JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Early Detection of Parkinson's Disease based on Ensemble Models

[1]M. Jyothirmayee, [2]Dr. G. Umamaheswara Reddy

[1]MTech Student, Department of ECE, S.V. University College of Engineering, Tirupati, Andhra Pradesh, India.

[2]Professor, Department of ECE, S.V. University College of Engineering, Tirupati, Andhra Pradesh, India

Abstract– In Parkinson's Disease (PD), a neurological condition that worsens over time, new symptoms gradually start to show up. The substantia nigra of the brain gradually loses neurons as a result of this. Gait impairment, voice impairment, writing changes, etc. are some of the signs that are used to identify Parkinson's disease (PD) more correctly. In this work, Multiple Fold Data (MFD) is ensembled with Gradient Boost Classifier, a machine learning classifier, to provide superior output. With 98.2% accuracy, this combination had the highest accuracy rate. An accurate Dataset from the UCI Machine Learning Repository was used to evaluate the approaches' efficacy.

Keywords: *Multiple Fold Data, Gradient Boost Classifier, XG Boost Classifier, Random Forest, precision, Parkinson's disease.*

1. INTRODUCTION

Parkinson's disease is a neurological illness that affects the brain cells that make dopamine, which causes movement impairment. People 50 years of age or older have been diagnosed with PD [1]. In India, where it affects about 1 million people, 60,000 new clinical diagnoses are reported each year. Shaking, difficulty moving, behavioral problems, amnesia, and sadness are just a few of the many symptoms of this illness. The predominant motor symptoms are collectively referred to as "Parkinsonism" or "Parkinsonian Syndrome" [2].

Additionally, PD raises the risk of experiencing depression, anxiety, panic attacks, and sleep issues. The voice and speech data must be correctly categorized in order to diagnose PD. [6]. One of the typical indicators that can be found by looking at the patient's speech data is changes in voice. As the patient's condition worsens over time, their voice starts to stammer [8].

More than 90% of individuals experience speech problems as their condition worsens. Dysphonia is the term used to describe the voice dysfunction that PD patients experience. There is no effective cure or treatment when the condition is advanced. So, it's best to diagnose the condition quickly. In addition to lowering the expense of the treatment, this could even save a life.

Due to developments in technology and the popularity of audio collecting equipment in daily life, reliable models that can turn this audio data into a diagnostic tool may be able to offer less expensive and more accurate diagnoses [3]. Using a voice dataset compiled from both people with and without Parkinson's disease, we here give proof to support this idea. This study examines the effectiveness of using supervised classification techniques, such as Random Forest classifier and XG Boost Classifier with Multiple Fold Data, to accurately classify patients with the disorder.
The remaining parts of the essay are organized as follows. The literature review is provided in section 2. Section 3 describes the suggested Voice Impairment Analysis. Experimental data and performance comparisons of the proposed and current techniques are presented in Section 4. The conclusion and the future scope are provided in Section 5.

2. LITERATURE REVIEW

A deep neural network classifier that combines a softmax classifier with a stacked auto-encoder was proposed by Abdullah et al. [6]. The DNN classifier is efficient in identifying vocal deficits in PD patients, improving PD diagnosis.

Noise is initially eliminated and then divided into time periods during the filtering of the voice signals. The second phase involves the extraction of various properties from each segment, followed by DNN classification using stacked autoencoders (SAE).

Groom et al. using the Parkinson's Telemonitoring Voice Dataset from the UCI Machine Learning (ML) Repository, al. [2] used deep neural networks to predict the severity of PD. They used the Python TensorFlow deep learning library to create the neural network for predicting the severity. The neural network contains three hidden layers, each with 10, 20, and 10 units, an output layer with 2 neurons, and an input layer with 16 units. The model's accuracy stands at 81.6667%.

The authors of [3] attempted to identify the PD group using different feature sets and Principal Component Analysis (PCA) and Online Feature Selection-based feature sets, which yield non-linear features on a dataset obtained from the Max Little University of Oxford. The authors used non-linear classifiers, Bagging classification, Regression tree (Bagging CART), Random Forest, and RPART. The accuracy of the classification using RF with PCA based on voice data was 96.83%.

Pereira et al. [4] proposed a unique technique in light of the decreased writing abilities. The researchers suggested using data collected from the six smart pen sensors to train two alternative CNN architectures, ImageNet and LeNet, to understand pen-based properties. The researchers discovered that ImageNet and Optimal Power Flow, with accuracy scores of 83.77% for spirals and meanders, respectively, had the greatest results.

A decrease in the amount of fluid dopamine produced by neurons is another indication of Parkinson's disease. The authors of [5] developed an automated deep-learning model that interprets a dataset of FP-CIT SPECT images that were obtained from the Parkinson's Progression Markers Initiative repository because dopamine transporter imaging, such as FP-CIT single-photon emission computed tomography (SPECT), can be used to determine it. First, SPECT images are supplied to the 3D convolutional layer as inputs. Additionally, 16 3D outputs are produced after passing through 7x7x7 convolutional filters, max-pooling, and the activation layer of Rectified Linear Units on the output layer (ReLU). This research served as the foundation for Oh et al.'s suggestion of a novel automatic detection method for PD CNN using EEG signals.

