Comparison of Machine learning models for Parkinson's Disease prediction

Tapan Kumar
Research Scholar,University School of ICT, GGSIP University, Delhi, India
tapanjha91@hotmail.com

Pradyumn Sharma
B.Tech 3rd Year, Dept. of CSE, IIT Jammu, Jammu, India
pradyumnsharma20@gmail.com

Prof. Nupur Prakash
Professor, University School of ICT, GGSIP University, Delhi, India
nupurprakash1@gmail.com

***Abstract*—Parkinson's Disease (PD) is a chronic degenerative disease that mainly affects the nervous system and motor controls in human beings. Early symptoms such as muscle stiffness, tremors, impaired balance and difficulty with walking are considerably less noticeable. Blood tests and Scans also do not provide sufficient evidence for early diagnosis. Hence it is very difficult for doctors to diagnose the onset of Parkinson's Disease. However, smearing of speech gives an early warning and can be effectively used for the prediction of PD. This paper, the voice recording samples of Parkinson's disease affected and healthy patients have been used for PD prediction. Thirteen predictive models using various Machine Learning techniques have been formulated using the University of California, Irvine (UCI) dataset. A comparative study of these predictive models has been carried out on the UCI dataset consisting of biomedical voice recording samples of healthy and Parkinson's Disease affected peoples. These predictive models have been trained and tested for their accuracy and efficiency. The performance analysis of the best five models has been presented in this paper, for accurate prediction of Parkinson's Disease at an early stage. The processing speed of these models has also been analysed, to assess their suitability for light weight mobile applications in the ubiquitous computing environment.**

Keywords— Parkinson's Disease, Predictive Models, voice pattern biometrics, motor disorders Introduction

I. INTRODUCTION

Neurodegenerative disorders are the results of the progressive tearing and loss of neurons in different areas of the nervous system and the human brain. Parkinson's Disease (PD) is a neurodegenerative disease. The people in the age group of 70 years and above are most vulnerable to Parkinson's Disease [1]. Such diseases spread all over the body gradually without early warning. PD generally affects the brain neurons which are responsible for overall body movements, reflexes and responses. The precise symptoms and causes of Parkinson's disease are unknown. Research work in this area and doctors claim that genetics, biochemical changes, the aging factor and environmental factors trigger the disease. The primary symptoms of Parkinson's disease are tremors, rigidity (muscle stiffness), slowing of reflex action on simple tasks, changes in facial expression and postural instability resulting in the risk of a fall. Other symptoms include difficulty in swallowing, chewing, or speaking, constipation, dementia, or other cognitive problems.

Currently, there are no particular blood tests and imaging techniques available that can diagnose the onset of Parkinson's disease accurately. This paper experiments with, thirteen machine learning models that can predict an early onset of Parkinson's disease using voice samples, which can be recorded using low cost smart phones also. These machine learning algorithms have been applied to a limited data set of voice recordings and their training and testing accuracy has been compared. A lightweight algorithm is considered to be suitable for mobile devices used by medical practitioners and healthcare workers anytime, anywhere. Therefore, the performance of the most efficient models, which are suitable for mobile applications, has been analysed for their compactness, prediction accuracy and speed.

II. LITERATURE SURVEY

In the last fifteen years a number of data mining and AI techniques have been used to predict neurodegenerative disorders like Parkinson's Disease and Alzheimer. In India, almost one percent of the population is affected from PD. Bonato et al. (2004) [2] gathered data using accelerometer and electromyography signals and proposed that the severity of motor neuron disorder can be recognized using data mining and artificial intelligence techniques.

