Parkinson's Disease Progression Prediction Using Voice-Based Telemonitoring and Machine Learning Regresssion Models

Yadav Vijayan

Supervisor: Dr. Ahmed Makki

Dublin Business School

This dissertation is submitted for the degree of

Master Of Science in Data Analytics

January 2026

Declaration

I, Yadav Vijayan hereby declare that the research project with the title “Parkinson’s Disease Progression Prediction Using Voice-Based Telemonitoring and Machine Learning Regression Models” is my own work and has been done under the supervision of Dr Ahmed Makki. Furthermore, this work has not been submitted to any other recognized university or institution for any kind of degree or diploma. Moreover, the sources used for information inside this work have been acknowledged accordingly.

Student Number: 20059508

Acknowledgement

I would like to take this opportunity to thank my project guide, Dr Ahmed Makki, for his patience, guidance, and support throughout this project. I would also like to take this opportunity to thank all those friends and family members who have been with me in this project and have supported me in this journey.

Abstract

Parkinson’s disease is major issue that affects people in different ways. If a person is affected by Parkinson’s disease they may not be able to function properly as their motor functions may be damaged due to the effects of the disease. So, Parkinson’s disease need to be identified early and treated properly to ensure that lives of people are affected by the disease. One way to identify the disease is to use machine learning and deep learning methods as these methods can learn data related to Parkinson’s disease and performs predictions. The UPDRS score is a measure to assess the level of Parkinson’s disease, the data related to these scores can be used for building a model that predicts the level of Parkinson’s disease. In the study proposed here the machine learning models are trained to predict the total UPDRS score and the Motor examination score. In the study proposed here, the machine learning and deep learning models used includes, Voting regressor, Long Short Term Memory(LSTM) and Gated Recurrent Unit(GRU). The models are trained using the data related to UPDRS scores. The models are trained and their performances were evaluated from the evaluation of the performances of the models it was seen that the best performance was shown by the Voting regressor. The predictions made by the Voting regressor interpreted using the SHAP and the features that were important for the prediction process were found.

Table of Contents

1 Introduction

1.1 Background

Parkinson's Disease is a progressive neurodegenerative disorder characterized by motor and non-motor features including, rigidity, tremor, bradykinesia, cognitive changes, postural instability and mood disorders. It was found in 2021, that 11.77 million people worldwide were suffering from Parkinson's Disease (Luo et al., 2025). The symptoms of Parkinson's Disease makes people weak and have a huge impact on patients' quality of life, involving the loss of daily living capacities and independence (Hoseinipalangi et al., 2023). Parkinson's Disease has a progressive nature and this creates the need for tools that assess the disease severity and progression to ensure timely interventions, optimized treatment strategies, and enhance patient outcomes.

The Unified Parkinson's Disease Rating Scale (UPDRS) is a commonly used clinical tool to evaluate the progression and severity of Parkinson's Disease (Mehta et al., 2021). It comprises four parts: Part I (Mentation, Behavior, and Mood), Part II (Activities of Daily Living), Part III (Motor Examination), and Part IV (Motor Complications of Therapy) (Bouça-Machado et al., 2022). Each part evaluates specific aspects of Parkinson's diseases, with Part III focusing on motor symptoms and the total UPDRS score providing a comprehensive measure of disease severity. The motor symptoms evaluated in Part III, like bradykinesia and tremor, are major effects of Parkinson's Disease, while non-motor symptoms, including mood disturbances and cognitive decline significantly add to disability (Vignoud et al., 2022 ; Chen and Liu, 2022). The total UPDRS score, summing all four parts, shows a comprehensive view of disease burden. An accurate assessment of motor and total UPDRS scores was needed for creating treatment plans for people suffering from Parkinson's disease , in a personalized manner, so that individualized treatment plans are designed for monitoring disease progression, and assessing therapeutic efficacy.

The ability to predict UPDRS scores has implications for optimizing medical treatment and patient care. Exact prediction of motor and total UPDRS scores allow the clinician to foreknow disease progression, modify medication schedules and administer non-pharmacological therapies like physiotherapy to reduce symptoms. For patients, accurate prediction can limit the frequency of clinical visits, which is a huge benefit for those with limited mobility or located in remote areas, it supports monitoring of Parkinson's Disease. In addition, accurate predictions of UPDRS scores allows researchers to study the disease progression and create new therapies.

1.2 Problem Statement

Even though UPDRS scores play an important role, the current techniques for assessing Parkinson's Disease severity are also fraught with complexity. The classic UPDRS is based

upon the clinical judgment of trained neurologists and suffers from being subjective in nature and requiring time to complete as well as experienced clinicians. These tests can only be performed in person, presenting logistical issues for patients with mobility problems or those living in remote locations. Further, the subjective character of clinical assessments adds a variability since different professionals might interpret symptoms differently on scores. External factors, like, the timing of medication doses can make assessments more complicated, as symptoms vary throughout the day (Farzanehfar, Woodrow and Horne, 2022).

Voice signal measures, which capture subtle variations in patterns associated with speech, a usually seen early symptom of Parkinson's Disease (Laganas et al. 2021). Speech defects, including reduced difficulties in articulation and vocal intensity, are related to disease severity and motor dysfunction, making voice analysis a relevant biomarker. Machine learning methods have been used for the prediction of different kinds of diseases (Jovovic et al., 2023). So, in the study proposed here a machine learning based model that predicts motor and total UPDRS scores, separately is proposed.

1.3 Aim

The aim of the study is to predict motor and total Unified Parkinson's Disease Rating Scale (UPDRS) scores, i.e., from 4 parts of the UPDRS scores(Mentation, Behavior, and Mood (Part I), Activities of Daily Living (Part II), Motor Examination (Part III), and Motor Complications of Therapy (Part IV)), the model predicts the score for Motor Examination and the sum of the scores of the 4 parts, scores using voice signal measures and demographic data.

1.4 Objectives

- Generate visual plots from a dataset that contains voice signal measures and demographic data, and analyze the hidden patterns.
- Train machine learning models to predict score for Motor Examination and the sum of the scores of the 4 parts of UPDRS.
- Use Explainable AI(XAI) techniques for finding the factors that are significant for the prediction of score for Motor Examination and the sum of the scores of the 4 parts of UPDRS.
- Assess the performance of the trained machine learning model by finding the values of the performance metrics associated with the model.

1.5 Research Questions

- What are the most significant acoustic features that correlate with motor and total UPDRS scores in Parkinson's patients ?

- Which regression model Voting Regressor, Long Short Term Memory(LSTM), and Gated Recurrent Unit(GRU) provides the most accurate and interpretable prediction of UPDRS scores, and what are the features that contribute to the prediction of score for Motor Examination and the sum of the scores of the 4 parts of UPDRS?

1.6 Report Structure

'Introduction' is the first section of the report. 'Literature review' contains the details of the existing literature related to the theory of UPDRS score prediction and the existing literature that focus on the prediction of UPDRS score using machine learning methods. 'Methodology' contains the discussion related the data used in the study and the machine learning algorithms that are used for building the UPDRS score prediction model and assessment of the prediction performance of the mode and the other algorithms considered in the study but not chosen for the study to build the UPDRS score prediction model. The 'Implementation' section contains the details on how the different methods defined for the study was utilized for implementing the model that performed the prediction of UPDRS scores. 'Results' section contains the details of the assessment of the prediction performance of the machine learning model built here and the results of the analysis of data related to different physiological elements. 'Discussion and evaluation' section contains the discussion of the insights related to the prediction of UPDRS scores, from this study and the discussion about the limitations associated with the study that was done here. 'Conclusion and future enhancements' is the summary of the different insights from the study, the methods used in the study and the results of the study along with the future enhancements that can be made to the study.

2 Literature Review

2.1 Theoretical foundation for the prediction of UPDRS scores

Assessing UPDRS traditionally depends on subjective clinical assessments, which are dependent on the clinician's ability, time-consuming and repeatedly needs in-person visits, creating challenges for patients that have mobility problems or those living in areas that are remote. Difference in the symptoms presented and external factors, like medication timing, further creates inconsistencies in scoring. For addressing these limitations methods that included non-invasive biomarkers, like voice signal measures, which capture speech impairments, an initial and prevalent symptom of Parkinson's disease (Motin et al., 2025). Speech characteristics, including reduced vocal intensity, pitch variability, shimmer and jitter, correlate with motor dysfunction and severity of the disease, making the analysis of voice a a promising tool for predicting UPDRS objectively (Grobe-Einsler et al., 2023).

2.2 Machine learning for the prediction of UPDRS scores

The Voting Regressor is an ensemble learning method that combines predictions from different multiple base regressors to generate a final output, usually by averaging or weighted averaging (Chen and Luc, 2022). Different kinds of machine learning models can be used as the base models like the RF which has multiple decision trees and aggregates their predictions, lowering overfitting via bagging (K, 2023). Its ability to capture non-linear relationships and high-dimensional data makes it apt for processing voice signal measures like shimmer and jitter, which shows complex patterns in Parkinson's disease.

DT Regressor divides data into hierarchical nodes based on feature thresholds, offering the ability to model non-linear relationships and interpretability (Agarwal et al., 2022). DT is prone to overfitting, however using it in voting regressor reduces the risk of overfitting by ensemble averaging.

SVR uses a hyperplane to model relationships in high-dimensional spaces, effectively handling relationships in small to medium-sized datasets that contain non-linear patterns (Roozbeh et al., 2023). It can handle outliers and can handle the variability in voice data.

Long Short-Term Memory (LSTM) networks are a kind of recurrent neural network (RNN) designed to model time-series and sequential data by solving the vanishing gradient problem (Al-Selwi et al., 2023). LSTMs use memory cells and gates (input, forget, and output) to retain and update information over long sequences, selectively, making them apt for processing temporal voice signal measures, like formant and pitch frequencies, which evolve over time in Parkinson's disease patients. LSTMs are apt for handling the inherently sequential voice signals, containing features like shimmer and jitter showcasing temporal dependencies. LSTMs can capture these patterns, allowing accurate modelling of speech impairments linked to motor dysfunction. They have the ability to handle long-term

dependencies, which makes sure that subtle changes in voice data are acknowledged which may signify the increase in the severity of Parkinson's disease,

GRU are a simplified version of LSTMs, utilising lesser gates(update and reset) to model sequential data with lower computational complexity (Can, Krishnamurthy and Schwab, 2020). GRUs balance performance and efficiency, making them apt for a small sized dataset, like the data collected in relation to Parkinson's disease. GRUs are effective at capturing temporal patterns in voice signals like, vocal intensity and articulation rate, which are significant for the prediction of UPDRS scores. GRUs have the ability to model sequential voice data with lesser parameters than LSTMs, possibly enhancing training efficiency without affecting accuracy (Feng et al., 2020). Their ability to handle noise in voice signals, which can arise from recording conditions, improves their suitability in handling real-world data.

The GRU and LSTM were chosen for the study because of their prediction performances than their ability to handle the data in the form of sequeneces.

2.3 Related Works

Machine learning methods like, Conjugate Gradient, Linear Least Square (LLS), Ridge Regression and Adam optimization algorithm were used for predicting total UPDRS in the study by (Hamzehei et al., 2023). The dataset used for this study contained data related to Parkinson's disease. The study also utilized cloud computing, specifically Google Cloud, to reduce time complexity and enhance accuracy and accessibility of the machine learning algorithms. Results indicated that the Adam optimization algorithm achieved the best R-squared score of 0.904410.

The features based on which the UPDRS part-III score, i.e. motor symptoms, is computed, was predicted based on data collected using a smartphone and machine learning methods in the study by (Guo et al., 2025). The study utilized data from a clinical trial involving early de novo Parkinson's disease patients, employing the PDAssist smartphone application to assess eight motor tasks: resting tremor, postural tremor, finger tapping, facial expressions, rigidity, speech, walking, and pronation/supination. These assessments were compared against UPDRS Part III scores. The study developed and evaluated machine learning models, including XGBoost, Extra Trees, Random Forest, LSTM, AdaBoost, and Bagging, for predicting UPDRS Part III scores based on smartphone data. The prediction model demonstrated acceptable performance in detecting mild Parkinson's disease symptoms, with accuracy ranging from 87% to 93% for resting tremor postural tremor, finger tapping, facial expressions, and postural stability, achieved by different machine learning models. However, this study did not predict the actual UPDRS score.

The deep learning Convolutional Neural Network (CNN) model was used for predicting the MDS-UPDRS-III scores based on the wearable-based gait data in the study by (Rehman et al., 2021). The data used in the study was the Gait-related data obtained from 70 people with Parkinson's disease. The performance of the model was evaluated based on mean absolute

error (MAE),), Intraclass correlation (ICC) and Pearson correlation values. The results of the study showed that the prediction model achieved an MAE of 6.29. However, this study only focuses on the prediction of only the MDS-UPDRS-III scores.

Clustering and prediction learning approaches were used for predicting Parkinson's Disease in the study by (Nilashi et al., 2022). The SVR model was used in the study. A real-world Parkinson's Disease dataset was used in this study. The results of the study showed that the SVR model achieved an MAE value of 0.721 and RMSE of 1.4942 for Motor-UPDRS and a RMSE value of 1.4526 and MAE of 0.689 for Total-UPDRS.

