management and patient care, providing accessible and reliable PD assessment for timely intervention and personalized treatment strategies.

Section 2 provides a review of relevant literature, while Section 3 covers the dataset and preprocessing techniques. Section 4 describes the research pipeline and data modelling. The obtained results and important contributing factors presented in Section 5. The results are discussed and compared with previous related studies in Section 6, and finally, Section 7 concludes the study and suggests avenues for future work.

2. Literature Review

Researchers in the field of Parkinson's disease have been actively working on the development of algorithms for PD prediction. A thorough literature review has been conducted to explore the existing research and studies about PD and its assessment techniques.

In a study on the effects of levodopa treatment on PD patients, evaluating essential lung capacity, maintaining prolonged vowel articulation, and calculating phonation ratio.. Following drug administration, the results demonstrated significant improvements in these parameters (De Letter et al., 2007). Additionally, another study found a reduction in fundamental frequency variability during text reading after the administration of levodopa (Skodda et al., 2011). However, a limited focus remains on predicting PD severity for effective treatment monitoring. Further research is required to address this aspect.

A novel prediction model was developed to estimate UPDRS-III scores using speech data from 42 PD patients. The model achieved high accuracy, with RMSE of 1.62 and 1.72 for males and females (Tsanas et al., 2010). Furthermore, researchers explored the potential of the UPDRS scale in creating a classification system to distinguish between healthy individuals and those with PD using speech signal analysis (Sakar et al., 2017).

One study investigated applying deep neural network and convolutional neural network with transfer learning to predict UPDRS scores. It explored feature engineering, utilizing established vocal features, and feature learning through modulation spectra transformations (Arias-Londoño & Gómez-García, 2020). Additionally, there is research into speech signal processing for assessing Parkinson's disease patients within a few hours after medication intake. Acoustic parameters are extracted to predict the Parkinson's disease rating scale score, enabling automatic monitoring of disease progression (Hemmerling & Wojcik-Pedziwiatr, 2022).

Various algorithms have been investigated for predicting PD progression. One study utilized the GMM-UBM algorithm to monitor disease progression, achieving a the Pearson's correlation of a value up to 0.60 concerning MDS-UPDRS-III labels (Arias-Vergara et al., 2016). Moreover, PD progression prediction has been explored using diverse approaches such as the Expectation Maximization (EM), principal component analysis (PCA), support vector regression (SVR), and the adaptive neuro-fuzzy inference system (ANFIS)., which led to a low mean absolute error of 0.4721 with the EM-PCA-SVR algorithm (Nilashi et al., 2016).

Cloud computing was utilized to enhance accessibility and decision-making support for physicians and IoT nodes. The ML algorithms demonstrated acceptable performance levels, with Motor UPDRS identified as a significant predictor of Total UPDRS (Hamzehei et al., 2023). Additionally, another investigation focused on using ML algorithms to analyze voice changes in PD patients at different disease stages. L-dopa therapy improved but did not fully restore voice in PD patients, and a new machine learning-derived score (LR value) allowed significant clinical-instrumental correlations (Suppa et al., 2022).

In related research, ML techniques, both unsupervised and supervised, have been utilized for Parkinson's disease diagnosis through UPDRS prediction (Nilashi et al., 2022). Clustering and prediction learning methods are compared, with SVR ensembles showing superiority in predicting motor and total UPDRS scores over other approaches. Another study focused on classifying PD using human voice signals, achieving high accuracies with various ML classifiers, with Random Forest performing the best (Ahmed et al., 2021). Further exploring PD patient characteristics may enhance its applicability in the medical field.

In addition to singular value decomposition (SVD), researchers have also investigated ensembles of adaptive neuro-fuzzy inference systems for UPDRS prediction, focusing on the motor part of UPDRS. One notable approach, the EM-SVD-ANFIS ensemble method, achieved remarkable results with minimal mean absolute errors of 0.480 (Nilashi et al., 2019). These collective efforts contribute significantly to advancing non-invasive PD assessment techniques and effectively monitoring disease severity to improve healthcare outcomes.

3. Dataset and Preprocessing

This research leverages a telemonitoring dataset focused on the progression of Parkinson's disease, sourced from Kaggle (*Parkinson's Disease Progression*, 2023). The dataset includes biomedical voice measurements obtained from 31 individuals during a six-month telemonitoring trial, with participants during their initial phases of Parkinson's ailment. Recordings were automatically captured in patients' homes, resulting in 5875 data instances with 22 numerical features representing speech attributes. Table 1 summarizes details about the dataset features.

Tablo 1. Summary of Features Present in the Dataset

Feature	Description
subject	Identifier for each subject ranging from 1 to 31.
age	The subject's age.
sex	The subject's gender.
test_time	The duration from the initiation of subject recruitment to their participation in the clinical trial.
motor_UPDRS	Motor UPDRS, a rating scale measuring motor indicators of Parkinson's disease.
total_UPDRS	Total UPDRS, a numeric rating scale measuring overall signs of Parkinson's disease.
Jitter %	Percentage of local variation in fundamental frequency, a measure of vocal stability.
Jitter(Abs)	Absolute jitter, a measure of absolute variation in fundamental frequency.
Jitter:RAP	Relative average perturbation, a average value of perturbation in consecutive periods.
Jitter:PPQ5	Five-point period perturbation quotient, a measure of perturbation during five consecutive periods.
Jitter:DDP	Difference between consecutive differences of fundamental frequency.
Shimmer	Local variation in amplitude, a measure of vocal amplitude variability.
Shimmer(dB)	Shimmer in decibels, a measure of amplitude perturbation in decibels.
Shimmer:APQ3	Three-point amplitude perturbation quotient, a measure of amplitude perturbation during three consecutive points.
Shimmer:APQ5	Five-point amplitude perturbation quotient, a measure of amplitude perturbation during five consecutive points.
Shimmer:APQ11	Eleven-point amplitude perturbation quotient, a measure of amplitude perturbation during eleven consecutive points.
Shimmer:DDA	Difference between consecutive differences of amplitude.
NHR	Noise-to-harmonics ratio measures the ratio between noise and harmonics in the speech signal.
HNR	Harmonics-to-noise ratio, a measure of the ratio between harmonics and noise in the speech signal.
RPDE	Recurrence period density entropy, a measure of the complexity of the speech signal.
DFA	Detrended fluctuation analysis, a measure of the long-term correlation properties of the speech signal.
PPE	Pitch period entropy, a measure of the variability in the pitch of the speech signal.

Upon careful examination, the dataset reveals a diverse range of ages among the subjects, spanning from 30 to 90 years. Most subjects cluster within the age group of 50 to 80 years, signifying the presence of an age demographic in the study. Furthermore, the test time, which denotes the duration since recruitment into the trial, varies between 10 to 180 units, indicating the longitudinal nature of the data collection process. To visually represent these distributions, histograms are presented in Figure 1, offering valuable insights into the distribution patterns of age and test time in the dataset.