By using three different data mining methods viz. sequential minimization optimization, logistic regression and decision stump for the prediction of PD, Yadav, Geeta, Yugal Kumar and Sahoo (2012) [3], obtained, the best score from the support vector machine model with 76% accuracy. Weitschek et al. (2014) [4], proposed a predictive model for Alzheimer Disease (AD), through the diagnosis of brain abnormalities, using Electroencephalography (EEG), a non-invasive and repeatable technique. To support the medical doctors in the correct diagnosis, they achieved an automatic patient's classification from the EEG biomedical signals involved in AD and Mild Cognitive Impairment (MCI). Using time-frequency transforms for pre-processing of EEG signals, the authors subsequently applied classification using machine learning. Kamal Nayan Reddy et al. (2016) [5] discussed the importance of non-motor systems over motor systems for the prediction of PD. The study was conducted around olfactory loss, sleep behavior distortion and rapid eye movement. The machine learning techniques, like Boosted Logistic Regression, Random forest, Bayes Net and Multilayer Perceptron were used for prediction of PD. An accuracy of 97.159% and area under the curve (AUC) of 98.9% was achieved using the Boosted logistic regression. Extreme Learning Machines (ELM) were proposed to predict Parkinson's disease by Chandrayan, et al. (2016) [6]. A comparative analysis has been done by them, which indicates, that unlike conventional Neural Networks, ELM does not require repetitive changes of hidden neurons. The simple architecture makes ELM a reliable choice than others for

978-1-7281-9656-5/20/$31.00 ©2020 IEEE

prediction. Kumar Tiwari (2016) [7] proposed a feature selection algorithm offering maximum relevance and minimum redundancy for predicting Parkinson's disease. He also found that random forest is a better technique with an overall accuracy of 90.3% in comparison to other ML approaches such as bagging, random subspace, support vector machines, etc. Sonu, S. R., et al. (2017) [8] implemented a JavaScript program to record the voice of the patient in a ".wav file". The best score is given by decision tree algorithm with the accuracy of 94% with feature selection for the prediction of PD. Jennifer He et al. (2017) [9] observed that the best feature for the prediction of Parkinson's disease is a fundamental frequency among all voice recording features. They tested a number of machine learning methods on Microsoft azure machine learning platform and found that the best score is given by two-class Boosted decision trees. Using serum samples from a clinically well-characterized longitudinally followed Michael J Fox Foundation cohort of Parkinson's disease patients, Ahmadi Rastegar et. al. (2019) [10] found that peripheral cytokines may have utility for aiding the prediction of Parkinson's disease progression using machine learning models. Using a decreased input feature space of Parkinson's tele-monitoring dataset, Shahid and Singh (2020) [11], proposed a deep neural network (DNN) model to predict the progression of PD. The accuracy of the model is determined through the mean absolute error, root mean squared error and coefficient of determination having values of 0.926, 1.422 and 0.970, respectively. The study carried out by Hemmerling and Wojcik-Pedziwiatr (2020) [12], involved recording of the vowel /a/ spoken by twenty seven people, five times each, while assessing the neurological state of patients suffering from PD. Subsequently, a software was developed to augment the work of the doctor, in order to provide a quantitative analysis of the treatment results. Huang, Guan Hua, et al. (2020) [13] identified the stages of Parkinson's disease by taking samples of the brain images from 6 healthy and 196 affected subjects. Several machine learning algorithms are used by Senturk, Zehra Karapinar (2020)[14] to diagnose Parkinson's disease. The Support Vector Machines with recursive feature elimination gave the highest accuracy of 93.84%. The present paper attempts to find the best prediction model which distinguishes a PD patient from a healthy patient. Thirteen machine learning based predictive models have been investigated on a dataset comprising biomedical voice measurements.

III. Design Consideration

In order to create a predictive model for the diagnosis of PD a dataset with preprocessed voice recordings has been used to apply various ML models.