Electroencephalogram (EEG) signals are widely employed for initial diagnosis, according to the author. In this investigation, EEG data from 20 participants with PD and 20 controls were used. The novel thirteen-layer CNN model developed by the researcher uses ReLU activation in hidden levels and softmax in the output layer. The proposed CNN model's accuracy, sensitivity, and specificity were 88.25%, 84.71%, and 91.77%, respectively.

In [7], non-motor traits such as olfactory loss, rapid eye movement (REM) sleep behavior disorder, and PPMI data were integrated. The author employed automated diagnostic machine learning techniques such as Multilayer Perceptron, Bayes Net, Random Forest, and Boosted Logistic Regression, with a stated accuracy of 97.16% and an area under the receiver operating characteristic curve of 98.9%.

The [8] asserts that non-invasive neuroimaging methods, such as magnetic resonance imaging and positron emission tomography, as well as invasive methods, such as electroencephalography, have proven useful in analyzing the pictures' functional brain activity to research neural functions. Among other things, they have taken advantage of Functional Connectivity, ReHo, and Amplitude of Low-Frequency Fluctuations. Swarm intelligence can be utilized to accelerate the performance of neural techniques that are already in use, according to published research [11]. Ensemble approaches can also be used to scale up performance by training on large datasets [12,13].

3. VOICE IMPAIRMENT ANALYSIS

3.1. Data Set Information

This dataset includes 195 samples of biological voice measures from 31 individuals, 23 of whom have Parkinson's disease and 8 of whom are in the control group. The data is formatted as Ascii Comma Separated Values. Each column in the chart denotes a certain vocal measure, and each row in the table represents one of the 195 voice recordings of these individuals (name, column). The primary goal of the data is to distinguish between those who are in excellent health and those who have Parkinson's disease (PD) based on the status column, which is set to 0 for healthy and 1 for the disease.

3.2. Dataset Description

Data for the Multidimensional Voice Program (MDVP) attribute dataset is shown in Table 1. In order to identify aberrant voice patterns in PD patients, it looks at a number of vocal traits, with Fhi being the highest or maximum frequency and Flo being the lowest or minimum frequency.

Table 1: Attribute information of the dataset

Attribute	Description
MDVP.Fo(Hz)	Average fundamental frequency of the voice
MDVP.Fhi(Hz)	Maximum fundamental frequency of the voice
MDVP.Flo(Hz)	Minimum fundamental frequency for vocals
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 ways to assess amplitude variation
NHR, HHR	Two measurements of the voice's noise tonal component ratio
Status	Subject's health status: (one) PD patient, (zero) healthy
RPDE, D2	Two indices of the complexity of nonlinear dynamics
DFA	Spread1, Spread2 of the signal fractal scaling exponent
PPE	Three methods of measuring fundamental frequency variation are nonlinear.

Table 2 displays the analytic dataset created by Max Little from the University of Oxford in collaboration with Denver, Colorado's National Centre for Voice and Voice, which recorded the voice signals.

The features that were extracted included two measures of the ratio of noise to tonal components in the voice (Noise to Harmonic Ratio (NHR) and Harmonic to Noise Ratio (HNR), three fundamental frequency types (high, low, and average), several measures of variation in fundamental frequency (jitter and its Absolute type), several measures of variation in amplitude (shimmer and its type), two nonlinear dynamically complexity measures (Recurrence), and more (spread1, spread2, PPE).

Table 2: Dataset description

	count	mean	std	min	25%	50%	75%	max
MDVP:Fo(Hz)	195.0	154.228641	41.390065	88.333000	117.572000	148.790000	182.769000	260.105000
MDVP:Fhi(Hz)	195.0	197.104918	91.491548	102.145000	134.862500	175.829000	224.205500	592.030000
MDVP:Flo(Hz)	195.0	116.324631	43.521413	65.476000	84.291000	104.315000	140.018500	239.170000
MDVP:Jitter(%)	195.0	0.006220	0.004848	0.001680	0.003460	0.004940	0.007365	0.033160
MDVP:Jitter(Abs)	195.0	0.000044	0.000035	0.000007	0.000020	0.000030	0.000060	0.000260
MDVP:RAP	195.0	0.003306	0.002968	0.000680	0.001660	0.002500	0.003835	0.021440
MDVP:PPQ	195.0	0.003446	0.002759	0.000920	0.001860	0.002690	0.003955	0.019580
Jitter:DDP	195.0	0.009920	0.008903	0.002040	0.004985	0.007490	0.011505	0.064330
MDVP:Shimmer	195.0	0.029709	0.018857	0.009540	0.016505	0.022970	0.037885	0.119080
MDVP:Shimmer(dB)	195.0	0.282251	0.194877	0.085000	0.148500	0.221000	0.350000	1.302000
Shimmer:APQ3	195.0	0.015664	0.010153	0.004550	0.008245	0.012790	0.020265	0.056470
Shimmer:APQ5	195.0	0.017878	0.012024	0.005700	0.009580	0.013470	0.022380	0.079400
MDVP:APQ	195.0	0.024081	0.016947	0.007190	0.013080	0.018260	0.029400	0.137780
Shimmer:DDA	195.0	0.046993	0.030459	0.013640	0.024735	0.038360	0.060795	0.169420
NHR	195.0	0.024847	0.040418	0.000650	0.005925	0.011660	0.025640	0.314820
HNR	195.0	21.885974	4.425764	8.441000	19.198000	22.085000	25.075500	33.047000
RPDE	195.0	0.498536	0.103942	0.256570	0.421306	0.495954	0.587562	0.685151
DFA	195.0	0.718099	0.055336	0.574282	0.674758	0.722254	0.761881	0.825288
spread1	195.0	-5.684397	1.090208	-7.964984	-6.450096	-5.720868	-5.046192	-2.434031
spread2	195.0	0.226510	0.083406	0.006274	0.174351	0.218885	0.279234	0.450493
D2	195.0	2.381826	0.382799	1.423287	2.099125	2.361532	2.636456	3.671155
PPE	195.0	0.206552	0.090119	0.044539	0.137451	0.194052	0.252980	0.527367
status	195.0	0.753846	0.431878	0.000000	1.000000	1.000000	1.000000	1.000000

