Parkinson's Disease detection using Classification

Sanket Pratik Sahoo
Computer Science and Engineering
Silicon Institute of Technology
Bhubaneswar, India
sanket123@gmail.com

K. Gurucharan Reddy
Computer Science and Engineering
Silicon Institute of Technology
Bhubaneswar, India
guru124@gmail.com

Dr. Satyananda Champati Rai
Computer Science and Engineering
Silicon Institute of Technology
Bhubaneswar, India
cr.satya@gmail.com

***Abstract*—A neurological disorder called Parkinson's disease, which impairs mobility and results in bradykinesia, rigidity, and tremors that become progressively more hazardous. The objective the idea is to use cutting-edge machine learning techniques to accurately identify Parkinson's disease from dysphonic speech disorders in Parkinson's patients. The goal of this research is to assess the efficacy of cutting-edge algorithms and determine their accuracy, F1-score, precision, and recall. To determine which machine learning algorithms had the best fit and accuracy overall, both parametric and non-parametric techniques were employed. This study evaluates the performance of methods including the Naive Bayes Algorithm, Logistic Regression, K-Nearest Neighbours, and Random Forest Algorithm in order to better classify Parkinson's disease. The performance is evaluated by pre-processing the data based on speech attributes and delivering it to the machine learning models for additional processing. Various performance metrics, including precision, recall, and F1 score, are generated for each of the four machine learning algorithms. The findings show that classification accuracy obtained by non-parametric models utilising the Random Forest and Naive Bayes algorithms is greater than that of parametric models.**

***Index Terms*—Random Forest, K-Nearest Neighbour, Logistic Regression, and Naive Bayes.**

I. Introduction

Parkinson's disease essentially affects normal bodily motions and posture and is a long-term neuro degenerative movement disorder. With the most prevalent symptoms like tremor, rigidity, slowness of movement, trouble walking, dementia, frequent falling, speech disorder, etc., the symptoms may normally worsen over time. In 1817, Doctor James described Parkinson's illness for the first time. [1].

Since there is now no treatment for the disease, there are numerous research projects in this area. As it is a progressive disease, only medicines may be administered to suppress the symptoms, which is why it affects males more frequently than women. In order to prevent symptoms from getting worse, patients must receive treatment and medication. [2].

Parkinson's disease is brought on by the death of dopamine-producing brain cells, which results in lower levels of dopamine, an important neurotransmitter for regulating movement and coordination. People who have Parkinson's disease consequently could have tremors, stiffness, slow movements, and issues with balance and coordination.Reduced dopamine chemical release in the middle brain is what causes the condition. As a result, certain body parts don't move their muscles. Parkinson's disease progresses faster as dopamine levels drop. With its rating factor, the Unified Parkinson's Disease Rating Scale (UPDRS) is a thorough measurement device used to gauge the disease's progression. Therefore, this UPDRS score is crucial for disease diagnosis. [5].

The major goal of this research is to categorise Parkinson's disease utilising characteristics of patients with PD and healthy controls' dysphonic voices. By contrasting parametric and nonparametric machine learning models, different performance metrics can be applied with cutting-edge machine learning algorithms.

The remainder of the essay is structured as follows. The literature reviews from relevant papers are found in Section II. The preprocessing procedures, planned methodology, and materials and methods associated with data set gathering and description are all included in Section III. A comparison chart and a result analysis are presented in Section IV. The conclusion of this work is presented in Section V, and section V also includes suggestions for future improvement.

II. Related Work

Parkinson's disease prediction employing a support vector machine training and test set of data.

a) : -jim-schwoebel/voice datasets are the name of the datasets.

b) : -Source: the University of Oxford

c) : -195 cases (48 participants without Parkinson's and 147 patients with Parkinson's)

d) : -22 characteristics (components that may be indicative of Parkinson's disease, including as frequency, pitch, and amplitude/period of the sound wave)

e) : -1 label (1 for Parkinson's, 0 for no Parkinson's)

We use a database and divide it into training and test datasets. Using two parametric and two non-parametric machine learning approaches along with a confusion matrix, we then select the algorithm that is most effective for detecting Parkinson's disease. Based on different stages of symptoms, such as healthy control, early stage, intermediate stage, and advanced stage, and their severity, Benmalek et al. discriminate against PD patients. The Patient Voice Analysis (PVA) dataset had 375 participants' voices, which the authors had employed in their study [4]. These features were used as input to train the linear, quadrant, and cubic SVM kernels of the machine learning SVM model. When utilising linear SVM and PCA, greater accuracy is achieved with 92.5%. According to Max A. Little et al. [6], by detecting dysphonia, the authors used a kernel support vector machine to discriminate healthy people from people with Parkinson's disease .

