International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Early Detection of Parkinson's disease using a Machine Learning-Based Framework for Differentiating the Disease's with Various Stages

Archana Panda
School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar, India
2081011@kiit.ac.in

Prachet Bhuyan
School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar, India
pbhuyanfcs@kiit.ac.in

Debasish Kumar Panda
devdesure@gmail.com

Abstract— The term "Parkinson's disease" (PD) is mostly brought on by a disturbance of a brain's dopamine-producing cells, a chemical which permits communication between brain cells. Brain cells that produce dopamine control movement, flexibility, and fluency. After 60% to 80% of these cells are gone, dopamine production is inhibited, and Parkinson's disease symptoms start to emerge. Researchers are concentrating their efforts on finding any early non-movement symptoms to stop the disease's development because it is thought that the disease starts many years before any evident movement-related indicators occur. Early correct diagnosis of the condition is crucial to halt the continuous advancement of Parkinson's disease and give people access to medications that can slow the disease. To do this, ongoing research into the premotor stage of Parkinson's disease is necessary.

Keywords- Classifier, Machine Learning, Early detection, and Parkinson's disease.

I. INTRODUCTION

The condition worsens and gets more incapacitating as it advances. "Neurodegenerative" describes the loss of brain cells. The human brain regularly produces dopamine in certain regions. These cells are found primarily in the ventral striatum. Via dopamine, the ventral striatum is linked to other parts of the brain that control movement. Humans can move in harmony thanks to dopamine. Parkinson's disease (PD) motor symptoms emerge after 60% to 80% of the body's endorphin cells are harmed. The nervous system, lower brain stem, and olfactory pathways are first impacted by Parkinson's disease. Movement irregularities, such as tremors and slowness of movement, sleep problems, and a reduction in scent, are believed to start several years before the disease manifests itself. To slow the spread of the disease, specialists are working as quickly as they can to find these non-movement symptoms. More than 90% of people with Parkinson's disease experience vocal impairment. Machine learning (ML) is becoming more and more popular for identifying illnesses because of its ease of use and high accuracy. In the literature review papers for ML in image processing, Parkinson's disease has been addressed using ML in image classification. This study analyses trials carried out after the condition has been diagnosed to measure the cognitive effects of Parkinson's disease and predict the severity of tremors in PD patients using ML applications.

SVM Classification methods, Naive Bayes, Decision Trees, and Neural Networks are included in this group. The study aims to compare the four available approaches for diagnosing Parkinson's.

II. RELEVANT WORK

The use of machine learning (ML) algorithms to recognize Parkinson's disease has been the subject of numerous studies. Islam et al.'s comparative investigation used a Feedforward Back Propagation Neural Network (FBANN), Random Tree (RT), SVM, and to efficiently detect Parkinson's illness. Each category underwent ten times of cross-validation. It functioned using the suggested model. To gauge the effectiveness of the model, SVM with radial basis function, “Artificial Neural Networks (ANN), and K-Nearest Neighbor (KNN)” were employed. The accuracy of the computer simulations was astounding. Researchers Shian Wu and Jiannjong Guo looked at whether Parkinson's disease patients and healthy individuals could be identified using various criteria, including “Factor Analysis, Logistic Regression, Decision Trees, and Artificial Neural Networks”. There are three ways to make a classification mistake: the selection. The framework Shirvan et al. proposed for diagnosing Parkinson's disease. The data were classified using the K-NN method. The most straightforward method for grouping similarities is K-NN.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

The most straightforward method for grouping similarities is K-NN. It is employed as a classifier when the data distribution information is insufficient. There are two sections to this method: Finding K near neighbors is step one. Step two involves using these neighbors to identify the class type. When compared to other research, it was shown that high accuracy for four optimum features, medium accuracy for seven optimized features, and low accuracy for nine optimized characteristics were reached. An outline of data mining strategies for categorization was given by Ramani and Sivagami.In their analysis's conclusion, they showed that the Random Tree algorithm successfully identified the disease and offered the highest level of accuracy 90%. The output of the network is categorized as healthy or PD using the K-Means Clustering technique. The findings demonstrate the model's high sensitivity, specificity, and accuracy in the identification and diagnosis of PD. To aid specialists in the diagnosis of PD, David et al. suggested approaches based on ANN and SVM. Performance-wise, the SVM outperforms the MLP. The SVM is highly accurate, with a 90% average. Two more characteristics are "sensitivity" and "negative predictive value," both of the high accuracy values.