Parkinson's disease subtypes and disease progression were predicted using unsupervised and supervised machine learning methods in the study by (Dadu et al., 2022). Longitudinal clinical data from the Parkinson's Disease Progression Marker Initiative (PPMI) was used for training the models and the models were validated using an independent cohort from the Parkinson's Disease Biomarker Program (PDBP). The study involved dimensionality reduction techniques like Non-negative Matrix Factorization (NMF) to create a multi-dimensional space capturing disease features and progression rate, followed by unsupervised clustering using Gaussian Mixture Models (GMM) to define subtypes. The study successfully distinguished three distinct PD subtypes: slow (PDvec1), moderate (PDvec2), and fast (PDvec3) progressors, achieving high accuracy in predicting disease progression 5 years after initial diagnosis with an average Area Under the Curve (AUC) of 0.92.

A data-driven framework that combines multimodal data with machine learning and statistical approaches to classify severity subtypes of Parkinson's disease in the study by (Park et al., 2025). The dataset included clinical characteristics, physical function and lifestyle data, gait parameters from motion analysis systems, and wearable sensors collected from 102 persons with Parkinson's disease. Three Parkinson's disease severity subtypes (mild, moderate, severe) with increasing clinical severity from clusters 1 to 3 were classified in this study. The machine learning models used in the study included, RF Regressor, Least Absolute Shrinkage and Selection Operator (LASSO) and Logistic Regression Model. The best performance in predicting MDS-UPDRS Part II and Part III was shown by the RF Regressor. The best performance in the prediction of severity subtypes of Parkinson's disease was achieved by the Logistic Regression Model as it achieved accuracy values ranging between 75% and 100% for the different modalities based on which prediction was done.

The support vector machine (SVM) was used for building a model that predicted depression in Parkinson's disease, in the study by (Byeon, 2020). The dataset used in the study contained the Parkinson's Disease Epidemiology data.. However, this study focused in the prediction of depression in Parkinson's disease and not on the prediction of UPDRS scores.

A model for unobtrusive and continuous prediction of UPDRS part-III was proposed in the study by (Hssayeni et al., 2021). An ensemble model containing three deep learning models was used in the study for prediction. Dual-channel Long Short-Term Memory (LSTM) Network, 1D Convolutional Neural Network-LSTMCNN-LSTM) Network, and 2D CNN-LSTM Network. The dataset used in the study contained gyroscope data from 24 Parkinson's

disease patients. The results of the study showed that the ensemble model achieved a mean absolute error (MAE) of 5.95 in predicting UPDRS part-III score.

A deep learning based, analysis framework to evaluate the UPDRS scores using a video-based analysis was proposed in the study by (Mehta, Asif, Hao, Bilal, Stefan, et al., 2021). An ensemble model was used in the study and the model contained LSTM, Temporal Convolutional Network (TCN) , Spatio-Temporal Graph Convolutional Networks (ST-GCN) and Resnet50. The dataset used for this study contained video recordings of 32 patients. The results of the study showed that the model proposed in the study achieved F1-scores of 0.78 for posture instability and gait disorders (PIGD) and 0.75 for bradykinesia. However, the data used for training the machine learning models in the study is video data and if the videos are of low quality then the prediction by the model will be improper.

A soft voting ensemble method that integrated base models like K-Nearest Neighbours (K-NN), Naïve Bayes (NB), Decision Tree (DT), Logistic Regression and SVM, for predicting Parkinson's disease in the study by (Aman and Chhillar, 2025). The dataset used was the Parkinson's Telemonitoring Voice Dataset from the UCI Repository. The data in the dataset was pre-processed for making it apt for training the machine learning models used in the study. Dimensionality reduction was done on the data for reducing the number of features in the dataset.. However, this study did not focus on the prediction of UPDRS and UPDRS part-III scores.

A DNN was used for predicting Parkinson's disease progression, particularly Motor and Total-UPDRS scores in the study by (Shahid and Singh, 2020). This model utilizes a reduced input feature space derived from a Parkinson's telemonitoring dataset. The dataset comprises 5875 voice measurements from 42 PD patients, including 16 biomedical voice features, age, gender, and test time duration. To address multicollinearity and reduce input feature space dimensionality, Principal Component Analysis (PCA) is employed on the dataset, with the first 8 principal components being used as input to the DNN model, capturing 96.23% of the total variance. The DNN model, consisting of four dense layers was optimized using Adam as the optimizer and L2 regularization to stop overfitting. The results of the study showed that for motor-UPDRS prediction, the R-squared, RMSE and MSE values were 0.970, 1.422 and 0.926., respectively, while for Total-UPDRS prediction the values were, 0.956, 2.221 and 1.334. This study used a deep learning model for the prediction of Motor and Total-UPDRS scores, however a model that combined the prediction of different machine learning methods were not used in this study.

The studies by ,Mehta, Asif, Hao, Bilal, Stefan, et al.(2021) and Hssayeni et al.(2021) showed that deep learning, a sub-branch of machine learning, based models are able to effectively analyze the data needed for predicting of UPDRS and UPDRS part-III scores. Machine learning models were used in the study by Byeon,(2020), however this study was focused on the prediction of depression in Parkinson's disease. Deep learning based models are better at handling complex data than machine learning based models(Shiri et al., 2023). The existing study by Aman and Chhillar (2025) used machine learning models and performed the prediction Parkinson's Disease severity using UPDRS. The study by Shahid

and Singh(2020), performed the prediction of both Motor and Total-UPDRS scores, however, a single deep learning model was used in this study. It was seen that the performance of models that combine the capabilities like the Voting regressor showed a better performance in predictive analysis than individual machine learning models (Kumar, Bilgaiyan and Mishra, 2023 ; Sankalpa, Kittipiyakul and Laitrakun, 2022). So, if a model like the voting regressor was used then the performance of the prediction model in the study by Shahid and Singh(2020) may be improved. However, there were limited existing studies that used a voting regressor for predicting the Motor and Total-UPDRS scores.

Study	**Dataset type (voice/video/wearables)**	**Target (UPDRS total vs III)**	**Metric used**	**Best score**
Hamzehei et al., 2023	voice	total	R-squared	0.904410
Guo et al., 2025	wearables	III	accuracy	87% to 93%
Rehman et al., 2021	wearables	III	MAE	6.29
Nilashi et al., 2022	voice	total vs III	MAE, RMSE	MAE 0.721 (III), MAE 0.689 (total)
Dadu et al., 2022	other (clinical)	progression (subtypes)	AUC	0.92
Park et al., 2025	wearables	II and III (subtypes)	accuracy	75% to 100%
Hssayeni et al., 2021	wearables	III	MAE	5.95
Mehta et al., 2021	video	III	F1-score	0.78
Shahid and Singh, 2020	voice	total vs III	R-squared	0.970 (III), 0.956 (total)

Table (1): summary of existing literature

Table(1), showed the summary of the literature studied and in the table the roman number ‘III’, represented motor examination score while total represented the total UPDRS score.

2.4 Research Gap

In the existing studies it was seen that machine learning and deep learning methods are effective in the prediction of Motor and Total-UPDRS scores. However, there were limited existing study's that used a model like the voting regressor for predicting the Motor and Total-UPDRS scores, and also interpreted the prediction by the prediction models. The voting regressor being an ensemble model may even show a performance better than the prediction shown by individual models that were used in the prediction of UPDRS scores.

In the study proposed here, a voting regressor that contains base machine learning models, RF regressor, DT regressor and Support Vector Regressor(SVR), LSDTM and GRU are used for the prediction of UPDRS and UPDRS part-III. The study involves both a voting regressor model and the prediction of Motor and Total-UPDRS scores. The predictions by the models built here, was done using SHAP.

3 Methodology

The methods used for building a model that predicts motor and UPDRS scores were discussed in this section. The reasons for choosing the different machine learning methods for this study is defined in this section.

3.1 Development methodology

A Software Development Lifecycle(SDLC) was needed so that the study can be done and the motor and UPDRS score prediction model can be built. The study proposed here involved, pre-processing the data related to UPDRS and using the data for training a machine learning based model that predicts motor and UPDRS scores. Several stages are defined for different SDLC methods. A method that can fit to the study proposed here can be used. The SDLC method apt for the study proposed here was CRISP-DM as it had stages that were needed for the study that was done here to build machine learning based motor and UPDRS score prediction models (Kannengiesser and Gero, 2023). The research methodology was design science and the execution framework was CRISP-DM.

The first stage of CRISP-DM was business understanding and this stage was used for identifying how the aim of the study can be achieved. This stage was needed in the study done here as the methods for implementing motor and UPDRS score prediction was needed. In this stage it was decided that the prediction of UPDRS and motor scores can be done by training a machine learning model using the data related to UDRS scores. The second stage of CRISP-DM was data understanding, which was the stage in which the data in the dataset used in the study was studied. This stage was needed in the study proposed here, because the dataset containing data related to UPDRS needed to be studied to find the hidden patterns in the data and for finding if the data contains any element that needs to be removed from the data. This stage was needed here, as the data related to UPDRS needed to be studied to understand if it needed to be transformed into a form apt for training the machine learning models. The underlying patterns in the data needed to be explored, so this also needed to be done. The underlying trends and` patterns related to UPDRS was analyzed by generating visual plots from the data in the dataset.

The data preparation stage in CRISP-DM involved preparing the data so that it is transformed into a form that is apt for the study it is being used for. Data preparation was needed in the study done here as the UPDRS data needed to be cleaned and made suitable for training machine learning models. The modeling stage in CRISP-DM defines how the data can be modeled. The modeling stage was needed in the study proposed here, as the data related to UPDRS needed to be used for training machine learning models. In this stage, the data related to UPDRS was used for training the Voting Regressor, LSTM, and GRU models. The evaluation stage in CRISP-DM defined the evaluation of the performance of the machine learning models that were trained to predict motor and UPDRS scores In the evaluation stage

for this study the trained machine learning models were evaluated and the values of the performance metrics associated with the machine learning models were found.

3.2 Data related to UPDRS scores

The Parkinson's Telemonitoring Dataset was used in the study (UCI Machine Learning Repository, 2025). This dataset contained data related to voice signals, so the features defined in the dataset also correspond to voice signals. The dataset contained demographic data, like age, gender, etc. The features that were used for both the prediction of total UPDRS and motor scores, were the same these were, 'age', 'Jitter(%)', 'Shimmer:APQ11', 'NHR', 'HNR', 'RPDE', 'DFA', and 'PPE'.

This data was chosen for the study because it contained the data related to voice signals that helped in building a machine learning model that predicted the total UPDRS and motor scores. There were columns in the dataset that were related to UPDRS and motor scores, which could be used as the target columns for training the machine learning models to perform the prediction of UPDRS and motor scores. The quality of the dataset was good but it still had some issues, so these issues needed to be solved using pre-processing techniques. Due to this the Parkinson's Telemonitoring Dataset was used in the study that was carried out here.

3.3 Modelling

The machine learning models used in the study were, Voting Regressor, LSTM, and GRU. The base models within the Voting Regressor, were RF, DT and SVM. These machine algorithms were trained to predict total UPDRS and motor scores. The Voting Regressor was used in the study over machine learning models like, XGBoost because of the lower risk of overfitting associated with the Voting Regressor model, when performing training using the data related to UPDRS scores, present in the dataset. The Voting Regressor has a very good ability to generalize data by averaging predictions and mitigating individual model biases. This makes the Voting Regressor model better than the XGBoost model as the XGBoost model needs extensive hyperparameter tuning and overfitting may be faced when dealing with biomedical data (Dalal et al., 2022). Voting regressor was more suitable for prediction of total UPDRS and motor score, than the Linear Regression, which assumes linearity, making it unsuitable for the complex non-linear data related to UPDRS scores (Yann Ling Goh, Chern Long Ng and Ling, 2023).

LSTM was chosen in this study because of its features that allows it to model long-term temporal dependencies in sequential voice recordings, important for obtaining subtle changes in jitter, shimmer, and pitch, which are early indicators of declining motor functions (Wang *et al.*, 2022). The LSTM was chosen over the standard RNN because of the presence of vanishing gradients (Al-Selwi *et al.*, 2023). The GRU is a lightweight version of LSTM. The GRU was chosen in this study over Transformer models because these models were used

when the size of the dataset was large, as the Transformer models are effective in handling data, which is large in size, so, the Transformer models were not needed in this study as the data that is being handled was very large in size. The Transformer models also needed high resources, so it was not used (X. Wang *et al.*, 2022).

4 Implementation

4.1 Tools Used

The tool used for building the total UPDRS and motor score prediction model was Python as a programming language was used for building the UPDRS and motor score prediction model. Python was used in this study because it was the programming language apt for implementing machine learning models. Python was apt because it had several libraries that can be used for implementing machine learning models. The presence of a large library of machine learning models makes sure that the different kinds of machine learning models used in the study can be implemented for performing the prediction of UPDRS and motor scores (Hachimi *et al.*, 2022). Python was chosen over other programming languages like, Java, because Java was not able to offer the large number of libraries for implementing machine learning models, which were offered by Python.

4.2 Exploratory Data Analysis(EDA)

EDA was done in the study carried out here to find the underlying patterns associated with the data in the UPDRS dataset. The different features in the dataset was used for generating visual plots and the hidden patterns in the data associated with UPDRS could be analysed from the visual plots that were generated. The EDA was done before pre-processing and some of the features that were removed during pre-processing like, 'Shimmer(dB)' was used in EDA.