III. MATERIALS AND METHODS

A. Data Collection

The National Centre for Voice and Speech in Colorado and Max Little built the UCI machine learning repository, which is where the voice dataset is found. This dataset includes 195 speech samples from 31 individuals, 23 of whom have Parkinson's disease and 8 of whom are healthy controls. Each patient's six voice samples from the data set have 24 properties.

Each individual provided an average of six phonations, ranging in length from one to 36 seconds. The dataset includes labelled data to distinguish between Parkinson's disease sufferers and healthy controls. 139 of the 195 voice samples are used to train the models, and 56 are used in the testing procedure.

fgure I contains a description of the dataset. The raw data is preprocessed using the system design model to identify Parkinson's illness.The raw data is preprocessed using the system design model to identify Parkinson's illness. figure II lists the characteristics that were taken from voice samples.

Number of Instances	195
Number of Attributes	24
Class Representation	Binary (0 – Control, 1 – PD)

Fig. 1. Data Description.

Frequency	Jitter	Shimmer	Voice Tone	Others
MDVP: Fo(Hz)	MDVP: Jitter(%)	MDVP: Shimmer	Harmonic to Noise Ratio	RPDE(Recurrent Period Density Entropy)
MDVP: Fhi(Hz)	MDVP: Jitter(Abs)	MDVP: Shimmer(dB)	Noise to Harmonic Ratio	DFA,PPE (Pitch Period Entropy)
MDVP: Flo(Hz)	MDVP: RAP	Shimmer: APQ3,APQ5,		
	MDVP:PPQ	MDVP:APQ		

Fig. 2. Voice Features for Parkinsos's Disease [6]

Another method is through clinical assessments, where trained medical professionals evaluate patients for symptoms of Parkinson's disease using standardized criteria. This may include testing for tremors, rigidity, and balance, among other symptoms.

Researchers may also collect data through brain imaging techniques, such as magnetic resonance imaging (MRI) and positron emission tomography (PET), which can help identify changes in the brain associated with Parkinson's disease.

In addition to these methods, researchers may also collect data from wearable devices that monitor movement and activity levels, providing insights into the daily lives of individuals with Parkinson's disease.

B. Pre-processing

Data cleaning, data integration, data reduction, and data transformation are common phases in the preprocessing of data. Data cleaning entails eliminating or fixing any mistakes, discrepancies, or missing data from the dataset. Several datasets are combined into one dataset by data integration. A subset of pertinent variables or observations is chosen in order to reduce the dataset's size. By standardising the size of the variables or encoding categorical categories, data transformation includes transforming the data into a more analytically-friendly format.

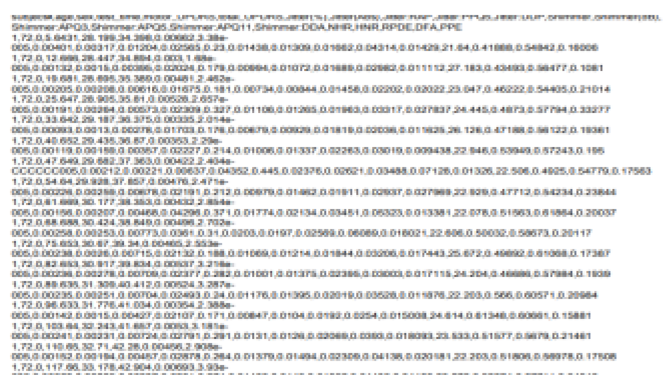


Fig. 1. Data Pre-Processing

After the dataset has been gathered, it is examined to determine the optimum classifier model in order to increase classification accuracy. The initial step is to improve and explain the raw data that was obtained using pre-processing procedures. The most crucial phase in data analysis and mining is data preparation, which is putting the raw data in an organised and comprehensible shape. Real-world data is typically redundant, inconsistent, incomplete, and likely to have noise and inaccuracies.Data pre-processing is a tried-and-true method for overcoming these problems in order to increase the data quality, as shown in Figure given below. This entails using common scalar operations with standardisation and normalisation feature scaling algorithms that employ zero score normalisation and min-max normalisation.variables.

