Parkinson's Disease Progression: An Analytical Study using Machine Learning Techniques

Devi Krishnan
dhkrishn@ncsu.edu
North Carolina State University
Raleigh

Rajshree Jain
rjain27@ncsu.edu
North Carolina State University
Raleigh

Sreeraksha Mavinhally Sreekantha
smavinh@ncsu.edu
North Carolina State University
Raleigh

Srujana Rachakonda
srachak@ncsu.edu
North Carolina State University
Raleigh

1 BACKGROUND AND INTRODUCTION

Problem Statement

Parkinson's disease is a neurodegenerative disorder that affects movement. The first step to living well with Parkinson's disease is to understand the disease and its progression. Hence all the available treatments aim to improve the symptoms of the disease. Since PD patients exhibit characteristic voice symptoms, voice recordings are a good way to diagnose the disease. The Unified Parkinson's Disease Rating Scale (UPDRS) has been one of the most commonly used tools for a comprehensive assessment of PD. The original scale for UPDRS consists of four parts, namely (i) mentation, behavior, and mood, (ii) activities related to daily life, (iii) motor examination, and (iv) treatment complications. Motor-UPDRS and Total-UPDRS are two important clinical scales of PD, with the ranges of each being 0 - 108 (0 being healthy and 108 indicating total motor impairment) and 0 - 176 (0 being healthy and 176 being total disability) , respectively.

This project studies the performance of various regression models on the prediction of total UPDRS metric in order to better monitor the progression of Parkinson's disease. The symptoms are not very noticeable in the early stages of the disease, and they develop slowly over years. Speech disabilities are therefore among the most indicative symptoms. In light of this fact, we also aim to identify an optimal value of UPDRS beyond which speech disorders are more significantly observed.

Related Work

A. Tsanas et al. first performed feature extraction using classical methods based on linear signal processing techniques, such as short-time autocorrelation, followed by peak-picking. Linear regression methods, i.e. classical least-squares regression, iteratively reweighted least-squares regression, and LASSO were used to find the relationship between the features and the total and motor UPDRS. The authors made use of classification and regression trees (CARTs) as well for comparison. Bayesian Information Criterion (BIC) and Akaike's Information Criterion were used for model selection. Cross-validation was used to test the performance of the algorithms. Correlation analysis indicated a strong association between speech and UPDRS. None of the measures individually appeared significantly correlated to either motor or total UPDRS. CART outperformed the linear predictors with a training MAE (mean absolute error) of 4.5 and testing MAE of 5.8 for motor UPDRS, and a training MAE of 6.0 and testing MAE of 7.5 for total UPDRS.

We referenced this paper because the authors collected the dataset and they came out with relatively good results for total and motor UPDRS using regression. Our project also uses regression, and these results establish a good baseline to compare our results against.

Betul Erdogdu Sakar has devised a method to determine the optimal UPDRS threshold value that can be discriminated by dysphonia measurements. They determined a UPDRS threshold value beyond which speech disorders can be significantly improved. For this they converted the UPDRS prediction problem into a binary classification problem for various motor UPDRS threshold values and then used KNN and SVM classification algorithms to determine patients with values of UPDRS less than or greater than the threshold. The correctness of this classification was determined using the Mathews Correlation Coefficient. They came up with a threshold value of around 15 and corresponding accuracy of 75.86 percent. Also, SVM classification showed to perform better.

We referenced this paper because the area of study corresponds to our second hypothesis and the authors have devised a good way to come up with finding a UPDRS threshold. We also plan to use different classification algorithms to determine the threshold and hence this paper can be a good

baseline.

S. Jain et al. devised a two-step predictive model; they first classified the stage of the disease, then used the predicted class as an attribute to predict the UPDRS score. Multivariate Linear Regression (MLR), Regression Tree, KNN, and Multilayer Perceptron Neural Network (MLP) were used to predict the total UPDRS. C4.5 decision tree algorithm was used to classify a patient's disease progression into one of two stages. 10-fold cross validation was used for assessing the accuracy of all the algorithms. The C4.5 algorithm predicted the stage correctly with an accuracy of about 99.35 percent. KNN had the least error of all the models with an RMSE of 1.19. The regression models were found to perform better after the classification step.