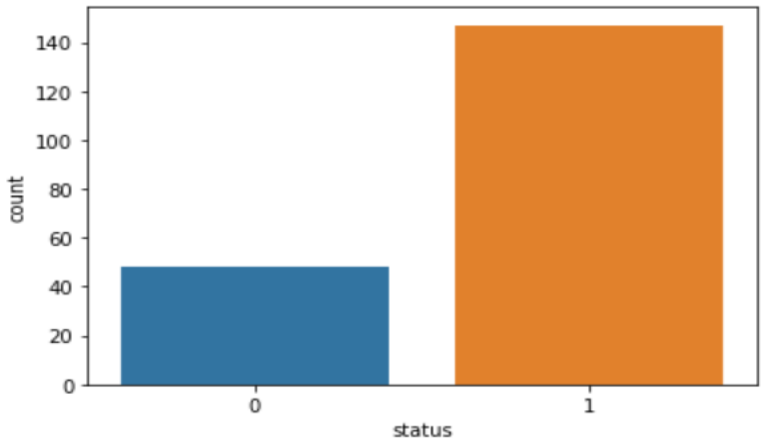

Figure 1: Value counts for two categorical 0 for healthy and 1 for PD

Figure 1 shows the data used to differentiate between people with Parkinson's disease (PD) and people in good health, with the "status" column set to 0 for healthy people and 1 for people with PD. Six to eight recordings are given to each patient. About 75% of the cases in the data set have Parkinson's disease, while 25% of the cases are healthy.

3.3. Feature Analysis

A collection of techniques for grading the relative weights of different qualities in the input data of a predictive model are together referred to as "feature importance analysis." Feature significance scores, which are crucial in a predictive modeling project because they offer insight into the data, information about the model, and information about the data, serve as the basis for dimensionality reduction and feature selection, which can enhance the efficiency and efficacy of a predictive model on the problem. A prediction model fitted to the dataset determines the most significant scores.

