Prediction of Unified Parkinson’s Disease Rating Scale (UPDRS) with Voice Measures

Group 2_Parkinsons: Chin Jun Hong, Lee Kah Win, Nur Aquilah Zulkifli, Zhao Dan

Abstract- The aim of this paper is to build a predictive model that can accurately predict the total UPDRS scores with at least 0.9 R Squared based on the voice measures. We utilized a dataset that recorded voice measures from 42 subjects with Parkinson Diseases, which consist of 5874 observations and 22 attributes. After performing data exploration, we noticed voice measures have a nonlinear relationship with UPDRS scores. Therefore, 3 nonlinear models are deployed, which are Support Vector Machine, Decision Tree Regressor, and Random Forest Regressor. After running all the models, the random forest regression model can best predict UPDRS based on voice measures (RQ2). This is because it has the highest R Squared (0.912), and has the lowest RMSE (3.39) and MAE (2.38) on the testing set. Besides, age is the highest feature importance in the prediction of UPDRS, which means that it is the most important predictor in the UPDRS score.

Index Terms - telemonitoring, Parkinson’s Disease, Support Vector Machine (SVM), Decision Tree Regressor (DTR) and Random Forest Regressor (RFR)

I. INTRODUCTION

PARKINSON disease (PD) is the second most common neurodegenerative disorder after Alzheimer‟s [1] and it rises steeply over the age of 50 [2]. Almost 10 million of the world population are living with Parkinson’s disease. To detect the severity of Parkinson’s disease, physical examination in a clinic is needed. However, these steps are often very costly [3] to the health system and time-consuming for the patients. It is very inconvenient for the patients with severe immobility to go to the clinic to determine their Unified Parkinson's Disease Rating Scale (UPDRS). And it is even difficult to check routinely. Therefore, the traditional method is very limited especially during the covid-19 pandemic. There is a new way called telemonitoring to record UPDRS scores rapidly and remotely with clinically useful accuracy about 5% prediction error by characterizing speech with signal processing algorithms, and statistically map these algorithms to UPDRS [4]. Speech signal processing algorithms offer an objective, potentially reliable method for assessing general voice disorders. In the context of PD, speech signals have been used to separate PWP and healthy controls (people with no PD symptoms) with very encouraging results [5]. The research questions in this project are as follows:

RQ1: Is there any relationship between UPDRS based on voice measures?
RQ2: Which regression model can best predict UPDRS based on voice measures?
RQ3: Which attribute has the highest feature importance in the prediction of UPDRS?

The aim of the project is to to build a predictive model that can accurately predict the total UPDRS scores with at least 0.9 R Squared based on the voice measures.

II. BACKGROUND AND LITERATURE REVIEW

The UPDRS was originally developed in the 1980s [6], which has become the most widely used clinical rating scale for PD [7] and the gold standard of PD clinical metric [3]. There are quite a number of papers that have already examined the effectiveness of voice measures in PD telemonitoring [8] [9] [10] [11]. Table 1 below summarize the literature review of voice measures toward Total UPDRS using regression coefficients [9].

TABLE I
REGRESSION COEFFICIENTS [9]

Measures	Total UPDRS	Relationship
Shimmer(dB)	0.61 ± 0.34	Positive
HNR	-1.09 ± 0.03	Negative
Log-Jitter(%)	2.46 ± 0.82	Positive
Log-Jitter(Abs)	-5.53 ± 0.30	Negative
Log-Jitter:PPQ5	2.03 ± 0.68	Positive
Log-Shimmer:APQ5	-23.17 ± 0.65	Negative
Log-Shimmer:APQ11	24.94 ± 0.61	Positive
Log-NHR	1.79 ± 0.26	Positive
Log-HNR	31.19 ± 0.73	Positive

Besides, the models that we used in this project will be literature reviewed, which are Linear Regression (LR), Support Vector Machine (SVM), Decision Tree Regressor (DTR), and Random Forest Regressor (RFR). The Table 2 below summarize the literature review of using linear regression toward the target “Total UPDRS”.

TABLE 2
LINEAR REGRESSION

Papers	Data	Results
[12]	Features (6): Jitter(Abs), Shimmer, NHR, HNR, DFA, PPE	Total UPDRS MAE: 8.47±0.27 Motor UPDRS MAE: 6.80±0.17
[13]	Log-Transformed Features (13): Jitter group (5), Shimmer group (6), NHR, HNR	Total UPDRS MAE: 8.43 ± 0.24 Motor UPDRS MAE: 6.62 ± 0.15
[14]	Features (19): Age, Gender, Test time, Jitter group (5), Shimmer group (6), NHR, HNR, RPDE, DFA, PPE	Total UPDRS RMSE: 9.93 Motor UPDRS RMSE: 7.61

Moreover, the Table 3 below shows the literature review of SVM in the context of predicting Total UPDRS.

