EasyChair Preprint

№ 4889

Early Prediction of Parkinson Disease Using Machine Learning and Deep Learning Approaches

Harshvardhan Tiwari, Shiji K Shridhar, Preeti V Patil,
K R Sinchana and G Aishwarya

January 12, 2021

EARLY PREDICTION OF PARKINSON DISEASE USING MACHINE LEARNING AND DEEP LEARNING APPROACHES

HARSHVARDHAN TIWARI*
*Centre for Incubation, Innovation, Research and Consultancy,
Jyothy Institute of Technology, Bengaluru, Karnataka, India
tiwari.harshvardhan@gmail.com

SHIJI K SHRIDHAR
Department of Information Science and Engineering,
Jyothy Institute Of Technology,Bengaluru, Karnataka, India
shijiks20@gmail.com.

PREETI V PATIL
Department of Information Science and Engineering,
Jyothy Institute Of Technology,Bengaluru, Karnataka, India
preetivp2004@gmail.com

SINCHANA K R
Department of Information Science and Engineering,
Jyothy Insititute of technology,Bengaluru, Karnataka, India
sinchanakr1207@gmail.com

AISHWARYA G
Department of Information Science and Engineering,
Jyothy Insititute of technology,Bengaluru, Karnataka, India
gaishwarya03@gmail.com

A B S T R A C T

Parkinson disease(PD), the second most common neurological disorder that causes significant disability, reduces the quality of life and has no cure. Nerve cells in this part of the brain are responsible for producing a chemical called dopamine. Dopamine acts as a message between the parts of the brain and nervous system that help control and co-ordinate body movements.As dopamine generally neurons in the parts begin to experience difficulty in speaking, writing, walking or completing other simple task.Approximately, 90% affected people with Parkinson have speech disorders. The average age of onset is about 70 years, and the incidence rises significantly with advancing age. However, a small percent of people with PD have "early-onset" disease that begins before the age of 50.More than 10 million people worldwide are living with PD. No cure for PD exists today, but research is ongoing and medications or surgery can often provide substantial improvement with motor symptoms.

Parkinson disease is one of the most serious diseases. Hence diagnosing it at an earlier stage could help prevent or reduce the effects. The machine learning classification algorithms are used to predict if a person has Parkinson disease or not, comparing different machine learning algorithm such as logistic regression, decision tree, k-nearest neighbour as well as some "Ensemble" learning techniques where we attempt to improve the accuracy by combining several models.The machine learning model can be implemented to significantly improve diagnosis method of Parkinson disease.In this study it indicates that the ensemble techniques Xgboost classification (Extreme gradient boosting) algorithm achieved the high test accuracy rate (95%) compared to other classification algorithm.The performance of the methods has been assessed with a reliable dataset from UCI Machine learning repository.

KEYWORDS: Parkinson's disease, Machine learning ,Parkinson's dataset,Classification algorithm,Xgboost

1. INTRODUCTION

Parkinson's disease (PD) is a progressive neurodegenerative condition leading to the death of the dopamine (**d**i-**o**rtho-**p**henyl-**a**lanine)-containing cells of the substantia nigra. There is no consistently reliable test that can distinguish PD from other conditions with similar clinical presentations. The diagnosis is primarily a clinical one based on a history and examination [1].Parkinson's disease is named after the British doctor who wrote the first book about the disease, in 1817, that made it an easily recognized entity. Parkinson called it, "The Shaking Palsy," or "paralysis agitans." In his day, the term "agitans" referred to tremors. "Palsy" meant weakness and "paralysis" meant paralyzed, so the condition was consider a disorder of weakness and tremors, which is not completely true, as we shall see.It is a chronic, progressive neurodegenerative disease characterized by both motor and non_motor features. The motor symptoms of PD are attributed to the loss of striatal dopaminergic neurons, although the presence of non motor symptoms supports neuronal loss in non dopaminergic areas as well. The term "parkinsonism"means "looks like parkison's disease." To neurologists this means that the person has a somewhat flexed posture, moves slowly, is stiff and usually walks slowly, with small steps and reduced or no arm swing. PD is the most common cause of parkinsonism, although a number of secondary causes also exist, including diseases that mimic PD and drug-induced causes.[2].There is no single test which can administered for diagnosis.Instead, doctors must perform a careful clinical analysis of the patient's medical history.Unfortunately, this method of diagnosis is highly inaccurate.A study from the National Institute of Neurological Disorders find that early diagnosis(having symptoms for 5 years or less) is only 53% accurate.This is not much better than random guessing,but an early diagnosis is critical to effective treatment.The number of people suffering from PD has increased rapidly worldwide. More than 10 million people worldwide are living with PD.It has 5 stages to it and affects more than 1 million individuals every year in India. . It affects about 5 lakh-one million Americans, or about 1% of people over the age of 60. Incidence of Parkinson's disease increases with age, but an estimated four percent of people with PD are diagnosed before age 50.Men are 1.5 times more likely to have Parkinson's disease than women[3]. Parkinson's disease can't be cured, but medication can help control the symptoms in PD patients. Medications may help PD affected people to manage problems with walking, movement and tremor. These medications increase or substitute for dopamine. In some more cases, surgery may be advised. Although there is large amount of research on PD, we still don't know what causes it. And we even have some trouble diagnosing it at times.[13]