We referenced this paper since the authors used MLR and MLP, which we are also implementing. Their results were good for the same data we will be using, so we know what range of values is the current norm for the same algorithms.

2 METHOD

Approach

We will be implementing various learning algorithms for feature selection, prediction of the total UPDRS and subsequent determination of optimal threshold.

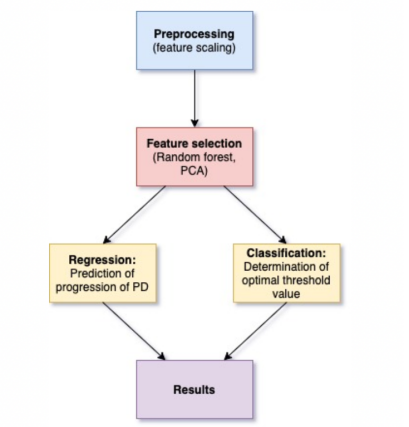

Figure 1: Architecture

1. Principal Component Analysis

Principal Component Analysis (PCA) is an unsupervised linear transformation technique that is widely used across different fields, most prominently for feature extraction and dimensionality reduction. PCA helps us to identify patterns in data based on the correlation between features. In a nutshell, it aims to find the directions of maximum variance in high-dimensional data and projects it onto a new subspace with equal or fewer dimensions than the original one.

In this project we applied PCA to identify the attributes that contributed the maximum variance (about 95%) and used the selected features for applying regression techniques. We did feature selection using both PCA and Random Forest, which is explained below, and we compared the feature selection techniques to see which gave us better results with each regression model.

2. Random Forest

Random Forest is a meta-estimator that aggregates the output of a number of decision trees. Decision trees are sensitive to the specific data on which they are trained and pose the risk of over-fitting. Random Forest helps us overcome this. It is an ensemble learning technique which implements bagging with decision trees, in which each tree is built from a sample drawn with replacement from the training set. Furthermore, when splitting each node during the construction of a tree, the best split is found either from all input features or a random subset of them. This randomness helps in achieving reduced variance.

In this project we decided to use Random Forest Regressors as a feature selection technique to identify the relatively important features with respect to the predictability of the target variable. The relative rank (i.e. depth) of a feature used as a decision node in a tree can be used for this purpose. Features used at the top of the tree contribute to the final prediction decision of a larger fraction of the input samples.

3. Multiple Linear Regression

Multiple linear regression is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression, is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable. By implementing linear regression we try to find a linear relationship between the target and attributes. Since the clinician's total UPDRS scores were linearly interpolated in the dataset, we concluded that it would be a good strategy to implement linear regression for predicting the UPDRS scores. After pre-processing the data and implementing the random forest algorithm for feature selection, we trained the linear regression model using gradient descent to predict the total UPDRS score while minimizing the mean absolute error. Towards this goal, we implemented our linear regression model from scratch.

4. Support Vector Regression

Support vector machines are supervised learning models that are used for binary classification problems. The algorithm finds the decision boundary (or hyperplane) that best

separates the two classes of data points when plotted on a plane. The best hyperplane is the one whose distance (margin) to the closest member of each class is largest. In case of linear data, classification is simply as described. In the case of non-linear data, the data needs to be mapped to higher dimensions, and this is done using a kernel. A kernel is a function that takes a low dimensional input space and maps it to a higher dimensional space.

SVR uses the same principles as SVM, with only a few differences. In SVR, a margin of tolerance is set with respect to the prediction of classes, called epsilon. SVR can be used for continuous values as well, making it useful for real-valued data. We decided to use SVR for estimating total UPDRS values since it has been used for the UPDRS prediction in related research with good results. We used it to establish a baseline result, with which we can compare the results of other predictors.

5. Multilayer Perceptron

A multilayer perceptron (MLP) is a class of feedforward artificial neural networks (ANNs). Neural network architectures have proven to be optimal models in recent years for classification and regression tasks. The perceptron is a linear classifier. These are simple computational units that have weighted input signals and produce an output signal using an activation function, which is a simple mapping of summed weighted inputs to the output of the neuron. Learning occurs in the perceptron by changing connection weights after each piece of data is processed, based on the amount of error in the output compared to the expected result. This is carried out through backpropagation.