A. Dataset for Parkinson's disease:

A dataset has been selected from the University of California, Irvine's public Machine Learning repository. This dataset is composed of a range of biomedical voice measurements from 31 people. Out of 31 subjects, the data related to 8 healthy and 23 Parkinson's disease affected patients has been used for training and testing different predictive models. The dataset contains 4290 samples (195 rows x 22 columns) of biomedical voice measurements [2]. It has 195 vowel voice records taken against 31 subjects, where 147 recordings are taken from patients suffering from Parkinson's disease and 48 recordings from healthy people. Each column of the data file corresponds to an individual voice recording and contains certain attributes and features which are used as inputs in the machine learning algorithm. The voice features are classified into 6 categories i.e. Amplitude, Pulse, Frequency, Voicing pattern, Pitch and Harmonicity.

B. The Input Variables

The input variables contain certain attributes and features of voice samples collected from 31 participants who consented for this study. The dataset contains 22 input variables used for training and testing the thirteen predictive models. These variables are classified into five voice features, like, the amplitude, frequency, harmonicity, complexity and signal scaling parameters. These voice samples are pre-processed so that they can be used as input variables to train the predictive models used in the diagnosis of PD. The efficiency and accuracy of each model is analysed.

The following parameters are considered as input variables for training the machine learning models:

- Six Amplitude parameters representing the local shimmer, local shimmer in dB, 3-point amplitude perturbation quotient, 5-point amplitude perturbation quotient, 11-point amplitude perturbation quotient, average absolute difference between the amplitude of consecutive periods,
- Eleven Frequency parameters representing the average vocal fundamental frequency, maximum vocal fundamental frequency, minimum vocal fundamental frequency, jitter in percentage, Absolute jitter, relative amplitude perturbation, period perturbation quotient, jitter cycles, two nonlinear measures of fundamental frequency variation and pitch period entropy.
- Two harmonicity parameters representing the Harmonic-to-Noise ratio and Noise-to-Harmonic ratio.
- Two Complexity parameters representing, the recurrence period density entropy measure and the Correlation dimension.
- One Signal scaling parameters representing the signal fractal scaling exponent of detrended fluctuation analysis.

C. The Output Variables

The 'Status' variable differentiates healthy people from PD patients. The single binary output variable given in Table 1, indicates the health status of the subject.

TABLE 1 : Health status (output variable)

Variable	Inference
Status (Diagnosis)	1- Parkinson's disease(Positive) 0- Parkinson's disease(Negative)

D. Data preprocessing

Data set contains raw and unstructured data, which has to be converted into an understandable format. Data pre-processing helps to remove inconsistency, incompleteness and the redundancy present in data sets. The data set has been pre-processed in Jupyter notebook (Anaconda IDE) using Python as programming language using the relevant library package.

While pre-processing the data correlation has been checked between the input variable. To do so, the custom delivering color-map and heat-map with a mask and the correct aspect ratio shown in Figure 1, has been generated using python in Jupyter notebook (Anaconda Environment). The scale is taken from - 0.75 to +1.00. It shows that some input variables have more correlation between them.

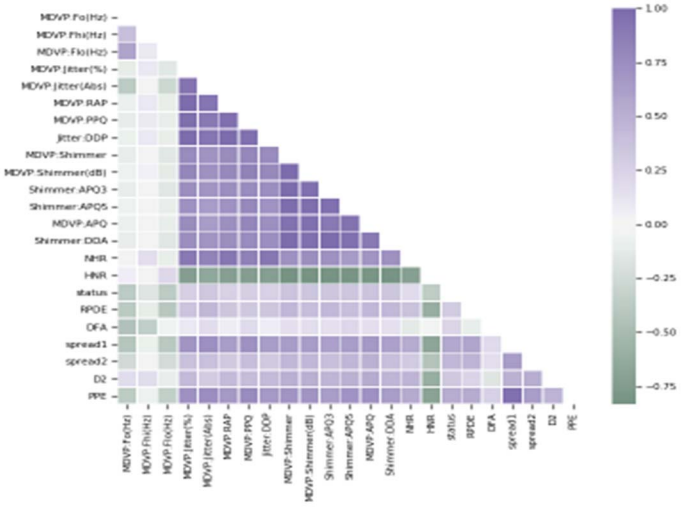

Figure 1: Heat-map

E. *Exploratory Data Analysis (EDA)*

To develop an understanding of data, the EDA has been conducted. It has been observed during the data analysis phase that data is imbalanced i.e. the observation per class is not equally distributed. The plotted graph shown in Figure 2 depicts, that data is imbalanced with the number of negative cases being on the higher side, thus, the ratio being 1:3 between negative and positive cases of Parkinson's disease.