III. METHODOLOGY

In this study, we created a machine learning-based empirical model using six classifier methods. In this work, PD was applied to 17 voice features from the Multi-Dimensional Voice Programme (MDVP). Only the most beneficial features were employed by employing PD, and a separate FS algorithm was used for each classification technique. From the initial feature set, these methods were utilized to produce additional subsets of features and classifications. The model's effectiveness was evaluated based on several factors. The study's flowchart is shown in the figure.

Figure 1: The flow diagram of the proposed decision assistance system

As part of data pre-processing, PD data is used to identify the main characteristics of the problems for classification. For classification accuracy to be improved, satisfactory attribute identification is essential. The effectiveness of the ML approach can be increased through dimensionality reduction.

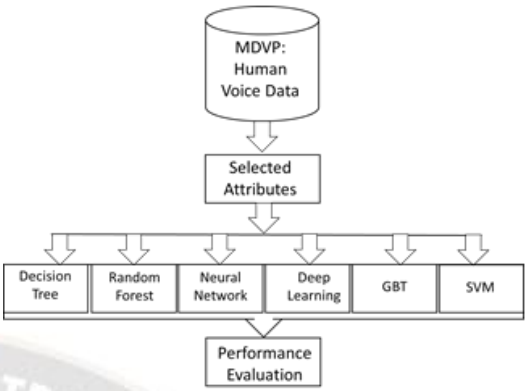

Figure 2: The performance evaluation flow chart

After preprocessing, the performance of six different classifier algorithms, "i) Decision Tree, ii) Random Forest, iii) Neural Network, iv) Deep Learning, v) Gradient Boost Tree (GBT), and vi) Support Vector Machine (SVM)" was assessed before the decision made to use an empirical model to detect Parkinson's disease early on, as shown in the flow chart above. Figure 3 explains this decision.

IV. PARKINSON'S DATASET

With the use of the Multi-Dimensional Voice Programme (MDVP), which extracts features from human speech, we analyze secondary data from the real world on the diagnosis of Parkinson's disease (PD). The dataset contains 16 features that were taken from the voices of 195 people who have Parkinson's disease. Each of the aforementioned elements is illustrated in the table below. Each row belongs to a specific person, and each column represents a unique aspect of their voice. The relevant information is contained in the table below. People's voices can be examined for sonic traits that can be used to diagnose Parkinson's disease and to differentiate between those who have the disorder and those who are well.

Table of Attributes

Sl. No.	Attribute Name	Meaning
1.	MDVP: Fo(Hz)	Average vocal fundamental frequency
2.	MDVP: Fhi(Hz)	Minimum vocal fundamental frequency
3.	MDVP: Flo(Hz)	Minimum vocal fundamental frequency
4.	MDVP: RAP	Amplitude perturbation
5.	MDVP: PPQ	Period perturbation quotient
6.	MDVP: APQ	11 points amplitude perturbation quotient
7.	MDVP: Jitter(%)	Jitter as percentage
8.	MDVP: Jitter(Abs)	Absolute jitter microsecond
9.	Jitter: DDP	Average absolute differences between cycles, divided by the average period
10.	MDVP: Shimmer	Local shimmer
11.	MDVP: Shimmer(dB)	Local shimmer in decibels

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

12.	Shimmera:APQ3	3 points amplitude perturbation quotient
13.	Shimmer : APQ5	5 points amplitude perturbation quotient
14.	Shimmer DDA	Absolute differences between the amplitude of consecutive periods
15.	NHR	Noise-to-harmonic ratio
16.	HNR	Harmonic-to-noise ratio
17.	Status	Health Status of Parkinson's Disease PD & Not PD

V. ALGORITHM

Parkinson's disease early diagnosis and research make use of machine learning algorithms. Six algorithms were employed in this investigation.