In our project, we used the MLP model for both or regression as well as classification (i.e. for the determination of optimal threshold) tasks. For regression, we train the model on the input features and then predict the total UPDRS value, whereas for the classification task, we binarize the motor UPDRS feature from the dataset (elaborated in Experimental Design) and then classify the input data patterns.

Novelty: Our proposed model uses the Random Forest Regressor technique for feature selection, which is relatively new since the syllabus covered only PCA for feature selection. Secondly, we decided to implement Multiple Linear Regression from scratch, something that the syllabus did not require us to do. We also used the Support Vector Regression model with the RBF kernel function. Finally, we also used a Multilayer Perceptron neural network, which is both outside the scope of our syllabus and not widely used so far for this particular prediction problem.
For the second hypothesis, we modify our regression dataset to fit our classfication task of determining an optimal threshold value of motor UPDRS beyond which speech disorders begin to appear. To this end, we make use of MLP again, and we evaluate the generated predictions using Matthew's correlation coefficient, which is also a novel evaluation metric with respect to the scope of the syllabus.

Rationale

The chosen approach for feature selection is Principle Component Analysis and Random Forests. Random Forest is chosen for feature selection as it uses tree based strategies and it naturally ranks features by how well they improve the purity of a node. The project uses Multiple Linear Regression with gradient descent as this is one of the proven techniques for a promising UPDRS score prediction. By implementing Linear regression we try to find a linear relationship between the target and the attributes. Since the clinician's total UPDRS and Motor UPDRS scores were linearly interpolated in the dataset, it would be a good technique to implement linear regression for predicting the UPDRS scores. We decided to use SVR for estimating total UPDRS values since it has been used for the UPDRS prediction in related research with good results. Multilayer Perceptron was chosen for predicting total UPDRS as it is a frequently used neural network for both regression and process modeling in related work. We are using Multilayer perceptron and Matthews Correlation Coefficient for predicting threshold value of speech disorder where MLP acts as a classifier. Multi Layer Perceptrons (MLP) are universal approximators which can be used to create mathematical models using regression analysis. As classification is a particular case of regression when the response variable is categorical, MLPs make good classifier algorithms.

3 EXPERIMENT

Dataset

In this project we have used the Parkinsons Telemonitoring Data Set from the UCI machine learning repository. We have used this multivariate dataset to achieve our regression task. This dataset was originally created in the University of Oxford in collaboration with 10 medical centers in US and Intel Corporation, who had developed a telemonitoring device to record the speech signals from people. The dataset has 22 attributes and about 5875 instances. To collect the dataset, 42 people with early stage Parkinsons disease were recruited to a trial of 6 month trial of the telemonitoring device so that their biomedical voice measurements are taken and for remote symptom progression monitoring. The attributes in the dataset are as follows:

(1) Subject Number - Integer that uniquely identifies each subject
(2) Subject Age - Subject age

(3) Subject gender - Subject gender '0' - male, '1' - female time interval from baseline recruitment date
(4) Motor UPDRS (Motor Unified Parkinson's Disease Rating Scale) - Clinician's motor UPDRS score, linearly interpolated
(5) Total UPDRS (Total Unified Parkinson's Disease Rating Scale) - Clinician's total UPDRS score, linearly interpolated
(6) Further are several measures of variation in fundamental frequency like Jitter(%), Jitter(Abs), Jitter:RAP, Jitter:PPQ5, Jitter:DDP
(7) Several measures of variation in amplitude are Shimmer, Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, Shimmer:APQ11, Shimmer:DDA
(8) Two measures of ratio of noise to tonal components in the voice are NHR, HNR
(9) RPDE - A nonlinear dynamical complexity measure
(10) DFA - Signal fractal scaling exponent
(11) PPE - A nonlinear measure of fundamental frequency variation