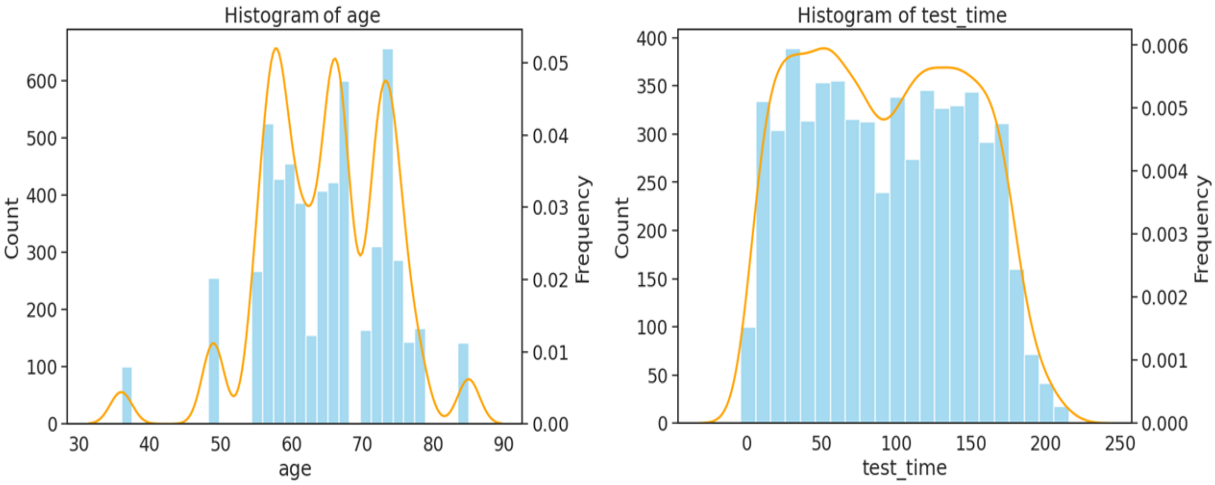

Figure 1. Age and Test Time Distribution of Subjects

The UPDRS is widely used to assess the severity of Parkinson's disease (Zimmerman et al., 2018). It comprises four parts, covering different aspects of the disease. Part I addresses cognitive conditions and emotional irregularities, Part II focuses on daily routines, Part III evaluates motor capabilities, with the final section evaluating therapy challenges (Holden et al., 2018). Each issue is assigned points ranging from a scale of 0 (absence of symptoms) to 4 (pronounced symptoms), with a maximum total score of 220. Higher UPDRS scores indicate more advanced disease stages (Hendricks & Khasawneh, 2021). Part III, known as UPDRS-III, includes the evaluation of speech, which is crucial for analyzing patients' speech-related symptoms (Costantini et al., 2023). In this study, UPDRS-III score can vary from 0 to 108.

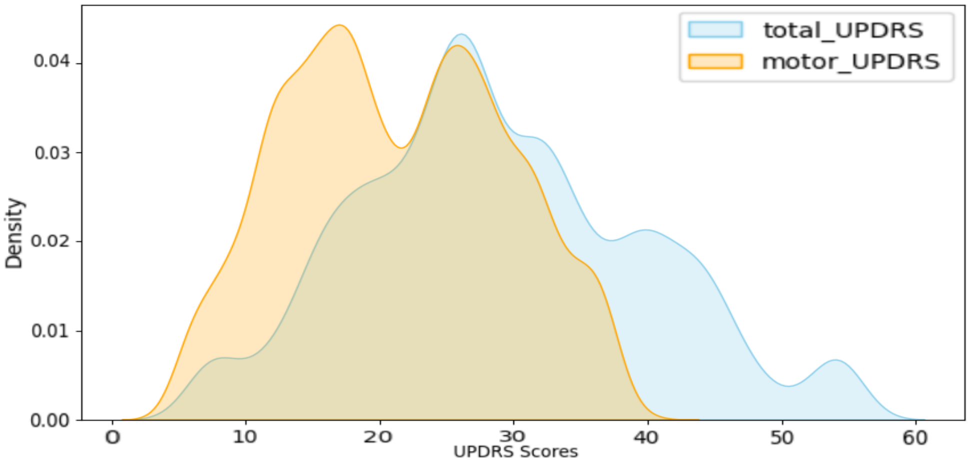

Figure 2. Kernel Density Plot for Motor UPDRS and Total UPDRS Scores

The kernel density plot for both motor UPDRS and total UPDRS scores is displayed in Figure 2. The plot demonstrates a central distribution for both variables, indicating that most subjects in the dataset exhibit scores concentrated around a central value. The scatter plot in Figure 3 offers valuable insights into the distribution patterns of Motor and Total UPDRS scores, aiding in the comprehension of disease progression. It reveals a significant linear correlation between motor and total UPDRS scores. The points align along, indicating a strong association between the two variables. As motor UPDRS scores increase, there is a corresponding rise in total UPDRS scores, highlighting the direct relationship between motor symptoms and disease progression. This finding reinforces the efficacy of voice inputs as a non-invasive and valuable tool for predicting both motor and total UPDRS scores.

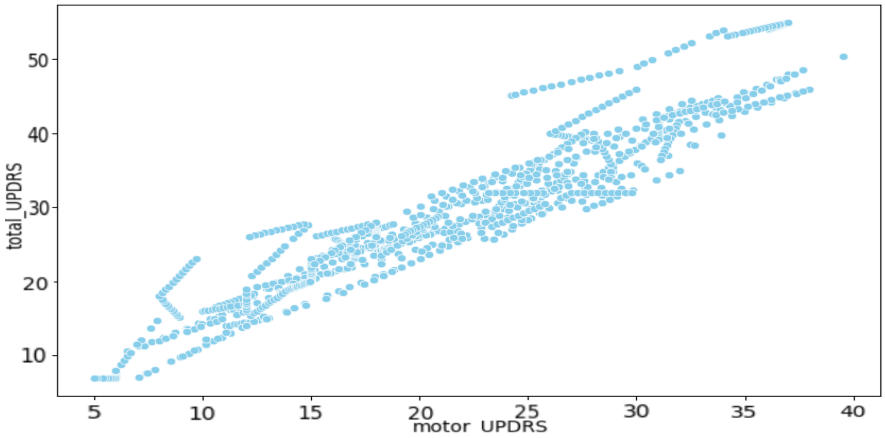

Figure 3. Scatter Plot for Motor UPDRS and Total UPDRS Scores

3.1. Feature Engineering

This study employs a feature engineering technique to create new informative attributes from the existing dataset. Four new features are generated: Jitter_Abs_Squared, Shimmer_ShimmerAPQ5_Ratio, Age_Test_Time_Ratio, and RPDE_Log. Jitter_Abs_Squared is a feature that represents the squared value of Jitter(Abs), which is a measure used in speech signal analysis. Jitter(Abs) quantifies the absolute variation in the fundamental frequency of the speech signal, specifically focusing on the cycle-to-cycle variations in the time between consecutive glottal closure instants. By squaring Jitter(Abs), we enhance its sensitivity to these absolute variations, providing a more robust indicator of the irregularities in the fundamental frequency of speech. This squared value can be particularly useful for detecting subtle changes and variations in voice patterns, making it a valuable feature in the context of voice-based assessments. It is displayed in Figure 4.

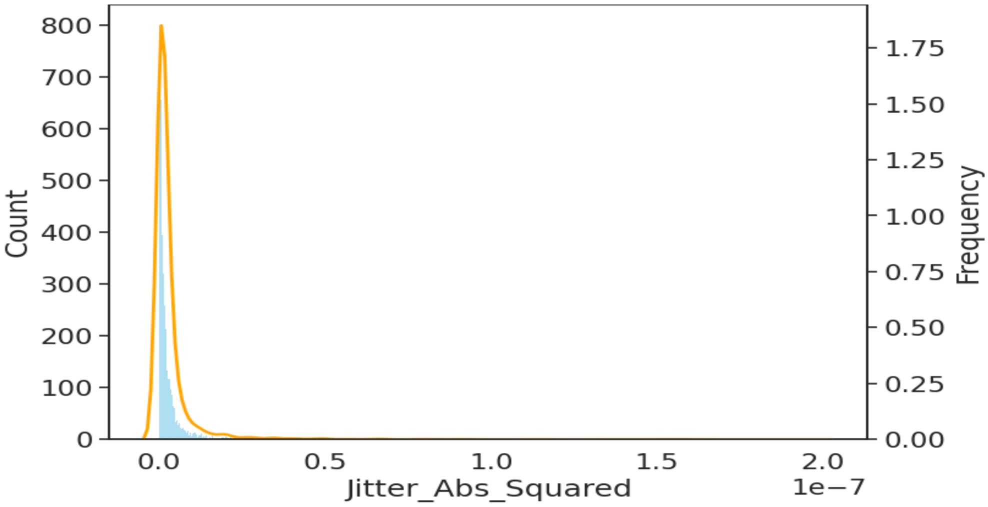

Figure 4. Histogram of Jitter_Abs_Squared Feature

The Shimmer_ShimmerAPQ5_Ratio feature is engineered by dividing the Shimmer measure by the ShimmerAPQ5 measure. Shimmer represents the variation in speech signal amplitude, reflecting voice quality, while ShimmerAPQ5 provides a more detailed assessment of amplitude perturbations in the speech signal, focusing on the five highest peaks in the amplitude waveform. The ratio of Shimmer to ShimmerAPQ5 quantifies the relationship between these two measures, indicating whether amplitude perturbations are concentrated or dispersed within the signal. A higher ratio suggests a more concentrated distribution of perturbations, potentially indicating specific voice irregularities, while a lower ratio implies a more uniform distribution, corresponding to a smoother voice quality. This ratio serves as a valuable acoustic feature for characterizing voice quality in the context of Parkinson's disease assessment. It is shown in Figure 5.