A. Decision Tree

As a general predictive analytic technique, Decision Tree analytics can be used in a wide range of industries. Decision trees are typically created utilizing an algorithmic technique that determines various splitting up a dataset in light of various characteristics. Undoubtedly one of the most popular and successful methods is sustained learning. Both classification and regression issues can be solved using decision trees. A practical model precisely forecasts the value of a dependent variable that can be created through means of data features. To remove contaminants, just two steps are needed in this method. Randomness and information. One requires a lot of entropy to be able to describe a sample accurately. We may assess the level of inequality in our data set using this index. Its value ranges from 0 to 1.
It is written as in mathematics,

$$Entropy = -\sum_{i=1}^{n} p_i * \log(p_i) \qquad (1)$$

$$Gini\ Index = 1 - \sum_{i=1}^{n} p_i^2 \qquad (2)$$

B. Random Forest

Decision tree algorithms are included in this group. By expanding bootstrap aggregation (bagging), you can utilize it to resolve regression and classification issues. Numerous decision trees are built using a bootstrap sample of the training dataset, which are all based on various sets of bootstraps. A bootstrap sample resamples a training dataset so that the same sample appears more than once. This model (CART) excludes classifying and regressing trees, which are common decision tree approaches.
In mathematics, it is expressed as,

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2 \qquad (3)$$

C. Neural Network

Inspired by biological neural networks in the nervous system or brain, neural networks were developed. It has generated a great deal of attention, and the business is continuing to research this branch of machine learning. The fundamental computational building block of a neural network is a neuron or node. It gathers data from other neurons and calculates the outcome. Each node or neuron is given a weight (w). This weight is assigned according to the relative importance of that particular neuron or node. In mathematics, it is expressed as,

$$f(b + \sum_{i=1}^{n} x_i w_i) \qquad (4)$$

x is the input to the neuron,
W is the weights,
n is the total number of inputs, and i is the counter.

D. Deep Learning

These technologies attempt to undercut human decision-making abilities by replicating important features of how the human brain works using "deep learning" ML and AI techniques. It is a crucial component of data science, which employs predictive modelling and statistics to build models based on data-driven methodologies. This human-like capacity for adaptation, learning, and algorithmic behavior must be fueled by some powerful results.
In mathematics, it is expressed as,

$$z = \sum_i w_i * x_i + b \qquad (5)$$

Weights (wi), input layer (xi), and bias (b)

E. Gradient Boost Tree

Application of Gradient Boosting to Decision Trees. The model becomes better as more learners are added using the gradient boosting method because each new learner fits the residuals of the one before it. A powerful learner is created after each step. The gradient-boosted decision trees algorithm employs decision trees as short-term learners. With the use of a loss function, the residuals are located.

$$\text{“}y = A_1 + A_2 + A_3 + (B_1 * x) + (B_2 * x) + (B_1 * x) + e_3\text{”} \qquad (6)$$

F. Support Vector Machine

The Support Vector Machine (SVM) is a method that may be applied to both regression and classification. Classification is the best strategy, even though we refer to it as a regression problem. Finding an N-dimensional hyperplane that can be used to categorize the data points is the goal of the SVM algorithm. The size of the hyperplane depends on the number of features. The hyperplane is essentially a line when there are just two input features. The hyperplane collapses to a two-dimensional plane when the input characteristics are higher than three. It's tough to envision having more than three features.
The SVM classifier is mathematically defined as,

$$h(x_i) = \begin{cases} +1\ if\ w\ .\ x + b \geq 0 \\ -1\ if\ w.x + b < 0 \end{cases} \qquad (7)$$

$$\left[\frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i(w\ .\ x_i - b))\right] + \lambda||w||^2 \qquad (8)$$

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

VI. RESULTS

We used data from a patient with PD to determine the relevance of the symptoms because the research sample size was insufficient for independent validation. Parkinson's disease movement symptoms are typically severe. Other ML-based classification models for diagnosis have been described with higher sensitivity and specificity than the current results. The following model, which details the accuracy of six different classifier techniques, is displayed for the aim of PD early detection.