In our project, out of these 22 attributes, we will be using 20 as 2 attributes (Total_UPDRS) and (Motor_UPDRS) are target variables. Further, although there are 22 attributes, we have mentioned 11 bullet points because a few attributes like Jitter, Shimmer, ratio of noise to tonal components in voice collectively have a bunch of related attributes. Unified Parkinson's Disease Rating Scale (UPDRS) is one of the most commonly used tools for comprehensive assessment of Parkinsons Disease. Severity and progression of PD symptoms as well as symptom fluctuations are typically evaluated using UPDRS. Motor-UPDRS and Total-UPDRS are two important clinical scales of Parkinsons Disease. More the value of UPDRS, more is the prominence of Parkinsons disease.

Hypotheses

We have explored our data and the past work done in this domain and have come up with the following hypotheses:

(1) *Prediction of the progression of Parkinson's disease* using non-invasive telemonitoring data. This is accomplished by predicting the Total UPDRS scores.
(2) *Determination of the optimal threshold value* that can be discriminated by dysphonia measurements for UPDRS. We propose to make use of the motor UPDRS scores from the dataset and try to identify an optimal value of UPDRS beyond which speech disorders are more significantly observed.

Experimental Design

Prediction of progression of PD: All of the implementation was done using Python 3. The libraries we used were Pandas, Scikit-learn, Numpy, Keras and Tensorflow.

Preprocessing: We decided on a random split of 70 percent for training data and 30 percent for testing data, as this is the norm. The total UPDRS was taken as the dependent variable and the rest of the attributes were taken as the independent variables. Preprocessing was carried out using Pandas and Scikit-learn in Python.

Feature Selection: Feature selection was performed using two different algorithms, PCA (principal component analysis) and Random forest, both in Scikit-learn. Before performing PCA, the data was first scaled in order to ensure that the features had a mean of zero and a standard deviation of one. Scaling is important here because PCA wants to maximize the variance, and if one feature's scale is higher, then it will influence the algorithm. PCA was applied on the transformed data to find the features that explained 95 percent of the total variance. The original data was then transformed using the principal components. For random forest, the number of trees chosen was 1000.

Prediction:

SVR Implementation Details: The chosen kernel type was RBF (radial basis function) as it was found to generate the lowest error rate out of the different kernel functions we tested on SVR (i.e. sigmoid, polynomial, and linear). The other parameters we experimented with were C (regularization parameter), gamma (kernel coefficient) and epsilon. The lowest error rate was generated in a model with C as 100, gamma as 0.1 and epsilon as 0.1. Epsilon defines our margin of tolerance within which no penalty is given to errors.

MLR Implementation Details: Our linear regression model used gradient descent for optimization. We have experimented with various learning rates and epoch values and the best MAE value was given with the foll0wing hyperparameters of learning rate = 0.0001 and epochs = 500.

MLP Implementation Details: We played around with different values for our hyper-parameters and the following values gave us the best results for our task at hand. Our model had four hidden layers with 500, 400, 300 and 200 neurons respectively. It used Sigmoid as the activation function. The loss function was chosen as Mean Squared Error. We took care of over-fitting by dropping layers. We made use of Adam Optimizer. 10% of the training data was considered for validation here.

Determination of Optimal Threshold:

Our goal is to find an optimum threshold for Motor UPDRS (response variable) beyond which speech impairment is noticeable. For this, we are binarizing the response variable based on a threshold value. We have tried out about 24 threshold values ranging from 10-34 on the Motor UPDRS

scale. We compare the predictions of our model with the new response variable and pick the threshold value that gives us best Matthew's Correlation Coefficient (MCC).

Feature Selection: Since optimal threshold is to be calculated on the basis of dysphonia features, we removed all other features from the data set such as follows:

(1) Subject#
(2) Age
(3) Sex
(4) Test-Time
(5) Total UPDRS
(6) Motor UPDRS (response variable)

We then used Random Forest Regressor, similar to our task in the above section. Our feature selection algorithm gave us the following list of 7 important features from 16 features:

(1) Jitter(Abs)
(2) Shimmer:APQ11
(3) NHR
(4) HNR
(5) RPDE
(6) DFA
(7) PPE

Prediction: We convert the original UPDRS prediction problem into a binary classification problem for various motor UPDRS threshold values. For each value of the threshold, if the motor UPDRS value of an input pattern is greater than the threshold value, it is labeled as positive, otherwise it is labeled as negative. We then use this as our new response variable and predict using our Multi Layered Perceptron model to classify the instances as above or below threshold classes.