TABLE 3
SUPPORT VECTOR MACHINE

Papers	Data	Results
[15]	Features (19): Age, Gender, Test time, Jitter group (5), Shimmer group (6), NHR, HNR, RPDE, DFA, PPE	Total UPDRS RMSE: 7.49
[16]	PCA - Features (19): Age, Gender, Test time, Jitter group (5), Shimmer group (6), NHR, HNR, RPDE, DFA, PPE	Total UPDRS R2: 0.831
[17]	Features (16): Jitter group (5), Shimmer group (6), NHR, HNR, RPDE, DFA, PPE	Total UPDRS MAE: 7.02 MSE: 82.67, R: 0.53

Furthermore, the Table 4 below shows the literature review of Random Forest Regression the context of predicting Total UPDRS.

TABLE 4
RANDOM FOREST REGRESSOR

Papers	Results
[18]	Predicted Parkinson Disease's UPDRS with 0.89 R Squared using 5 Parent RF (500 Trees)
[19]	Predicted Parkinson Disease's MDS-UPDRS III with R Squared 0.79

Lastly, Table 5 below shows the literature review of Decision Tree Regressor in the context of predicting Total UPDRS.

TABLE 5
DECISION TREE REGRESSION

Papers	Data	Results
[16]	PCA - Features (19): Age, Gender, Test time, Jitter group (5), Shimmer group (6), NHR, HNR, RPDE, DFA, PPE	Total UPDRS R2: 0.803

So, by using machine learning modelling, it can help to determine the severity of PD by using the data collected from the patient itself and the result for total UPDRS scores can be easily obtained. Therefore, the main objective of this study is to build an effective and reliable predictive model that can accurately predict the total UPDRS scores based on the voice measures to identify the person's severity and progression of PD.

III. METHODOLOGY

This section describes the methodology that has been employed. The section is divided into several subsections namely dataset description, data treatment and data exploration.

A. Dataset Description

The dataset was collected from a self-administered, and non-invasive speech test which enables patients and medical staff to track PD symptom progression at home for a duration of 6 months in 2009 [3]. It was created by Athanasios Tsanas and Max Little of the University of Oxford, in collaboration with 10 medical centers in the US and Intel Corporation who developed the telemonitoring device to record the speech signals. The dataset contains 5874 instances with 22 attributes including demographics and voice measures of 42 patients. 65% of the patients are male and others are female. All of them age from 36 to 85 years old. The data is in CSV format, and each row of the CSV file contains an instance corresponding to one voice recording. Every patient contributes around 150 recordings in the dataset.

The target is identified. The name of it is "total_UPDRS" which stands for total unified PD rating scale. "total_UPDRS" is a continuous numerical attribute with a mean of 28.74. Other statistical measurements include range (47.99), standard deviation (10.98) and median (27.28). According to medical research, higher value of total UPDRS indicates more severe PD symptoms. As for the features, though they are of different data types, 90% of the features are numerical data and the rest are categorical data. The majority of the features are voice measures, and all of the 16 measures can be categorized in this study as follow: (1) Jitter group - five measures of variation in fundamental frequency, (2) Shimmer group - six measures of variation in amplitude, (3) NHR, HNR - two measures of ratio of noise to tonal components in the voice, (4) RPDE - a nonlinear dynamical complexity measure, (5) DFA - signal fractal scaling exponent, (6) PPE - a nonlinear measure of fundamental frequency variation.

B. Data Treatment

Two attributes in the dataset have numerous missing values namely “motor_UPDRS” and “NHR”. Attribute “motor_UPDRS” has more than 11% missing values, so it is unsuitable to be the target, the attribute is dropped, and the attribute “total_UPDRS” is selected as the target. Attribute “NHR” has about 7% missing values. Since the attribute “NHR” has right-skewed distribution, data imputation is performed by filling the missing values with the median of “NHR”. The outliers are detected by using several boxplots. Based on the boxplots, 15 out of 16 voice measures contain outliers which represent the measurement errors. The outliers are removed by using Interquartile (IQR) Rule. The instances with values fall outside the lower and upper bound of boxplots are removed. Besides, the instances with negative “test _time” values are also removed because the “test_time” value should be at least 0. After that, data binning is performed to divide the attribute “age” into 5 groups. Then, all the features are normalized in the range of 0 and 1.

C. Data Exploration

The dataset is analyzed to gain insight from it. The univariate analysis is performed on both numerical and categorical features. For numerical features, the central tendencies and the statistical dispersions are calculated as shown in Table 8. Besides, histograms are used to visualize the voice measures. 13 out of 16 voice measures have right-skewed distribution, the rest have normal distribution. Fo the right-skewed distribution attributes, they are transformed into normal distribution by using log transformation. For categorical features, the sample count of each category is calculated as shown in Table 8.

The bivariate analysis is performed to determine the relationship between the features and the target. The correlation between numerical features and the target are analyzed, and the top 10 correlations are recorded in Table 6. The highest correlation is only -0.184 which is very low. Scatterplots are used to examine it further. Linear line and non-linear line are fitted separately into each scatterplot as shown in Fig. 1. The results show that the non-linear line fits the data better, indicating that the features have non-linear relationship with the target. Furthermore, the correlation within Jitter and Shimmer group are high, which is at least 0.848, indicating that these features are conveying the very similar information. Therefore, to remove the redundancy, the attributes “Jitter(%)” and “Shimmer:APQ11” are selected from Jitter and Shimmer group respectively because they have the highest correlation with the target compared to others.