Histograms were generated for all the features defined in the UPDRS dataset. The generation of histograms based on the different features in the dataset helps in understanding the nature of the data associated with the different features in the UPDRS dataset. These histograms generated are given in figure(1).

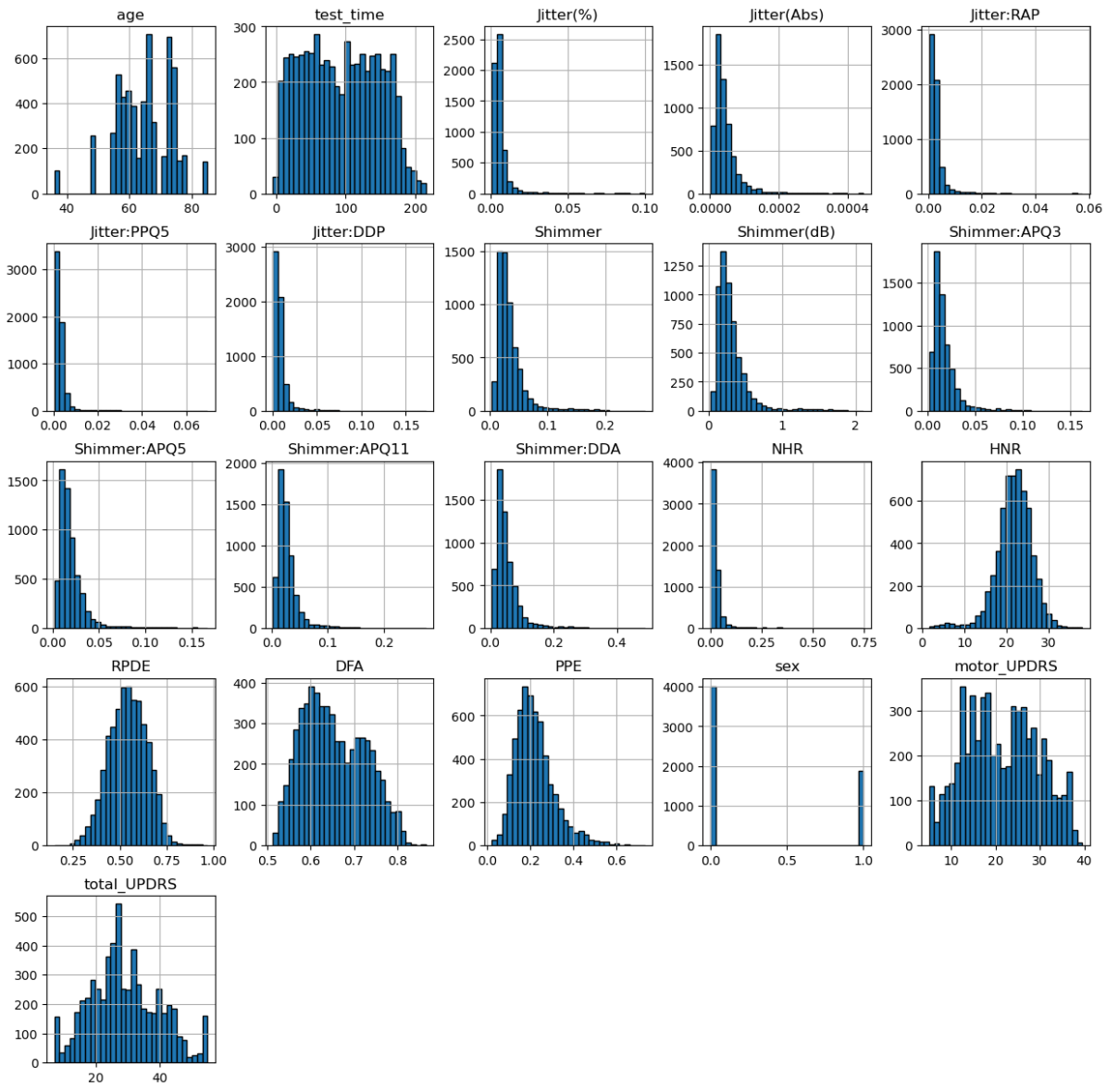


Figure (1): the histograms associated with different features

From figure(1), it can be seen that the ‘Age’ feature has a distribution that is slightly right-skewed, showing a higher frequency of data related to younger people in the dataset. The ‘Test Time’, feature is uniformly distributed. The two features ‘Jitter(%)’ and ‘Jitter(Abs)’, associated with jitter showed a right skew, specifying that lower jitter values are more common, with gradually decreasing frequencies as values increase. High jitter levels may indicate poorer voice quality. The feature ‘Shimmer’, which represents a variability in amplitude, also shows a right-skewed distribution.

Features like ‘NHR’ (Noise-to-Harmonic Ratio) and ‘HNR’ (Harmonic-to-Noise Ratio) showed distributions has a large number of normal voice samples with balanced harmonic structures. The feature ‘Total UPDRS’ has a complex distribution. The distribution of the data samples based on the feature ‘sex’ shows that the representation of genders in the dataset was balanced. The nature of the distributions associated with the features may reveal some insights related to the prediction of total UPDRS and motor scores.

A number visual plots were generated based on different features in the dataset against each other. The features in the dataset that were used for the generation of the visual plot were, ‘motor_UPDRS’, ‘total_UPDRS’, ‘Jitter(%)’, ‘Shimmer(dB)’, ‘HNR’, ‘RPE’, ‘PDPF’, ‘DFA’ and ‘PPE’.

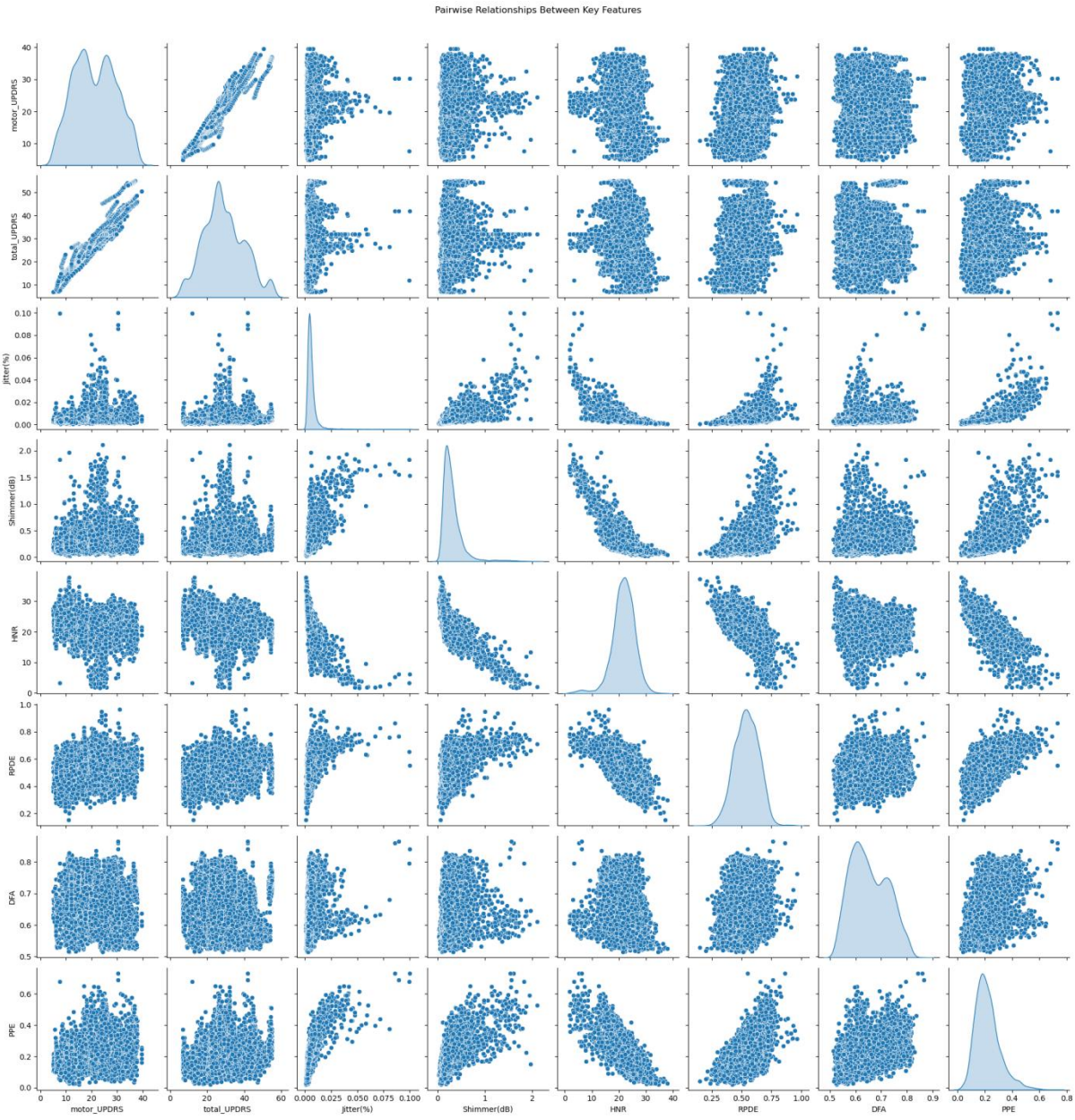


Figure (2): plots of the different features against each other

From the scatter plots in figure(2), it was seen that there were positive correlations between the features motor_UPDRS' and 'total_UPDRS', showing that with the increase in Motor Examination score, the total UPDRS may also increase, this relationship indicates that the overall ratings of Parkinson's disease severity is affected by the changes in motor functions. It was seen that the relationship between 'Jitter(%)' and 'motor_UPDRS' showed a trend of having higher jitter levels corresponding to the higher motor impairment scores, this may indicate that the variations in voice frequency may act as an indicator of motor control issues.

The diagonal plots in figure(2), the one between the features 'Shimmer(dB)' and 'HNR' show distinct distributions. Some possible outliers in the data were seen for features like, 'DFA' and 'PPE', these outliers were removed from the data when it was standardised during the pre-processing of the UPDRS data.

A box plot was generated using the data associated with the different features in the dataset. The box plot was generated to get a comprehensive overview of the distribution and central tendency of various features in the dataset.

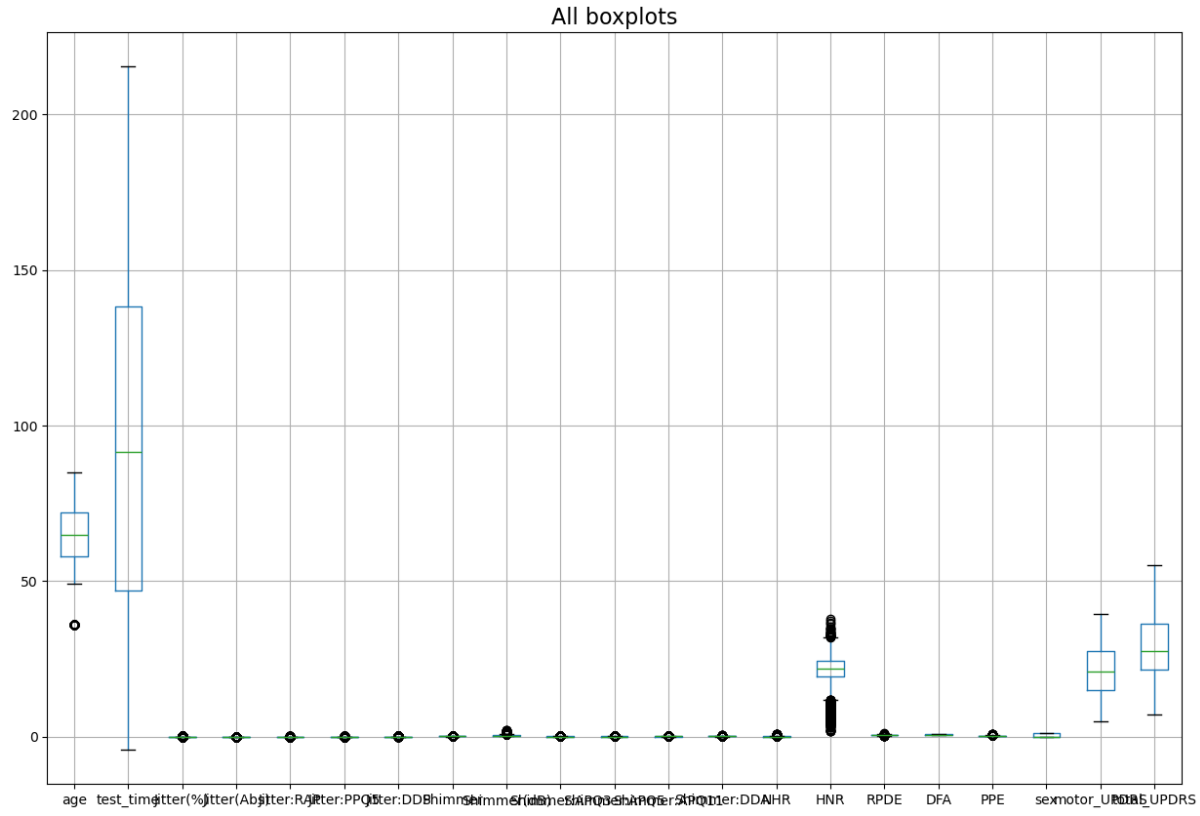


Figure (3): Box plot based on several features in the UPDRS dataset

From figure(3), it was seen that the boxplot showed the distributions of key features, including demographic factors and a range of voice signal measures. The boxplot generated based on the data associated with the features in the UPDRS dataset showed that there were

outliers in the data. The features that show high variability can impact the prediction of total UPDRS and motor scores. For example, extreme values can be shown by some voice signal measures, representing unique cases or possible noise in the dataset. This helps in finding if there are outliers in the data as these outliers affect the prediction of UPDRS and motor scores.