The interval of the UPDRS threshold value (i.e. 10 to 34) that has been evaluated was determined so that each of the classes contains at least 10 of the total number of samples. Splitting input patterns based on thresholds can pose a problem of generating imbalanced class sizes. In such situations Accuracy might give us skewed results. To handle this we are using Matthew's Correlation Coefficient (MCC) along with accuracy for evaluation.

Evaluation Metrics:

Matthew's Correlation Coefficient . MCC metric is a balanced measure which can be used even if the classes are of very different sizes. It outputs a value between 1 and + 1. The formulation of the MCC metric is given below:

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

4 RESULTS

Prediction We have implemented our prediction models with both the results of both our feature selection techniques mentioned above. However, we noticed that the results are significantly better for random forest when compared with PCA. Although we are still working on the different combinations of models and feature selection techniques. The results that we have so far are listed in Table 1.

(1) **Multivariate linear Regression with Gradient Descent:** Prediction of Parkinson's disease progression by predicting Total vs using Multivariate Linear Regression with gradient descent gave us the following values. We can see that with learning rate = 0.0001 and epochs = 200 we are obtaining a Testing MAE of 7.90 with Random forest regressor pre-processing model (Figure 2) and with learning rate = 0.05 and epochs = 100 we are obtaining a Testing MAE of 7.64 with PCA pre-processing model (Figure 3). These values are comparable to the baseline results provided by the linear regression library model.

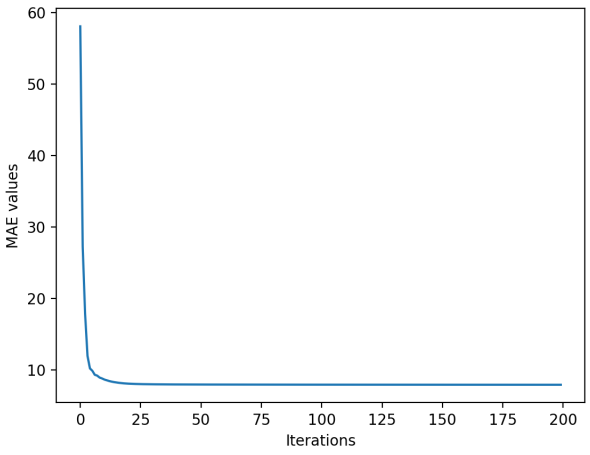

Figure 2: MLR with Random Forest feature selection; learning rate = 0.0001, epochs = 200

(2) **Support Vector Regression:** Prediction of Total UPDRS values using PCA and Random Forest Regressor preprocessing techniques gave us the following results. SVR with Random Forest Regressor gives a Testing MAE of 2.783 and SVR with PCA gives a Testing MAE of 1.715, which is a much better value than baseline model (i.e. CART [1]).
(3) **Multilayer Perceptron:** Prediction of Total UPDRS values using PCA and Random Forest Regressor preprocessing techniques gave us the following results. MLP with Random Forest Regressor gives a Testing

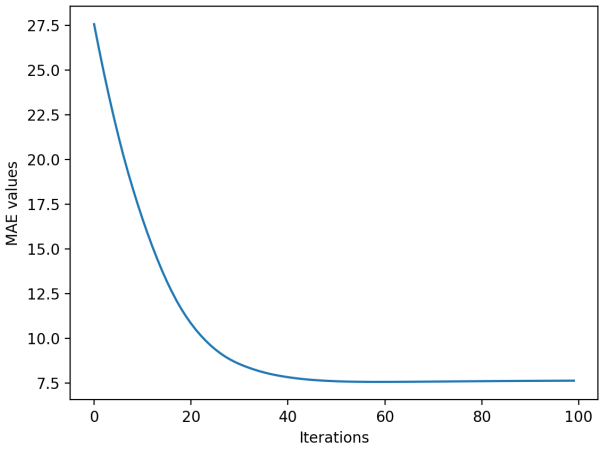

Figure 3: MLR with PCA feature selection; learning rate = 0.05, epochs = 100

MAE of 0.415 and SVR with PCA gives a Testing MAE of 0.573 which gives the best results compared to other models that have been implemented.