Empirical Model

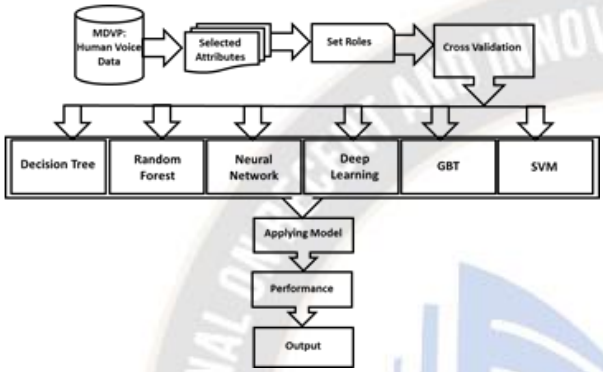

Figure 3: Early PD identification using an empirical model

ML	Accuracy	CE	Kappa	WMP	RMSE	PD_CP	Non_PD_CP
Decision Tree	82.56	17.44	0.51	76.8	0.399	86.93	66.67
Random Forest	87.18	12.82	0.628	84.85	0.289	88.61	81.08
Neural Network	82.56	17.44	0.47	80.88	0.352	84.24	73.33
Deep Learning	76.41	23.59	0.081	76.79	0.388	91.06	51.39
GBT	86.67	13.33	0.182	84.2	0.354	91.72	72.01
SVM	82.05	17.95	0.427	82.49	0.371	82.94	76.01

The model mentioned above has a number of phases and operators. In this model, we first used the selected attribute operator to extract the significant characteristics from the aforementioned dataset, and then we used the set-role operator to choose a categorical variable from the significant attributes. After that, we applied a cross-validation operator to the results of the six algorithms: i) Decision Tree, ii) Random Forest, iii) Neural Network, iv) Deep Learning, v) Gradient Boost Tree (GBT), and vi) Support Vector Machine (SVM)" to check accuracy and decision making. Finally, we obtained the findings shown below, which very effectively compare the precision of the various employed techniques and detected PD.

Results:

ML	PD	Non_PD
Decision Tree	86.93%	66.67%
Random Forest	88.61%	81.08%
Neural Network	84.24%	73.33%
Deep Learning	91.06%	51.39%
GBT	91.72%	72.01%
SVM	82.94%	76.01%

Model Accuracy and Statistics

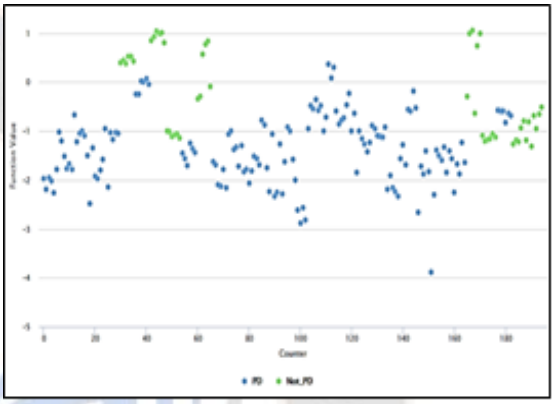

Figure 4: PD and Non-PD Scatterplots are shown in Figure 4. from total dataset.

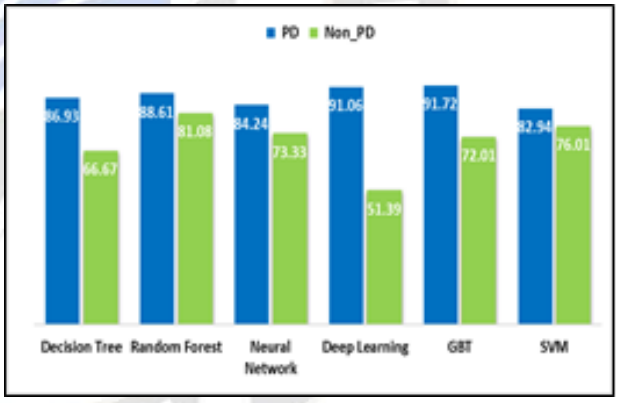

Figure 4: 5 shows a bar plot of all algorithms for PD and Non-PD.