TABLE 6
TOP 10 CORRELATIONS BETWEEN FEATURES AND TARGET

Features	Correlation	Features	Correlations
1. Age	0.323	6. Shimmer(dB)	0.135
2. HNR	-0.184	7. DFD	-0.134
3. Shimmer: APQ11	0.173	8. Shimmer:APQ5	0.123
4. PPE	0.163	9. Jitter(%)	0.116
5. RPDE	0.153	10. Jitter:PPQ5	0.112

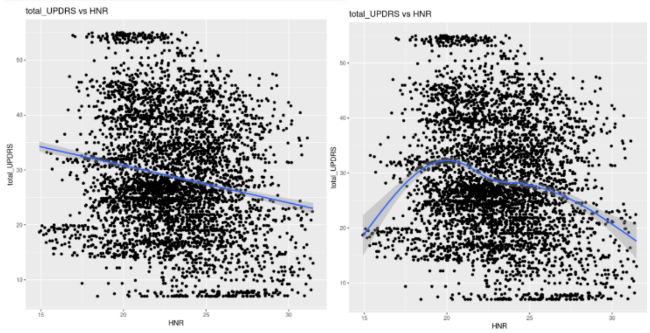


Fig. 1. Scatterplots of total_UPDRS vs HNR with linear and non-linear line

For categorical features, we use ANOVA test to determine if the features can be used to predict the target. Table 7 shows the results of ANOVA test. Both features have high F-value, and p-value smaller than the significance level (0.05), indicating that the means of the different groups are significantly different from each other. Boxplots are used to examine it further, and it shows the similar result as the ANOVA test. Therefore, both features are suitable to be the predictors for the target.

TABLE 7
ANOVA TEST OF CATEGORICAL FEATURES WITH TARGET

Feature	Df	Mean Sq	F value	PR(>F)
Sex	1	11157	94.43	<2e-16*

Out of 22 attributes, nine (9) are selected as the features and one (1) as the label for the prediction problem. In general, two (2) of the selected features are demographic features and another seven (7) are clinical features. All features are numerical except “sex” which are categorical features. Table 8 summarizes the features and label

TABLE 8
THE SELECTED FEATURES AND LABEL

Feature	Type	Value/Statistics	
sex	Nominal categorical	0 – male (3034) 1 – female (1403)	
age	Discrete numerical	Range: 49 Median: 65	Std: 9.028 Mean: 64.44
Jitter(%)	Continuous numerical	Range: 0.0090 Median: 0.0043	Std: 0.0016 Mean: 0.0045
Shimmer:APQ11	Continuous numerical	Range: 0.0443 Median: 0.0197	Std: 0.0085 Mean: 0.0209
NHR	Continuous numerical	Range: 0.0422 Median: 0.0162	Std: 0.0092 Mean: 0.0169
HNR	Continuous numerical	Range: 16.57 Median: 22.91	Std: 2.964 Mean: 23.03
RPDE	Continuous numerical	Range: 0.5125 Median: 0.5195	Std: 0.0909 Mean: 0.5189
DFA	Continuous numerical	Range: 0.3146 Median: 0.6357	Std: 0.0682 Mean: 0.6455
PPE	Continuous numerical	Range: 0.3361 Median: 0.1854	Std: 0.0606 Mean: 0.1893
total_UPDRS	Continuous numerical	Range: 47.99 Median: 27.28	Std: 10.98 Mean: 28.74

This study has 42 subjects, each contributes numerous records. Sample count for sex shows how many records belong to the category.

IV. DATA MODELLING TECHNIQUES AND RESULTS

This section starts with the data modeling methods employed in this study. Then, the results of each model are discussed in detail, and the best model is selected at the end of this section.

A. Data Modeling Methods

Three evaluation metrics are selected to evaluate the regression models which are R-squared, root mean squared error (RMSE) and mean absolute error (MAE). R-squared is the main evaluation metric to select the best model while performing hyperparameters tuning. The target performance to be achieved is 0.9 R-squared which is the same as the aim of this study. The dataset (4437 instances) is divided into 2 parts, 80% of training data (3550 instances) and 20% of testing data (887 instances). 4-fold cross validation is used for all models for hyperparameter tunings to avoid overfitting. Why use 4-fold cross validation? This is to ensure that the internal test set in cross validation has approximately the same number of instances (887 instances) as the hold-out test set as illustrated in Fig. 2. After that, 4 regression models are created to predict the total UPDRS namely Linear Regression, Decision Tree, Support Vector Machine and Random Forest. Finally, the best model is selected based on the evaluation metrics and the target performance.