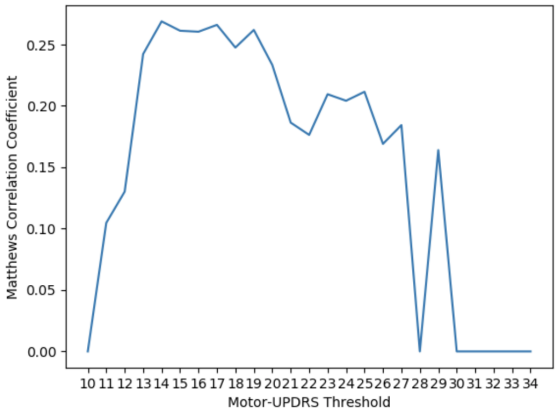

Figure 4: Matthew's correlation coefficient vs. UPDRS Threshold

Figure 4 plots the Motor UPDRS threshold values against MCC. MCC is a measure of the quality of binary classification, and higher the value of MCC, better is the model. We can observe from the graph that motor UPDRS threshold values between 14 -17 provide a stable classification model. Figure 5 shows the prediction accuracy for various threshold values of motor UPDRS.

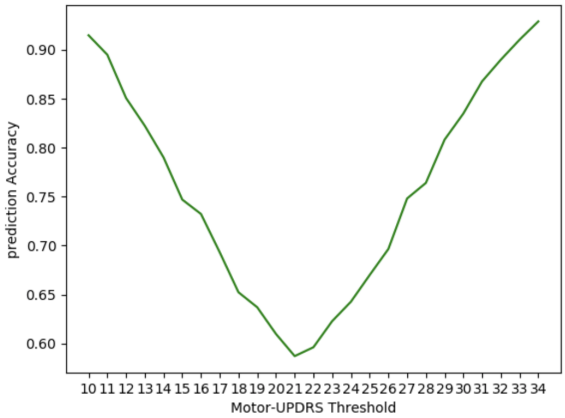

Figure 5: Prediction Accuracy vs. UPDRS Threshold

Discussion

We have used Multiple Linear Regression, Support Vector Regression and Multiple Layer Perceptron model to determine the progression of Parkinson's disease by predicting the Total UPDRS scores.

Using the Multivariate Linear Regression model we were able to find a relationship between, the independent variable (Total UPDRS) and the dependent variables (attributes). Having tested the algorithm with various values of learning rates and epochs we could observe that learning rate = 0.0001 and epochs = 200 tend to give an MAE value that is comparable to the MAE value of given by the baseline linear regression model. Hence, we can establish that linear regression with gradient descent fares well. Further, more we could observe that the value of MAE decreased exponentially with the increase of number of epochs and also the value of MAE varied on changing the learning rate. Hence, it is important to have appropriate hyper parameter values good predictions. Also, we can observe from the results an improvement in the MAE results when we changed the feature selection model from Random Forest to PCA. We believe that although the Linear Regression model gives an MAE comparable to the baseline model the model can be modified to give more accurate results.

Further, we could observe that on using SVR as another linear model we get better results. The testing MAE with Random Forest as 2.78 and with PCA it is 1.715. On the other hand, based on results and experimentation of previous research [1] and [2] we implemented MLP as a non linear regression model and could draw a comparison between our results and previous results. The MAE value on using the MLP model drastically reduced to 0.57 with PCA as the feature selection model and 0.41 with the Random Forest as feature selection model. The CART non linear algorithm

Table 1: Results

Preprocessing Technique	Model	MAE
PCA	SVR	1.715
PCA	MLR	7.64
PCA	MLP	0.573
Random Forest	SVR	2.783
Random Forest	MLR	7.90
Random Forest	MLP	0.415

[2] which has been implemented in previous works gives an MAE of 7.52 and hence we can say that MLP is definitely better model to predict Parkinson's disease progression.