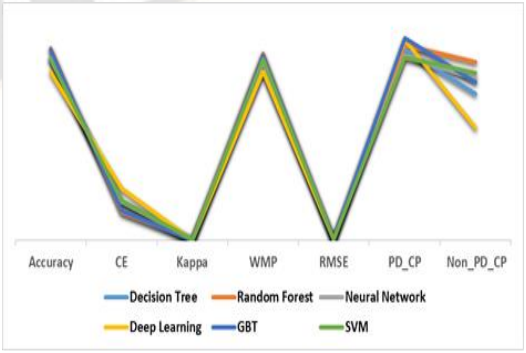

Figure 6: Diagram showing model accuracy across all methods

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

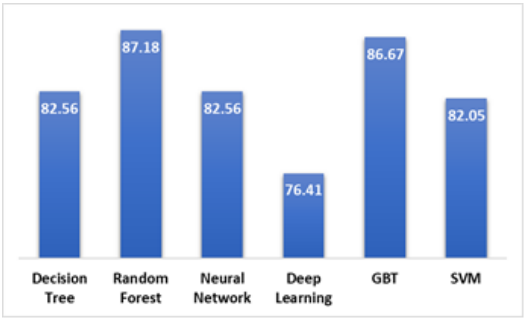

Figure 7: Model accuracy bar graph utilizing all algorithms

According to the aforementioned findings and a comparison of the six methods utilized, with a Kappa of 0.62 and an RMSE of 0.289, as well as class precisions of true PD and true Non-PD of 88.61% and 81.08%, respectively, we can state that the Random Forest method had the highest accuracy. According to the Model Accuracy Table 3, these numbers are quite good and have the maximum accuracy, while other algorithms have lower accuracy than the Random Forest Classifier algorithm. With this outcome, we may conclude that our empirical model is reliable and effectively detects PD early on. The following trees are displayed according to the PD and Non-PD conditions that were found.