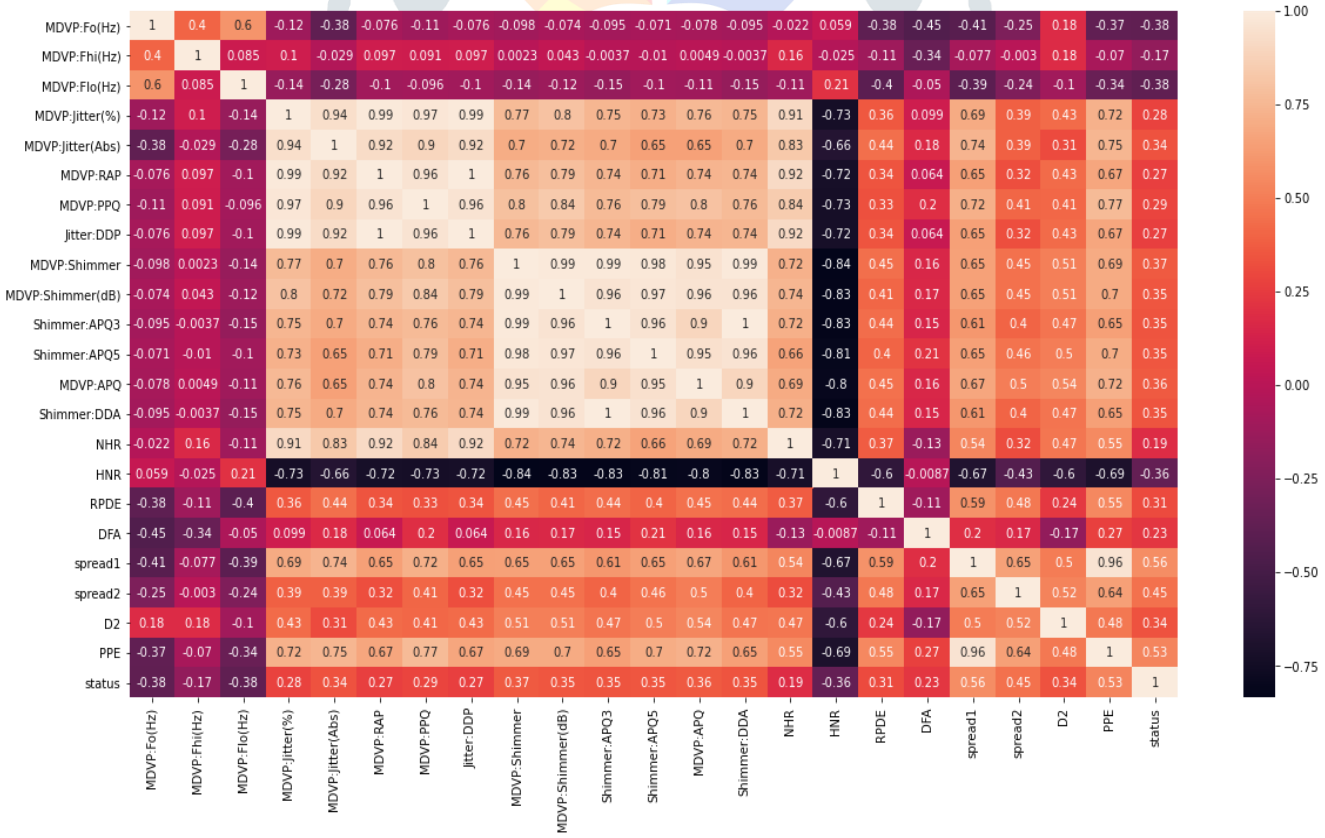

Figure 2: Parkinson's disease features and their correlation

Figure 2 displays the correlations between the various features in the Parkinson's disease dataset.

3.4. METHODOLOGY

In this study, a Gradient Boost Classifier with Multiple Fold Data is used to build a model, along with Python packages like scikit-learn, NumPy, and pandas. The following steps are necessary to implement the suggested model:

1) The model is loaded with data.
2) Appropriate labels are used for extracted features.
3) In addition, the features are resized.
4) The Dataset is partitioned.
5) The Gradient Boost Classifier is constructed.
6) After that, the model's accuracy, specificity, and sensitivity are determined.

The system listed below can identify Parkinson's signs in the human body.

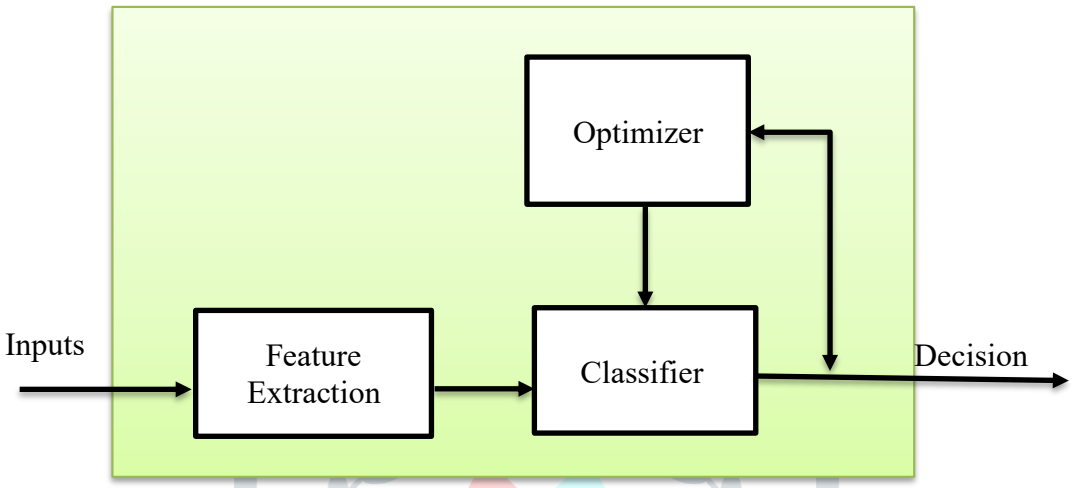

Figure 3: Block Diagram of Parkinson's Disease Detection

The block diagram for diagnosing Parkinson's disease is depicted in Figure 3. It consists of numerous different blocks, including feature extraction, a classifier, and an optimizer. Feature extraction is the process of transforming unprocessed data into the inputs needed for a particular Machine Learning algorithm.

The machine learning algorithm's Optimizer, which sets up and trains the model, is an essential component. Identification and classification of items are two processes that go hand in hand. This paper employs a gradient boost classifier with MFD. For problems like classification and regression, among others, a machine learning technique called gradient boosting is used. It offers a prediction model in the form of a group of feeble prediction models, most frequently decision trees. As a result, when a decision tree is a weak learner, gradient-boosted trees outperform random forests. A gradient-boosted trees model is built piecemeal, just like previous boosting methods.