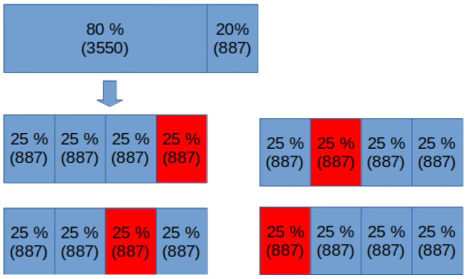
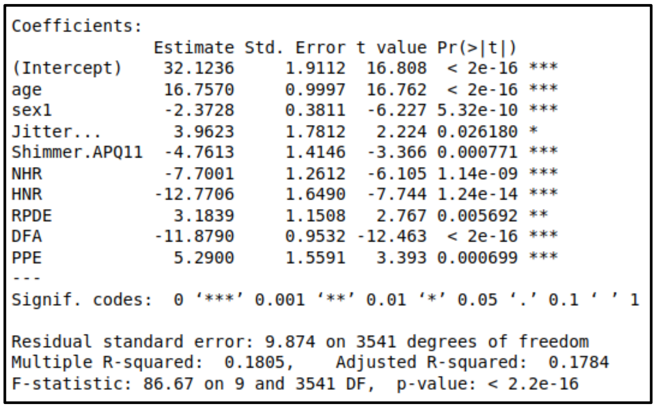


Fig. 2. Train test split and 4-fold cross validation

B. Model 1: Linear Regression

First model created is Linear Regression. Fig. 3. shows the summary of Linear Regression model. The summary shows that all features are significant because all have a p-value less than 0.05. Based on the coefficients, age has the greatest impact on total UPDRS. If the age of a patient increases by 1, the total UPDRS will increase by 16.7570. The attribute sex has lowest impact on total UPDRS. Female patients (sex1) will have 2.3728 less total UPDRS compared to male patients. Attributes Shimmer (APQ11), NHR, HNR and DFA are negatively correlated with total UPDRS whereas attributes age, Jitter (%), RPDE and PPE are positively correlated with total UPDRS. However, Linear Regression model does not fit well with the data because it produces 0.1805 R-squared only.