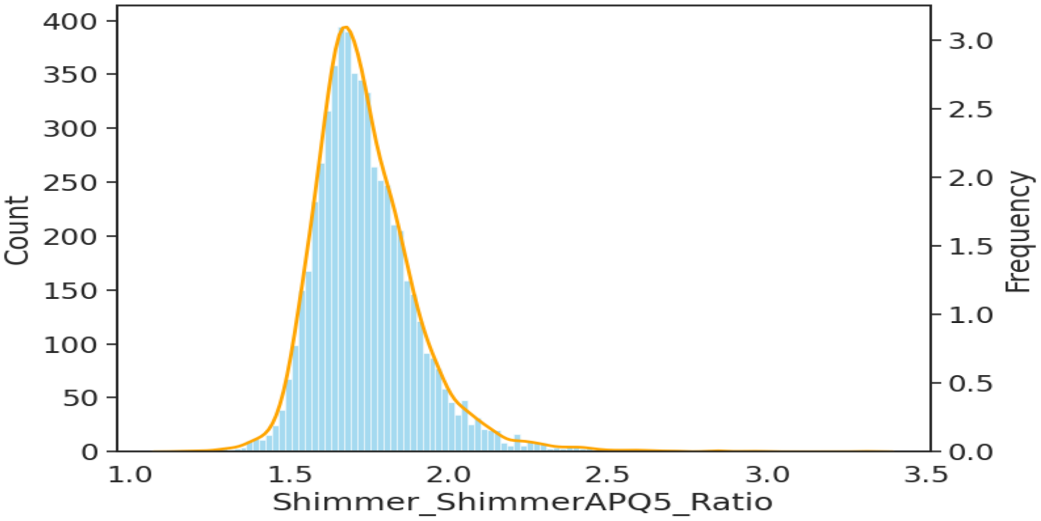

Figure 5. Histogram of Shimmer_ShimmerAPQ5_Ratio Feature

The Age_Test_Time_Ratio engineered feature is an metric that captures the temporal aspect of Parkinson's disease progression in relation to age. This ratio is calculated by dividing the age of an individual by the time elapsed since the initial diagnostic test was conducted. It aims to assess whether the rate of PD progression is influenced by age. A higher ratio suggests that the disease progresses more slowly with increasing age, potentially indicating a milder form of PD or a slower rate of degeneration. Conversely, a lower ratio may imply that younger individuals experience a more rapid disease progression. Understanding this relationship can provide valuable insights into the interplay between age and PD severity. This feature is visualized in Figure 6.

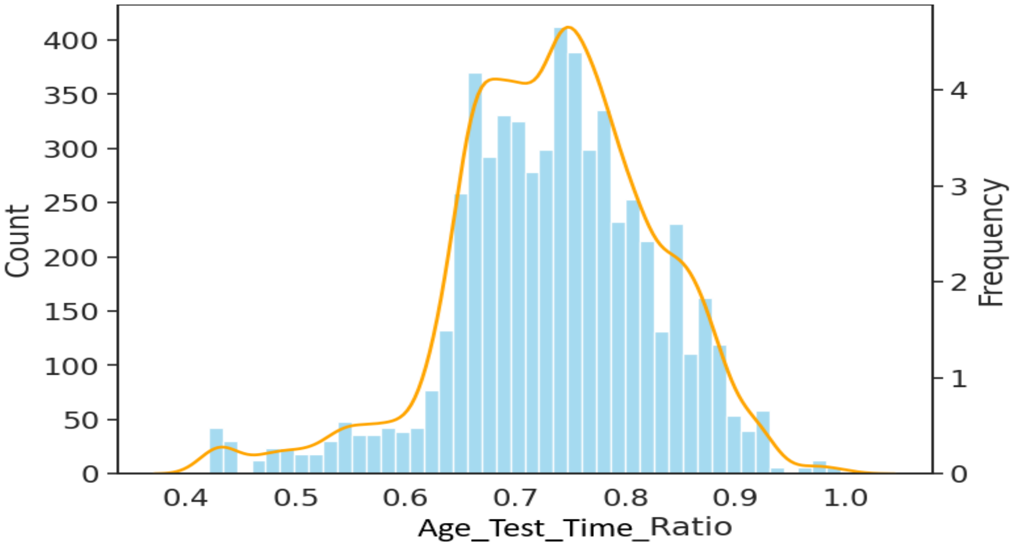

Figure 6. Histogram of Age_Test_Time_Ratio Feature

Finally, a logarithm transformation is applied to RPDE to address non-linearity, resulting in the RPDE_Log feature, as shown in Figure 7. It is derived from the Recurrence Period Density Entropy (RPDE), a nonlinear dynamical analysis applied to speech signals. RPDE quantifies the irregularity and complexity of speech patterns, which can be indicative of changes in vocal control associated with PD. To obtain "RPDE_Log," the logarithm of the RPDE values is calculated. Taking the logarithm of RPDE can help enhance the interpretability of the feature and make it more amenable to modeling. A higher "RPDE_Log" value suggests a greater degree of irregularity and complexity in the speech patterns, which may reflect voice abnormalities associated with Parkinson's Disease. Conversely, a lower "RPDE_Log" value indicates a more regular and predictable speech pattern. By thoughtfully crafting these new features and analyzing their distributions, we gain a deeper understanding of the dataset, leading to more accurate and reliable predictions of Parkinson's disease progression.

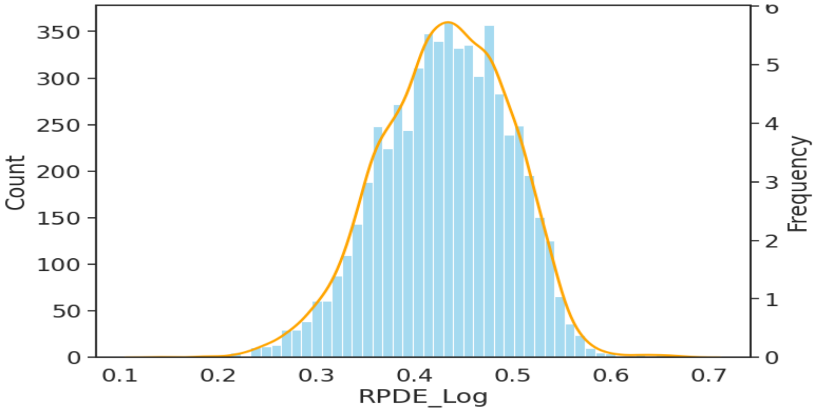

Figure 7. Histogram of RPDE_Log Feature

3.2. Dataset Creation and Modification

Feature engineering and oversampling techniques are used to create three new datasets to address challenges related to class imbalance and diverse demographics. These datasets aim to mitigate biases and enhance the predictive model's performance.

3.2.1. Original Imbalanced Dataset

The original imbalanced dataset comes with more females than males. Figure 8 visually presents the correlation between UPDRS scores and features in the original dataset. As depicted in the plot, most features exhibit a loose correlation with the UPDRS scores.

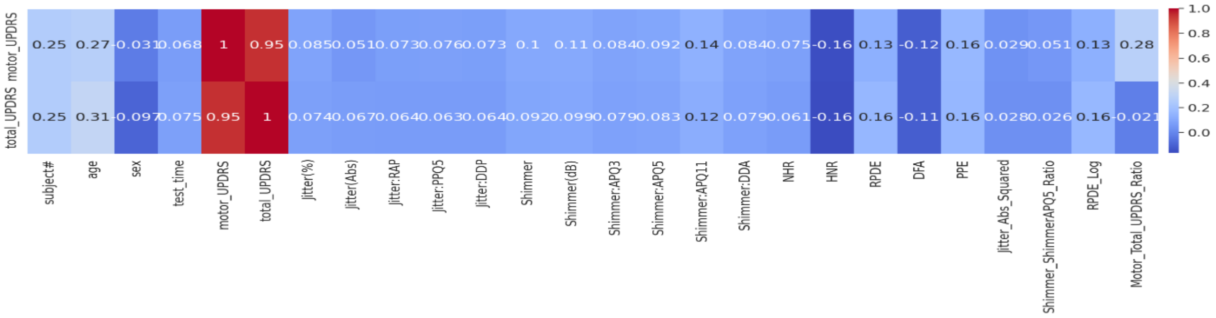

Figure 8. UPDRS Scores Correlation with Features in Original Dataset

3.2.2. Oversampled Balanced Dataset

The first dataset is an oversampled balanced dataset created using the Synthetic Minority Over-sampling Technique (SMOTE). By creating artificial instances for the underrepresented category, a balanced distribution is achieved, which improves the model's ability to generalize and make accurate predictions for both classes. Before oversampling, the distribution was 4008 females and 1867 males. After applying SMOTE, the distribution became 4008 females and 4008 males. Figure 9 illustrates the correlation between UPDRS scores and features in the balanced dataset. The plot reveals that most features exhibit a loose correlation with UPDRS scores. Notably, HNR, Sex, and DFA correlate negatively with UPDRS scores.