Ensemble learning is a branch of machine learning in which a number of models are trained to produce a single, superior outcome. The XG Boost Classifier belongs to this branch. Gradient Boost and Multiple Fold Data are combined to create an ensemble model that fits our training dataset. The models are fitted using the scikit-learn API and the model. Efficacy Fit () The dataset is created with the help of the scikit-learn module, and an XGB Classifier is used to identify the presence of Parkinson's disease in individuals using a number of criteria.

3.5. EVALUATED PARAMETERS

The Confusion Matrix is used to determine performance metrics including Sensitivity, Specificity, and Accuracy. Figure 4 demonstrates how the data is divided into two categories: Actual and Predicted. Additionally, classification is completed, and the values are divided into four cases as mentioned below.

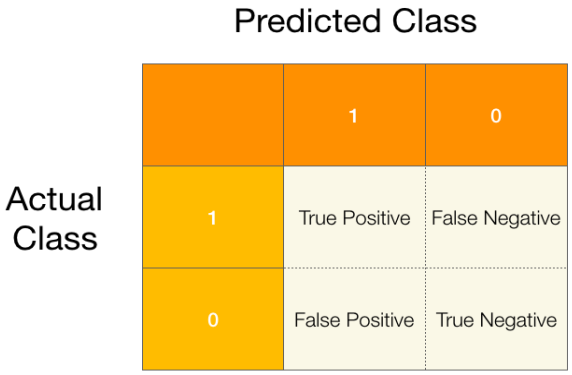

Figure 4: Confusion Matrix

True Positive (TP): The person has the disease and the test is POSITIVE.
True Negative (TN): The person does not have the disease and the test is NEGATIVE.
False Positive (FP): The person does not have the disease and the test is POSITIVE.
False Negative (FN): The person has the disease and the test is NEGATIVE.

Sensitivity: It is sometimes referred to as recall. the capacity of a test to accurately locate people who have a disease. It is stated as,

$$Sensitivity = \frac{TP}{(TP+FN)} \quad \text{-------(1)}$$

Specificity: The precision with which a test may identify those without the disease. It is stated as,

$$Specificity = \frac{TN}{(TN+FP)} \quad \text{-------(2)}$$

Accuracy: It is the most crucial indicator for determining the model's effectiveness. The percentage of accurate predictions to all other guesses is known as accuracy.

It is denoted mathematically as,

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad \text{-------(3)}$$

Precision: the proportion of all positive cases the model returned that were correct positive occurrences. It is stated as,

$$Precision = \frac{TP}{(TP+FP)} \quad \text{-------(4)}$$

4. RESULTS AND DISCUSSION

A precision-recall curve is a graph that contrasts Precision (y-axis) and Recall (x-axis) for different threshold values. It is utilized to assess how well the binary classification algorithm function. Producing the precision-recall curves depends on the parameters of Precision and Recall as well as an independent diagnosis that separates the study participants into two distinct groups: a diseased group and a non-diseased group. The average precision (AP) that occurred is known thanks to the Precision-Recall Curve that was obtained.

By averaging all recall values between 0 and 1, the precision, or AP, is determined. When the value is higher, the classifier performs better. The ROC curve is a plot between the False Positive Rate and True Positive Rate. Operating characteristics are the terms used to describe both of the factors, which serve as the features that define the ROC curve. The Area Under the Curve (AUC) demonstrates how well the classifier performs. Performance improvement brings the ROC AUC closer to 1. Calculations are also done for the values of sensitivity, specificity, and accuracy.

- **EXISTING RESULTS**

The Receiver Operating Characteristic (ROC) curve and Precision-Recall curve produced by the Random Forest Classifier are shown in Figures 5 and 6. One can see that the value of AP is 0.99 from the Precision-Recall curve. A better PR curve and higher average precision are both a result of making more accurate forecasts.