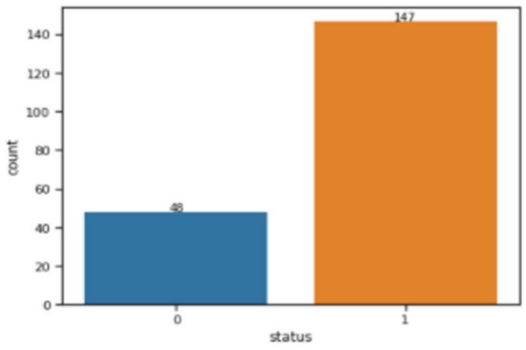

Figure 2: Imbalance datasets

Machine learning algorithms give better results, if the number of samples in each class is equal, as these algorithms are designed to maximize accuracy and reduce errors. Therefore, the dataset has been tested on thirteen different Machine learning models to overcome the imbalance which is present in the dataset due 147 samples of positive cases (recorded by patients suffering from PD) against 48 samples of negative cases (recorded by healthy patients)

IV. Comparison of ML Models

The selected dataset has been used to train and test thirteen different models using python in Jupyter notebook (Anaconda Environment). The detailed observation is give in the Table 2 as follows.

Table 2: Comparison of results

Models	*Train Accuracy*	*Test Accuracy*	*AUC*
logit reg	85.29	89.83	82.441472
knn	86.76	89.83	85.200669
naive bayes	77.94	89.83	78.595318
perceptron	75.74	79.66	53.846154
decision tree	100.00	93.22	95.652174
random forest	100.00	94.92	93.979933
xgb	82.35	89.83	82.441472
lgb	92.65	91.53	89.046823
gradient boost	97.06	93.22	87.374582
bagging	98.53	91.53	83.528428
adaboost	83.09	89.83	82.441472
hard voting	100.00	93.22	90.133779
soft voting	97.06	93.22	84.615385

Out of 13 different machine learning models, five models that offer the best test accuracy are, Random Forest classifier (94.92%), Decision Tree (93.22%), Gradient Boosted Trees (93.22%), Hard Voting (93.22%) and Soft voting (93.22%).

While applying different machine learning models on the UCI dataset, it has been noticed that in case of the XGBoost (XGB) classifier, a weighted feature versus F-score graph is generated which is shown in fig 4. The XGBoost is an improved version of gradient boosted decision trees designed for speed and performance. The XGB model selected 12 out of 22 features as shown in Figure 3.

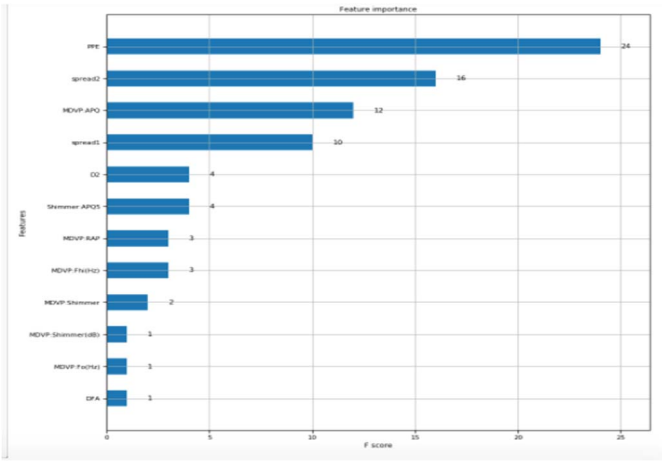

Figure 3: Features Vs. F-Score

The XGB classifier assigned maximum weight as 24 to Pitch Period Entropy (PPE) and minimum weight as 1 to Detrended Fluctuation Analysis (DFA). It shows that out of 22 features only 12 features are important for predicting PD. In addition to this, PPE is the most important feature for the XGB model, which means that the variation in the fundamental frequency component in the voice of the subject, dominates the prediction. Even after feature selection by XGB Model, it has test efficiency of 89.83%, as shown in Figure 4.