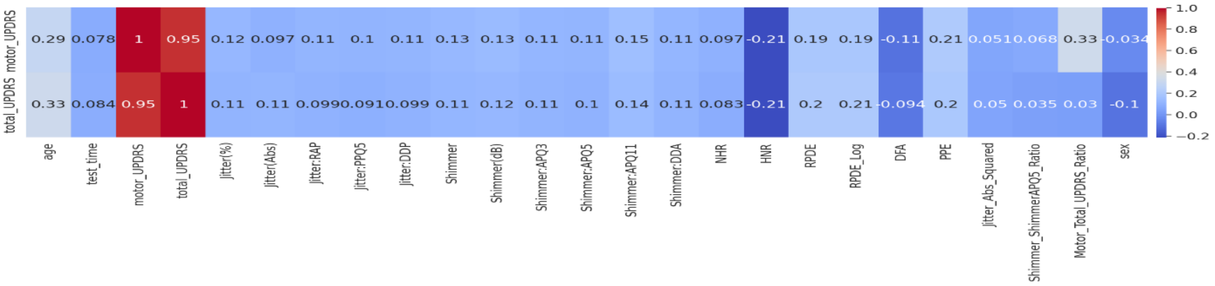

Figure 9. UPDRS Scores Correlation with Features in Balanced Dataset

3.2.3. Only-Female Dataset

A subset comprising only female subjects to examine the effects of gender-specific data is extracted. This dataset helps to understand potential variations in PD symptomatology between genders and assess the model's performance in predicting UPDRS scores for female subjects exclusively. The correlation between UPDRS scores and features in the Only-Female Dataset is illustrated in Figure 10. The features exhibit a loose correlation, with some showing negative correlation patterns. The observed loose correlations indicate that the relationship between UPDRS scores and the features in the Only-Female Dataset may not be strong.

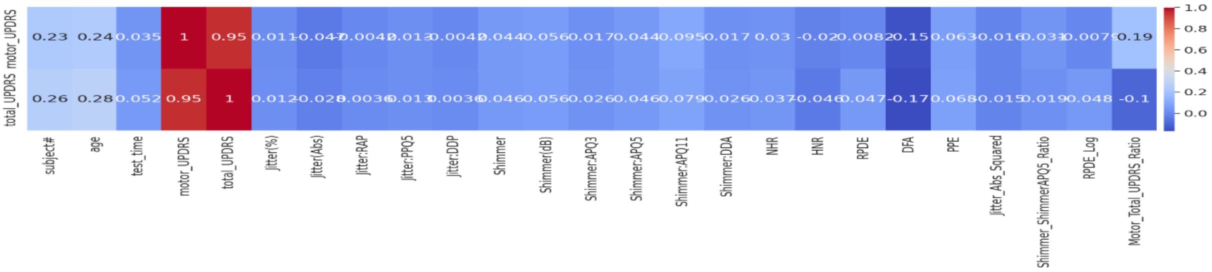

Figure 10. UPDRS Scores Correlation with Features in Only-Female Dataset

3.2.4. Only-Male Dataset

Similar to the previous subset, an only-male dataset to focus on male subjects is created. Analyzing this subset allows to explore differences in symptom severity and progression between male and female patients. Figure 11 illustrates the correlation between UPDRS scores and features in the Only-Male dataset. Among all the datasets, the features in this dataset exhibit stronger correlations. Notably, the HNR feature is the only one with a negative correlation with UPDRS scores in the Only-Male dataset.

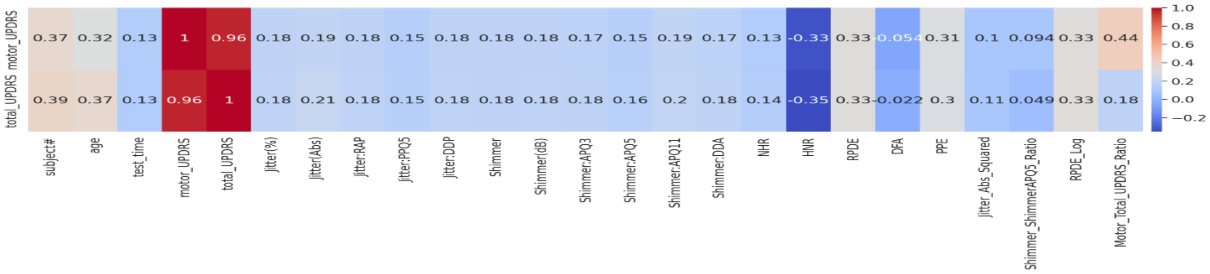

Figure 11. UPDRS Scores Correlation with Features in Only-Male Dataset

These new and original datasets provide a comprehensive framework for evaluating and comparing model performance in predicting UPDRS scores using voice inputs. Valuable insights can be gained into the potential applications of voice-based methods for non-invasive PD assessment by investigating the various datasets.

4. Data Modelling

In this study, the authors explore the potential of voice inputs as a non-invasive approach for estimating total UPDRS and motor UPDRS scores. Various techniques are used to address challenges related to class imbalance and diverse demographics. Figure 12 shows the research pipeline that has been used for data modelling. In addition to the original imbalanced dataset, three new datasets were created: oversampled balanced, only-female, and only-male datasets. Further, the authors employed an ensemble-based stacking model, utilizing random forest and extreme gradient boosting as base models and the gradient boosting regressor as the meta-regressor. To rigorously evaluate the model's performance, two different methods were employed for training and testing: a traditional 70-30 split for training and testing, and 5-fold cross-validation.

In the first method, each dataset was divided into a 70-30 split for training and testing, providing insights into how the model performs on a standard training-testing setup. In the second method, 5-fold cross-validation was used to assess the model's robustness and generalization across different subsets of the data. This dual approach offered a comprehensive understanding of the model's performance. The results demonstrated promising performance and robustness in predicting UPDRS scores, showcasing the efficacy of voice inputs for PD assessment.

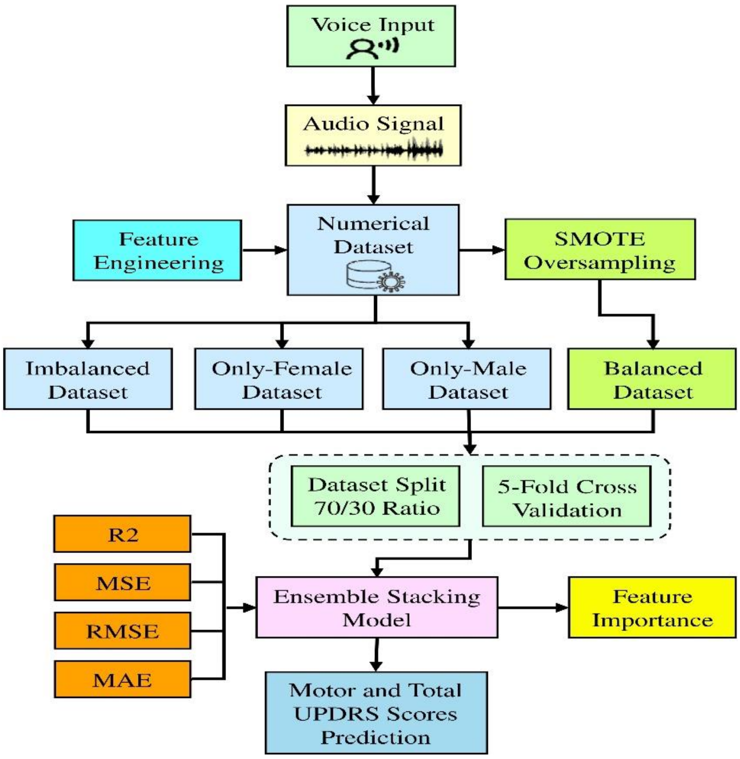

Figure 12. Research Pipeline for Data Modelling

Furthermore, the feature importance analysis provided insights into crucial contributors influencing predictions. To evaluate and compare model performance, authors used a range of performance measures, including, R-squared (R2), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). These performance metrics offered valuable insights into the model's accuracy and predictive capabilities for both training and cross-validation methods.