1) *Standardization*

Data standardisation, sometimes referred to as Z-score normalisation, is the process of transforming the values of the dataset's variables so that they have a mean of 0 and a standard deviation of 1. This is accomplished by dividing the result by the variable's standard deviation after deducting each variable's mean from each observation.

Data standardisation is still advantageous even if there are outliers in the data. The values in this are centred using the mean and standard deviation. The mean and standard deviation values are used to complete the standardisation process in Figure given below. The data standardisation has prepared the attributes for the next phase of feature scaling. 13 attributes were chosen among the 23 in this work, and standardised feature scaling was done in R Studio.

Formula for Standardization:

$$x' = \frac{x - \mu}{\sigma} \quad (1)$$

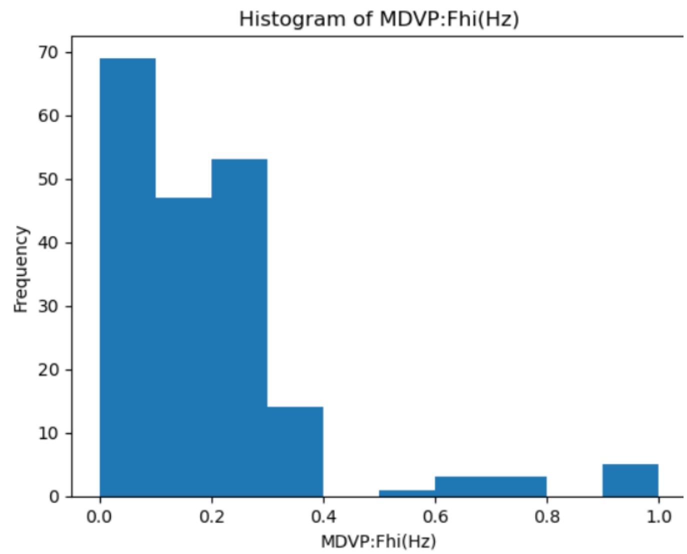
	MDVP.Fo.Hz.	MDVP.Fhi.Hz.	MDVP.Flo.Hz.	MDVP.Jitter...	MDVP.Jitter.Abs.	MDVP.RAP	MDVP.PPQ	Jitter.DDP	MDVP.Shimmer
1	119.992	157.302	74.997	0.00784	7.0e-05	0.00370	0.00554	0.01109	0.04374
2	122.400	148.650	113.819	0.00968	8.0e-05	0.00465	0.00696	0.01394	0.06134
3	116.682	131.111	111.555	0.01050	9.0e-05	0.00544	0.00781	0.01633	0.05233
4	116.676	137.871	111.366	0.00997	9.0e-05	0.00502	0.00698	0.01505	0.05492
5	116.014	141.781	110.655	0.01284	1.1e-04	0.00655	0.00908	0.01966	0.06425
6	120.552	131.162	113.787	0.00968	8.0e-05	0.00463	0.00750	0.01388	0.04701
7	120.267	137.244	114.820	0.00333	3.0e-05	0.00155	0.00202	0.00466	0.01608
8	107.332	113.840	104.315	0.00290	3.0e-05	0.00144	0.00182	0.00431	0.01567
9	95.730	132.068	91.754	0.00551	6.0e-05	0.00293	0.00332	0.00880	0.02093
10	95.056	120.103	91.226	0.00532	6.0e-05	0.00268	0.00332	0.00803	0.02838
11	88.333	112.240	84.072	0.00505	6.0e-05	0.00254	0.00330	0.00763	0.02143
12	91.904	115.871	86.292	0.00540	6.0e-05	0.00281	0.00336	0.00844	0.02752
13	136.926	159.866	131.276	0.00293	2.0e-05	0.00118	0.00153	0.00355	0.01259
14	139.173	179.139	76.556	0.00390	3.0e-05	0.00165	0.00208	0.00496	0.01642
15	152.845	163.305	75.836	0.00294	2.0e-05	0.00121	0.00149	0.00364	0.01828
16	142.167	217.455	83.159	0.00369	3.0e-05	0.00157	0.00203	0.00471	0.01503
17	144.188	349.259	82.764	0.00544	4.0e-05	0.00211	0.00292	0.00632	0.02047
18	168.778	232.181	75.603	0.00718	4.0e-05	0.00284	0.00387	0.00853	0.03327
19	153.046	175.829	68.623	0.00742	5.0e-05	0.00364	0.00432	0.01092	0.05517
20	156.405	189.398	142.822	0.00768	5.0e-05	0.00372	0.00399	0.01116	0.01995
21	153.848	165.738	65.782	0.00840	5.0e-05	0.00428	0.00450	0.01285	0.03810
22	153.880	172.860	78.128	0.00480	3.0e-05	0.00232	0.00267	0.00696	0.04137
23	167.930	193.221	79.068	0.00442	3.0e-05	0.00220	0.00247	0.00661	0.04351
24	173.917	192.735	86.180	0.00476	3.0e-05	0.00221	0.00258	0.00663	0.04192
25	163.656	200.841	76.779	0.00742	5.0e-05	0.00380	0.00390	0.01140	0.01659
26	104.400	206.002	77.968	0.00633	6.0e-05	0.00316	0.00375	0.00948	0.03767
27	171.041	208.313	75.501	0.00455	3.0e-05	0.00250	0.00234	0.00750	0.01966