To classify PD and healthy people the usage of speech signals is an effective technique for diagnosing PD from speech impairments. In Literature, different machines learning based classification techniques have been proposed to classify PD and healthy people from speech signals, and are reported in the study. Machine learning (ML) is frequently used for medical disease diagnosis recently because of its implementation convenience and high accuracy ML has also been used for the treatment of PD in the literature [4]. Tsanas et al. [5] used a data set consisting of 263 speech samples from 43 people and 76.7 % of dataset were PD, the leftover data set was healthy. They utilized an updated version of the data set that was utilized in [5]. Little et

al. [6] present an assessment of measures for the identity of PD subjects from healthy by detecting dysphonia. They diagnosed 23 PD and 8 healthy people and their data set recorded vowels and used a Support Vector Machine (SVM) for classification and achieved classification accuracy 91.4 %. Moreover; classifiers such as Ada boost, SVM, K-NN, multilayer perceptron (MLP), and Naïve Bayes (NB) were applied for classification PD and healthy subjects.They showed that phonation is the most convenient task for PD detection. K-NN, Multilayer Perceptron (MLP), Optimum Path Forest, and Support Vector Machines (SVM) were the evaluated classifiers in the study. Voice features were reduced using artificial neural networks for the ML based diagnosis of PD in [7].

2. RELATED WORK

It is important to predict clinical tasks for health base systems. Recently, a wide range of speech signal processing algorithms(dysponia measures) aiming to predict PD symptom severity using speech signals have been introduced. There have been several studies reported focusing on the diagnosis of Parkinson disease.In, [6] Little et al. present an assessment of measures for the identity of PD subjects from healthy by detecting dysphonia. They diagnosed 23 PD and 8 healthy people and their dataset recorded vowels and used a Support Vector Machine (SVM) for classification and achieved classification accuracy 91.4 %. Different from the work of Little et al., Sakar et al. [9] designed voice experiments with sustained vowels, words, and sentences from PD patients and controls. The paper reported that sustained vowels had more PD-discriminative power than the isolated words and short sentences. The study result achieved 77.5% accuracy by using SVM classifier. In [8], Das made a comparison of classification score for diagnosis of PD between artificial neural networks (ANN), Regression and Decision Trees. The ANN classifier yielded the best results of 92.9%.In [10], The new method compared to hitherto methods outperforms the state-of-the-art in terms of both predictions of accuracy (98.46%) and area under receiver operating characteristic curve (0.99) scores applying rotation-forest ensemble *k*-nearest neighbour classifier algorithm.In[11] ,study proposes a method in early detection and diagnosis of PD by using the Multilayer Feed forward Neural Network (MLFNN) with Back-propagation (BP) algorithm.The result shows that network can be used in diagnosis and detection of PD due to the good performance, which is 83.3% for sensitivity, 63.6% for specificity, and 80% for accuracy.In[14]study it indicates that the linear logistic regression and sparse multinomial logistic regression achieved a highest accuracy 100% and sensitivity, specificity of 0.983 and 0.996. In other related study, the Multi-Layer Perceptron (MLP)with Back-Propagation learning algorithm are used to classify to effective diagnosis Parkinsons disease(PD).In[19] study supervised learning algorithm such as deep neural networks and achieved a highest accuracy 83% provided by the machine learning models exceed the average clinical diagnosis accuracy of non-experts (73.8%) and average accuracy of movement disorder specialists (79.6% without follow-up, 83.9% after follow-up) with pathological post-mortem examination as ground truth.