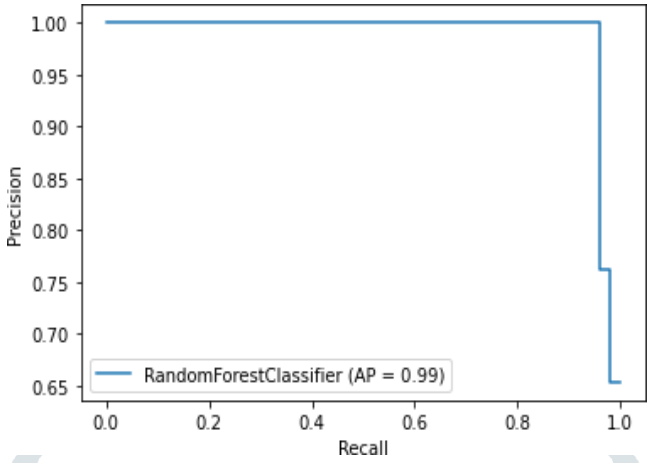

Figure 5: Precision-Recall Curve

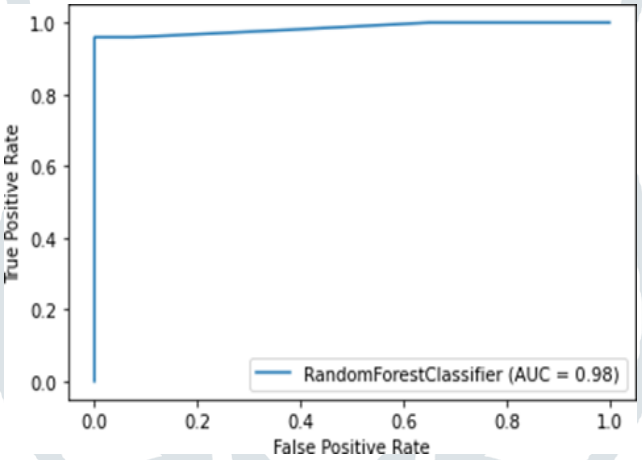

Figure 6: ROC Curve

The ROC curve produced by the RF Classifier displays an AUC of 0.98, which is quite near the maximum value of 1. The classifier accurately separates all of the positive and negative class points as a result.

```
print(TP/float(TP+FN))
print(metrics.recall_score(Bal_y_test,RF_y_predicted))
```

```
0.9523809523809523
0.9591836734693877
```

Specificity:

```
print(TN/float(TN+FP))
```

```
1.0
```

Figure 7: Sensitivity and Specificity (RF)

```
With 99% confidence,we can say that the mean test accuracy is between 0.917 and 0.908, based on 50 sample accuracies.
```

Figure 8: Accuracy (RF)

Figures 7 and 8 display the Sensitivity, Specificity, and Accuracy values that were acquired using the Random Forest Classifier.

PROPOSED RESULTS

Gradient Boost Classifier with Multiple Fold Data Precision-Recall Curve and ROC Curve are shown in Figures 9 and 10.

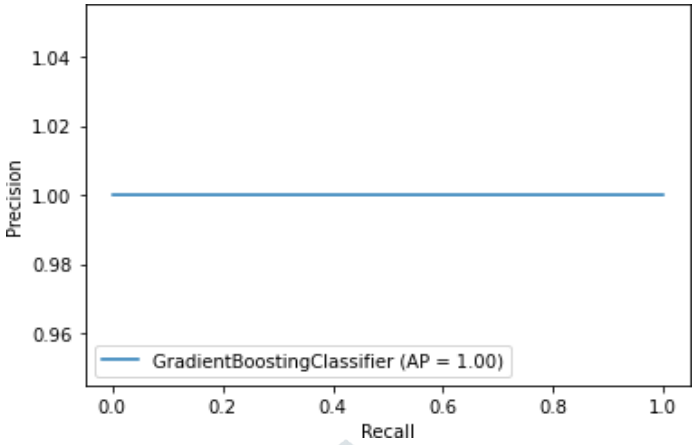

Figure 9: Precision-Recall Curve

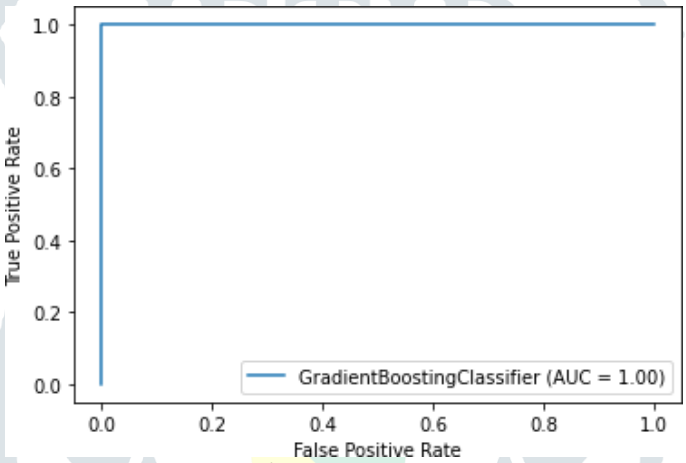

Figure 10: ROC Curve

The average precision is exactly equal to the highest possible score of 1, as seen in the aforementioned curve. The ROC curve represents a comparison of the False Positive Rate and True Positive Rate. When the AUC is 1, the classifier is able to accurately distinguish between every Positive and every Negative class point. All Negatives would be predicted to be Positives if the AUC was 0, and all Positives would be predicted to be Negatives. It is noted that AP in the Precision-Recall curve and AUC in the ROC curve, however, are equal to 1. So, in comparison, the classifier's performance is good. We can see an improvement in Sensitivity, Specificity, and Accuracy as a result of the strong performance.

The performance study of shimmer, jitter, and their fundamental frequencies is presented as seen in figures 11 to 14. Maximum, minimum, and average vocal fundamental frequencies are the three different types of fundamental frequency. Performance analysis is performed to investigate the effects of basic frequencies, shimmer, and jitter on the probability of Parkinson's disease occurrence.