Fig. 2. Standardization

2) *Normalization*
A data preparation method called normalisation is used to scale the values of a variable to a particular range of 0,1 or -1,1. In order to make the data more comparable and understandable, normalisation aims to change the data such that it lies between a predetermined range, often between 0 and 1.Since the attributes in the data set utilised in this study have a different range of numeric values, this normalisation step is necessary to use. Additionally, it improves data integrity and reduce data redundancy. The feature values are scaled in Figure given below using the min-max normalisation formula used in R Studio with the minimum value of 0 and the highest value of 1. The classifier models receive these normalised attribute values for further processing.

Formula used for Normalization:

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2)$$



C. Proposed Model

Figure 4 shows the system design model for the suggested task. After preprocessing, the dataset is split into training and validation sets and fed into cutting-edge machine learning classification algorithms, such as the Naive Bayes algorithm, Logistic regression, K-Nearest Neighbour, and Random forest, in order to compare the classification accuracy of parametric (Naive Bayes, Logistic regression) and nonparametric (KNN, and Random forest) models. The implementation tool is the R Studio tool. In identifying Parkinson's disease from healthy people using speech features and machine learning approaches, this work emphasises the usefulness and efficiency of parametric and non-parametric models. The method of performance analysis used in earlier works of literature review proved ineffective.

Given that it doesn't require additional expertise, parametric modelling is beneficial in understanding and analysing the findings.

because it needs a set quantity of characteristics or criteria. All parametric models—linear regression, naive bayes, logistic regression, and linear SVM—require a fixed number of features. Additionally, it will function properly with overfitting data. Error minimization strategies perform better when non-parametric modelling is used. nonetheless, does not function well when there are more and more data.It is quite helpful in building the mapping function. All nonparametric models, including KNN, Decision Trees, Support Vector Machines, and Random Forest, increase the amount of the data as the number