The AUC refers to the area under the Receiver Operator Characteristic (ROC) curve as shown in Figure 5. It is a curve drawn with a True Positive rate on the Y-axis and False Positive rate on X-axis with varying the threshold. The ideal value of AUC is 1. The area under the curve measures the tendency of the model to classify cases it is uncertain about as positive or negative. In disease prediction, a false negative is generally regarded as more damaging, as compared to a false positive. False-negative diagnosis results in neglect of the disease, whereas, false positive results in advanced diagnostics, which, may at a later stage reveal the absence of disease.

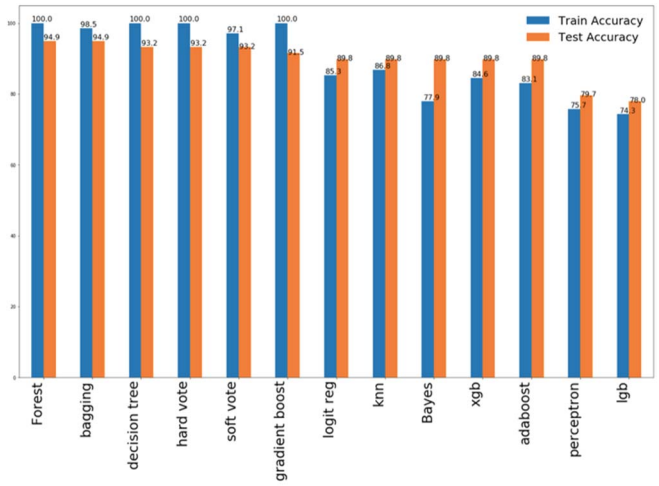

Figure 4: Comparison of Train-Test Results of Models

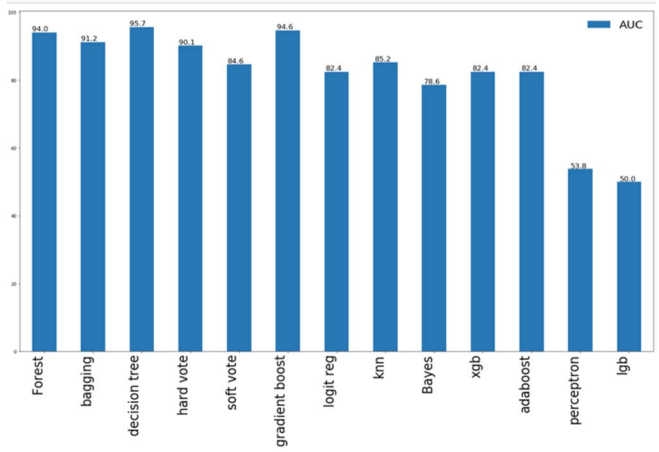

Figure 5: Comparison of overall AUC of Models

V. Performance Analysis

If training time is used as a metric for determining the complexity of a model, the model with the least training time is Stochastic Gradient Descent (SGD), which is justified as it approximates the gradient of the entire dataset by a small subset of data. Also, it is seen that hard and soft voting classifiers take significant training time, which is expected, as they train three different models (logistic regression classifier, random forest classifier with n-estimator=50 and Adaboost classifier) in the particular implementation. Random Forest Classifier, Decision Tree, Light Gradient Boost (LGB) Model and Bagging classifier appear to provide the best performance over train accuracy, test accuracy, AUC and training time, with slight variations in metrics. LGB and Bagging provide faster training time, with reduced accuracy, whereas, Random Forest and Decision Tree provide better accuracy with an increase in training time.

