Parkinson's Disease Detection Using Machine Learning

M.S. Roobini[1], Yaragundla Rajesh Kumar Reddy[2], Udayagiri Sushmanth Girish Royal[3], Amandeep Singh K[4], Babu.K[5]

[1,2,3,4,5]Sathyabama Institute of Science and Technology, Chennai-600127, India.

roobinims@gmail.com, yaragundlarajesh5@gmail.com[2], pacharlanarayana1957@gmail.com[3], amandeepsingh.cse@sathyabama.ac.in[4], babu.cse@sathyabama.ac.in[5]

Abstract— **Firstly, Parkinson delineates Parkinson's sickness as a neurologic syndrome, it affects the central system, as a result, the patients face difficulty talking, strolling, tremor throughout the motion. Parkinson's sickness patient generally encompasses a low-volume noise with a monotone quality, this method explores the classification of audio signals feature dataset to diagnose Parkinson's sickness (PD), the classifiers we tend to use during this system area unit from Machine Learning. Our model tends to utilize provision regression and XGboost classifiers, and therefore the audio feature dataset from the UCI dataset repository. The system has achieved a much better end up in predicting the palladium patient is healthy or not, XGBoost provided the height accuracy of 96% and therefore the Matthews parametric statistic (MCC) of 89%.**

Keywords— Parkinson's Disease (PD), Extreme Gradient Boosting (XGBoost), Matthews Correlation Coefficient (MCC), Logistic regression (LR)

I. INTRODUCTION

Parkinson's sickness is represented as a neurodegeneration problem that is the death of Dopastat generating cells [1]. The loss of dopaminergic neurons within the neural structure decreases the accomplishable rate of communication [2]. Parkinson's sickness affects the central system that ends up in the impact on the motor system, the most metal symptoms square measure tremors, rigidity, and movement disorders. The those that square measure having Parkinson's sickness largely ninetieth of them have a speech impairment, solely three-dimensional to four-dimensional of the metal patient receives therapy and additionally, only 1 of the foremost necessary issue for metal is age, the patient with metal square measure most of them square measure aged between 45-60[3]. As the speech of metal patients have modification within the frequency specter in their voice as a result of the loss of the management of the limb, that decreases the frequency of the audio. So, the low-frequency region offers necessary information to differentiate the speech impairments in metal. Unified Parkinson sickness rating scale (UPDRS) is employed to search out the severity of the metal by facilitating clinical experience and experience[4].Here it was performed a feature choice for the audio options dataset created by GHB very little of the University of Oxford[5], the high

prediction has been achieved with classification accuracy, the formula predicts varied accuracy for varied variables that square measure relevant on the opposite attributed gift within the feature dataset, as feature plays a necessary role the dataset that we tend to take from UCI repository contains twenty-one options and applied a Pearson's constant coefficient of correlation parametric statistic on a feature to see the coefficient correlation among options. This section presents the comparative determination endeavors, here each model-based and model-free techniques algorithms square measure used for predicting Parkinson's sickness. Model-based approach heavily depends upon the prior applied math statements, like the relation between variables.

The most typically used model-based tool is logistical regression [6], which measures the result on a binary scale (e.g., healthy/not), here the classification method applied supported the estimation chances. Whereas model-free strategies like XGBoost [7] adapt to the intrinsic information characteristic with a prior model. We tend to use the XGBoost formula for classification, XGBoost formula like constant learning or preparation, doesn't guarantee optimized classification/regression. However, once trained and maintained, XGBoost. learning technique has more potential than logistical regression in resolving universe issues. The previous report of mistreatment XGBoost technique to diagnoses Parkinson's sickness square measure determines in step with their psychological feature standing [8]. XGBoost provided associate degree accuracy of ninety-six for classification the dataset associates degreed logistical regression provided an accuracy of seventy-nine, this technique that predicts metal has been developed that compares the accuracy of LR and XGBoost on the train and take a look at the dataset. It utilized co-efficient correlation to search out the correlation among options, on comparison it only if XGBoost performed higher than LR with an accuracy of ninety-six [9].