Its a challenging problem for Medical community and it was characterized by tremor, PD occurs due to the loss of dopamine in the brains thalamic region that results in involuntary or oscillatory movement in the body and the feature selection algorithm along with biomedical test valued the diagnose Parkinson disease.Therefore, this work proposes a new approach to classifying PD tremor and improving the patients quality´s live on treatment, using machine learning classification algorithms.

3. MATERIALS and METHOD

3.2 DESCRIPTION OF DATASET:

ATTRIBUTE	DESCRIPTION
MDVP:Fo (Hz)	Average vocal fundamental frequency
MDVP:Fhi (Hz)	Maximum vocal fundamental frequency
MDVP:Flo (Hz)	Minimum vocal fundamental frequency
MDVP:Jitter(%) MDVP:Jitter(Abs) MDVP:RAP MDVP:PPQ Jitter:DDP	several measures of variation in fundamental frequency.
MDVP:Shimmer MDVP:Shimmer(dB) Shimmer:APQ3 Shimmer:APQ5 MDVP:APQ Shimmer:DDA	Several measures of variation in amplitude.
PDE,D2	Two nonlinear dynamical complexity measures
NHR, HNR	Two measures of ratio of noise to tonal components in the voice
DFA	Signal fractal scaling exponent
Spread1, spread2,PPE	Three nonlinear measures of fundamental frequency variation.
Status	Health status of the subject (one) - Parkinson's, (zero) - healthy.

Table 1: UCI Parkinson's disease dataset

The data set which is used for analysis is created by Max Little of the University of oxford, in collaboration with the National Centre for Voice and speech, Denver, Colorado, who recorded the speech signals.This data set consists of a range of biomedical voice measurement with 195 samples of features from 31 people, 23 with Parkinson's disease (PD) and 8 of them are the control group. The data is in ASCII CSV format. A series of features was extracted which can be categorized as : name of the patients, three types of fundamental frequency(high, low and average),several measures of variation in fundamental frequency(jitter and its type)several measures of variation of amplitude(shimmer and its type),two measures of ratio of noise to tonal components in the voice(NHR and HNR),Two nonlinear dynamically complexity measures(RPDE,D2),DFA is a signal fractal scaling exponent,and Three nonlinear measures of fundamental frequency variation(spread1,spread2,PPE).The main aim of the data is to discriminate healthy people from those with PD, according to "status" column which is set to 0 for healthy and 1 for PD.There are around 6 recording per patients.The data set has about 75% of cases suffering from Parkinson disease and 25% of cases which are healthy.[12]

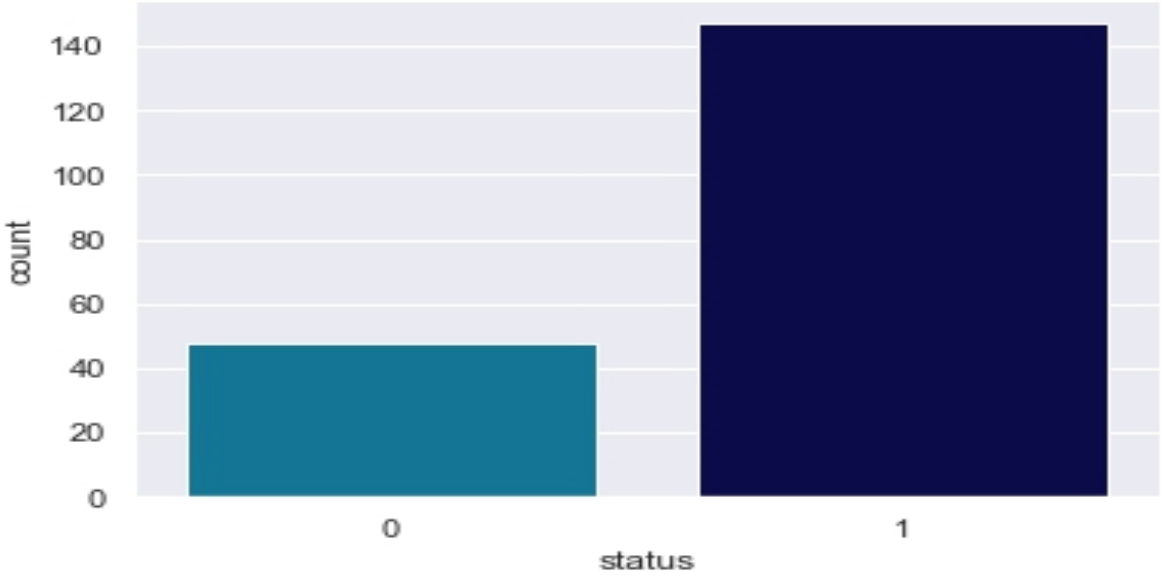


Fig 1: Show value counts for two categorical 0 for healthy and 1 for parkinson's

3.3 FEATURE IMPORTANT ANALYSIS

Feature importance refers to a class of techniques for assigning scores to input features to a predictive model that indicates the relative importance of each features when making a predict. Feature important analysis provide insight into the dataset.The relative scores can highlight which features can be more relevant to the target.Feature importance analysis can provide insight into