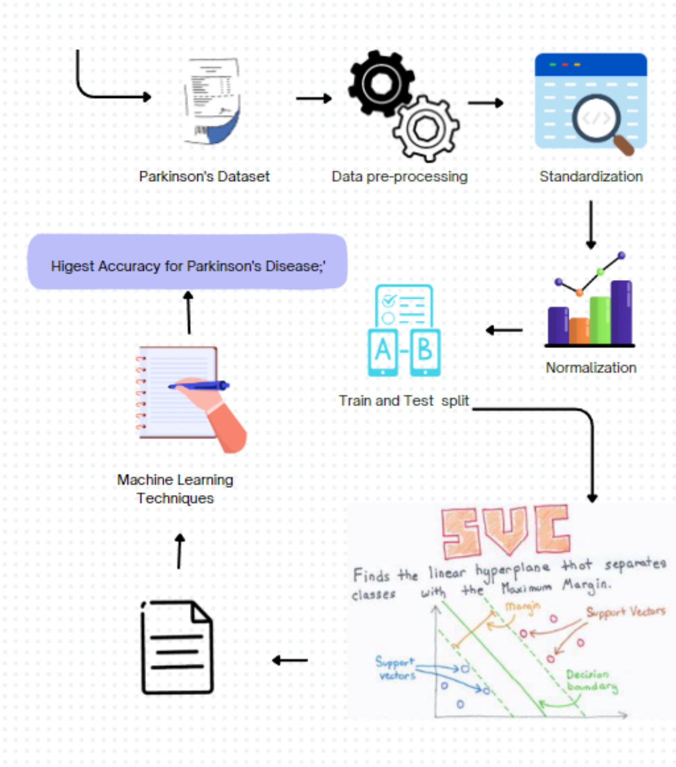


Fig 4. work flow of The Model

of parameters grows. Figure 5 illustrates the classification and evaluation of these using two parametric model algorithms and two nonparametric model algorithms. In predicting Parkinson's disease using speech features, a comparison is conducted.

1) *Parametric Model:-*

a)*Naive Bayes Algorithm*

The Bayes theorem is the foundation for the parametric machine learning technique known as naive Bayes. Through the computation of conditional probabilities and posterior probabilities, it is utilised in numerous classification applications. The posterior probability P(Y—X) can be computed by making a frequency table for each column for the class label. Additionally, the likelihood table is used, and the class with the highest probability determines the outcome of the prediction. This Naive Bayes algorithm offers 71.5% classification accuracy for the used speech dataset, as illustrated in Figure given below:-

b)*Logistic Regression*

Machine learning's "logistic regression" classification method is used to address binary classification challenges. A type of linear regression is the modelling of the probability of a binary result as a function of the input attributes. The outcome of logistic regression is a probability between 0 and 1, which expresses the chance that the input example belongs to the positive class (commonly expressed as 1).Machine learning's classification technique, logistic regression, is used to tackle binary classification jobs. A type of linear regression is the modelling of the probability of a binary result as a function of the input attributes. The logistic regression produces a probability value between 0 and 1 that expresses the likelihood of the input.

2) *Non-Parametric Model:-*

a)*K-Nearest Neighbors Algorithm*

The non-parametric K-Nearest Neighbours approach , we can be utilised in the classification process. This technique uses closely spaced data points that are similar. The first k nearest training data points are in the input. The nearest neighbours are identified using the Euclidean distance function. By using k number (k=5) produces reliable predictions because a lower k value may affect outliers in the model. Figure given below shows that test accuracy on the k most similar training items was 87% after the K-Nearest Neighbours algorithm was fitted to the training set.

b)*Random Forest Algorithm*

It is an ensemble learning technique that may be used with both classification and regression techniques. Several decision trees are used by this classifier on different voice dataset subsets. It leverages majority voting estimates obtained from each decision tree rather than depending on a single decision tree. Out of the entire M variables that are accessible for each node, K variables have been randomly selected. The method determines the prediction of each decision tree for each new data instance, and it then allocates that new data instance to the class with the majority of voters. The Random Forest was used to get the accuracy, and 79.4%.

D. Evaluation Metrices

The type of Performance metrices are:- 1. Precision , 2. Recall , 3.Accuracy , 4. F1-Score The ratio of correctly predicted positive outcomes to all positive outcomes is known as precision. The prediction model performs better the higher the precision.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

The recall is the proportion of accurately predicted positive values to all positive values. The predictive model performs better the higher the recall.

By computing the weighted average of Precision and Recall, the F-Score seeks to strike a balance between the two values.

A key performance matrix is accuracy. By dividing the entire number of data points by 100, it gives the percentage of precisely anticipated data. It works properly if the data points for each label are distributed equally. The accuracy is

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\text{F1-Score} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

True Positive

the ratio of accurate predictions to all other predictions, and it is expressed as a percentage between 0% and 100%.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

IV. RESULTS AND DISCUSSION

The results of both parametric and non-parametric models are evaluated based on various performance measures given in equation (3), (4), (5) and (6). As a result of Parkinson's data speech feature classification, naive Bayes and logistic regression of parametric models produce 61.5% and 79.03% classification accuracy, respectively.