II. LITERATURE SURVEY

Surendra Bikram Thapa, Surabhi Adhikari [7] proposed that PD became a significant downside for a large fraction of individuals higher than sixty. The sickness affects the system and might result in death. We don't have any cure for the sickness that makes diagnosing harder. bound symptoms like unintelligible speech. Parkinsonism (PD)

affects the patient's ability to maneuver and is caused generally attacked by loss of nerve cells within some part of the brain known as substantia nigral. As mentioned earlier, bound symptoms seem ahead of effects within the motor system. Here used easy drawing options for building AN mil model for Pd. options were elite employing a detection formula and pulse deterioration to attain ninety-two% of accuracy. Yuxin Lin, Nuo Xu, Ringo Lam, and co-authors [5] examine the bio-electronic stimulation medical aid to eight patients laid low with Parkinsonism via the resting tremor signals. The signals area unit depicted within the frequency domain via Fourier transforms and area unit classified into 2 teams by the ratios of the frequencies within the bands 3-8Hz. Parkinsonism is the second commonest neurodegenerative disorder within the world, with countless folks, laid low with it. The incidences of shaking palsy within our area unit concern zero.8 per one hundred folks per annum within the 0-29 cohort and three per one hundred folks per annum within the 30-49 cohort. The diagnosis of Parkinsonism is predicated on a patient's physical, behavioral, and spirit still because of the chemical make-up of their brain. The bio-electronic stimulation medical aid to treat Parkinsonism is used, however, there's no conclusive proof whether or not or not it works. It aims to build AN analysis of the medical aid on patients laid low with the sickness. The resting tremor is one in every of the foremost classical tremors, that is AN involuntary, measured, and measured A movement that may affect the complete body. A paper by Shrinidhi Kulkarni, Neenu George Kalayil, and co-authors [4] proposed a technique for police work Parkinsonism through human smell signatures. The values of those parts area unit accustomed verify whether or not an individual is laid low with the sickness at AN early stage.

A Central system (CNS) could be a unit of the brain that transmits signals from the brain to the muscles in response to external stimuli. The projected Parkinsonism system focuses on motor and non-motor symptoms, that area unit early visible signs of the sickness. Adriano A. G. Siqueira, Juan C. Pérez-Ibarra [15] During this paper, we tend to propose an associate rule that used linear classifiers to discover the transition between gait section and event in a period. mechanical phenomenon sensors square measure sturdy, cheap, small, simple to position, and information concerning human gait that may be accustomed to determine gait patterns and discover events.Iqra Nissar, Tawseef Ayoub Shaikh, cake Ghulam crowned head Badshah, and Waseem AhmadMir[13]. Voice analysis is one of every of the foremost vital tools within the designation of Parkinson's syndrome. Non-invasive treatments for voice analysis are on the market that improves the standard of a patient's life. Deep learning and telemonitoring techniques are used as a part of the analysis of metal designation. The results obtained from totally different analyses counsel that computing is turning into a strong learning tool with a lot to supply to information scientists and neurologists is caused by the loss of bound neuron clusters that manufacture neurotransmitters like Intropin and monoamine neurotransmitters. Speech and vocal issues are among the primary symptoms that may be seen. Voice mensuration will be accustomed to track the progression of Parkinson's syndrome (PD). Voice mensuration is non-invasive and non-contact. varied drug therapies are on the market that supply vital mitigation of symptoms, particularly at its earlier stages. Selecting the proper classification model is a vital issue that must be maneuvered within the designation of Parkinson's syndrome. This paper tries to explore the various machine learning and AI techniques that are used for the designation of metal.

The conclusion is conferred in section four. Max A. very little et al steered a unique technique for classifying the topics into metal and management subjects. the information for his or her study was collected from thirty-one individuals (23 of whom were mental patients). This article [11] Hanbin Zhang and their co-authors has been accepted for publication in an exceedingly issue of IEEE Reviews, however not emended. Personal use is allowable, Patients struck by brain disorder (PD) enjoy early detection and treatment. Pd could be a progressive unwellness wherever the symptoms grow bit by bit however doesn't seem to be obvious at the first stage. A smartphone with constitutional sensors will give access to predict the danger of developing brain disorder (PD) in an exceedingly non-clinical atmosphere There is no comprehensive survey of mHealth on Pd detection. Some researchers solely specialize in symptoms like Lieber. Daniel Palacios-Alonso and their co-authors [10] Speech has relevancy as a conveyance mechanism to observe medical specialty evolution victimization well-known options extracted from sustained phonations or running.