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    32.1236     1.9112  16.808  < 2e-16 ***
age            16.7570     0.9997  16.762  < 2e-16 ***
sex1           -2.3728     0.3811  -6.227 5.32e-10 ***
Jitter...       3.9623     1.7812   2.224 0.026180 *  
Shimmer.APQ11  -4.7613     1.4146  -3.366 0.000771 ***
NHR            -7.7001     1.2612  -6.105 1.14e-09 ***
HNR           -12.7706     1.6490  -7.744 1.24e-14 ***
RPDE            3.1839     1.1508   2.767 0.005692 ** 
DFA           -11.8790     0.9532 -12.463  < 2e-16 ***
PPE             5.2900     1.5591   3.393 0.000699 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.874 on 3541 degrees of freedom
Multiple R-squared:  0.1805,	Adjusted R-squared:  0.1784 
F-statistic: 86.67 on 9 and 3541 DF,  p-value: < 2.2e-16
```

Fig. 3. Summary of Linear Regression

There are four main assumptions that should be met before using Linear Regression which are**:** 1. linear relationship between features and target, 2. all features are independent, meaning that there is no multicollinearity among features, 3. data homoscedasticity**,** meaning that data should have constant variance, and 4. normality of residuals, meaning that residuals should follow normal distribution. Fig. 4 shows the diagnostic plots for Linear Regression model to determine if the model has met all the assumptions. The Residuals vs Fitted plot indicates that there is a linear relationship between features and target but a lot of points are not being predicted correctly due to large residual values. Besides, the upper and lower tails of Normal Q-Q plot are off from the dashed line, indicating that some data is still not following the normal distribution although log transformation has been performed on skewed attributes. Scale-Location plot shows that there is heteroscedasticity in the data whereas Residuals vs Leverage plot indicates that there are a lot of influential points in the data.

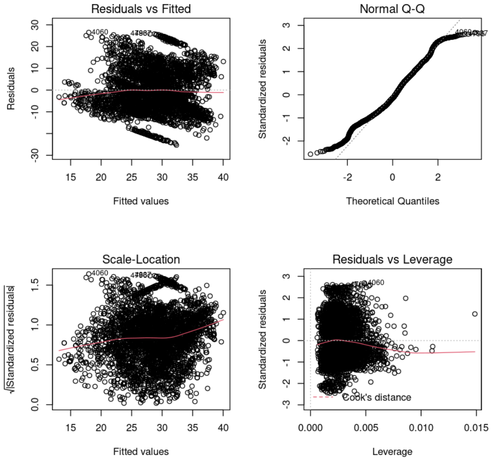


Fig. 4. Diagnostic plots for Linear Regression

After checking the Variance Inflation Factor (vif) of each feature, all features have a value less than 5, indicating that all features are independent. Table 9 shows the training and testing errors of Linear Regression. The results show that Linear Regression suffers from underfitting with only 0.1805252 and 0.1416984 training and testing R-squared value respectively, which is far away from the target performance (0.9 R-squared). The other metrics also show the same, both RMSE and MAE of Linear Regression are high. Therefore, Linear Regression is not suitable for this data because it violates most of the assumptions, and it produces large errors.

TABLE 9
TRAINING AND TESTING ERRORS OF LINEAR REGRESSION

Metrics	Train	Test
R-Squared	0.1805252	0.1416984
RMSE	9.860108	10.50919
MAE	8.147032	8.638832

Fig. 5 shows the residual plot of Linear Regression. It indicates that Linear Regression cannot predict most of the instances correctly, especially the points in the lower and upper end.

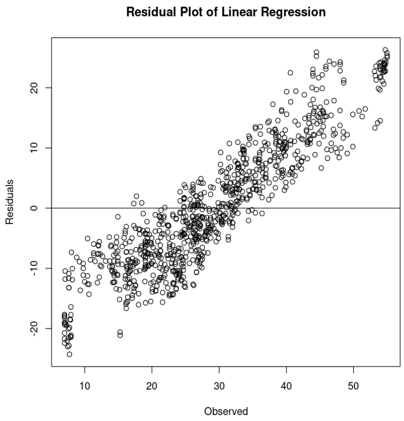


Fig. 5. Residual plot of Linear Regression

C. Model 2: Decision Tree Regression

The complexity parameter (cp) is used to control the size of the decision tree and to select the optimal tree size. To set up, "Rsquared" metric is used with cp initial value of 0.0001, each step of additional 0.0002 until value reaches 0.002. To visualise the hyperparameter tuning of the decision tree, the graph below shows Rsquared against CP. The Fig. 6 shows (cp=0.0007) is the best parameter value. When further increase the cp value, there is no sign that the R-Squared will go up.

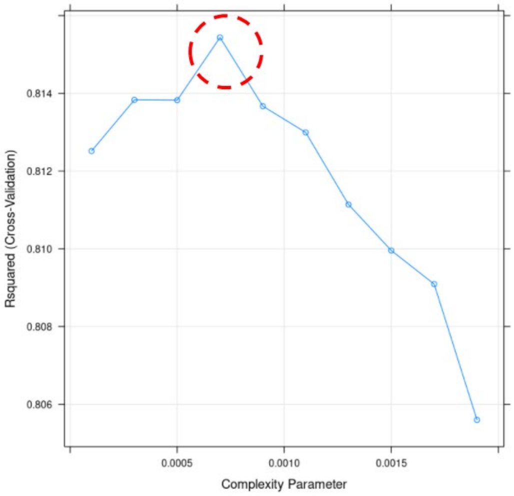


Fig. 6. Hyperparameter tuning for DTR