4.1. Synthetic Minority Oversampling Technique (SMOTE)

SMOTE is a representation of data augmentation approach to address the class imbalance in the dataset. It produces synthetic samples for the minority class by interpolating between neighbouring instances, effectively balancing the class distribution (Chawla et al., 2002). For each minority-class data point, SMOTE identifies its k nearest neighbors. It then constructs new synthetic instances by randomly selecting one or more of these neighbors and generating data points along the line segments connecting them. This process expands the representation of the minority class, effectively balancing the class distribution (Hassan & Yousaf, 2022). By oversampling the minority class, SMOTE enhances the representation of underrepresented instances, improving model performance and mitigating the bias towards the majority class. This approach ultimately leads to more accurate and reliable predictions, particularly in contexts where the minority class holds significant importance (Elreedy et al., 2023). Implementing SMOTE allows a comprehensive evaluation of voice-based methods for predicting Parkinson's disease progression, contributing to more accurate and reliable results.

4.2. K-fold Cross Validation

K-fold cross-validation is a robust technique used to evaluate the performance of predictive models, ensuring they generalize well to unseen data (Bradshaw et al., 2023). In this approach, the dataset is divided into 'k' subsets of equal size. The model is then trained 'k' times, each time using 'k-1' of the subsets as the training data and the remaining subset for validation. This process iterates 'k' times, with each subset serving as the validation data exactly once. The final performance metric, whether it's accuracy, mean squared error, or any other relevant measure, is calculated as the average across these 'k' iterations (White & Power, 2023). This method provides a more reliable estimate of a model's performance as it assesses its consistency across different data subsets, helping to identify potential overfitting or underfitting issues (Kaliappan et al., 2023).

4.3. Ensemble Stacking Model

This study employs an innovative ensemble-based stacking model to predict Parkinson's disease progression. The model combines the strengths of different base models, including random forest and extreme gradient boosting, to improve predictive accuracy. The gradient boosting regressor is used as the meta-regressor to combine the outputs of the base models and make the final prediction.

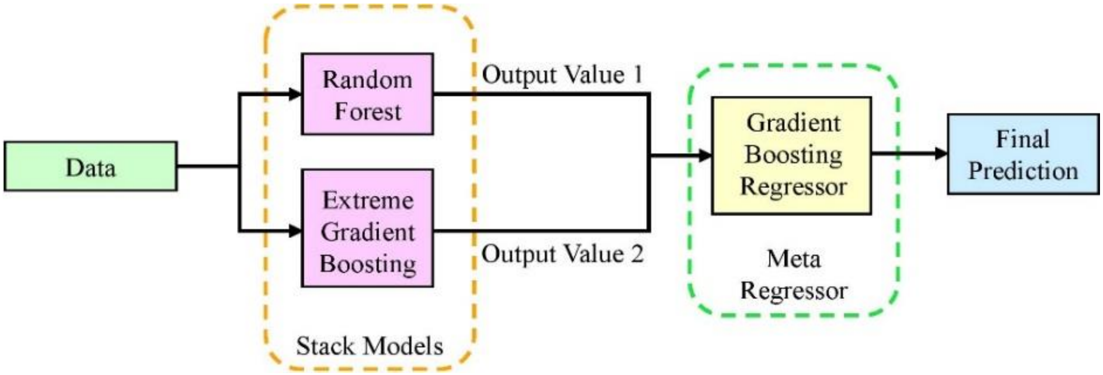

Figure 13. Visualization of the Ensemble Stacking Model Architecture

This stacking approach allows the model to learn from the diverse predictions of the base models and leverage their complementary strengths, resulting in enhanced performance and robustness. Figure 13 visualizes the architecture of the stacking model, showcasing the flow of information and how the predictions from the base models are combined to make the final prediction.

4.4. Performance Metrics

In this section, the authors evaluate the performance of the ensemble-based stacking model using various performance metrics. The following subsections outline each metric and briefly explain their significance.

4.4.1. R-squared (R2)

R-squared, known as coefficient of determination, measures the capability of the model to explain the variance in UPDRS scores. A higher R2 value indicates better learning of the model on the data. R2 is calculated as:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{1}$$

where SS_{res} is the added sum of the squared residuals and SS_{tot} is the total squares sum.

4.4.2. Mean Squared Error (MSE)

MSE is a statistical measure metric used to average the squared difference between the predicted and actual UPDRS scores. A lower MSE value suggests improved accuracy and indicates how well the model's predictions align with the true values. It is calculated as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

where n represents the total number of data points in the dataset. The y_i is the actual value of the target variable for the i-th data point, and $\hat{y}_i$ is the predicted value generated by the model for the i-th data point.

4.4.3. Root Mean Squared Error (RMSE)

RMSE is a critical metric used to evaluate the model's prediction accuracy. It represents the square root of MSE and serves as a valuable tool for comprehending the magnitude of prediction errors. The formula for it is:

$$RMSE = \sqrt{MSE} \tag{3}$$

4.4.4. Mean Absolute Error (MAE)

The MAE signifies the average absolute difference between the UPDRS predicted and actual scores. Similar to RMSE, a lower MAE value signifies better model performance. The MAE is calculated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{4}$$

5. Results and Features Importance

This section unveils the outcomes of two distinct methodologies: a conventional 70-30 split for training and testing, and an extensive 5-fold cross-validation. These results provide a comprehensive evaluation of the employed model's performance. Furthermore, authors explore feature importance, identifying the pivotal factors that shape the model's predictions. This concise analysis offers valuable insights into the model's accuracy and influential features.

5.1. 70/30 Training and Testing Method

The ensemble stacking model showcases the promising performance, highlighting the potential of voice inputs for non-invasive PD assessment. Performance metrics are employed to evaluate and compare model performance on various datasets. The ensemble-based stacking model achieved impressive performance, with R2 values ranging from 0.991 to 0.995 across different datasets. Additionally, the model showcased its effectiveness in minimizing prediction errors, as indicated by very low MSE and RMSE values. The low MAE values further demonstrate the model's precision. Moreover, the high R2 values signify the model's ability to explain variance in UPDRS scores.

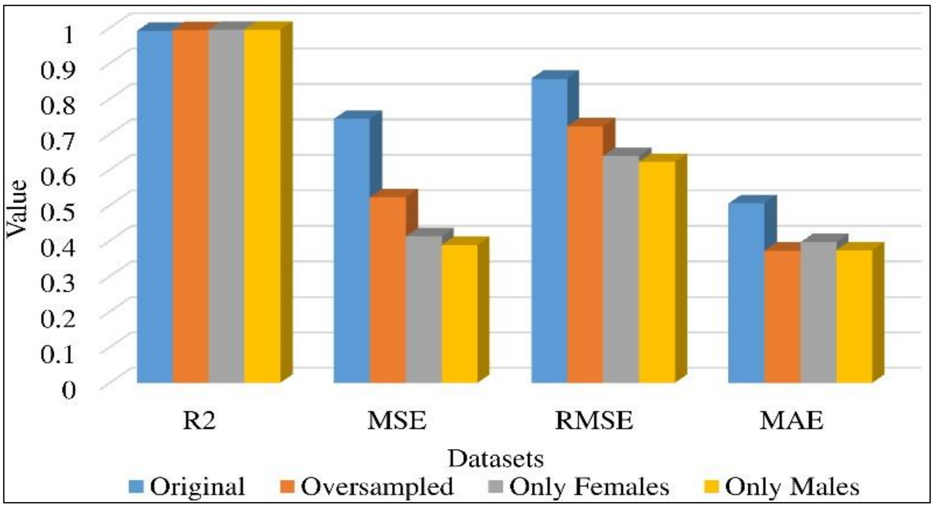

Figure 14. Plot of Mean Performance Metrics for All Datasets using 70/30 Training and Testing Method

The outcomes of this study highlight the potential of the proposed approach in predicting UPDRS scores and its robustness. As depicted in Figure 14 and summarized in Table 2, the model demonstrates accuracy and reliability across diverse datasets. Moreover, employing 5-fold cross-validation across all datasets offers insights into the model's performance consistency. However, it's important to note that the results observed in the cross-validation may not reach the same heights as those achieved through this 70/30 ratio dataset split training and testing method. A comprehensive analysis of these findings is presented in the following subsection, shedding light on the stability and generalization of the model.