Analysis and watches take into account, however, the disease influences speech. Speech is often used as a customary for following atomic number 46 within the same approach that proof shows that speech will facilitate live motor symptoms. it's necessary to form multidisciplinary groups to supply help to patients with progressive medical specialty impairment. A new system has been planned for analyzing drawing patterns in patients affected by shaking palsy. It is created by Inje University, an Asian country that combines 2 differing kinds of computer science (AI) techniques. Fig. 1(a), Fig. 1(b), Fig. 1(c), shows the shaking palsy (PD) is typically determined in patients at totally different stages of the malady. historically sketching or handwriting was performed. Newspapers are understood manually by the interpreters. a number of the common options gift within the sketches. They determined that there was a big distinction within the correlation issue between the options and severity of the palladium at totally different stages. The study was done by Kotsavasiloglou[9] et. al. mistreatment the flight of the pen tip on the pad moreover as numerous machine learning algorithms.
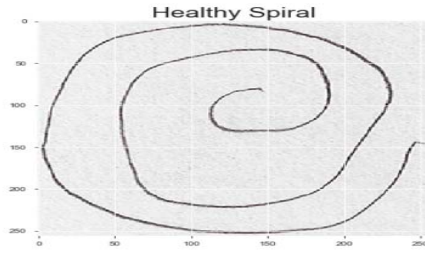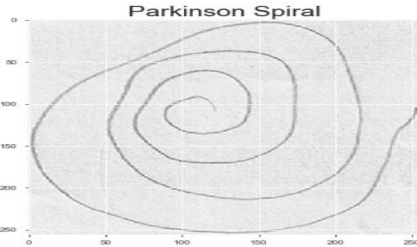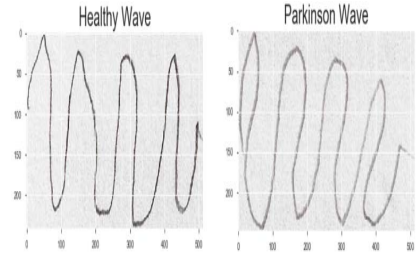
Fig. 1(a)

Fig. 1(b)

Fig. 1(c)

Dr. Pooja Raundale, Chetan Thosar, Shardul Rane [8] proposed that we've prompted a technique for the prediction of shaking palsy severity exploitation machine learning and deep neural networks. Neural networks were found to be the foremost systematic classifier and regression rule compared to the machine learning regression rule and the call tree. the target of this analysis was to predict if the victim has to shake palsy or not employ a keyboard writing check, and if he has then we tend to conjointly foretell the severity of the malady. Aims to develop a software system capable of following the prevalence and intensity of. neurological disease at 1-second resolution by analyzing the info gathered from shaking palsy (PD) patients. Results show that despite. different approaches to dynamic pattern classification, dynamic support-vector. and hidden Andre Markoff models were. Respectively equally economical in maintaining dynamic pattern observation. below 10%.

III.   METHODOLOGY

In this system, we tend to apply 2 machine learning algorithms that measure supplying Regression and XGBoost. we tend to enforce this model to search out the simplest model among them for the datasets.

A. LOGISTIC REGRESSION:

Logistic Regression is wont to justify the link between one or a lot of freelance variables and one dependent binary variable, the variable should be binary (e.g., zero or 1) [12]. They shouldn't be a high correlation among the prediction,

this could be assessed by a matrix, wherever the result has 2 categories, supplying regression starts with a special model setup than regression rather than modeling Y as a function of X directly, we have a tendency to model the chance that's adequate category one, gives X. First, abbreviation P(X)=P(Y=1/X). Then the supplying model is:

$$p(x) = \frac{\exp\exp\left(\beta^T x\right)}{1 + \exp\exp\left(\beta^T x\right)}$$

i.e., the plot $(e^a/1 + e^a)$ as a

$$\log\log\left(\frac{p(x)}{+p(x)}\right) = \beta^T x \quad \text{--------------------(1)}$$

Where

$$\beta^T x = \beta_0 + \beta_1 x_{i2} + \beta_1 x_{i3} + \ldots + \beta_p x_{in}$$

.