Therefore, (cp=0.0007) is chosen and applied to the decision tree model. The decision tree is visualized as in Fig. 7, which was set with a "maxdepth=8".

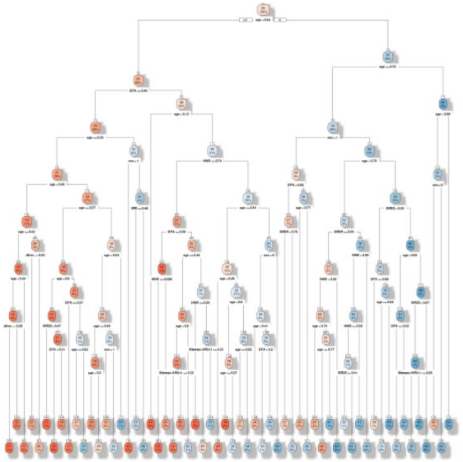


Fig. 7. Visualizations of DTR

After that, we continue to apply the best tuned parameter to extract the feature importance score in descending order as shown in Table 10 below. We noticed that age has the highest score, which is the most useful in predicting the target, total_UPDRS.

TABLE 10
FEATURE IMPORTANCE BY DECISION TREE REGRESSOR

Features	Importance Score
Age	279945
DFA	71312
HNR	61390
Sex	60989
RPDE	52621
Shimmer.APQ11	42585
PPE	26434
Jitter	23347
NHR	21091

Furthermore, RMSE, MAE and R Squared are computed for both training and testing sets as shown in Table 11. We noticed that both RMSE and MAE in the testing set are slightly higher, and R Squared is slightly lower than the training set. This indicates that the decision tree is slightly overfitting.

TABLE 11
TRAINING AND TESTING ERRORS OF DECISION TREE REGRESSION

Metrics	Train	Test
R-Squared	0.8667124	0.8314797
RMSE	3.976571	4.657112
MAE	2.895605	3.196446

Lastly, residuals against observations are plotted as shown in Fig. 8. When the observation values increase, the residuals are pretty symmetrically distributed, tending to cluster towards the middle of the plot. This indicates that it is a good fit for regression.

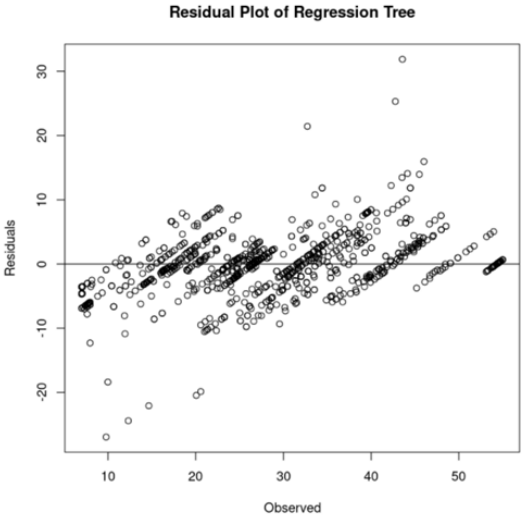


Fig. 8. Residual Plot of DTR

D. Model 3: Support Vector Machine

The third model created is Support Vector Machine (SVM). Hyperparameters tuning is performed by adjusting the parameter cost (C). Cost values represent the cost of constraints violation of SVM, the higher the value, the lower the generalization ability of the model. The C values tested are [0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128]. The SVM kernel used in this study is Radial Basis Function (RBF). This is because the features have a nonlinear relationship with the target, so the RBF kernel will outperform the linear kernel. Fig. 9 shows the R-squared values of SVM over different values of C. Based on the graph, the optimum C is 32 with 0.5580343 R-squared value. This is because further increase or decrease the C value will only result in decrease of R-squared value.

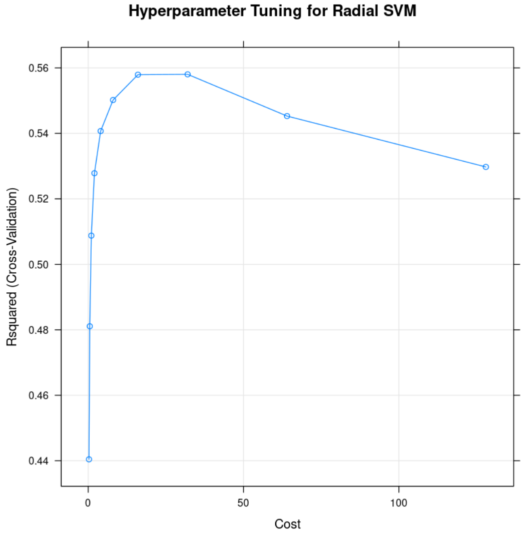


Fig. 9. Hyperparameter tuning for RBF SVM

Therefore, the final parameters for RBF SVM after performing hyperparameter tuning are:

- Kernel: Radial Basis Function (RBF)

- Sigma: 0.09639375
- C: 32

After fitting data to RBF SVM model with best hyperparameters, it produced the results as in Table 12. The results show that the RBF SVM has overfitted with the training data because the testing R-squared (0.5794922) is much smaller than the training R-squared (0.7028793). The other metrics also show the same, RMSE and MAE of training data is much smaller than the testing data. Furthermore, this model also did not achieve the target performance (0.9 R-squared), it has only moderate fit with the data based on the R-squared value (0.57979422). Therefore, this model is not suitable for this data as well.

TABLE 12
TRAINING AND TESTING ERRORS OF RBF SVM

Metrics	Train	Test
R-Squared	0.7028793	0.5794922
RMSE	5.940168	7.383144
MAE	3.981698	5.468278

Fig. 10 shows the residual plot of RBF SVM. It indicates that RBF SVM is only moderately fit with the data where most of the residual values are within ± 10. However, this model also has the difficulty to correctly predict the values in the lower and upper end.

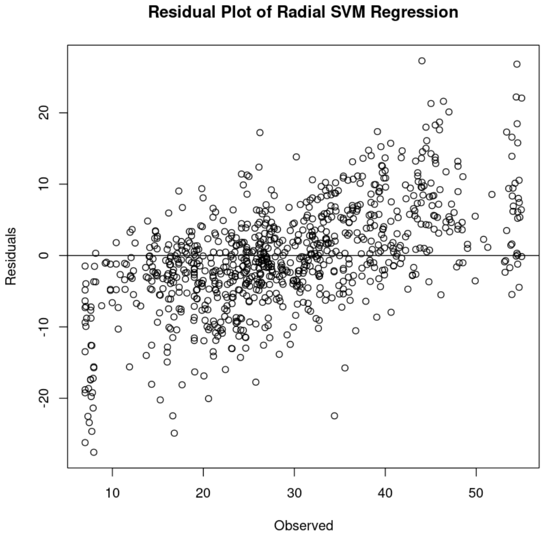


Fig. 10. Residual Plot of RBF SVM

E. Model 4: Random Forest Regression