The processing speed analysis graph of all thirteen machine learning algorithms over UCI datasets has been generated in Jupyter notebook (Anaconda IDE) using Python shown in Figure 6. The processing speed analysis of the best five models, such as Random Forest classifier (206.6 ms), Decision Tree (4.5 ms), Gradient Boosted Trees (152.8ms), Hard Voting (314.6ms), and Soft voting (263ms) shows that, the Decision-Tree has the lowest processing time of 4.5 ms and Hard voting model exhibits the highest processing time of 314.6ms. Therefore, the model with the lowest processing speed is most suitable for mobile applications.

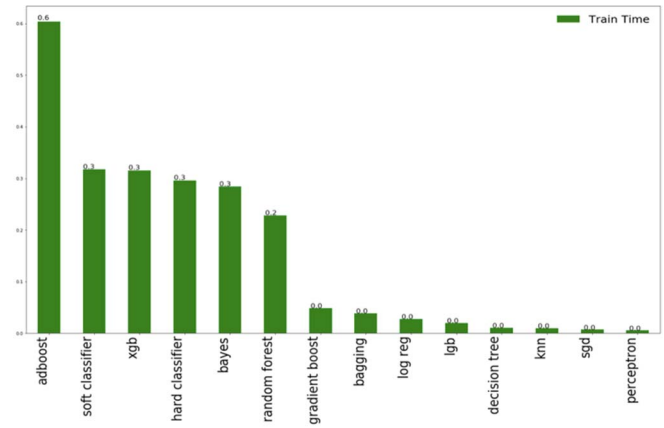

Figure-6: Comparison of processing speed

To show the concern precision and recall in one number the F-Score has been plotted against all thirteen applied models as shown in Figure 7. Overall, it shows the harmonic mean of the model's precision and recall. The F-Score of Random Forest and Bagging models are found to be the same, whereas the F-Score of the rest models slightly differs.

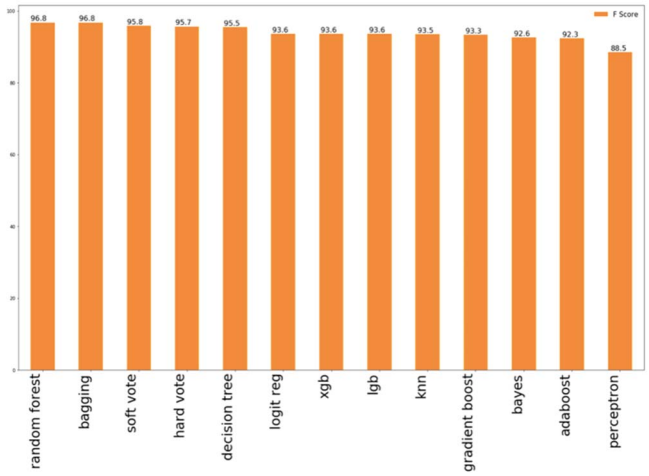

Figure-7: Overall F-Score of Model

VI. CONCLUSION

Use of Machine Learning based Predictive models can make tasks like diagnosing Parkinson's disease more automated, efficient and accurate. Machine learning based models also identify patterns in the data in which may not be noticeable to human beings. It has been observed in the results, that, three models i.e. random forest, decision tree and, hard voting achieved 100% training accuracy, whereas, test accuracy was 6-7% lesser. It shows that the models were slightly over-fitted and decreasing the complexity (bias) of the models may further boost the test accuracy of the models. Further, it has been observed, that, the bagging classifier is also over-fitted, with a training accuracy of 98.5%, however, the test accuracy is only 91.5%. Therefore, decreasing the number of estimators in bagging classifier (n_estimators) may also boost accuracy. The problem of over-fitting as seen in a few models is also resulting due to a small dataset. The number of samples (194x22) are limited, due to only 31 participants volunteering for the study. If the number of participants, are increased, resulting in larger dataset of voice samples, the performance of the predictive models can further be improved.