4.3 Data Preparation

The data in the UPDRS dataset was studied and it was seen that the data in the dataset was of high-quality. Machine learning based studies usually includes a data cleaning or data preparation phase where the data is transformed to a form that is of high quality and apt for training machine learning models to perform the prediction that is required. The dataset used in the study carried out here contained data that did not have any irrelevant data like missing values or duplicate values, this meant that tasks like the removal of the missing and duplicate values were not needed in the study. However, during the analysis of the data in the dataset containing data related to UPDRS scores, it was seen that several features in the dataset did not have any significant contribution to the prediction of total UPDRS and motor scores. These columns or features in the data that were irrelevant to the prediction of UPDRS scores, were removed from the data. The columns removed were, 'Jitter(Abs)', 'Jitter:RAP', 'Jitter:PPQ5', 'Jitter:DDP','Shimmer', 'Shimmer(dB)', 'Shimmer:APQ3', 'Shimmer:APQ5', 'Shimmer:DDA','test_time', and 'sex'.

One issue that the data in the UPDRS dataset has is that the data was not standardised. The data in the dataset not being in a particular standard was a problem as the training of the machine learning models will be improper if the non- standardised data was used for training the machine learning models. It was understood that the data in the dataset was not standardised because the values of the different features in the UPDRS dataset were in different numerical ranges, and the numerical range of one feature was very large numerically compared to another feature. This difference in the distribution range associated with the different features in the UPDRS dataset was a problem because during training the machine learning model will consider the feature having values from the lower numerical range as irrelevant compared to the features having values from higher numerical range, which leads to the loss of crucial information during the training of the machine learning models to perform the prediction of the total UPDRS score and the motor score. So, the data in the dataset needed to be standardised.

The data in the UPDRS dataset was standardised after the Exploratory Data Analysis(EDA) was done. The data was standardised using the 'StandardScaler' technique (*StandardScaler*, 2025). The 'StandardScaler' technique was applied on the data in the UPDRS dataset and the data in the dataset was standardised.

4.4 Training Voting regressor, LSTM and GRU

The data related to UPDRS scores was used for training the machine learning models. The machine learning models were trained for the prediction of, the total UPDRS scores and Motor Examination score.

UPDRS prediction

For training the models to perform the prediction of the total UPDRS score, the features and the target variable 'total_UPDRS' were separated. The data was then separated to a training set, containing 80% of the data, while the testing set contained 20% of the data. The training set was used for training the models. For this training the 'motor_UPDRS' feature was not used.

The Voting Regressor was trained, beginning with the initialisation of the base models used in the study. The base models used were, RF Regressor, DT Regressor and Support Vector Regressor(SVR). The RF Regressor was initialised using the parameters 'n_estimators' and 'random_state'. The parameter 'n_estimators' defined the number of tress in the RF, it was assigned the default value 100 because this value provided a balance between the time taken an performance as a higher value for 'n_estimators' would have increased the training time of the model, meanwhile if the value for 'n_estimators' was too low, like, a value below 100, then the RF Regressor was not able to show the best performance (*RandomForestRegressor*, 2025). The parameter 'random_state' was used for ensuring that the training of the Voting Regressor was reproducible(*RandomForestRegressor*, 2025).

The DT Regressor was defined using the parameter 'random_state' for ensuring reproduciblility (*DecisionTreeRegressor,* 2025). The SVR was initialised with the value of the parameter 'kernel' set as 'rbf', this value of the kernel was used because using this kernel allowed the SVR to handle complex relationships between the different features in the data related to biomedical voice signal patterns (*SVR*, 2025).

The LSTM model used in the study had three layers. The first layer in the LSTM model was the LSTM layer, which was defined with 64 units, i.e., the parameter 'units' was assigned the value 64, this was done to make sure that complex temporal dynamics in voice features are captured by the units as a lower number of units may not be able to capture the complex relationships, while a higher number can lead to huge computational overhead, so a moderate value 64 was used (TensorFlow,2025). The activation function was set as 'tanh', because this was the default value for the activation function(TensorFlow,2025). The data related to the UPDRS scores was passed to the first layer of the LSTM. The second layer of the LSTM model was initialised with 'unit' set as 32 and 'activation' set as 'relu'. The 'unit' value was set as 32 because reducing the values of the units as each layer progresses towards the output ensured the capture of all the complex features in the data (Raj, Nayak and Kalyani, 2020). The activation function used in the study was 'ReLU', this activation function was used for this layer, as this activation function was able to model nonlinear relationships (Bai, 2022). The next layer was also a Dense layer with 'activation' set as 'relu' and 'unit' set as 8. The

output layer of the LSTM model is a Dense layer with 'activation' set as 'linear' and 'unit' set as 1. This was the layer from the which the out of the LSTM was obtained. The 'activation' was set as 'linear' because the UPDRS scores needed to be generated and they have a continuous nature (Jagtap and Karniadakis, 2023). The parameter 'unit' was set as 1 because it defined the output that is generated from the layer and, here the model performs the prediction of total UPDRS score which is a single value, so as only a single value was obtained from the output layer, 'unit' was set as 1. The LSTM was compiled with optimizer set as 'Adam' and loss set as 'Mean Squared Error (MSE)'. Adam was chosen because it performed learning at an adaptive rate (Tian and Parikh, 2022). MSE was chosen as the loss function because it penalized large errors more heavily, which make it an apt loss function for building a deep learning model for the prediction of UPDRS scores, which is a medical application that has low margin for error (Zhang *et al.*, 2021). The LSTM was trained and saved.

The GRU model was built as a Sequential model, this was done to ensure that the features in the data was captured as a sequence, and for the UPDRS scores such a way of capturing data was needed (Corsini, Yang and Apruzzese, 2021). The input layer of the GRU model is such that it specifies that each sample is treated as a single time-step with multiple features, this structure allows the GRU to analyse feature interactions in a sequential manner (Tuan, Chiu and Wang, 2022). The GRU model used in the study also followed the funnel structure with reducing number of units over the layers, similar to the LSTM, for capturing the complex features in the data. The 'units' was set as 64 and the default 'tanh' was used as the activation function(TensorFlow, 2025). The Dense layer followed by the GRU layer, for which the 'unit' was set as 32 and activation function was 'relu'. The next Dense layer had activation function as 'relu' and 'unit' as 8. The final layer of the GRU model was a Dense layer with 'unit' value as 1, because of only a single output(the total UPDRS score), and activation set as 'linear'. The model was trained with the 'adam' optimizer and loss function asset as 'MSE'.

Motor Examination score

The 'motor_UPDRS', column in the dataset was chosen as the target variable for the prediction of Motor Examination score, as the column contained the motor scores. The Voting regressor used the same base regressors, DT regressor, RF regressor, and SVR, that was used as the base regressors in the prediction of total UPDRS score. The GRU and LSTM were also built in the same way the models were built for the prediction of total UPDRS score. The Voting regressor, LSTM and GRU were trained with the target variable as 'motor_UPDRS', and saved. For this training the 'total_UPDRS' feature was not used.

4.5 Interpretation of the models using SHAP

The trained voting regressor, LSTM and GRU models, were interpreted using SHAP. SHAP was chosen as the method for interpreting the prediction of both the total UPDRS scores and

motor Examination score. SHAP was chosen in this study for the interpretation of the UPDRS score prediction and the Motor Examination score prediction by the Voting regressor and deep learning models, LSTM and GRU. SHAP was used in the study because it was based on cooperative game theory, which ensured that the feature contributions were fair and additive (Mayer, 2022). This allows SHAP to provide both global and local explanations, this meant that the SHAP technique was able to show the important of features related to UPDRS scores over the entire dataset(global), and show the feature importance associated with a specific prediction made by the predictions models built in this study (Misheva *et al.*, 2022).

SHAP was chosen over LIME for interpreting the prediction of the models that predict total UPDRS scores and the Motor examination scores because LIME was able to provide only local interpretations in the prediction by the voting regressor and the deep learning models (Gaudel *et al.*, 2022).

The SHAP technique was used and the trained voting regressor and deep learning models were interpreted.

4.6 Performance Evaluation

For evaluating the performances of the UPDRS score and motor Examination score models built here, the data in the testing set related to UPDRS scores was used in the study. The testing set that contained 20% of the entire data in the dataset was used for testing the performance of the models because if the data used for training machine learning or deep learning models was used again for testing the model then the model may overfit and a proper assessment of the performance cannot be done. This overfitting was due to the models already being familiar with the data when they were being trained.

The performances in the prediction of total UPDRS score and motor examination scores, by the Voting regressor and deep learning models were evaluated based on the values of the performance metrics defined below:

- **R-Squared score:** This is a statistical measure that evaluates the regression data fit in a model. This metric defines the difference between predicted values and actual values (*r2_score*, 2025). This metric should be high for a model that shows good performance.
- **Root mean squared error(RMSE):** RMSE is the direct measure of prediction error (*mean_squared_error*, 2025). This metric should be low for a model that shows good performance.

- **Mean Absolute Error(MAE) :** MAE is defined as the average sum of the absolute differences between the actual value and the predicted value (*mean_absolute_error*, 2025). This metric should be low for a model that shows good performance.

4.7 Ethical Considerations

The dataset used in the study was the Parkinson's Telemonitoring Dataset. The Parkinson's Telemonitoring Dataset was available from a free and public source. The dataset contained data related to UPDRS scores. The dataset did not contain any data related to the personal details of the people to whom the UPDRS scores may be associated with. This meant that no sensitive data was used in this study. The data used in the study like, the data in the Parkinson's Telemonitoring Dataset and the code for building the model that predicts total UPDRS score and UPDRS Motor-III score were stored securely and this protection of data ensured that the GDPR laws are not broken. The programming language used in the study was Python and the use of Python libraries did not lead to the copyright laws being broken as Python was a free and open-source language. The Python code used in the study was the code written for building the total UPDRS score and UPDRS Motor-III score prediction models and the code was not taken from the work of any other person or any existing work. The data used in the study is the data related to UPDRS scores and this data does not have any information that considers the different social groups in society. The absence of the data associated with social groups meant that this study did not have any biases against any social groups.

5 Results

5.1 Performance Evaluation

The performance metrics scores were found for both the prediction of total UPDRS score and Motor examination scores, by the Voting regressor, LSTM and GRU.

Model	**R^2 Score**	**RMSE**	**Mean Absolute Error (MAE)**
VotingRegressor	0.9883	0.4401	0.3266
GRU	0.6729	0.5696	0.4295
LSTM	0.6729	0.5696	0.4251

Table (2): the performances of the models in the prediction of the total UPDRS scores

From table(2), it can be seen that the best performance in the prediction of the total UPDRS score was shown by the VotingRegressor as the model achieved the highest value for R-squared score and the lowest values for RMSE and MAE, which were the best scores compared to the GRU and LSTM.

Model	**R^2 Score**	**RMSE**	**Mean Absolute Error (MAE)**
VotingRegressor	0.7921	0.4481	0.3295
LSTM	0.6442	0.5862	0.4185
GRU	0.7298	0.5108	0.3761

Table (3): the performances of the models in the prediction of the Motor examination scores

From the table(3), it can be seen that the best performance in the prediction of the Motor examination score was shown by the VotingRegressor as it achieved the best values for R-Squared score, RMSE and MAE. It was seen that the R-Squared score was the highest for the VotingRegressor while the RMSE and MAE were the lowest compared to the LSTM and GRU, which meant that the performance was better.