B. XGBoost:

XGBoost is a boosting algorithmic program, it's applied math learning methodology and derived from gradient boosting call tree [16], it's higher performance and improvement. the explanation why we tend to use XGBoost is its sensible potency and feasibleness, XGBoost permits dense and distributed matrix because the input and a numeric vector use number ranging from zero for classification, we will add a range of iteration to the model A dataset with of n samples and d options of each sample then s_k is that the prediction from call tree.

$$D = \{(x1, y1)\}(|D| = n_1 X1 \in R_\alpha^d, y_1 \in R \text{----------(2)}$$

Mathematically, our model in the form

$$y^\wedge = \phi\left(x_i\right) = \sum_1^k s_k(x_i), s_k \in s \quad \text{---------(3)}$$

(3)

Where k is the n of trees, s function in function space s. The objective function is given by

$$L(\emptyset) = \sum_i l\left(y^\wedge, y_1\right) + \sum_k \Omega\left(F_K\right)\text{---------- (4)}$$

Here, l is the loss function and it measures the prediction error.

C. Evaluation Metrics:

The success of the classifier measured supported Accuracy, Sensitivity, Specificity, and MCC. The formula for accuracy is quantitative relation of properly classified instance to the whole instance

$$\text{Accuracy} = \frac{Tp + TN}{TP + Fp + TN + FN} \quad \text{-------(5)}$$

Here TP, FP, FN are truly positive, true negative, false positive, false negative respectively, and Sensitivity and Specificity are:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad \text{------------(6)}$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad \text{------------(7)}$$

MCC (Matthews Correlation Coefficient) shows the quality of binary classification. Correlation Coefficient between the predicted and binary classification is MCC

$$\text{MCC} = \frac{Tp_x TN - FP_x FN}{\sqrt{(TP+Fp)(TP+FN)(TN+FP(TN+FN)}} \quad (8)$$

D. DATA PREPROCESSING:

MinMaxScaler, normalizer is methodology in scikit-learn are preprocessing ways, supported our options values we tend to choose the strategy, as we all know machine learning rule can perform higher and quicker once options are relative or similar scale, we recommend MinMaxScale () for preprocessing because it subtracts the minimum price in feature and divides with it vary, the distinction of most and minimum is varied MinMaxScale come to the default vary zero to 1.

E. FEATURE SELECTION

As we know, options play a vital role in classification [17], there square measure completely different approaches in feature choice and supported the edge worth and benchmark formula we tend to verify the optimality of feature within the dataset. parametric statistic options choice is that the most generally used parameter, as a result of feature choice is predicated on their correlation issue among the options [18], Let suppose f1 and f2 square measure 2 related to options than to search out Pearson's parametric statistic (p)

$$\rho xy = \frac{(x,y]}{x \sim y \sim} \quad (9)$$

Where, cov(x,y) are covariance of variable x and y

$$R = \frac{n(\sum xy) - (\sum_x)(\sum y)}{\sqrt{\left[n\Sigma x^2 - (\sum x)^2\right]\left[n\Sigma y^z - (\sum y)^2\right]}} \quad (10)$$