The hyperparameter of mtry (number of variables randomly sampled as candidates at each split) is used to tune Random Forest Regressor (RFR) with metric "R Squared". The mtry values tested are [2, 3, 4, 5, 6, 7, 8, 9]. To visualize the hyperparameter tuning of the random forest regressor, the graph below shows R-Squared against mtry (Randomly Selected Packets). Fig. 11 shows (mtry=8) is the best parameter value. After increasing further, the R-Squared value (as highlighted in the red circle) starting to drop.

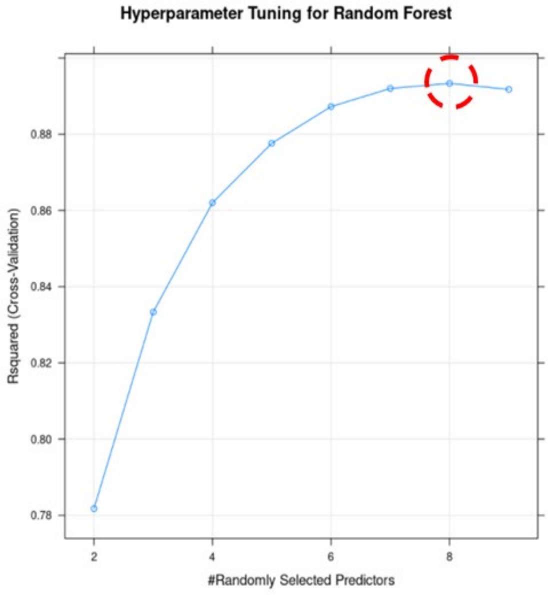


Fig. 11. Hyperparameter tuning for RFR

Fig. 12 below shows an inverse relationship between Error and number of trees. The error is dropping significantly when the number of trees increases from 0 to 50. In the random forest regressor, 500 trees are chosen as they have the lower error.

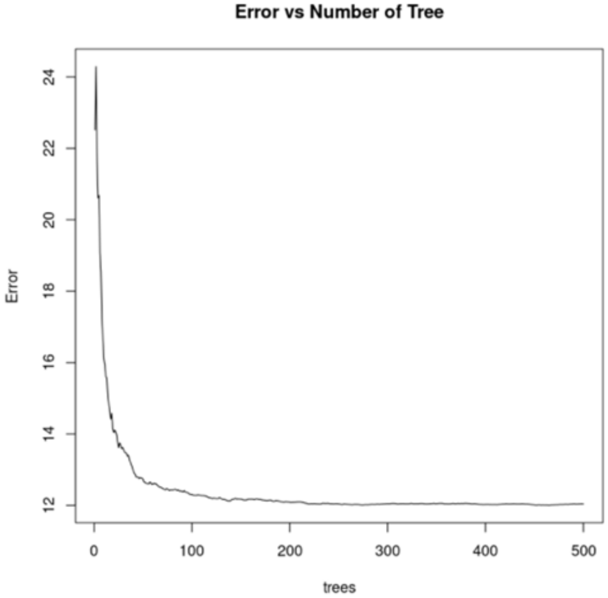


Fig. 12. Error Against Number of Tree

After tuning, the best hyperparameter of (mtry=8) is chosen and applied in order to compute the feature importance score. Table 13 shows the feature importance score in descending order. Age is still the most important attribute in predicting the target.

TABLE 13
FEATURE IMPORTANCE BY RANDOM FOREST REGRESSOR

Features	Feature Importance
Age	262871
DFA	40643
Sex	35806
RPDE	26895
HNR	21901
Shimmer.APQ11	11085
PPE	8540
Jitter(%)	5985
NHR	5108

Furthermore, RMSE, MAE and R Squared are computed for both training and testing sets as shown in Table 14. Random Forest Regressor has the highest R squared in both training and testing set in this experiment. Besides, we noticed that both RMSE and MAE in the testing set are slightly higher, and R Squared is slightly lower than the training set. This indicates that the random forest regressor is slightly overfitting.

TABLE 14
TRAINING AND TESTING ERRORS OF RANDOM FOREST REGRESSION

Metrics	Train	Test
R-Squared	0.983014	0.9115383
RMSE	1.460298	3.391643
MAE	1.027004	2.381224

Lastly, residuals against observations are plotted as shown in Fig. 13. When the observed values increase, the residuals are pretty symmetrically distributed, tending to cluster towards the middle of the plot. This indicates that it is a good fit for regression.

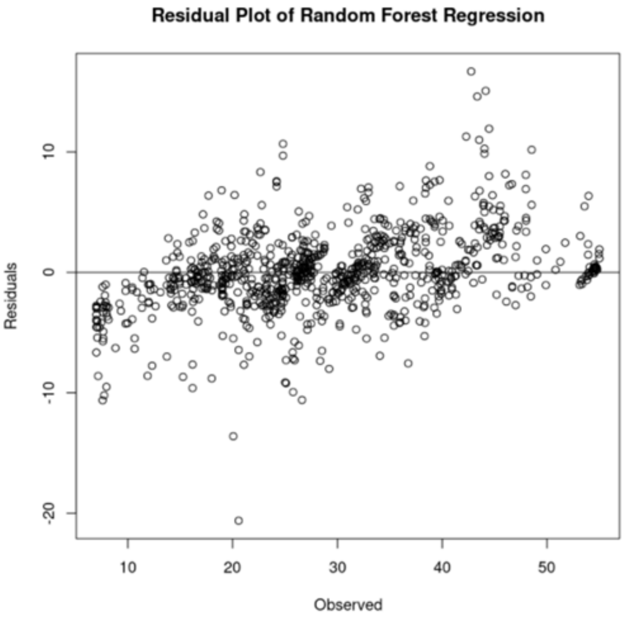


Fig. 13. Residual Plot of RFR

V. DISCUSSION

The best model in this experiment is the random forest. It has the highest R Squared, lowest RMSE and MAE for both training and testing sets. Moreover, it has achieved R squared more than 0.90 to fulfill the objective of the experiment. Table 15 and Table 16 show the training and testing set errors of all models respectively.

TABLE 15
TRAINING SET ERRORS OF ALL MODELS

Metrics	R-Squared	RMSE	MAE
Linear Regression	0.1805252	9.860108	8.147032
Decision Tree	0.8667124	3.976571	2.895605
SVM	0.7028793	5.940168	3.981698
Random Forest	0.9830140	1.460298	1.027004

TABLE 16
TESTING SET ERRORS OF ALL MODELS

Metrics	R-Squared	RMSE	MAE
Linear Regression	0.1416984	10.509188	8.638832
Decision Tree	0.8314797	4.657112	3.196446
SVM	0.5794922	7.383144	5.468278
Random Forest	0.9115383	3.391643	2.381224

VI. CONCLUSION

From the data exploration, we have identified voice measures that have a nonlinear relationship with UPDRS (RQ1). Therefore, 3 nonlinear models are deployed, which are SVM, Decision Tree Regressor, and Random Forest Regressor. After running all the models, the random forest regression model can best predict UPDRS based on voice measures (RQ2). This is because RFR has the highest R Squared (0.912), and has the lowest RMSE (3.39) and MAE (2.38) on the testing set. Besides, Age is the highest feature importance in the prediction of UPDRS. The aim of the project is fulfilled, which successfully builds a predictive model that can accurately predict the total UPDRS scores with at least 0.9 R Squared based on the voice measures.

VII. REFERENCES