Tablo 2. Summary of Mean Performance Metrics using 70/30 Training and Testing Method

Performance Metrics	Datasets			
	Original Imbalanced	Oversampled Balanced	Only Females	Only Males
R2	0.991	0.994	0.995	0.995
MSE	0.743	0.522	0.412	0.388
RMSE	0.857	0.722	0.639	0.622
MAE	0.505	0.372	0.396	0.373

5.2. 5-fold Cross Validation Method

The ensemble stacking model demonstrates its potential for non-invasive PD assessment across different datasets. However, when compared to the previous results obtained using the 70/30 dataset split training and testing method, some variations are observed. In the original imbalanced dataset, the model achieves an R2 value of 0.781, which is lower than the previous method's R2 value of 0.991. Similarly, the oversampled balanced dataset records an R2 value of 0.873, whereas the previous method achieved 0.993. Notably, the performance drops significantly in the "Only Females" dataset, with a R2 value of 0.436 compared to the previous 0.995. However, in the "Only Males" dataset, the R2 value remains relatively high at 0.865. These variations are also reflected in other performance metrics, including MSE, RMSE, and MAE, suggesting that the choice of dataset and validation method can influence model performance.

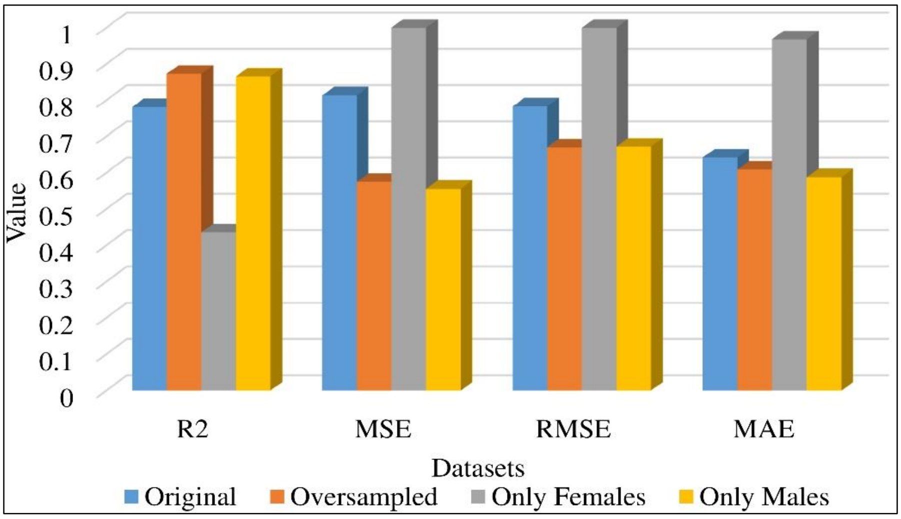

Figure 15. Plot of Mean Performance Metrics for All Datasets using 5-fold Cross Validation Method

Figure 15 and Table 3 serve as valuable references to summarize and visualize the key findings discussed in this subsection. They provide a concise and informative overview of the model's performance across different datasets and highlight the variations observed when compared to the previous results obtained using an alternate validation method. These visual aids aid in comprehending the nuances of the model's performance and the impact of dataset selection on predictive performance.

Tablo 3. Summary of Mean Performance Metrics using 5-fold Cross Validation Method

Performance Metrics	Datasets			
	Original Imbalanced	Oversampled Balanced	Only Females	Only Males
R2	0.781	0.873	0.436	0.866
MSE	0.814	0.575	1.675	0.556
RMSE	0.783	0.670	1.134	0.672
MAE	0.642	0.609	0.968	0.588

5.3. Features Importance

The authors have undertaken an extensive and thorough analysis aimed at discerning the pivotal factors that exert significant influence over the prediction of both motor and total UPDRS scores. This comprehensive investigation, as visually depicted in Figures 16 and 17, has illuminated the primary determinants of UPDRS score predictions. Notably, age, the ingeniously engineered feature known as Age_Test_Time_Ratio, the temporal aspect of test time, alongside the acoustic attributes of Jitter and Shimmer, have emerged as the standout features, playing a central role in the intricate prediction model. These findings provide invaluable insights into the complex interplay of variables impacting UPDRS score predictions, thereby enhancing our understanding of this predictive process.

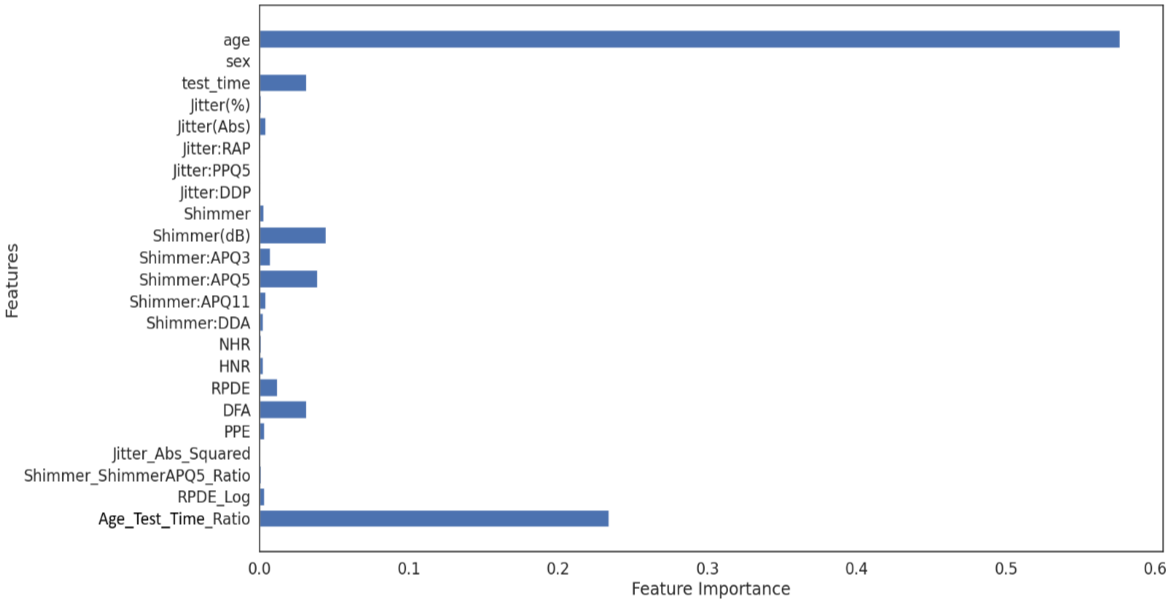

Figure 16. Feature Importance for Motor UPDRS Score Prediction

DFA, HNR, RPDE, and PPE also emerged as essential features in the prediction process. These findings align with the previous correlation plots, which indicated that the selected features had a loose correlation with each other, and their individual contributions played pivotal roles in the predictive performance of the model. The prominence of age and test time further suggests that the temporal aspect of the disease may influence the progression of Parkinson's disease.