These models can also be used for the prediction of other diseases related to motor disorders like Alzheimer Disease and can be used for voice pattern analysis for vocal biometrics. Similar machine learning models are being studied in Bangladesh to diagnose Tuberculosis disease by analysing the pattern of cough of a patient recorded using a smart phone.

VII. FUTURE SCOPE

In this paper, a comparison of various machine learning models has been carried out for Parkinson's Disease Prediction. However, using deep learning methods has not been explored for PD prediction to a great extent. In the future, the work can be extended by using the deep learning framework with auto-encoders to reduce the number of features and to extract the most important from them. Also, the UCI dataset used in this work is small and not so complex, therefore, the auto-encoder may not learn well.

The auto-encoders learn well with a large and complex dataset to give better results.

Future work needs to be done in the area of developing compact, light-weight models with higher processing speed to ensure their suitability for mobile applications in the ubiquitous computing environment.

VIII. REFERENCES

[1] "Parkinson's Disease Information Page", NINDS, 30 June 2016. Archived from the original on 4 January 2017. Retrieved 18 July 2016.

[2] Bonato Paolo, et al. "Data mining techniques to detect motor fluctuations in Parkinson's disease.", in Proc. of *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4766-4769, 2004.

[3] Yadav, Geeta, et.al, "Predication of Parkinson's disease using data mining methods: A comparative analysis of tree, statistical and support vector machine classifiers.", in Proc. Of the *National Conference on Computing and Communication Systems*, 2012.

[4] Giulia Fiscon, Emanuel Weitschek. et al. "Alzheimer's disease patients classification through EEG signals processing.", in Proc.of the *IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pp 1-4, 2014.

[5] Kamal Nayan Reddy, Challa, et al. "An improved approach for prediction of Parkinson's disease using machine learning techniques.", in Proc. of the *International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, 2016.

[6] Agarwal Aarushi, Spriha Chandrayan, and Sitanshu S. Sahu. "Prediction of Parkinson's disease using speech signal with Extreme Learning Machine", in Proc. of the *International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pp 1-4, 2016.

[7] Arvind Kumar Tiwari, "Machine learning based approaches for prediction of Parkinson's disease", *Machine Learning Applications*: An International Journal (MLAIJ), vol. 3. Pp. 33-39, (2016)

[8] Sonu, S. R., et al. "Prediction of Parkinson's disease using data mining.", in Proc. of the International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS). Pp. 1082-1085, 2017.

[9] Akshaye Dinesh and Jennifer He, "Using Machine Learning to diagonose Parkinson's Disesase from Voice Recording", in Proceedings of IEEE MIT Undergraduate Research technology Conference (URTC), 2017

[10] Ahmadi Rastegar, D., Ho, N., Halliday, G.M. *et al.* "Parkinson's progression prediction using machine learning and serum cytokines", NPJ *Parkinson's disease*, 5, 14 (2019).

[11] Shahid, Afzal Hussain, and Maheshwari Prasad Singh. "A deep learning approach for prediction of Parkinson's disease progression." *Biomedical Engineering Letters* (2020): 1-13.

[12] Hemmerling, Daria, and Magdalena Wojcik-Pedziwiatr. "Prediction and Estimation of Parkinson's Disease Severity Based on Voice Signal." Volume 34, Issue 5, Pages 651-814,*Journal of Voice* (2020).

[13] Huang, Guan-Hua, et al. "Multiclass machine learning classification of functional brain images for Parkinson's disease stage prediction." *Statistical Analysis and Data Mining: The ASA Data Science Journal* 13.5, Pages 508-523 (2020).

[14] Senturk, Zehra Karapinar. "Early diagnosis of Parkinson's disease using machine learning algorithms." Medical Hypotheses Volume: 138,Article:109603(2020).