the model.Most important scores are calculated by a predictive model that has been fit on the dataset.In fig 2.it describes how different features in the parkinson's disease dataset correlated to each other.

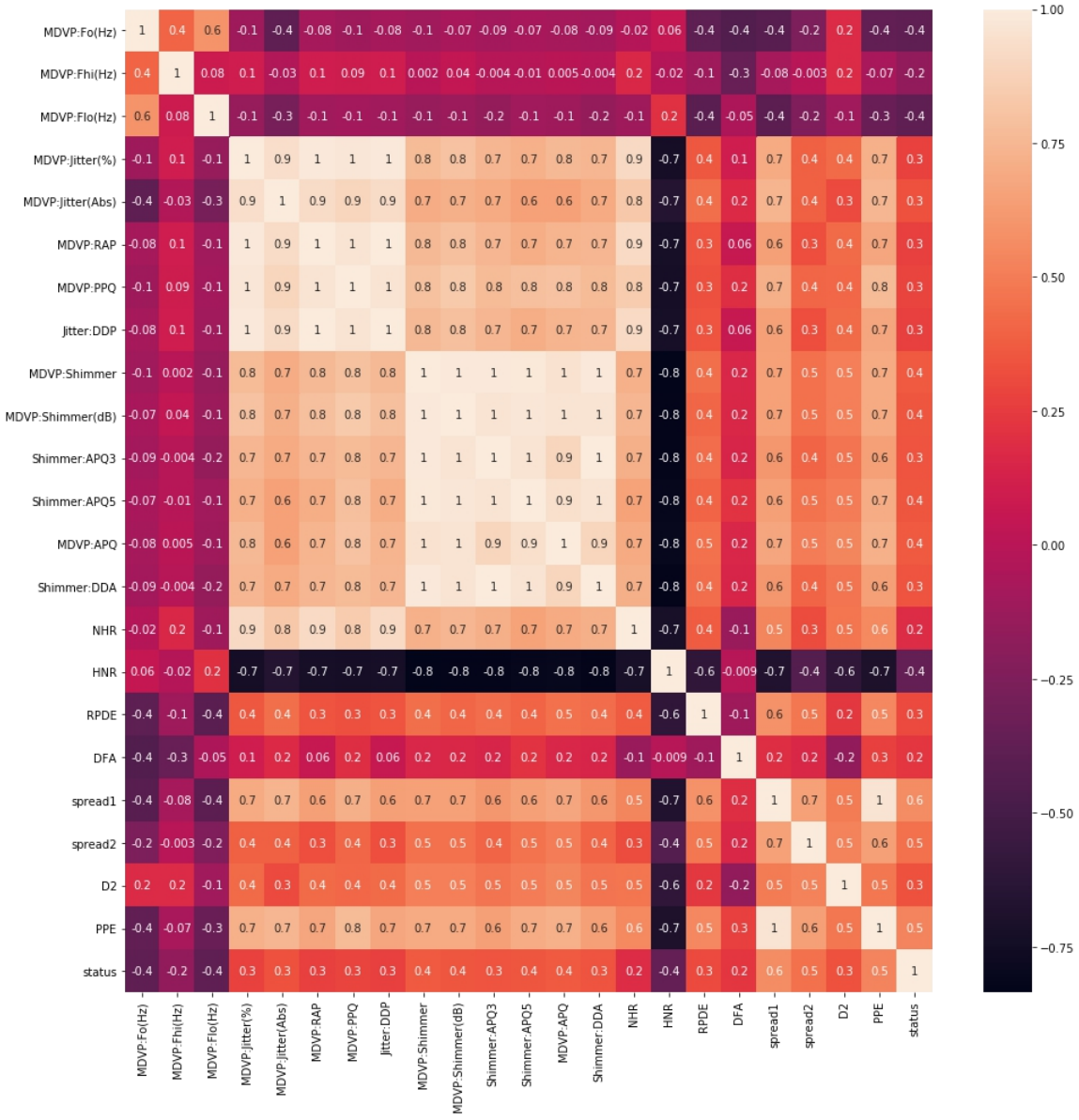


Fig 2: Correlation between features of parkinson's disease

3.4 PREDICTION TECHNIQUES

LOGISTIC REGRESSION:

Logistic regression is another techniques borrowed by machine learning from the field of statistics.It is the go-to method for binary classification problem(problems with 2 class values). Logistic regression is a type of predictive model that can be used when the target variable is a categorical variables. Logistic regression yields as good performance as machine learning model to predict the risk of major chronic diseases with low incidence and simple clinical predictors.Logistic regression measures the relationship between the categorical dependent variables by estimating the probabilities, the dependent variable must be binary in nature, e.g. 0 or 1. Formula used to calculate the logistic regression :

Y=1/[1+ e^-(beta0 + beta1x)].

DECISION TREE:

Decision tree algorithm is supervised learning algorithm which is used for the classification as well as regression problems.Decision tree is one of the predictive modelling approaches used in statistics,data mining and machine learning. Decision tree classifier, the input is split into sub-spaces based upon certain functions. It helps in reaching a conclusion based upon conditional control statement. In decision tree, there are two nodes, which are the decision node and leaf node.Decision nodes are used to make any decision and have multiple branches,whereas leaf nodes are the output of those decisions and do not contain any further branches. The goal is to create a model that predicts the value of target variable by learning simple decision rules inferred from the data features.

SVM(SUPPORT VECTOR MACHINE):

Support vector machine(SVM) is a popular classification technique.Support vector machine(SVM) is a supervised machine learning algorithm which can be used for both classification or regression challengers. It is mostly used in classification problems.SVM is highly preferred by many as it produces significant accuracy with less computation power. The main aim of SVM is to find a hyper plane in a N-dimensional space(N- number of features) that distinctly classifies the data points.The SVM classifier is a frontier which best segregates the two classes(hyper plane/line).The goal of SVM is to divide the dataset into classes to find a maximum marginal hyper_plane(MMH).Its high generalization ability makes it to be used in many fields of classification successfully.

KNN (K-Nearest Neighbour)

K-Nearest Neighbour is one of the simplest machine learning algorithm based on supervised learning techniques which can be used for both classification or regression challengers. It is mostly used in classification problems.The K is a important parameter in creating a KNN classifier. KNN algorithm assumes the similarity between the new data and available data and put the new data into the category that is most similar to the available categories.Based on the similarity KNN algorithm stores all the available data and classifies a new data point .This means when new data appears then it can easily classified into a well suite category by using KNN algorithm.KNN algorithm is robust to noisy training data and it can be more effective if the training data is large.

STEPS

Load the data

- Initialize the value of k.
- Fitting the KNN algorithm to the training set.
- To predict a data Calculate the distance between test data and each row of training data.Here Euclidean distance is used.
- Sort the calculated distance based on the distance value.
- Test the accuracy of the result
- Visualizing the test-set result

BAGGING CLASSIFIER:

Bagging classifier is also known as Bootstrap aggregating.It is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression.This classifier is almost similar to Random forest but the differ in the sense that in Random forest only subset of features are selected at random out of the total and the best split feature from the subset is used to split each node in tree unlike in bagging it consider all the features available for splitting at the node.

XGBOOST Classifier:

XGBOOST is widely used algorithm in machine learning,whether the problem is classification or regression problem.Its known for good performance as compared to all other machine learning algorithms.It stands for extreme gradient boosting algorithm and it is based on decision tree. It has an immensely high predictive power which makes it the best choice for accuracy and making the algorithm almost 10x faster than existing gradient booster techniques.Xgboost is also known as Regularized boosting technique.