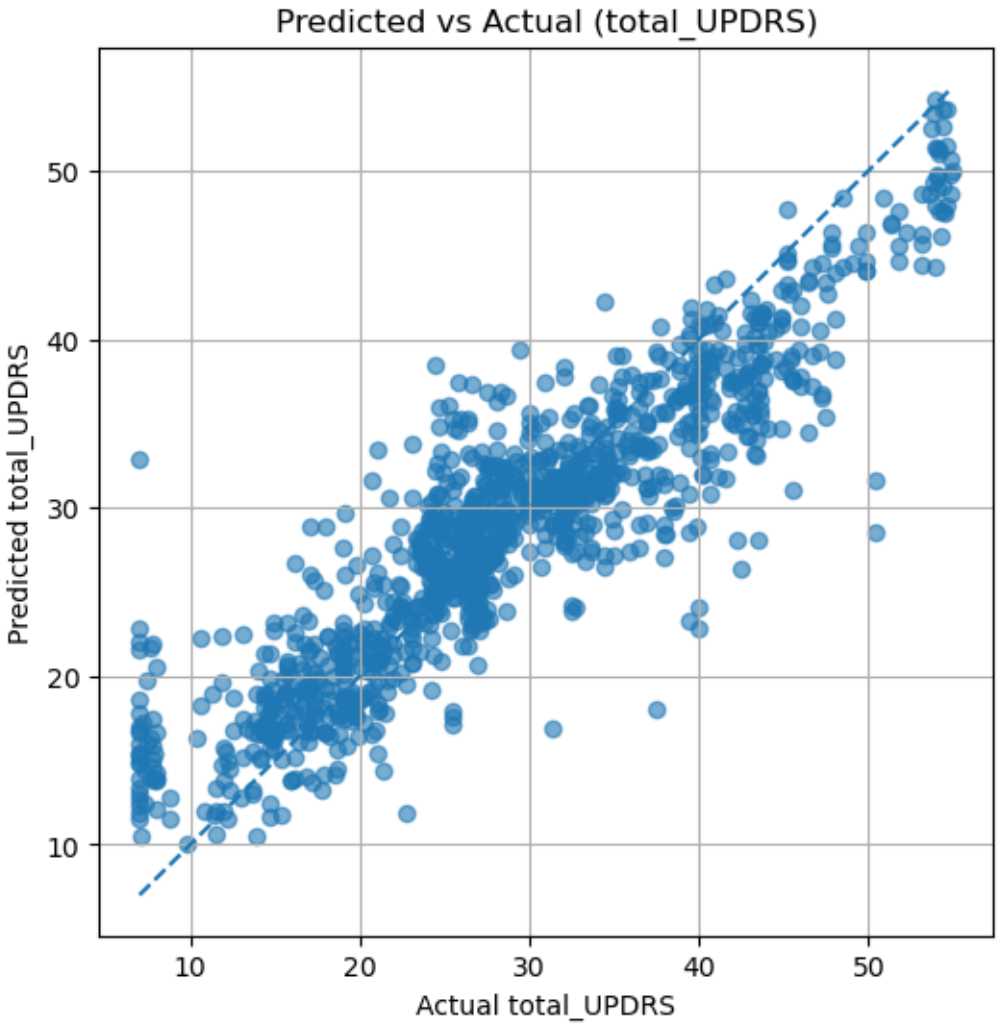


Figure (4): predicted vs actual performance for total UPDRS score by Voting regressor

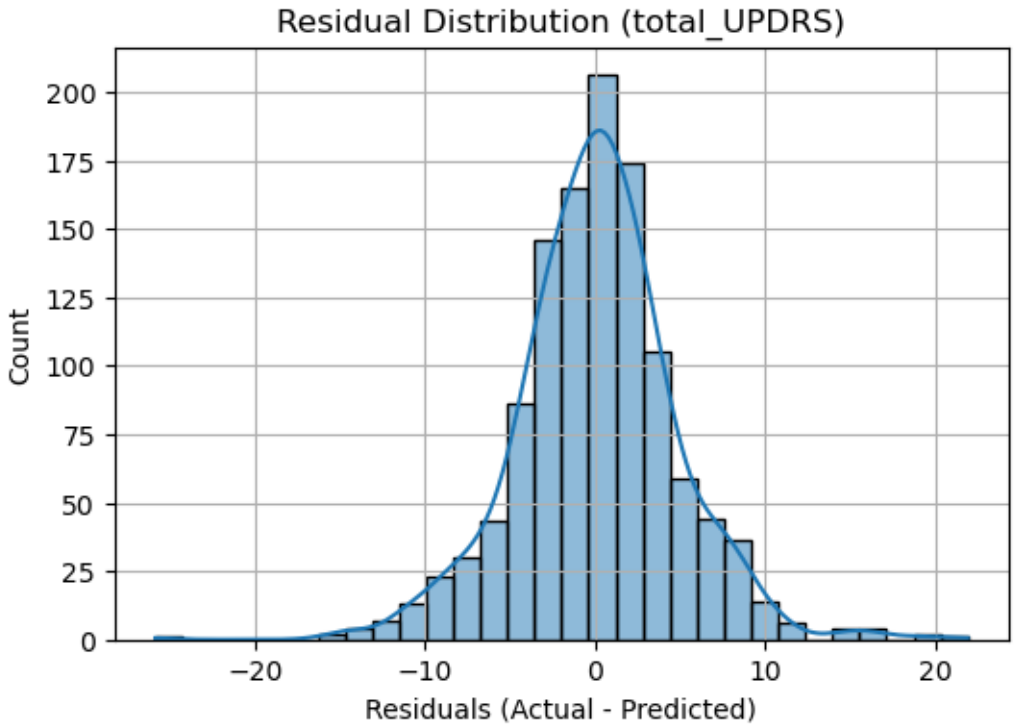


Figure (5): residual distribution in the prediction of total UPDRS score

Figure(4) and (5), shows the performance of the Voting regressor in the prediction of total UPDRS score.

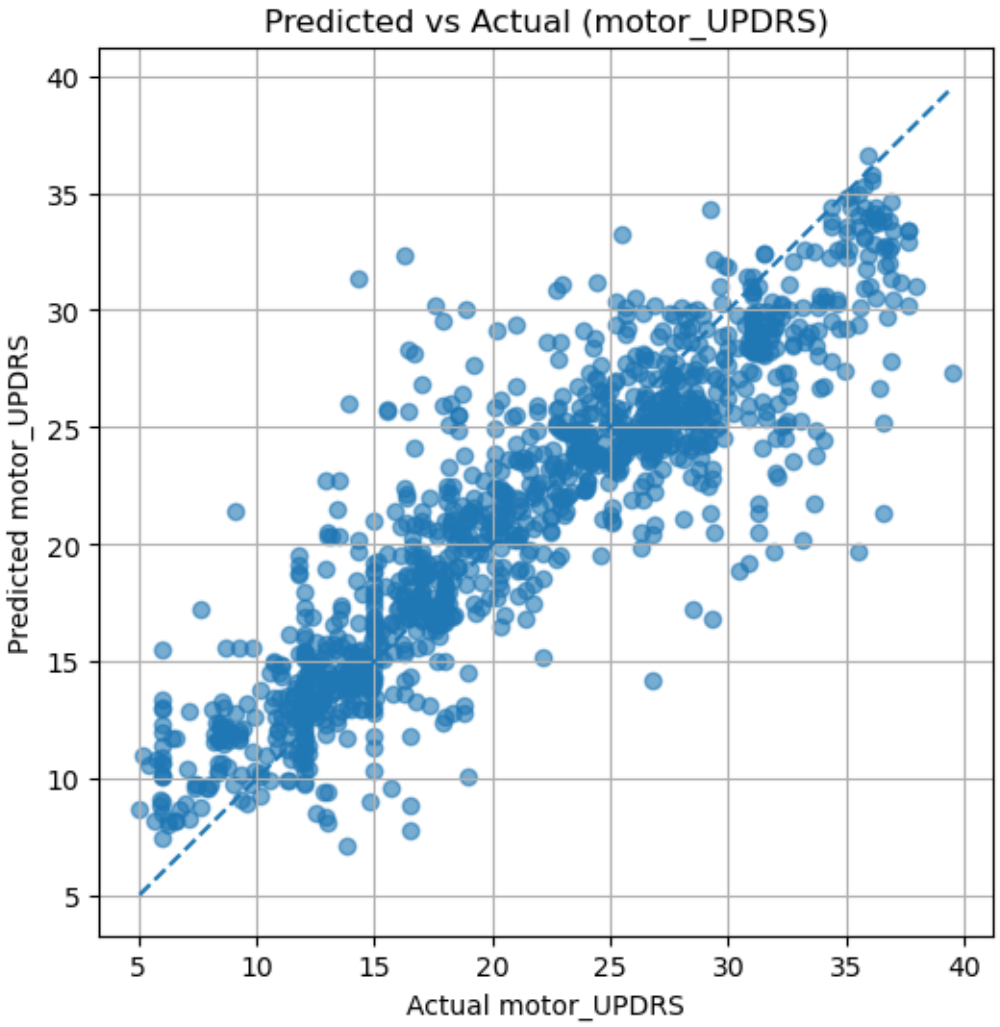


Figure (6): predicted vs actual performance for motor UPDRS score by Voting regressor

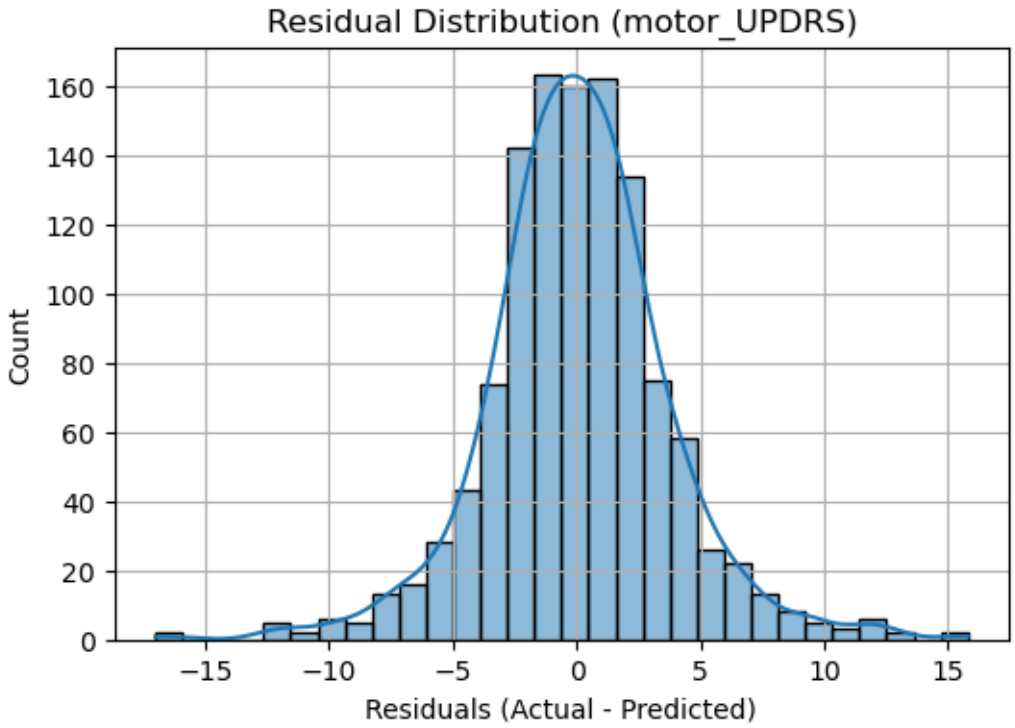


Figure (7): residual distribution in the prediction of total UPDRS score

Figure(6) and (7), shows the performance of the Voting regressor in the prediction of motor UPDRS score.

5.2 Interpretation of model prediction

The VotingRegressor was interpreted both for the prediction of the total UPDRS score and the Motor examination score. The VotingRegressor model was chosen as the model to be interpreted because it was the model that showed the best performance in both the prediction of the total UPDRS score and the Motor examination score, so the SHAP method was used and both predictions made by the VotingRegressor was interpreted.

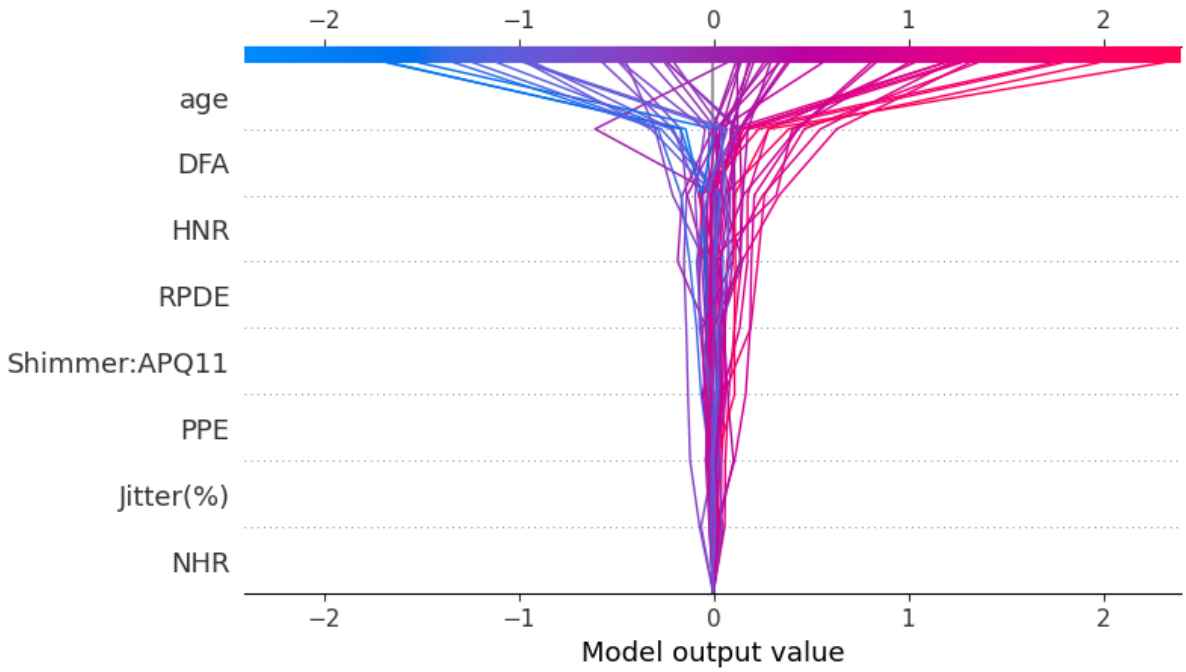


Figure (8): the feature importance associated with each features interpreted for the prediction of total UPDRS score

Figure(8) shows the importance associated with each features in the data related to UPDRS scores in the prediction of total UPDRS scores by the Voting regressor. It can be seen that the feature ‘Age’ has a strong impact on the prediction of total UPDRS score by the Voting regressor as it had a concentration of points on the positive side of the X-axis. The feature ‘DFA’ had a mixed impact on the prediction of total UPDRS score as it showed a mixed distribution, in which points were scattered around both sides of the X-axis. This shows that ‘DFA’ can have differing impacts on the prediction by the Voting regressor, showcasing a complex relationship where individuals with high ‘DFA’ might have both higher and lower UPDRS scores, depending on other factors. The feature has some positive SHAP values, meaning that higher shimmer levels correspond to the better prediction of the total UPDRS scores. This is in-line with the insight from studies that acoustic features hugely affect motor function in people suffering from Parkinson's disease. It was seen that the features, HNR, RPDE, PPE, NHR, and Jitter(%), showed a range of influences in the prediction of the total UPDRS score. For example, while some features like HNR and RPDE has positive impacts on the prediction of total UPDRS score, other features have lesser impact on the prediction, showing that they have differing impacts on the prediction of total UPDRS score.