1. M.C. de Rijk, L.J., Launer, K. Berger, M.M. Breteler, J.F. Dartigues, M. Baldereschi, L. Fratiglioni, A. Lobo, J. Martinez-Lage, C. Trenkwalder, A. Hofman: “Prevalence of Parkinson‟s disease in Europe: a collaborative study of population-based cohorts”, Neurology, Vol. 54, pp. 21–23, 2000)
2. Elbaz A, Bower JH, Maraganore DM, McDonnell SK, Peterson BJ, Ahlskog JE, Schaid DJ, Rocca WA. Risk tables for parkinsonism and Parkinson's disease. Journal of clinical epidemiology. 2002 Jan 1;55(1):25-31.
3. Tsanas A, Little M, McSharry P, Ramig L. Accurate telemonitoring of Parkinson’s disease progression by non-invasive speech tests. Nature Precedings. 2009 Oct 29:1-.

4. Ebersbach G, Baas H, Csoti I, Müngersdorf M, Deuschl G. Scales in Parkinson's disease. Journal of neurology. 2006 Sep;253(4):iv32-5.

5. Little M, McSharry P, Roberts S, Costello D, Moroz I. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. Nature Precedings. 2007 Jul 9:1-.Fahn S,Elton RL, UPDRS Program Members. Unified Parkinson's disease rating scale. In: Fahn S,Marsden CD,Goldstein M,Calne DB, editors. Recent developments in Parkinson's disease, Vol. 2. Florham Park, NJ: Macmillan Healthcare Information; 1987. p 153–163, 293–30

6. Fahn SR. Unified Parkinson's disease rating scale. Recent development in Parkinson's disease. 1987.

7. Ramaker C, Marinus J, Stiggelbout AM, Van Hilten BJ. Systematic evaluation of rating scales for impairment and disability in Parkinson's disease. Movement disorders: official journal of the Movement Disorder Society. 2002 Sep;17(5):867-76.

8. Midi I, Dogan M, Koseoglu M, Can G, Sehitoglu MA, Gunal DI. Voice abnormalities and their relation with motor dysfunction in Parkinson's disease. Acta Neurologica Scandinavica. 2008 Jan;117(1):26-34.

9. Tsanas A, Little M, McSharry P, Ramig L. Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests. Nature Precedings. 2009 Oct 29:1-.

10. Erdogdu Sakar B, Serbes G, Sakar CO. Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson's disease. PloS one. 2017 Aug 9;12(8):e0182428.

11. Yoon H, Li J. A novel positive transfer learning approach for telemonitoring of Parkinson's disease. IEEE Transactions on Automation Science and Engineering. 2018 Nov 2;16(1):180-91.

12. Tsanas A, Little M, McSharry P, Ramig L. Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests. Nature Precedings. 2009 Oct 29:1-.

13. Tsanas A, Little MA, McSharry PE, Ramig LO. Enhanced classical dysphonia measures and sparse regression for telemonitoring of Parkinson's disease progression. In2010 IEEE International Conference on Acoustics, Speech and Signal Processing 2010 Mar 14 (pp. 594-597). IEEE.

14. Varghese BK, Amali D, Devi KS. Prediction of parkinson's disease using machine learning techniques on speech dataset. Research Journal of Pharmacy and Technology. 2019;12(2):644-8.

15. Varghese BK, Amali D, Devi KS. Prediction of parkinson's disease using machine learning techniques on speech dataset. Research Journal of Pharmacy and Technology. 2019;12(2):644-8.

16. Nilashi M, Ahmadi H, Shahmoradi L, Mardani A, Ibrahim O, Yadegaridehkordi E. Knowledge Discovery and Diseases Prediction: A Comparative Study of Machine Learning Techniques. Journal of Soft Computing and Decision Support Systems. 2017 Sep 3;4(5):8-16.

17. Eskidere Ö, Ertaş F, Hanilçi C. A comparison of regression methods for remote tracking of Parkinson's disease progression. Expert Systems with Applications. 2012 Apr 1;39(5):5523-8.

18. Fawagreh K, Gaber MM. Resource-efficient fast prediction in healthcare data analytics: A pruned Random Forest regression approach. Computing. 2020 Jan 9:1-2.

19. Butt AH, Rovini E, Fujita H, Maremmani C, Cavallo F. Data-driven models for objective grading improvement of parkinson's disease. Annals of Biomedical Engineering