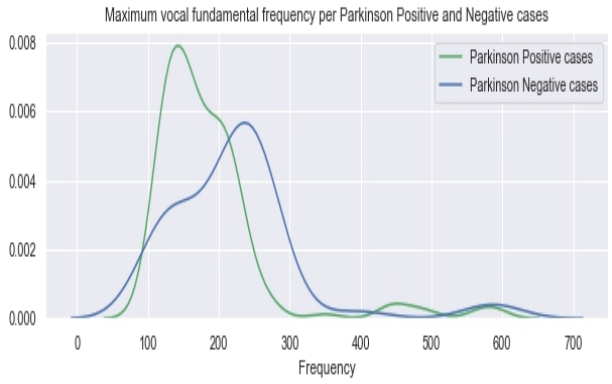
4. RESULT

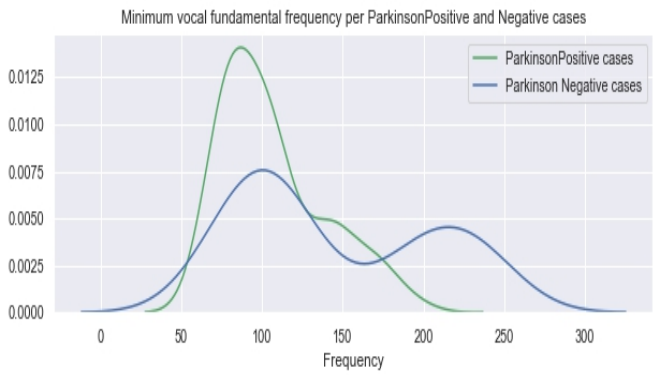
In this work we have used various prediction models for the prediction of parkinson disease using a reliable dataset from UCI machine learning repository where 195 samples of features from 31 people, 23 with Parkinson's disease (PD) and 8 of them are the control group. In the **Table 2**,we can observe that xgboost has achived highest test accuracy rate of 0.95 (95%) and training accuracy rate 0f 1.00 (100%).

CLASSIFICATION TECHNIQUES	TRAINING ACCURACY RATE	TEST ACCURACY RATE
LOGISTIC REGRESSION	0.88	0.79
DECISION TREE	1.00	0.90
SVM(SUPPORT VECTOR MACHINE)	0.89	0.92
KNN(K- NEAREST NEIGHBOUR)	0.94	0.95
XGBOOST	1.00	0.95
BAGGING	0.99	0.92

Table 2: Result of different prediction techniques

In the below fig. the performance analysis on fundamental frequency ,shimmer and jitter is done.The fundamental frequency are categorized into 3 types maximum, minimum and average vocal fundamental frequency .The performance analysis is done to know the role of fundamental frequencies ,shimmer and jitter in the prediction of parkinson's disease.





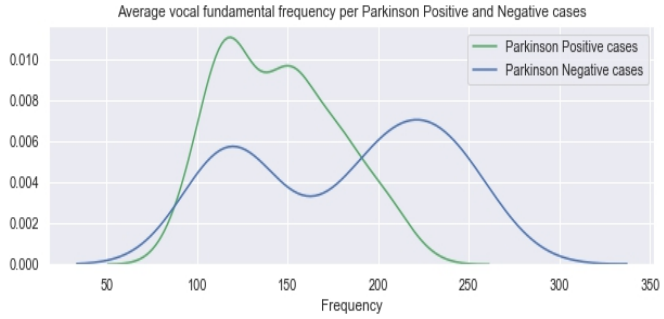


Fig 3:Distribution plot of Maximum ,Minimum,Average ,vocal fundamental frequency per Parkinson Positive and Negative cases.