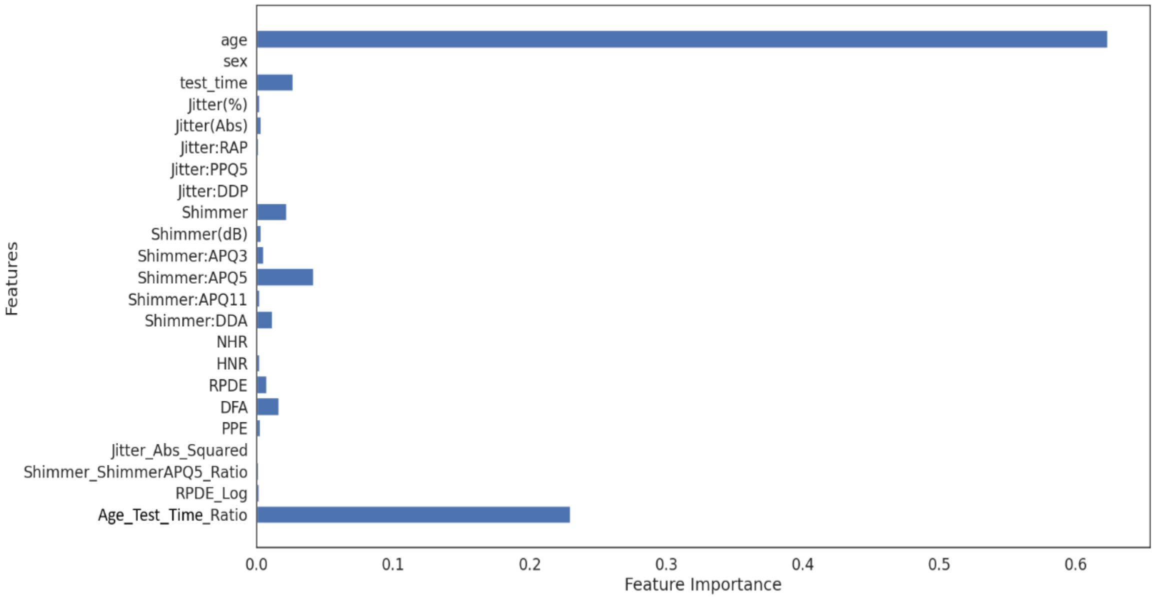

Figure 17. Feature Importance for Total UPDRS Score Prediction

6. Discussion

The stark contrast between the results obtained from the two validation methods, namely the 70/30 dataset split and 5-fold cross-validation, is an intriguing aspect of this study. Under the 70/30 dataset split, the ensemble stacking model exhibited remarkable performance, achieving R2 value levels ranging from 0.991 to 0.995 across various datasets. These results were complemented by low values of MSE, RMSE, and MAE values, which underscore the model's accuracy and precision in predicting UPDRS scores. This suggests that the 70/30 dataset split method provided a conducive environment for the model to excel.

On the other hand, the 5-fold cross-validation results portrayed a decline in model performance, with R2 value levels ranging from 0.781 to 0.873 across the same datasets. The corresponding MSE, RMSE, and MAE values exhibited variations, indicating that the model's performance may be sensitive to the choice of validation method. While the 5-fold cross-validation results are not as superior as those obtained from the 70/30 split, they still demonstrate the model's substantial potential for PD assessment.

Tablo 4. Comparison of Results with Similar Studies on Parkinson's Prediction using Voice Signals

Study	Training & Testing Method	Dataset Description	Prediction Model	Performance
(Nilashi et al., 2016)	10-fold cross validation.	16 distinct voice signal features.	Support vector regression with expectation maximization and principle component analysis.	0.457 MAE, and 0.984 R2.
(Sakar et al., 2017)	70/30 testing and training ratio.	16 distinct voice signal features.	Support vector machine.	96.43% accuracy with Matthews Correlation Coefficient (MCC) score of 0.77.
(Nilashi et al., 2019)	Clusters based training and testing.	16 distinct voice signal features.	Adaptive neuro-fuzzy inference with expectation maximization and singular value decomposition.	0.681 RMSE, 0.485 MAE, and 0.961 R2.
(Ahmed et al., 2021)	80/20 testing and training ratio.	16 distinct voice signal features.	Random forest regressor.	97.101% accuracy.
(Nilashi et al., 2022)	Clusters based training and testing.	16 distinct voice signal features.	Support vector regression and hypergraph partitioning algorithm ensemble.	2.697 RMSE, 1.853 MAE, and 0.908 R2.
(Hemmerling & Wojcik-Pedziwiatr, 2022)	5-fold cross validation.	Custom phonatory analysis voice signal dataset comprising of 13 distinct features.	Random forest regressor.	0.416 RMSE, 0.559 MAE, and 0.961 R2.
(Rajeswari & Nair, 2022)	80/20 testing and training ratio.	16 distinct voice signal features.	Convolutional neural network and long short-term memory network.	85% accuracy.
(Hamzehei et al., 2023)	75/25 testing and training ratio.	Age, sex, and 19 distinct voice signal features.	Adam optimization algorithm.	10.907 MSE, and 0.904 R2.
(Alshammri et al., 2023)	70/30 testing and training ratio.	Custom dataset with 22 distinct voice signal features.	Multilayer perceptron.	98.31% accuracy.
Proposed	70/30 testing and training ratio.	Age, sex, and 19 distinct voice signal features along with 4 engineered features.	Ensemble stacking model.	Averaged 0.993 R2, 0.516 MSE, 0.71 RMSE, and 0.411 MAE values accross all datasets.
	5-fold cross validation.			Averaged 0.739 R2, 0.905 MSE, 0.815 RMSE, and 0.702 MAE values accross all datasets.

Table 4 provides a comparative analysis of the proposed approach's performance against similar studies on Parkinson's prediction using voice signals. Notably, the ensemble stacking model, when subjected to the 70/30 dataset split, outperforms many existing studies in terms of performance metrics such as, high R2, low MSE, RMSE, and MAE values. This suggests the effectiveness of the proposed approach, especially in leveraging both voice signals and engineered features for accurate PD assessment. However, it's essential to acknowledge that the 5-fold cross-validation results, while not as impressive as the 70/30 split, still position the model as a viable tool for PD prediction. These divergent results emphasize the importance of selecting an appropriate validation method when applying voice-based models for PD assessment.

7. Conclusion and Future Work

This study explored the potential of voice inputs as a non-invasive approach for predicting total UPDRS and motor UPDRS scores in Parkinson's disease patients. Leveraging an ensemble-based stacking model, comprising random forest and extreme gradient boosting as base models and gradient boosting regressor as the meta-regressor. Additionally, the study employed two distinct validation methods, the 70/30 dataset split and 5-fold cross-validation, to assess model performance under varying conditions. The ensemble stacking model yielded impressive R2 values ranging from 0.991 to 0.995 across all datasets using 70/30 dataset split method. The performance of the ensemble-based stacking model was consistently maintained across various datasets, substantiating its reliability. While the 70/30 dataset split demonstrated superior results, the 5-fold cross-validation results, although not as robust ranging from 0.781 to 0.873 across the same datasets, indicates that model's performance is sensitive to different validation methods. These results highlight the model's potential for PD assessment. Various performance metrics, including MSE, RMSE, and MAE, further validate the robustness of the model's predictions. Importantly, the proposed approach offers the possibility of remote and continuous PD assessment through telemonitoring, allowing for real-time monitoring and early detection of symptom changes.

Future work could focus on incorporating additional demographic and clinical data to enhance the model's predictive capabilities and explore the integration of advanced machine learning techniques with more robust validation methods for more precise and personalized Parkinson's disease assessment and management. The findings of this research hold significant promise for improving the quality of life for Parkinson's patients and advancing the field of voice-based medical assessment.

References

Ahmed, I., Aljahdali, S., Khan, M., & Kaddoura, S. (2021). Classification of Parkinson Disease Based on Patient's Voice Signal Using Machine Learning. Intelligent Automation & Soft Computing, 32(2), 705–722. https://doi.org/10.32604/iasc.2022.022037

Alshammri, R., Alharbi, G., Alharbi, E., & Almubark, I. (2023). Machine learning approaches to identify Parkinson's disease using voice signal features. Frontiers in Artificial Intelligence, 6. https://doi.org/10.3389/frai.2023.1084001

Arias-Londoño, J. D., & Gómez-García, J. (2020). Predicting UPDRS Scores in Parkinson's Disease Using Voice Signals: A Deep Learning/Transfer-Learning-Based Approach (pp. 100–123). https://doi.org/10.1007/978-3-030-65654-6_6

Arias-Vergara, T., Vasquez, J., Orozco, J. R., Vargas-Bonilla, J., & Noeth, E. (2016). Parkinson's Disease Progression Assessment from Speech Using GMM-UBM (p. 1937). https://doi.org/10.21437/Interspeech.2016-1122

Bradshaw, T. J., Huemann, Z., Hu, J., & Rahmim, A. (2023). A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging. Radiology: Artificial Intelligence, 5(4), e220232. https://doi.org/10.1148/ryai.220232