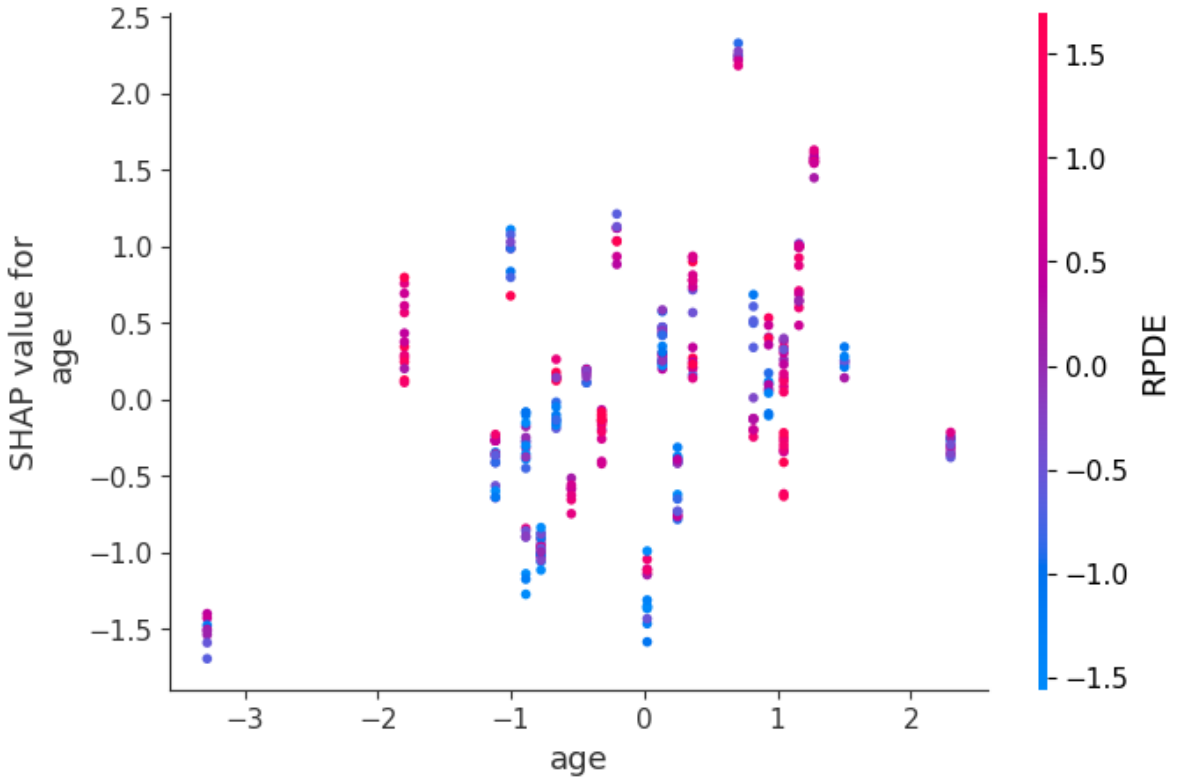


Figure (9): SHAP value for age in the prediction of total UPDRS scores

From figure(9), it can be seen that there was no distinct linear correlation between age and SHAP values in the scatter plot generated based on the interpretation of the prediction of total UPDRS score by the Voting regressor. From the scatter plot it was seen that both higher and lower SHAP values are present across various ages, which can mean that the influence of age might be moderated by the voice signal features and other demographic factors. This might mean that older people who show specific characteristics in their voice signals may have different implications for their UPDRS scores compared to their younger counterparts.

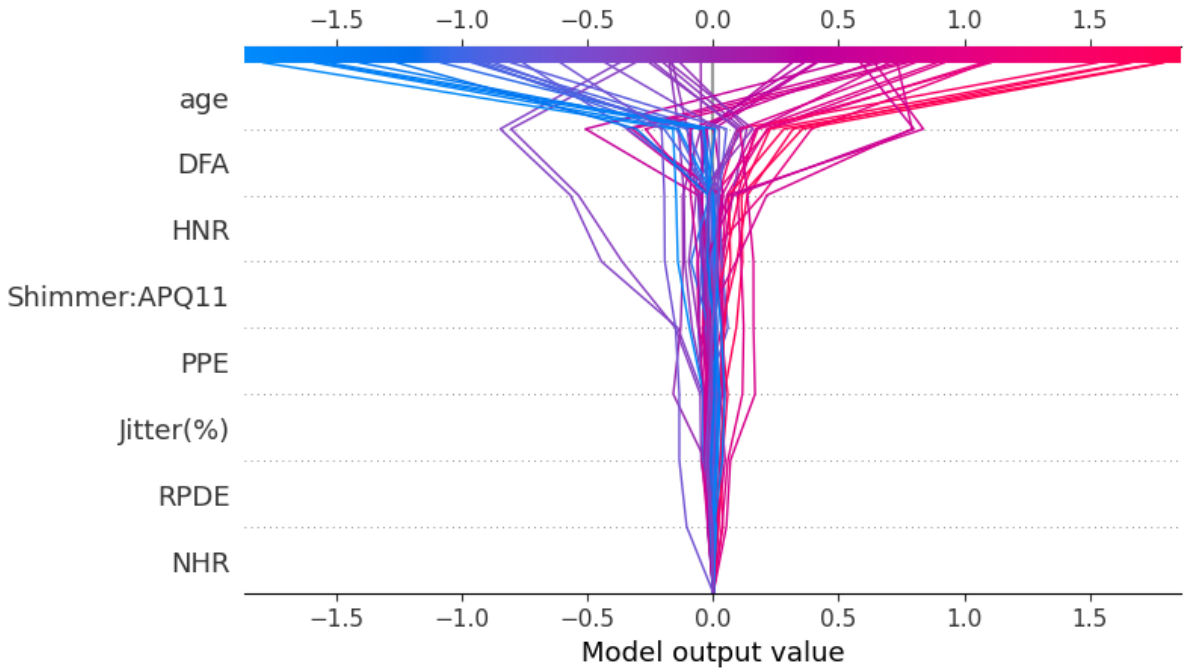


Figure (10): feature importance associated with the features in the prediction of motor examination score

Figure(10), shows the importance associated with the different features in the prediction of motor examination score. The feature 'age' has importance in the prediction of motor examination scores, however, the concentration of SHAP values around zero shows that it has a mixed contributions to the predictions, i.e., while 'age' is a relevant feature, the impact of the feature differs across observations, which shows that the feature is having a complex relationship in the prediction of motor examination scores. The feature 'DFA' showed positive SHAP values, which indicates that higher DFA scores are related to better prediction of motor examination scores. This gain shows that effect of motor issues affecting vocal abilities. The feature 'Shimmer:APQ11' has both positive and negative SHAP values, which indicates that it has a complex relationship with the prediction of motor examination scores. Features like HNR (Harmonics-to-Noise Ratio), RPDE (Recurrence Period Density Entropy), PPE (Pitch Period Entropy), NHR (Noise-to-Harmonics Ratio), and Jitter(%), have differing effects on the prediction of motor examination scores as the SHAP values varied showing the different impacts on prediction.

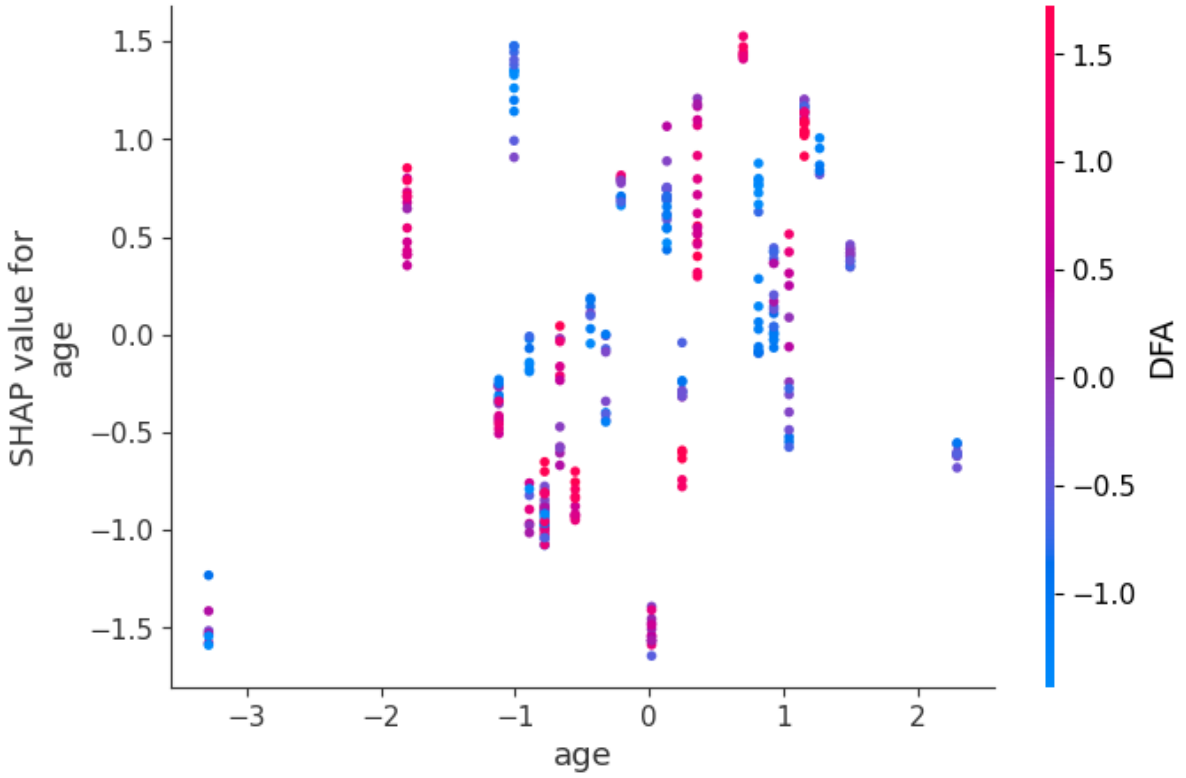


Figure (11): the SHAP values associated with 'age' in the prediction of motor examination scores

Figure(11), shows the visual plot generated by SHAP for representing SHAP values associated with the feature 'age'. From the plot it was seen that the SHAP values for 'age' range from approximately, 1.5 to 1.5, showing that 'age' has a diverse impact on the predicted Motor examination scores. It was seen that the feature 'DFA' had positive impact on the feature 'age', this indicates that with some voice characteristics, like higher DFA, can show different patterns in the prediction of Motor examination scores based on 'age'. The plot shows that 'age' can affect the prediction of Motor examination scores both positively and negatively. Higher SHAP values for age can indicate that older people may have high predicted motor examination scores, especially when the values of 'DFA' in the data was high. The plot shows that even though 'age' is a significant factor, its influence is lowered voice signal measure defined by the 'DFA'.

5.3 Prototype Desktop app for the prediction of total UPDRS score and UPDRS-III motor scores

The prototype desktop app was built using The 'Tkinter' library in Python. The trained Voting regressor was integrated with the desktop app, so that the prediction of total UPDRS score and UPDRS-III motor scores, were done. The prediction here was the based on the features associated with UPDRS score. These values associated with the features had to be

given as input to the trained Voting regressor for making it predict the total UPDRS score and UPDRS-III motor scores. So, the Voting regressor model needed to receive the inputs associated with features related to UPDRS data. For achieving this the interface of the prototype desktop app was designed with input boxes. The input boxes for providing the data related to features of UPDRS score. The input boxes corresponded to the entire features in the dataset, however, the values associated with features that were used for training the model, were only selected from the inputs, i.e., the inputs associated with, 'age', 'Jitter(%)', 'Shimmer:APQ11', 'NHR', 'HNR', 'RPDE', 'DFA', and 'PPE', were chosen from the entire value of features given as the input. Both the total UPDRS score and UPDRS-III motor scores needed to be predicted and displayed to the users.