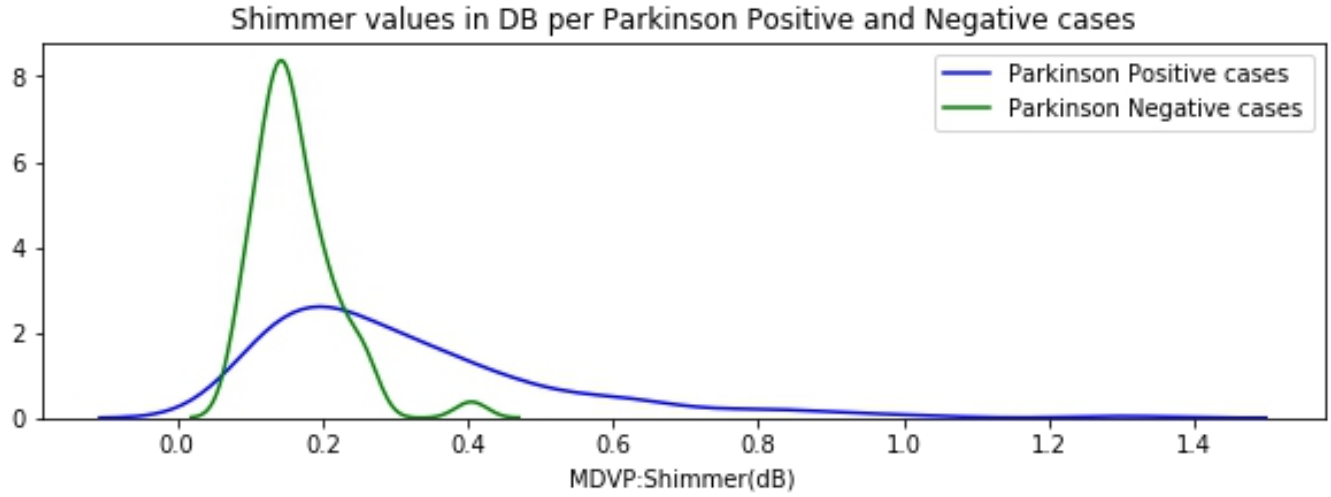


Fig 4:**Distribution plot of shimmer values in DB Parkinson Positive and Negative cases.**

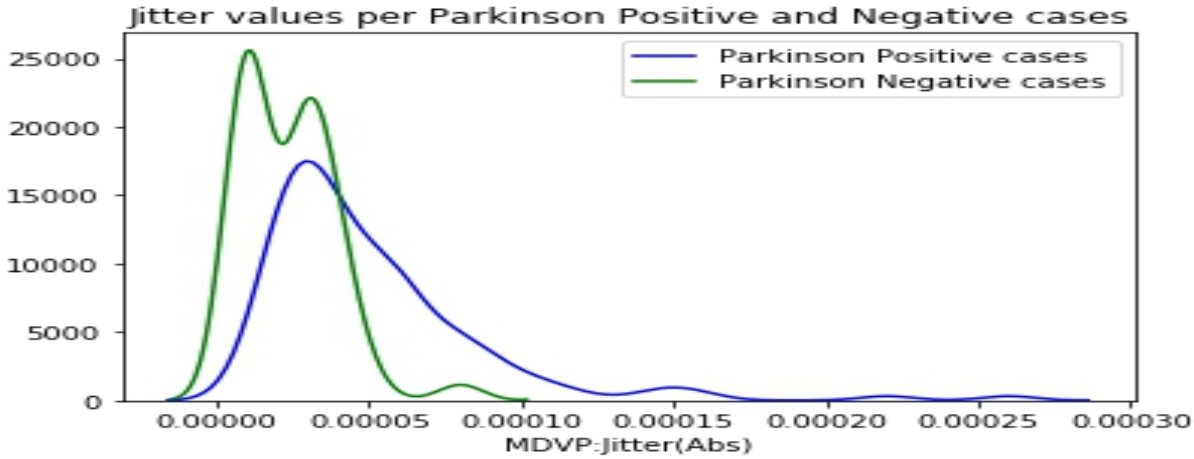


Fig 5: Distribution plot of Jitter values per Parkinson Positive and Negative cases.

5. CONCLUSION:

In this paper it deals with the application of six classification algorithms on the acquired data set. The algorithm such as Logistic Regression,Support vector machine(SVM),Decision tree, K-Nearest Neighbour (KNN), and XGBOOST(Extreme gradient boosting) are used to predict the outcome whether the person is healthy or parkinson disease effected based on the voice input parameters.Then lead to comparison of results to one another.From the result the conclusion obtained is that the ensemble techniques gives effective results compared to base classification algorithm.The ensemble techniques such as xgboost ,bagging Classifier gives more accurate prediction.The base classification algorithm such as Logistic Regression, Support vector Machine (SVM), Decision tree and K-Nearest Neighbour(KNN).Ensemble classifier models are evaluated based on the metrics such as .Accuracy achieved by the seven classifiers lies within the range 70-100%. Extreme Gradient Boost(XGBOOST) achieved high accuracy compared to other algorithm. It perform with impressive accuracy of training set is 100% and test accuracy rate 95%.Overall this study supports the use of these models for detecting Parkinson's disease.There is lot of scope to improve the technology as the diagnosis can be done in several means.

REFERENCES

[1] **National Institute for Health and Care Excellence (NICE). Parkinson's disease:** diagnosis and management in primary and secondary care. NICE clinical guidelines 35. June2006.Available http://www.nice.org.uk/guidance/cg35/resources/guidanceparkinsons-disease-pdf. Accessed April 28, 2015.

[2]. **Parkinson J. An Essay on the Shaking Palsy.** London: Sherwood, Neely, and Jones; 1817:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4517533/pdf/ptj4008504.pdf

[3].https://parkinsonsnewstoday.com/parkinsons-disease-statistics/#:~:text=An%20estimated%20seven%20to%2010,who%20are%2080%20and%20older.