Fig. 2 shows the correlation between options, it differentiates the sturdy positive and indirect correlation among options, that we tend to elect a couple of vital options, MDVP: Shimmer (DB), MDVP: APQ, Shimmer: DDA, MDVP: RAP, MDVP: PPQ, NHR MDVP: Jitter (%), Jitter (Abs), Jitter: DPP, MDVP: Shimmer, Shimmer APQ3, Shimmer APQ5, these square measure options from the dataset for the classification and also the classification is created on these selected options for higher accuracy.
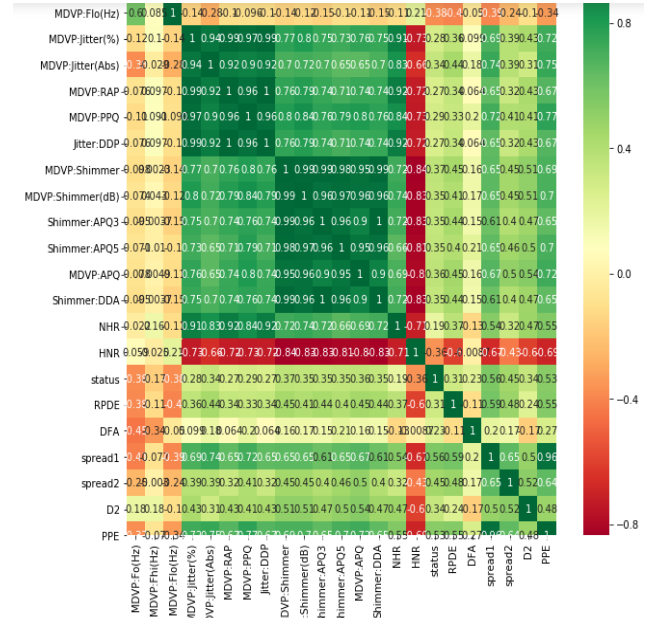

Fig. 2 Cross-correlation among features

IV. RESULTS

By examination of the output of the system, the most classification rate is achieved by XGBoost than LR with AN accuracy of ninety-six and MCC of eighty-nine, whereas LR achieved solely seventy-nine accuracies and MCC of forty seconds. Fig. 3 shows the bar graph of the classifier's accuracy and Fig. 2 shows that only twelve options squares

measure the foremost necessary and characteristic among different options gift within the dataset.
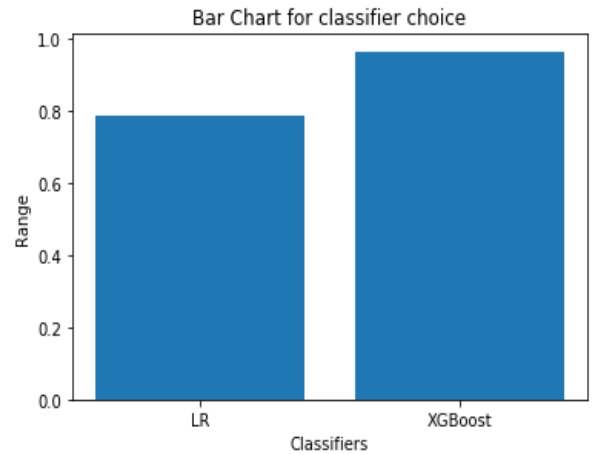

Fig. 3 Classifiers accuracy

Fig. 4 shows the K-Nearest Neighbors accuracy. KNN just classifies a data point based on its few nearest neighbors. It's important to find an optimal value of K, such that the model is able to classify well on the test data.
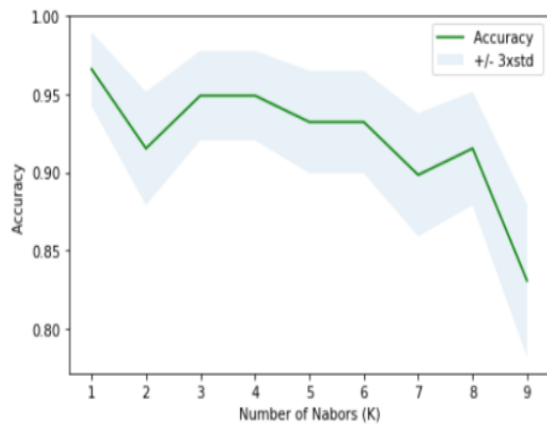
Fig. 4 KNN accuracy