The prediction needed to come for the Voting classfier and for performing the prediction a button was added to the interface of the prototype desktop app. The button was programmed in a way that the data related to the features of UPDRS given as input in the different input fields, were fetched and passed to the Voting regressor, based on the input the Voting regressor makes predictions and the desktop app was designed in a way that the total UPDRS score and UPDRS-III motor scores, were displayed on the interface of the desktop app prototype. The interface also has fields were the values of the features can be provided as comma-separated values, which are also used for prediction based on the click of the button. The interface of the desktop app prototype built in the study is given in figure(12).

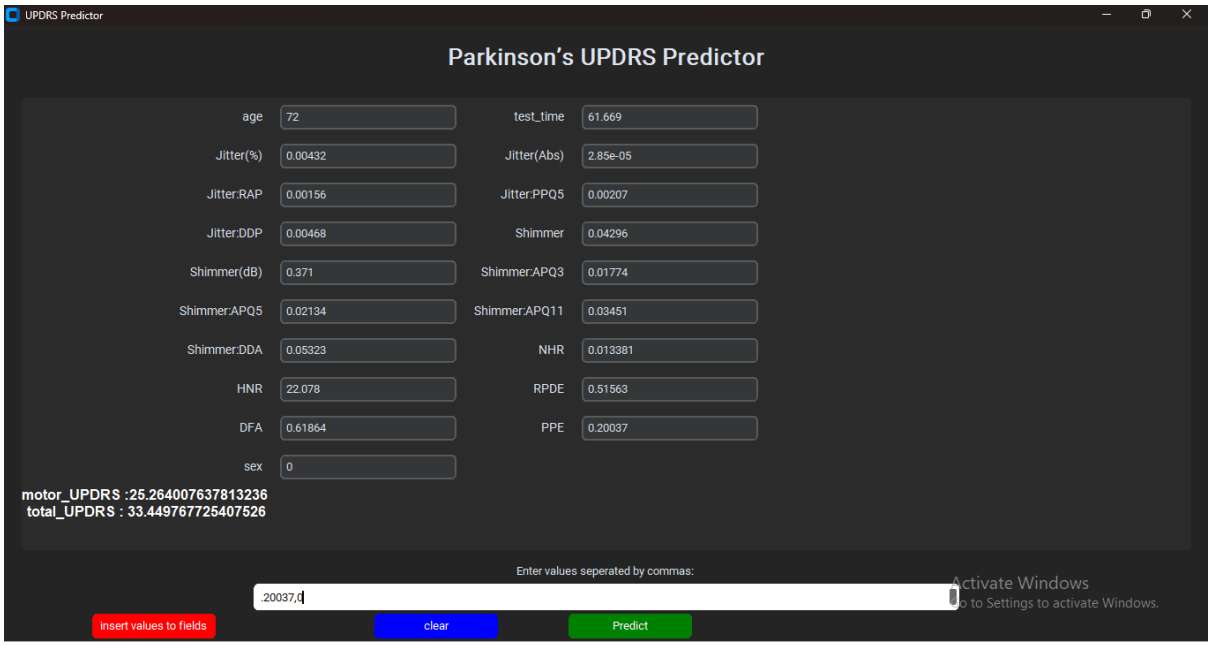


Figure (12): The prototype deskt6op app for the prediction of UPDRS scores

This prototype was a demo tool and cannot be used for clinical use.

6 Discussion and Evaluation

6.1 Discussion

Regression models that predicted the total UPDRS scores and the Motor examination or UPDRS-III scores, were built successfully in this study. From the evaluation of the performance of the trained regression models it was seen that the best performance was achieved by the Voting regressor, for both the prediction tasks. The performance of Voting regressor in predicting both total UPDRS scores and UPDRS-III scores can be compared with the models used for predicting these scores, in existing studies.

Model	**R-squared score**
Hamzehei et al.(2023)	0.904410
Shahid and Singh(2020)	0.956
Voting regressor	0.9883

Table (4): the comparison of performances of existing studies and the Voting regressor in the prediction of total UPDRS scores

Model	**MAE**
Rehman et al.(2021)	6.29
Nilashi et al.(2022)	0.721
Hssayeni et al.(2021)	5. 95
Voting regressor	0.3295

Table (5): the comparison of performances of existing studies and the Voting regressor in the prediction of total UPDRS-III scores

In table(4), it can be seen that the performance of the Voting regressor built here was compared with the performances of the models proposed in existing studies, that focused on the prediction of UPDRS score. From table(4), it can be seen that the best R-squared value was shown by the Voting regressor built here, as it achieved the highest value for R-squared score. There was a study by Mehta, Asif, Hao, Bilal, Stefan, et al.(2021) that focused on the prediction of the total UPDRS scores, however, this study was based on video data. The Voting regressor was not used in any of the studies mentioned in table(4).

From table(5),it can be seen that the performance of the best performing model built, here , the Voting regressor in predicting the motor examination score, was compared with the performances of existing models that focused on the prediction of motor examination score. The comparison of the performances of the models was done based on the values of MAE associated with the models. The MAE value should be ideally low and from the table(5), it was seen that the best performance was shown by the Voting regressor built here as that model had the lowest MAE value, which meant that the prediction performance of the model was very good.

Even though it can be seen that the performance of the Voting regressor was better than the models in the existing literature considered here, based on the values of R-Squared score and MAE, a direct comparison cannot be done with the results of the study that was done here with the results of the existing literature, as the methods(machine learning algorithms) and data used in the study proposed here were different form the methods and data used in existing literature.

6.2 Evaluation

The objectives of the study were achieved by generating visual plots based on the data related to UPDRS scores from the dataset using Python libraries. The Voting regressor, LSTM and GRU were trained using the data related to UPDRS scores and the models were trained to predict the total UPDRS score and the UPDRS-III motor score. The SHAP XAI method was used for interpreting the prediction of total UPDRS score and the UPDRS-III motor score by the Voting regressor, and the features that have the most impact on the prediction of both scores, were found. The performance of the trained machine learning models were assessed by finding values for the metrics like, R-Squared, MAE and RMSE.

The data in the UPDRS dataset contained irrelevant and unwanted elements like unwanted feature and feature values present in different numerical scales. The data with these unwanted elements cannot be used for training the machine learning model to perform the prediction of UPDRS scores and UPDRS-III motor score. This difficulty in the study was solved by cleaning the data and pre-processing the data for transforming the data into a form apt for training the deep learning model to perform the prediction.

The limitation of the current study is that the performance of the Voting regressor may have been lowered as the base models used for the Voting regressor, used in this study were conventional machine learning models, however, just like, the LSTM and GRU models used in this study, deep learning models could have used as the base classes of the Voting regressor. This may have resulted in the performance of the total UPDRS prediction and the prediction of the Motor examination score, improving, even if the currently built model shows a good performance in the prediction of UPDRS prediction and the prediction of the Motor examination score. The LSTM and GRU were chosen in this study for handling continuous data, however, this data was converted to non-sequential or continuous data before the model was trained, however, the LSTM and GRU were strong predictors.

The threats to validity of this study include, the small number of data samples that were used in the study for building the machine learning models, and the generalizability of the model built in the study proposed here.

7 Conclusion and Future enhancements

The aim of the study proposed here was to build models that predicted the total UPDRS score and the Motor examination score or the UPDRS-III motor score. The prediction was done using machine learning and deep learning models. The machine learning and deep learning models used in the study included the, Voting regressor, LSTM and GRU. The base models used in the Voting regressor were, the RF regressor, DT regressor and SVR. The Parkinson's Telemonitoring Dataset which contained the data related to UPDRS scores was used in the study for training the models for predicting both total UPDRS score and UPDRS-III motor score. The data related to UPDRS scores in the dataset was used for generating visual plots which were analysed for understanding the hidden patterns associated with UPDRS score and the different features that were defined for the data in the data in the dataset. The data related to UPDRS score was cleaned and transformed to be made apt for training the machine learning and deep learning models. The models were trained successfully and all three models were able to predict total UPDRS score and UPDRS-III motor score, successfully. The performance of the models were evaluated for both the prediction of total UPDRS score and UPDRS-III motor score. The performance of the models were evaluated based on the values of performance metrics like, R-squared score, RMSE and MAE.

From the values of the performance metrics it was seen that the best performance in the study was shown by the Voting regressor as it achieved the best values for all the performance metrics that were compute, here. The Voting regressor showed the best performance in both the prediction of total UPDRS score and UPDRS-III motor score. The predictions by the Voting regressor was interpreted using SHAP and the features that were important for the prediction of the total UPDRS score and UPDRS-III motor score, were found, these features included age and acoustic features, which reveal that the vocal functions of a person are affected by the problems associated with the motor functions of a person. The tool used for building the total UPDRS score and UPDRS-III motor score prediction model was Python. The Voting regressor which showed the best performance was integrated to a prototype desktop app, this desktop app showcased the prediction of total UPDRS score and UPDRS-III motor score.

The research questions of the study were :

- What are the most significant acoustic features that correlate with motor and total UPDRS scores in Parkinson's patients ?

  The most significant acoustic feature that is related to motor and total UPDRS scores in Parkinson's patients is 'Shimmer'. The interpretation of the prediction made by the models revealed that 'Shimmer' is affected by any issues with motor functions. The impact that the feature 'Shimmer' has on the prediction of motor and total UPDRS scores can be seen in the plot generated by SHAP in figures(8) and (10).

- Which regression model Voting Regressor, Long Short Term Memory(LSTM), and Gated Recurrent Unit(GRU) provides the most accurate and interpretable prediction of UPDRS scores, and what are the features that contribute to the prediction of score for Motor Examination and the sum of the scores of the 4 parts of UPDRS?

  It was seen that the regression model that showed the best performance in the prediction of score for Motor Examination and the sum of the scores of the 4 parts of UPDRS, was the Voting regressor as it achieved the best values for the different performance metrics that were computed, here for assessing the performances of the models. The comparison between the performances of the different regression models was shown in tables(4) and (5). From the interpretation of the prediction made by the Voting Regressor for both total UPDRS score and Motor examination score, it was seen that the features which had the best impact on the prediction of total UPDRS score and the prediction of UPDRS-III motor score, were features like, 'age' and 'Shimmer', which can be seen in figures(8) and (10).

In the future, several changes can be made to the study that was done to improve the study. The base regressors of the Voting regressor in the current study were conventional machine learning models. These conventional machine learning models can be replaced with deep learning models in future studies, this replacement of the conventional machine learning models with deep learning models may improve the performance of the Voting regressor, in both kinds of prediction that the model performing. In the current study it was seen that the best performance was shown by the Voting regressor, an ensemble model, in future studies, other ensemble models can be considered like, Hybrid models or Stacking classifiers. The use of the ensemble models in the study can help in building models that may show better performances in the prediction of total UPDRS score and UPDRS-III motor score than the Voting regressor, which currently achieved the best performance in the study done here. In future studies, true sequence modelling can be considered, in the current study, the sequence data was converted to non-sequence data, which can avoided in future studies.

References

Agarwal, A. *et al.* (2022) 'Hierarchical Shrinkage: improving the accuracy and interpretability of tree-based methods,' *arXiv (Cornell University)* [Preprint]. https://doi.org/10.48550/arxiv.2202.00858.

Al-Selwi, N.S.M. *et al.* (2023) 'LSTM inefficiency in Long-Term Dependencies Regression Problems,' *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 30(3), pp. 16–31. https://doi.org/10.37934/araset.30.3.1631.

Aman, N. and Chhillar, R.S. (2025) 'Machine learning analysis of the Telemonitoring Voice Dataset for enhanced Parkinson's disease severity prediction,' *Research Square (Research Square)* [Preprint]. https://doi.org/10.21203/rs.3.rs-6186623/v1.

Bouça-Machado, R. *et al.* (2022) 'Measurement tools to assess activities of daily living in patients with Parkinson's disease: A systematic review,' *Frontiers in Neuroscience*, 16. https://doi.org/10.3389/fnins.2022.945398.

Byeon, H. (2020) 'Development of a depression in Parkinson's disease prediction model using machine learning,' *World Journal of Psychiatry*, 10(10), pp. 234–244. https://doi.org/10.5498/wjp.v10.i10.234.

Can, T., Krishnamurthy, K. and Schwab, D.J. (2020) *Gating creates slow modes and controls phase-space complexity in GRUs and LSTMs*, *Proceedings of Machine Learning Research*, pp. 476–511. https://proceedings.mlr.press/v107/can20a/can20a.pdf.

Chen, J. and Liu, J. (2022) 'Management of nonmotor symptoms in Parkinson Disease,' *Journal of Innovations in Medical Research*, 1(5), pp. 18–33. https://doi.org/10.56397/jimr/2022.12.03.

Chen, S. and Luc, N.M. (2022) 'RRMSE Voting Regressor: A weighting function based improvement to ensemble regression,' *arXiv (Cornell University)* [Preprint]. https://doi.org/10.48550/arxiv.2207.04837.

Corsini, A., Yang, S.J. and Apruzzese, G. (2021) *On the Evaluation of Sequential Machine Learning for Network Intrusion Detection*, pp. 1–10. https://doi.org/10.1145/3465481.3470065.