[4]. Early diagnosis of Parkinson's disease using machine learning algorithms
Zehra Karapinar Senturk Duzce University, Engineering Faculty, Department of Computer Engineering, 81620 Duzce, Turkey

[5]. A. Tsanas, et al., "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," IEEE Transactions on biomedical engineering, vol. 59, pp. 1264-1271, January 2012.

[6] M. A. Little, et al., "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," IEEE Transactions on biomedical engineering, vol. 56, pp. 1015-1022, 2009.

[7] Parisi L, RaviChandran N, Manaog ML. Feature-driven machine learning to improve early diagnosis of Parkinson's disease. Expert Syst Appl 2018;110:182–90.

[8] R. Das, "A comparison of multiple classification methods for diagnosis of Parkinson disease," *Expert Systems with Applications*, vol. 37, no. 2, pp. 1568–1572, 2010.

[9] B. E. Sakar, M. E. Isenkul, C. O. Sakar et al., "Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings," IEEE Journal of Biomedical and Health Informatics, vol. 17, no. 4, pp. 828–834, 2013.

[10] An ensemble of *k*-nearest neighbours algorithm for detection of Parkinson's disease
Murat GökPublished online: 19 Jun 2013.

[11] **Application of Neural Networks in Early Detection and Diagnosis of Parkinson's Disease** 1Rashidah. Funke Olanrewaju, 1Nur Syarafina Sahari, 2Aibinu A. Musa and 3Nashrul Hakiem 1Department of Electrical and Computer Engineering, Faculty of Engineering, International Islamic University Malaysia, Kuala Lumpur, Malaysia. 2Federal University of Technology Minna, Niger State Nigeria. 3Department of Informatics Engineering Faculty of Science & Technology, UIN Syarif Hidayatullah Jakarta.

[12] UCI Machine Learning Repository : PARKINSON DATA SET https://archive.ics.uci.edu/ml/datasets/parkinsons.

[13] **Performance Analysis of Classification algorithms on Parkinson's Dataset with Voice Attributes .**T.Swapna and Y.Sravani Devi Department of Computer Science and Engineering, G. Narayanamma Institute of Technology & Science, Hyderabad, India.International Journal of Computer Applications (0975 – 8887) Volume 119 – No.3, June2015.

[14] New machine-learning algorithms for prediction of Parkinson's disease Indrajit Mandal* and N. Sairam School of Computing, SASTRA University, Thanjavur – 613401, Tamil Nadu, India (Received 24 February 2012; final version received 20 July 2012).

[15] Accuracy Improvement for Predicting Parkinson's Disease Progression Mehrbakhsh Nilashi, Othman Ibrahim & Ali Ahani .*Scientific Reports* volume 6, Article number: 34181 (2016)

[16] Parkinson's disease: A review
Authors: Divya Madathiparambil Radhakrishnan,All India Institute of Medical Sciences NewDelhi and Vinay Goyal,All India Institute of Medical Sciences. March 2018,Neurology India 66(Supplement):S26-S35,DOI: 10.4103/0028-3886.226451.

[17] Prediction of Parkinson's Disease using Machine Learning Techniques on Speech dataset:Basil K Varghese, Geraldine Bessie Amali D*, Uma Devi K S ,School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India*Corresponding Author.Volune - 12,Issue - 2, Year - 2019.

[18] A deep learning approach for prediction of Parkinson's disease progression Afzal Hussain Shahid1 Maheshwari Prasad Singh1 Received: 26 September 2019 / Revised: 20 February 2020 / Accepted: 7 April 2020 © Korean Society of Medical and Biological Engineering 2020.

[19] **Parkinson's Disease Diagnosis Using Machine Learning and Voice**

*Timothy J. Wroge*1 *, Yasin Ozkanca* 2 *, Cenk Demiroglu*2 *, Dong Si*3 *, David C. Atkins*4 *and Reza Hosseini Ghomi*4 1. Department of Bioengineering, University of Pittsburgh, Pittsburgh, Pennsylvania, USA 2. Department of Engineering, Ozyegin University, Istanbul, Turkey

3. Division of Computing and Software Systems, University of Washington, Seattle, Washington, USA

4. Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, Washington, USA timothy.wroge@pitt.edu, {yasin.ozkanca, cenk.demiroglu}@ozyegin.edu.tr, {dongsi, datkins, rezahg}@uw.edu