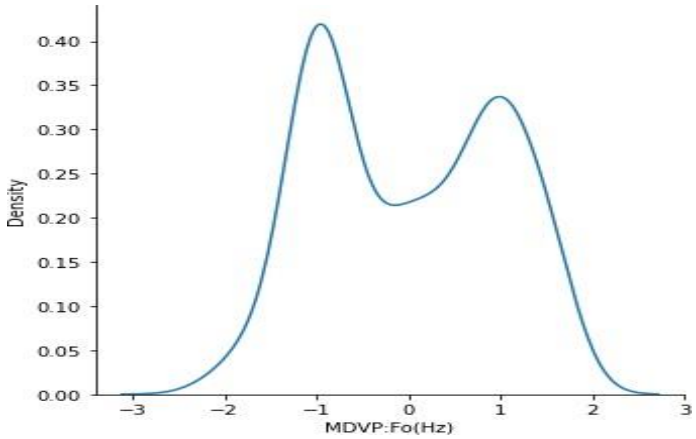

Figure 11: Performance Analysis of Fundamental Frequency of MDVP.Fo

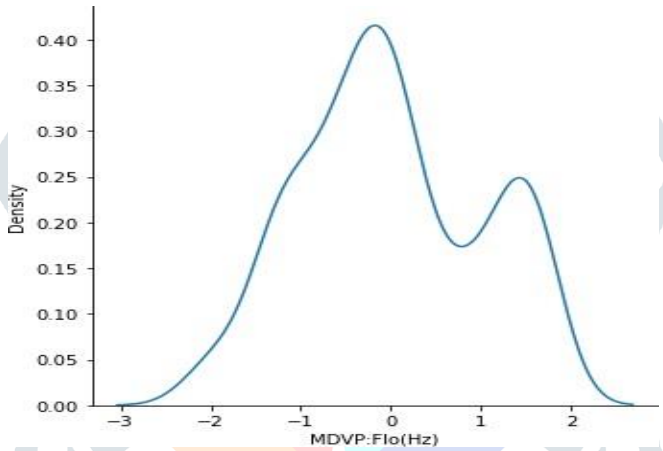

Figure 12: Performance Analysis of Fundamental Frequency of MDVP.Flo

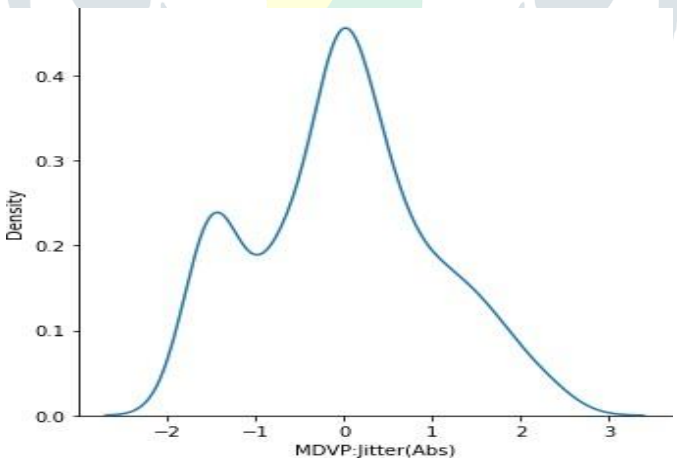

Figure 13: Performance Analysis of Fundamental Frequency of Jitter

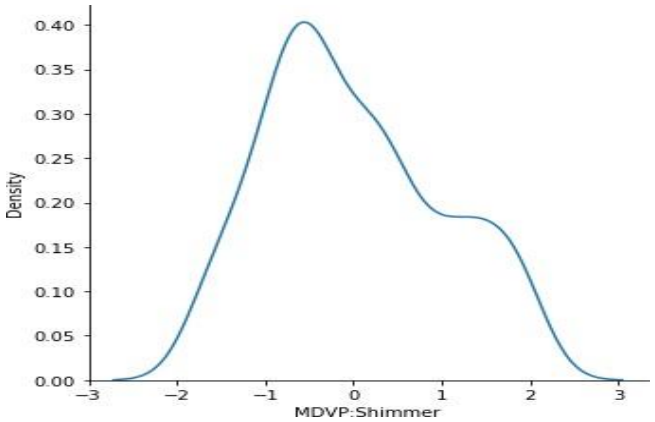

Figure 14: Performance Analysis of Fundamental Frequency of Shimmer

Acoustic features of speech signals such as jitter and shimmer are brought on by erroneous vocal fold vibrations. The roughness, breathiness, or hoarseness they produce in a speaker's voice is experienced by listeners. Fundamentally, jitter and shimmer can be assessed to find vocal problems even if they are present to some extent in all-natural speech.

```
Sensitivity
1.0
1.0
Specificity
0.9375
```

Figure 15: Sensitivity and Specificity (MFD)

```
With 99% confidence,we can say that the mean test accuracy is between 0.982 and 0.915, based on 50 sample accuracies.
```

Figure 16: Accuracy (MFD)

The results for Sensitivity, Specificity, and Accuracy achieved with the use of an ensemble of the XG Boost Classifier and MFD are shown in Figures 15 and 16.