Dadu, A. *et al.* (2022) 'Identification and prediction of Parkinson's disease subtypes and progression using machine learning in two cohorts,' *Npj Parkinson S Disease*, 8(1). https://doi.org/10.1038/s41531-022-00439-z.

Dalal, S., Seth, B., Radulescu, M., Secara, C. and Tolea, C. (2022). Predicting Fraud in Financial Payment Services through Optimized Hyper-Parameter-Tuned XGBoost Model. *Mathematics*, 10(24), p.4679. doi:https://doi.org/10.3390/math10244679.

Farzanehfar, P., Woodrow, H. and Horne, M. (2022) 'Sensor measurements can characterize fluctuations and wearing off in Parkinson's disease and guide therapy to improve motor, non-motor and quality of life scores,' *Frontiers in Aging Neuroscience*, 14. https://doi.org/10.3389/fnagi.2022.852992.

Feng, R. *et al.* (2020) 'Projected minimal gated recurrent unit for speech recognition,' *IEEE Access*, 8, pp. 215192–215201. https://doi.org/10.1109/access.2020.3041477.

Gaudel, R. *et al.* (2022) 's-LIME: Reconciling Locality and Fidelity in Linear Explanations,' in *Lecture notes in computer science*, pp. 102–114. https://doi.org/10.1007/978-3-031-01333-1_9.

Grobe-Einsler, M. *et al.* (2023) 'SARAspeech—Feasibility of automated assessment of ataxic speech disturbance,' *Npj Digital Medicine*, 6(1). https://doi.org/10.1038/s41746-023-00787-x.

Guo, W.-H. *et al.* (2025) 'Early detection of Parkinson's disease: Machine learning-based prediction of UPDRS Part III scores in de novo patients using smartphone assessments,' *Journal of Parkinson S Disease* [Preprint]. https://doi.org/10.1177/1877718x251359494.

Hachimi, C.E. *et al.* (2022) 'Data Science Toolkit: An all-in-one python library to help researchers and practitioners in implementing data science-related algorithms with less effort,' *Software Impacts*, 12, p. 100240. https://doi.org/10.1016/j.simpa.2022.100240.

Hamzehei, S. *et al.* (2023) 'Predicting the total Unified Parkinson's Disease Rating Scale (UPDRS) based on ML techniques and cloud-based update,' *Journal of Cloud Computing Advances Systems and Applications*, 12(1). https://doi.org/10.1186/s13677-022-00388-1.

Hoseinipalangi, Z. *et al.* (2023) 'Systematic review and meta-analysis of the quality-of-life of patients with Parkinson's disease,' *Eastern Mediterranean Health Journal*, 29(1), pp. 63–70. https://doi.org/10.26719/emhj.23.013.

Hssayeni, M.D. *et al.* (2021) 'Ensemble deep model for continuous estimation of Unified Parkinson's Disease Rating Scale III,' *BioMedical Engineering OnLine*, 20(1). https://doi.org/10.1186/s12938-021-00872-w.

Jovovic, I. *et al.* (2023) *Disease Prediction Using Machine Learning Algorithms*, pp. 1–4. https://doi.org/10.1109/it57431.2023.10078464

K, A. (2023) 'Regression modeling approaches for red wine quality prediction: individual and ensemble,' *International Journal for Research in Applied Science and Engineering Technology*, 11(6), pp. 3621–3627. https://doi.org/10.22214/ijraset.2023.54363.

Kannengiesser, U. and Gero, J.S. (2023) 'MODELLING THE DESIGN OF MODELS: AN EXAMPLE USING CRISP-DM,' *Proceedings of the Design Society*, 3, pp. 2705–2714. https://doi.org/10.1017/pds.2023.271.

Kumar, B.K., Bilgaiyan, S. and Mishra, B.S.P. (2023) 'Software Effort Estimation through Ensembling of Base Models in Machine Learning using a Voting Estimator,' *International*

Journal of Advanced Computer Science and Applications, 14(2). https://doi.org/10.14569/ijacsa.2023.0140222.

Laganas, C. *et al.* (2021) 'Parkinson's disease detection based on running speech data from phone calls,' *IEEE Transactions on Biomedical Engineering*, 69(5), pp. 1573–1584. https://doi.org/10.1109/tbme.2021.3116935.

Luo, Y. *et al.* (2025) 'Global, regional, national epidemiology and trends of Parkinson's disease from 1990 to 2021: findings from the Global Burden of Disease Study 2021,' *Frontiers in Aging Neuroscience*, 16. https://doi.org/10.3389/fnagi.2024.1498756.

Mayer, M. (2022) 'SHAP for additively modeled features in a boosted trees model,' *arXiv (Cornell University)* [Preprint]. https://doi.org/10.48550/arxiv.2207.14490.

mean_absolute_error (2025). https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html.

mean_squared_error (2025). https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html.

Mehta, D. *et al.* (2021) 'Towards Automated and Marker-less Parkinson Disease Assessment: Predicting UPDRS Scores using Sit-stand videos,' *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3836–3844. https://doi.org/10.1109/cvprw53098.2021.00425.

Mehta, D., Asif, U., Hao, T., Bilal, E., Stefan, V.C., et al. (2021) Towards Automated and Marker-less Parkinson Disease Assessment: Predicting UPDRS Scores using Sit-stand videos. https://arxiv.org/abs/2104.04650.

Misheva, B.H. *et al.* (2022) *eXplainable AI in Credit Risk Management*. https://doi.org/10.24818/ida-cl/2022.35.

Motin, M.A. *et al.* (2022) 'Parkinson's Disease Detection Using Smartphone Recorded Phonemes in Real World Conditions,' *IEEE Access*, 10, pp. 97600–97609. https://doi.org/10.1109/access.2022.3203973.

Nilashi, M. *et al.* (2022) 'Predicting Parkinson's disease progression: Evaluation of ensemble methods in machine learning,' *Journal of Healthcare Engineering*, 2022, pp. 1–17. https://doi.org/10.1155/2022/2793361.

Park, H. *et al.* (2025) 'Using machine learning to identify Parkinson's disease severity subtypes with multimodal data,' *Journal of NeuroEngineering and Rehabilitation*, 22(1). https://doi.org/10.1186/s12984-025-01648-2.

r2_score (2025). https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html.

Rehman, R.Z.U. *et al.* (2021) 'Predicting the Progression of Parkinson's Disease MDS-UPDRS-III Motor Severity Score from Gait Data using Deep Learning,' *2021 43rd Annual International Conference of the IEEE Engineering in Medicine &Amp; Biology Society (EMBC)*, pp. 249–252. https://doi.org/10.1109/embc46164.2021.9630769.

Roozbeh, M. *et al.* (2023) 'Generalized support vector regression and symmetry functional regression approaches to model the High-Dimensional data,' *Symmetry*, 15(6), p. 1262. https://doi.org/10.3390/sym15061262.

Sankalpa, C., Kittipiyakul, S. and Laitrakun, S. (2022) 'Forecasting Short-Term electricity load using validated ensemble learning,' *Energies*, 15(22), p. 8567. https://doi.org/10.3390/en15228567.

Shahid, A.H. and Singh, M.P. (2020) 'A deep learning approach for prediction of Parkinson's disease progression,' *Biomedical Engineering Letters*, 10(2), pp. 227–239. https://doi.org/10.1007/s13534-020-00156-7.

StandardScaler (2025). https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

TensorFlow. (2025). *tf.keras.layers.GRU | TensorFlow v2.13.0*. [online] Available at: https://www.tensorflow.org/api_docs/python/tf/keras/layers/GRU.

Tian, R. and Parikh, A.P. (2022) 'Amos: An Adam-style Optimizer with Adaptive Weight Decay towards Model-Oriented Scale,' *arXiv (Cornell University)* [Preprint]. https://doi.org/10.48550/arxiv.2210.11693.

Tuan, Y.-L., Chiu, Z.-Y. and Wang, W.Y. (2022) 'Dynamic latent separation for deep learning,' *arXiv (Cornell University)* [Preprint]. https://doi.org/10.48550/arxiv.2210.03728.

UCI Machine Learning Repository (2025). https://archive.ics.uci.edu/dataset/189/parkinsons+telemonitoring.

Vignoud, G. *et al.* (2022) 'Video-Based Automated Assessment of Movement Parameters Consistent with MDS-UPDRS III in Parkinson's Disease,' *Journal of Parkinson S Disease*, 12(7), pp. 2211–2222. https://doi.org/10.3233/jpd-223445.

Wang, S.-S. *et al.* (2022) 'Continuous speech for improved learning Pathological voice disorders,' *IEEE Open Journal of Engineering in Medicine and Biology*, 3, pp. 25–33. https://doi.org/10.1109/ojemb.2022.3151233.

Wang, X. et al. (2022) LightSeq2: Accelerated Training for Transformer-Based Models on GPUs, pp. 1–14. https://doi.org/10.1109/sc41404.2022.00043.

Xu, H. *et al.* (2025) 'Non-invasive detection of Parkinson's disease based on speech analysis and interpretable machine learning,' *Frontiers in Aging Neuroscience*, 17, p. 1586273. https://doi.org/10.3389/fnagi.2025.1586273.

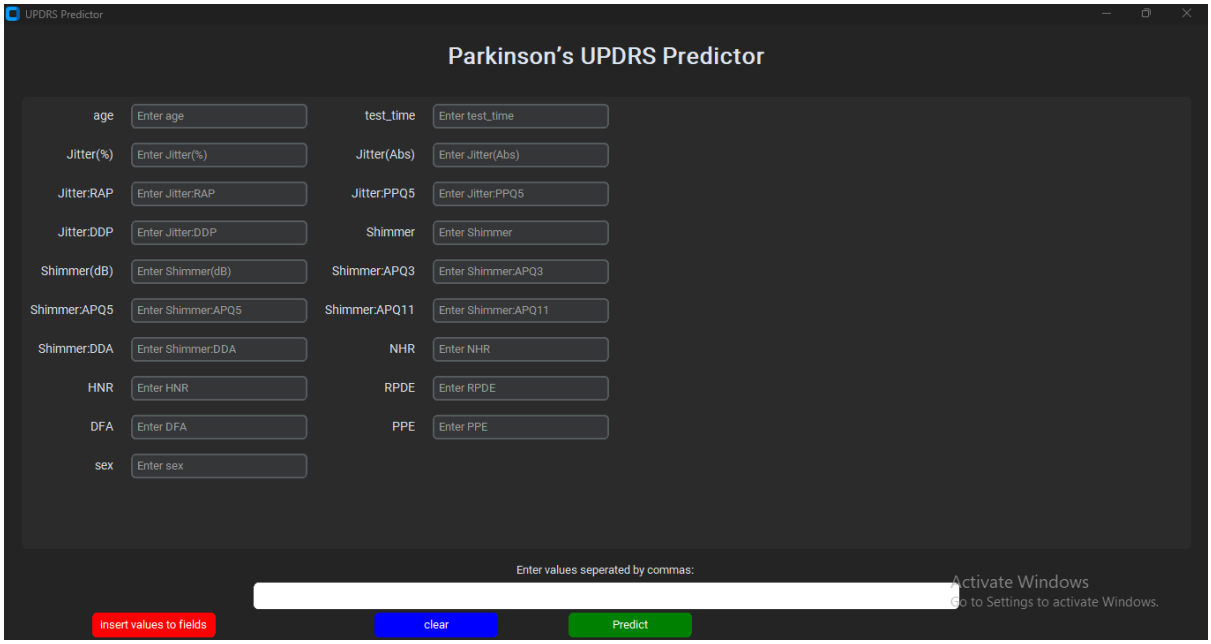
Yann Ling Goh, Chern Long Ng and Ling, R. (2023). Productivity Prediction of Garment Employees using Multiple Linear Regression. *International Journal of Advanced Natural Sciences and Engineering Researches*, [online] 7(4), pp.163–168. doi:https://doi.org/10.59287/ijanser.644.

Zhang, Y. *et al.* (2021) 'Mean square cross error: performance analysis and applications in non-Gaussian signal processing,' *EURASIP Journal on Advances in Signal Processing*, 2021(1). https://doi.org/10.1186/s13634-021-00733-7.

Appendices

Appendix A

The interface of the prototype desktop app.



Appendix B

The code for building the models for prediction.

```
from sklearn.ensemble import RandomForestRegressor, VotingRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.svm import SVR
from sklearn.metrics import r2_score, mean_absolute_error, root_mean_squared_error

# Initialize models
rf = RandomForestRegressor(n_estimators=100, random_state=42)
dt = DecisionTreeRegressor(random_state=42)
svm = SVR(kernel='rbf')

# Create Voting Regressor (ensemble of RF, DT, and SVM)
voting = VotingRegressor([('rf', rf), ('dt', dt), ('svm', svm)])

# Fit the model
voting.fit(X_train, y_train.ravel())  # converting to 1D

# Predict
y_pred_scaled = voting.predict(X_test)
y_pred_voting = y_pred_scaled.reshape(-1, 1)

# Evaluation
print("VotingRegressor: Result")
print("R2 Score:", r2_score(y_test, y_pred_voting))
print("RMSE:", root_mean_squared_error(y_test, y_pred_voting))
print("Mean Absolute Error:", mean_absolute_error(y_test, y_pred_voting))
```