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

Costantini, G., Cesarini, V., Di Leo, P., Amato, F., Suppa, A., Asci, F., Pisani, A., Calculli, A., & Saggio, G. (2023). Artificial Intelligence-Based Voice Assessment of Patients with Parkinson's Disease Off and On Treatment: Machine vs. Deep-Learning Comparison. Sensors, 23(4), Article 4. https://doi.org/10.3390/s23042293

De Letter, M., Santens, P., De Bodt, M., Van Maele, G., Van Borsel, J., & Boon, P. (2007). The effect of levodopa on respiration and word intelligibility in people with advanced Parkinson's disease. Clinical Neurology and Neurosurgery, 109(6), 495–500. https://doi.org/10.1016/j.clineuro.2007.04.003

De Miranda, B. R., & Greenamyre, J. T. (2017). Etiology and Pathogenesis of Parkinson’s Disease. https://doi.org/10.1039/9781782622888-00001

Disease, M. D. S. T. F. on R. S. for P. (2003). The Unified Parkinson’s Disease Rating Scale (UPDRS): Status and recommendations. Movement Disorders, 18(7), 738–750. https://doi.org/10.1002/mds.10473

Elreedy, D., Atiya, A. F., & Kamalov, F. (2023). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. Machine Learning. https://doi.org/10.1007/s10994-022-06296-4

Hamzehei, S., Akbarzadeh, O., Attar, H., Rezaee, K., Fasihihour, N., & Khosravi, M. R. (2023). Predicting the total Unified Parkinson’s Disease Rating Scale (UPDRS) based on ML techniques and cloud-based update. Journal of Cloud Computing, 12(1), 12. https://doi.org/10.1186/s13677-022-00388-1

Hassan, A., & Yousaf, N. (2022). Bankruptcy Prediction using Diverse Machine Learning Algorithms. 2022 International Conference on Frontiers of Information Technology (FIT), 106–111. https://doi.org/10.1109/FIT57066.2022.00029

Hemmerling, D., & Wojcik-Pedziwiatr, M. (2022). Prediction and Estimation of Parkinson’s Disease Severity Based on Voice Signal. Journal of Voice, 36(3), 439.e9-439.e20. https://doi.org/10.1016/j.jvoice.2020.06.004

Hendricks, R. M., & Khasawneh, M. T. (2021). An Investigation into the Use and Meaning of Parkinson’s Disease Clinical Scale Scores. Parkinson’s Disease, 2021, e1765220. https://doi.org/10.1155/2021/1765220

Holden, S. K., Finseth, T., Sillau, S. H., & Berman, B. D. (2018). 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, 5(1), 47–53. https://doi.org/10.1002/mdc3.12553

Kaliappan, J., Bagepalli, A. R., Almal, S., Mishra, R., Hu, Y.-C., & Srinivasan, K. (2023). Impact of Cross-Validation on Machine Learning Models for Early Detection of Intrauterine Fetal Demise. Diagnostics, 13(10), Article 10. https://doi.org/10.3390/diagnostics13101692

McGregor, M. M., & Nelson, A. B. (2019). Circuit Mechanisms of Parkinson’s Disease. Neuron, 101(6), 1042–1056. https://doi.org/10.1016/j.neuron.2019.03.004

Nilashi, M., Abumalloh, R. A., Minaei-Bidgoli, B., Samad, S., Yousoof Ismail, M., Alhargan, A., & Abdu Zogaan, W. (2022). Predicting Parkinson’s Disease Progression: Evaluation of Ensemble Methods in Machine Learning. Journal of Healthcare Engineering, 2022, e2793361. https://doi.org/10.1155/2022/2793361

Nilashi, M., Ibrahim, O., & Ahani, A. (2016). Accuracy Improvement for Predicting Parkinson’s Disease Progression. Scientific Reports, 6(1), Article 1. https://doi.org/10.1038/srep34181

Nilashi, M., Ibrahim, O., Samad, S., Ahmadi, H., Shahmoradi, L., & Akbari, E. (2019). An analytical method for measuring the Parkinson’s disease progression: A case on a Parkinson’s telemonitoring dataset. Measurement, 136, 545–557. https://doi.org/10.1016/j.measurement.2019.01.014

Parkinson’s Disease Progression. (2023). https://www.kaggle.com/datasets/thedevastator/unlocking-clues-to-parkinson-s-disease-progressi

Pastor-Sanz, L., Pansera, M., Cancela, J., Pastorino, M., Waldmeyer, M. T. A., Pastor-Sanz, L., Pansera, M., Cancela, J., Pastorino, M., & Waldmeyer, M. T. A. (2011). Mobile Systems as a Challenge for Neurological Diseases Management – The Case of Parkinson’s Disease. In Diagnostics and Rehabilitation of Parkinson’s Disease. IntechOpen. https://doi.org/10.5772/16729

Polverino, P., Ajčević, M., Catalan, M., Bertolotti, C., Furlanis, G., Marsich, A., Buoite Stella, A., Accardo, A., & Manganotti, P. (2022). Comprehensive telemedicine solution for remote monitoring of Parkinson’s disease patients with orthostatic hypotension during COVID-19 pandemic. Neurological Sciences, 43(6), 3479–3487. https://doi.org/10.1007/s10072-022-05972-6

Rajeswari, S. S., & Nair, M. (2022). Prediction of Parkinson’s disease from Voice Signals Using Machine Learning. Journal of Pharmaceutical Negative Results, 2031–2035. https://doi.org/10.47750/pnr.2022.13.S07.294

Rizek, P., Kumar, N., & Jog, M. S. (2016). An update on the diagnosis and treatment of Parkinson disease. CMAJ, 188(16), 1157–1165. https://doi.org/10.1503/cmaj.151179

Sakar, B. E., Serbes, G., & Sakar, C. O. (2017). Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson’s disease. PLOS ONE, 12(8), e0182428. https://doi.org/10.1371/journal.pone.0182428

Skodda, S., Grönheit, W., & Schlegel, U. (2011). Intonation and Speech Rate in Parkinson's Disease: General and Dynamic Aspects and Responsiveness to Levodopa Admission. Journal of Voice, 25(4), e199–e205. https://doi.org/10.1016/j.jvoice.2010.04.007

Stamate, C., Magoulas, G. D., Kueppers, S., Nomikou, E., Daskalopoulos, I., Jha, A., Pons, J. S., Rothwell, J., Luchini, M. U., Moussouri, T., Iannone, M., & Roussos, G. (2018). The cloudUPDRS app: A medical device for the clinical assessment of Parkinson's Disease. Pervasive and Mobile Computing, 43, 146–166. https://doi.org/10.1016/j.pmcj.2017.12.005

Suppa, A., Costantini, G., Asci, F., Di Leo, P., Al-Wardat, M. S., Di Lazzaro, G., Scalise, S., Pisani, A., & Saggio, G. (2022). Voice in Parkinson's Disease: A Machine Learning Study. Frontiers in Neurology, 13. https://www.frontiersin.org/articles/10.3389/fneur.2022.831428

Trudelle, P. (2006). Instructions for the Unified Parkinson Disease Ratings Scale (UPDRS). Kinésithérapie, La Revue, 6. https://doi.org/10.1016/S1779-0123(06)74622-8

Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2010). Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. Journal of The Royal Society Interface, 8(59), 842–855. https://doi.org/10.1098/rsif.2010.0456

Tsanas, A., Little, M. A., & Ramig, L. O. (2021). Remote Assessment of Parkinson's Disease Symptom Severity Using the Simulated Cellular Mobile Telephone Network. IEEE Access, 9, 11024–11036. https://doi.org/10.1109/ACCESS.2021.3050524

Van Den Bergh, R., Bloem, B. R., Meinders, M. J., & Evers, L. J. W. (2021). The state of telemedicine for persons with Parkinson's disease. Current Opinion in Neurology, 34(4), 589. https://doi.org/10.1097/WCO.0000000000000953

White, J., & Power, S. D. (2023). k-Fold Cross-Validation Can Significantly Over-Estimate True Classification Accuracy in Common EEG-Based Passive BCI Experimental Designs: An Empirical Investigation. Sensors, 23(13), Article 13. https://doi.org/10.3390/s23136077

Zimmerman, M., Morgan, T. A., & Stanton, K. (2018). The severity of psychiatric disorders. World Psychiatry, 17(3), 258–275. https://doi.org/10.1002/wps.20569