Fig. 5 Shows the results of Data Pre-Processsing MinMaxScaler, normalizer is methodology in scikit-learn are preprocessing ways, supported our options values we tend to choose the strategy, as we all know machine learning rule can perform higher and quicker once options are relative or similar scale, we recommend MinMaxScale () for preprocessing because it subtracts the minimum price in feature and divides with it vary, the distinction of most and minimum is varied MinMaxScale come to the default vary zero to 1.
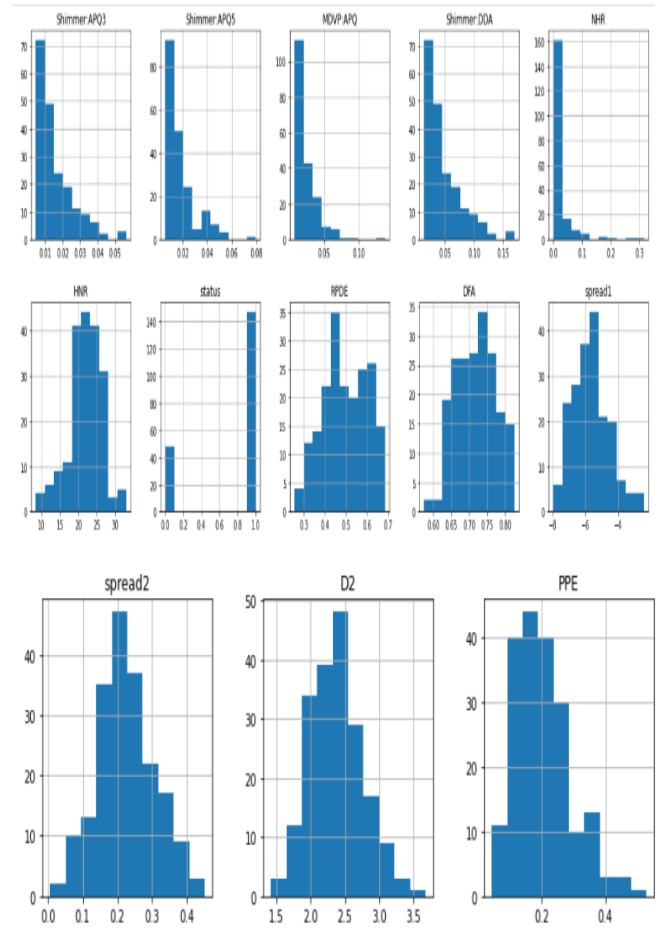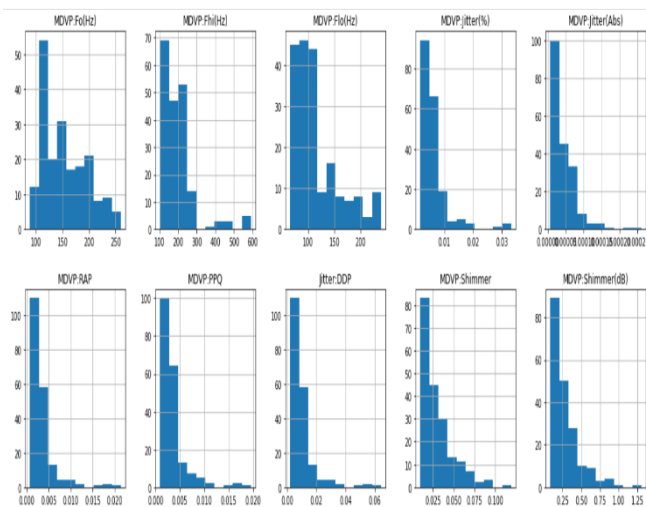

Fig. 5: Data Preprocessor

Fig. 6 Shows the Accuracy of different algorithms which we have used to get the best accuracy to predict parkinson's disease.

	Algorithm	Accuracy	MCC
0	KNN	0.932203	0.824775
1	Decision Tree	0.821429	0.572418
2	Logistic Regression	0.796610	0.425447
3	XGBoost	0.928571	0.781736

Fig. 6: Accuracy

V. CONCLUSION

Vocal-based Parkinson's malady detection is to investigate that rule provides the high accuracy of prediction for the Parkinson's malady dataset, here the classification accuracy was studied and compared, with sensible performance and quick implementation XGBoost achieved a high accuracy

with ninety-six. this technique provides the comparison between machine learning classifiers of LR and XGBoost in atomic number 46 malady identification with high dimensional knowledge.