Table 3: Performance Comparison Table

ML CLASSIFIERS	SAMPLES	SENSITIVITY	SPECIFICITY	TEST ACCURACY (%)
Unregularized Decision Tree	50	0.91	0.85	86.7
Regularized Decision Tree	50	0.95	0.898	86.9
Random Forest	50	0.95	1.0	91.7
XG Boost Classifier (Proposed)	50	1.0	0.875	96.1
Gradient Boost with MFD (Proposed)	50	1.0	0.937	98.2

Table 3 shows that when compared to other machine learning classifiers such as the XG Boost Classifier, Decision Tree Classifier, and Random Forest Classifier, the proposed Gradient Boost Algorithm with Multiple Fold Data achieves a high accuracy rate of 98.2%.

5. CONCLUSION

In this study, a Gradient Boost classifier combined with Multiple Fold Data is used to create a machine learning model for the voice-based detection of Parkinson's disease in a person. The proposed and existing models' performance is evaluated in terms of classification accuracy. The recommended model, Gradient Boost with Multiple Fold Data, outperformed the existing Random Forest and Decision Tree machine learning classifiers in the diagnosis of PD disease using high-dimensional data, as shown by the findings, and was effective in obtaining high accuracy of 98.2%.

Different methods will be employed in the future to predict Parkinson's disease using additional symptoms including handwriting difficulties, vision problems, etc. Additional research can be done utilizing various criteria to categorize patients and determine the various phases of Parkinson's disease.

REFERENCES

[1] Shu Lih Oh, Yuki Hagiwara, U. Raghavendra, Rajamanickam Yuvaraj, N. Arunkumar, M. Murugappan, U. Rajendra Acharya. “A deep learning approach for Parkinson’s disease diagnosis from EEG signals.” Computer Aided Medical Diagnosis. The Natural Computing Applications Forum (2018) August.

[2] Srishti Grover, Saloni Bhartia, Akshama, Abhilasha Yadav, Seja K. R. “Predicting Severity of Parkinson's Disease Using Deep Learning”. International Conference on Computational Intelligence and Data Science (ICCIDS 2018). Procedia Computer Science 132 (2018) 1788–1794.

[3] Satyabrata Aich, Kim younga, Kueh Lee Hui, Ahmed Abdulhakim Al Absi, Mangal Sain. “A Nonlinear Decision Tree based Classification Approach to Predict the Parkinson’s disease using Different Feature Sets of Voice Data.” International Conference on Advanced Communications Technology (ICACT). ISBN 979-11-88428-01-4. ICACT (2018) February 11-14.

[4] Clayton R. Pereira, Silke A. T. Weber, Christian Hook, Gustavo H. Rosa, Joao Papa. “Deep Learning-aided Parkinson’s Disease Diagnosis from Handwritten Dynamics.” Conference Paper. (2017) October 17.

[5] Hongyoon Choia, Seunggyun Haa, Hyung Jun Ima, Sun Ha Paekd, Dong Soo Leea. “Refining diagnosis of Parkinson's disease with a deep learning-based interpretation of dopamine transporter imaging.” NeuroImage Clinical 16 (2017) 586–594.

[6] Abdullah Caliskan, Hasan Badem, Alper Basturk, Mehmet Emin Yuksel. “Diagnosis of the Parkinson’s Disease by using Deep Neural Network Classifier.” IU-JEEE Vol. 17(2), (2017), 3311-3318.

[7] Kamal Nayan Reddy Challa, Venkata Sasank Pagolu, Ganapati Panda, Babita Majhi. “An Improved Approach for Prediction of Parkinson’s Disease using Machine Learning Technique”. International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)-(2016) October.

[8] Vijay Khare, Neha Mehra, Shamim Akhter, Shaurya Singh, Anshul Singh, Dr. Chakresh Kumar Jain. “Analysis and Identification of Parkinson’s disease based on fMRI.” International Journal of Engineering Technology, Management and Applied Sciences (IJETMAS). (2017) January.

[9] Soheil Esmaeilzadeh, Yao Yang, Ehsan Adeli. “End-to-End Parkinson Disease Diagnosis using Brain MR-Images by 3D-CNN.” CoRR, abs/1806.0.

[10] Wu, Wu, Defeng, et al. “Prediction of Parkinson's disease tremor onset using a radial basis function neural network based on particle swarm optimization.” International journal of neural systems 20.02 (2010): 109- 116.

[11] Ashish, Tripathi, Sharma Kapil, and Bala Manju. “Parallel bat algorithm-based clustering using MapReduce.” Networking Communication and Data Knowledge Engineering. Springer, Singapore,

(2018). 73-82.

[12] Tripathi, Ashish Kumar, Kapil Sharma, and Manju Bala. "A novel clustering method using enhanced grey wolf optimizer and MapReduce." Big data research 14 (2018): 93-100.

[13] Tripathi, Ashish Kumar, Kapil Sharma, and Manju Bala. "Dynamic frequency based parallel k-bat algorithm for massive data clustering (DFBPKBA)." International Journal of System Assurance Engineering and Management 9.4 (2018): 866-874.