Further, we can also observe that for the linear models it looks like PCA was a better feature selection algorithm. However, MLP with Random Forest results outperforms many prediction techniques.

Coming to the results of our second hypothesis, we predicted MCC values for 24 different values of the threshold ranging from 10 to 34. We used the MLP algorithm as a classifier on the classification data-set we obtained by binarizing the Motor UPDRS value which we considered as our target variable. On performing several rounds of the above steps we found a threshold of 15 which has the highest corresponding MCC value. This result confirms the results from past work [6], where KNN and SVM models have been used to determine the threshold values. We observe that there is an accuracy of 75 percent when MCC is 15. The accuracy plot in Fig. (3) might lead us to believe that the threshold value should be chosen as 10 because it has an accuracy value of 90 percent. But, doing that would lead to false inferences as the data set on choosing the threshold would be highly imbalanced, with huge number of positive cases of Parkinson's disease. But since MCC is a good metric to evaluate algorithms applied on highly imbalanced datasets, we can say that a threshold of 15 can be a value above which, speech disorders can be significantly observed. Also, we could observe a similar trend of accuracy and MCC metric as seen in [4] for KNN and SVR.

5 CONCLUSION

Our study indicates that we can determine the progression of Parkinson's disease with a minimal MAE error of 0.4155 which is given by our MLP model. Since this is our testing MAE, we should be able to predict progression of Parkinson's in a new patient on the basis of vocal measurements accurately. Being able to predict the progression with such low error would be helpful as it can reduce the healthcare costs and clinical visits for a patient significantly and also lead to more accurate diagnosis. If only we can find a fast and economic way to collect these dysphonia measurements of patients in real-time. It would be of great value to the health care industry.

One of the major problems of Parkinson's disease is that it is very hard to diagnose at earlier stages. According to our results, speech impairment symptoms are shown from a threshold value of 15 on the UPDRS scale (0-108) that means that constant monitoring of the patient using these non-invasive testing methods can help in early diagnosis and treatment of symptoms of Parkinson's disease. It also implies that speech disorders are early indicators of Parkinson's disease.

6 VIRTUAL MEETING SCHEDULE

We had the following virtual meetings over Zoom ("All" includes all four members):

Table 2: Meeting Attendance

Date	Time	Attendees
March 24th	4:30 - 6:00 p.m.	All
March 27th	4:00 - 5:30 p.m.	All
April 4th	4:00 - 5:30 p.m.	All
April 10th	4:00 - 5:30 p.m.	All
April 14th	5:00 - 7:00 p.m.	All
April 17th	4:00 - 6:00 p.m.	All
April 22nd	4:00 - 6:00 p.m.	All
April 23rd	5:00 - 7:00 p.m.	All

GitHub Project Link

https://github.com/srujana13/ParkinsonsDiseaseProgression

REFERENCES

[1] Tsanas, Athanasios, et al. "Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests." IEEE transactions on Biomedical Engineering 57.4 (2009): 884-893.

[2] Nilashi, Mehrbakhsh, et al. "An analytical method for measuring the Parkinson's disease progression: A case on a Parkinson's telemonitoring dataset." Measurement 136 (2019): 545-557.

[3] Jain, Shreya, and Sujala Shetty. "Improving accuracy in noninvasive telemonitoring of progression of Parkinson's Disease using two-step predictive model." 2016 Third International Conference on Electrical, Electronics, Computer Engineering and their Applications (EECEA). IEEE, 2016.

[4] B. E. Sakar, C. O. Sakar, G. Serbes and O. Kursun, "Determination of the optimal threshold value that can be discriminated by dysphonia measurements for unified Parkinson's

Disease rating scale," 2015 IEEE 15th International Conference on Bioinformatics and Bioengineering (BIBE), Belgrade, 2015, pp. 1-4

[5] Linear Regression: https://mubaris.com/posts/linear-regression/

[6] Parkinson's disease: https://en.wikipedia.org/wiki/Parkinson%27s_disease