REFERENCES

[1] Niya Romy Markose, Priscilla Dinkar Moyya, Mythili Asaithambi, "Analysis of tremors in Parkinson's Disease using accelerometer," IEEE,4th July 2021.

[2] Oliver Y.chen, Florian lipsmeier, Huy Phan, and John prince, "Building a Machine-learning Framework to Remotely Assess Parkinson's Disease Using Smartphones," IEEE, May 16th, 2020.

[3] Mohamed Shaban, "Deep Convolutional Neural Network for Parkinson's Disease Based Handwriting Screening," IEEE, Sep 12th, 2020.

[4] Shrinidhi Kulkarni, Neenu George Kalayil, Jinu James, Sneha Parsewar and Revati Shriram, "Detection of Parkinson's Disease through Smell Signatures," IEEE, Sep 10th, 2020.

[5] Yuxin Lin, Bingo Wing-Kuen Ling, Nuo Xu, Ringo Lam, and Charlotte Ho, "Effectiveness Analysis of Bio-electronic Stimulation Therapy to Parkinson's Disease via Discrete Fourier Transform Approach," IEEE, May 17th, 2021.

[6] Yuqi Qiu, "Efficient Pre-diagnosis Approach for Parkinson's Disease with Machine Learning," IEEE, Nov 2nd, 2020.

[7] Surendrabikram Thapa, Surabhi Adhikari, Awishkar Ghimire, and Anshuman Aditya, "Feature Selection Based Twin-Support Vector Machine for the Diagnosis of Parkinson's Disease," IEEE, June 20th, 2021.

[8] Dr. Pooja Raundale, Chetan Thosar, Shardul Rane, "Prediction of Parkinson's disease and severity of the disease using Machine Learning and Deep Learning algorithm," IEEE, Aug 12th,2021.

[9] Sabyasachi Chakraborty, Satyabrata Aich, Jong-Seong-Sim, Eunyoung Han, Jinse Park, Hee-Cheol Kim, "Parkinson's Disease Detection from Spiral and Wave Drawings using Convolutional Neural Networks: A Multistage Classifier Approach," IEEE, June 4th, 2020.

[10] Daniel Palacios-Alonso, Guillermo Melendez-Morales, Agustin Lopez-Arribas, Carlos Lazaro-Carrascosa, Andres Gomez-Rodellar, and Pedro Gomez-Vilda "MonParLoc: A Speech-Based System for Parkinson's Disease Analysis and Monitoring, IEEE, Oct 27th, 2020.

[11] Hanbin Zhang, Chen Song, Aditya Singh Rathore, Ming-Chun Huang, Yuan Zhang, Wenyao Xu, "mHealth Technologies towards Parkinson's Disease Detection and Monitoring in Daily Life," IEEE, May 31, 2020.

[12] Carlo Ricciardi, Marianna Amboni, Chiara De Santis, Gianluca Ricciardelli, Giovanni Improta, Sofia Cuoco, Marina Picillo, Paolo Barone, "Machine learning can detect the presence of Mild cognitive impairment in patients affected by Parkinson's Disease," IEEE, Aug 15th, 2020.

[13] Iqra Nissar, Waseem Ahmad Mir, Azharuddin, Tawseef Ayoub Shaikh, "Machine Learning Approaches for Detection and Diagnosis of Parkinson's Disease-," IEEE, July 07, 2021.

[14] Protima khan1, MD. fall Kader, s. m. brazil Islam, Aisha b. Rahman, MD. Shahriar Kamal, Misbah Uddin to, and Kyung-sup Kwak, "Machine Learning and Deep Learning Approaches for Brain Disease Diagnosis: Principles and Recent Advances," IEEE, March 11th, 2021.

[15] Juan C. Pérez-Ibarra, Adriano A. G. Siqueira, Member, IEEE, and Hermano I. Krebs, Fellow, "Identification of Gait Events in Healthy and Parkinson's Disease Subjects using Inertial Sensors: A Supervised Learning Approach," IEEE, Aug 25, 2020.