1) Comparison between Naive Bayes Algorithm and Logistic Regression:-:

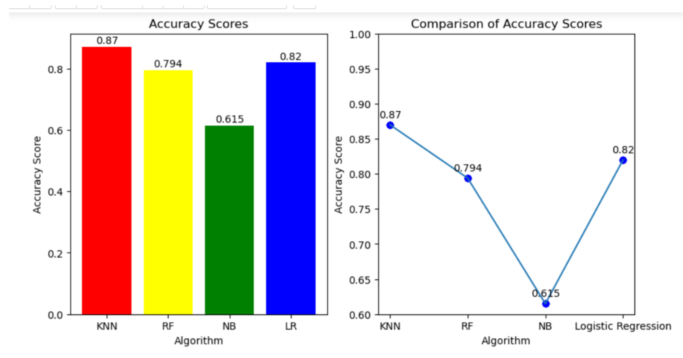
	accuracy	precision	recall	f-measure
Naive Bayes	0.615	0.866	0.615	0.647
LR	0.820	0.829	0.615	0.647

However, the nonparametric model uses two algorithms, KNN and Random forest, to classify Parkinson's disease, with predictions of 87.27% and 90.2%, respectively. Based on voice features, nonparametric models perform better in terms of classification accuracy for diagnosing Parkinson's illness. Figure 8 shows the classification accuracy of the KNearest Neighbour algorithm, while Table below provides performance metrics for non-parametric models.

2) Comparison between Knn and Ran dom forest:-:

	accuracy	precision	recall	f-measure
Knn	0.769	0.802	0.769	0.781
RF	0.820	0.928	0.838	0.881

The main objective, which is to compare the classification accuracy of parametric models with non-parametric models for classifying Parkinson's disease with healthy controls, is illustrated in given figure below.



V. CONCLUSION

For separating Parkinson's disease from healthy controls and Parkinson's disease, the accuracy of parametric models and non-parametric models is The benchmarking speech dataset results using state-of-the-art machine learning methods of parametric and non-parametric models lead to the conclusion that both parametric and non-parametric model algorithms perform well depending on the kind and size of the dataset. This study demonstrates that non-parametric model approaches like Random Forest and K-Nearest Neighbour algorithm offer greater classification accuracy when compared to Logistic Regression and Random Forest of parametric models. The non-parametric model in this study provides the best match to the voice dataset.

VI. FUTURE WORK

Future research might concentrate on enhancing the precision of Parkinson's disease identification by utilising machine learning methods and Ada boost. This may be done by merging several machine learning techniques with additional biomarkers, such as gait analyses, eye movements, and EEG recordings.

approaches for building more reliable models. Future research might also concentrate on creating more approachable and available tools for managing and diagnosing Parkinson's disease utilising machine learning approaches.

ACKNOWLEDGMENT

We would like to sincerely thank everyone who has helped with our effort on utilising classification approaches to identify Parkinson's disease.

First and foremost, we would want to express our gratitude to our project mentor for his direction, assistance, and knowledge during the project. Their comments and recommendations were really helpful in creating our study and technique.

We also like to express our gratitude to the research participants who gave their time and data freely for this study. We are grateful to them for agreeing to take part in our study because their involvement was essential to its success..

We would like to thank the many research papers and studies that contributed to the background information needed for our investigation. These sites were crucial in advancing our knowledge of Parkinson's disease and methods for categorization.

We want to thank everyone who has helped with our initiative from the bottom of our hearts once more. Our study is intended to support continuing attempts to use classification methods to identify Parkinson's disease.

References

[1] Parkinson's disease: present and upcoming difficulties, Langston, J. William. pp. 443–450 in Neurotoxicology 23.4-5 (2002).

[2] A survey of Parkinson's disease patients: the most annoying symptoms and preferred coping strategies, Uebelacker, Lisa A., et al. Journal of Parkinson's disease 4.4, 717–723, 2014.

[3] Varghese, Basil K., D. Amali, and K. S. Devi. "Prediction of Parkinson's Disease using Machine Learning Techniques on Speech dataset." Research Journal of Pharmacy and Technology 12.2 (2019): 644-648.

[4] Voice Analysis (PVA) Synapse ID: syn2321745 https://www.synapse.org

[5] "Features selection for building an early diagnosis machine learning model for Parkinson's disease," by Abu Bakr Soliman et al. AIPR 2016, the third international conference on artificial intelligence and pattern recognition. IEEE, 2016.

[6] "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," by Max A. Little et al. IEEE Transactions on 56.4 (2009): 1015–1022. Biomedical Engineering.