Discovering Parkinson's disease

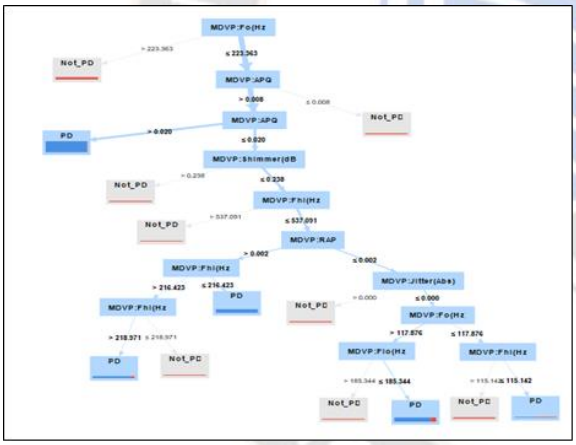

Figure 8: Tree Plot for PD Detection

Non-PD

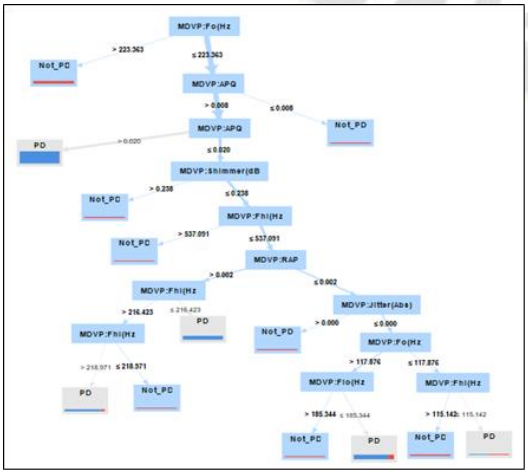

Figure 9: Tree Plot of Non-PD Detection

Scatter plots of PD and Non-PD as per MDVP

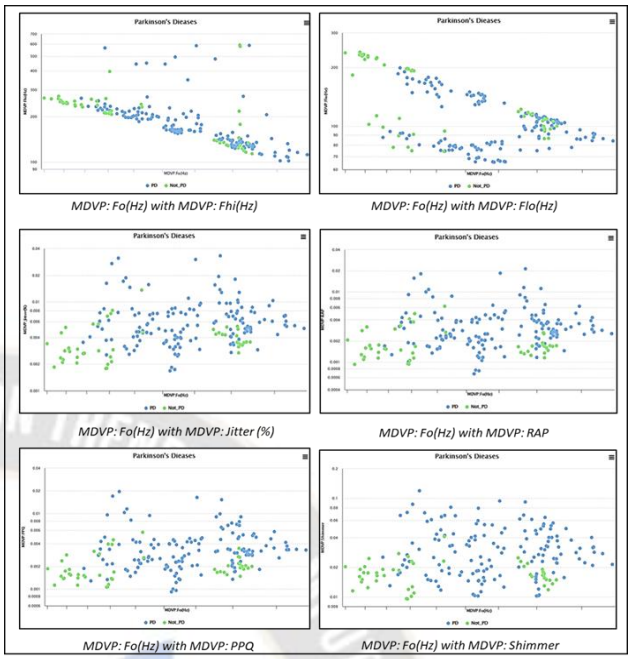

Figure 10: Scatter plots of PD and Non-PD as diagnosed by MDVP

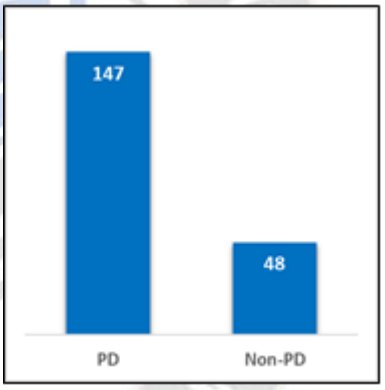

Figure 11: PD and Non-PD detected based on MDVP

Figure 10 above shows the scatter plot of PD and Non-PD, which was generated applying the features of MDVP: the human voice dataset and a machine learning-based framework. PD and Non-PD are marked by blue and green, respectively, dotted lines. The PD% is higher than the Non-PD percentage, as can be observed. Early Parkinson's disease identification has been shown to be quite effective. The essential features of the comparison line graph are shown in Figure 12.

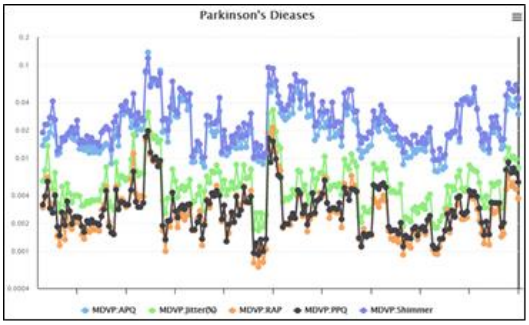

Figure 12: A comparison line graph showing the key characteristics of MDVP

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Ensemble Model Being Proposed

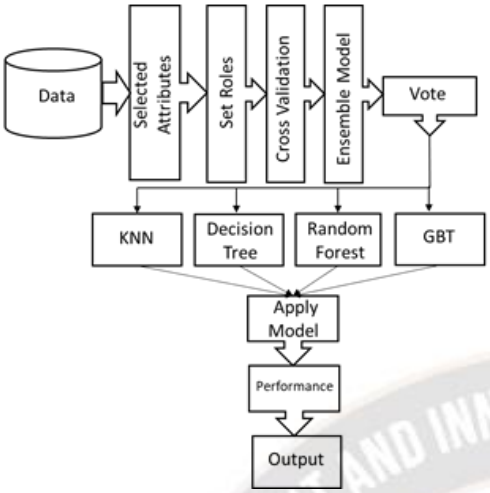

Figure 13: Proposed Ensemble Model

The above-proposed model was created by us using a number of operators and phases for the ensemble approach, and it is a well-designed system to generate enriched ML outcomes. In this model, we first used the selected attribute operator to extract the significant characteristics from the aforementioned dataset, and then we used the set-role operator to choose a categorical variable from the significant attributes. Then, in order to achieve better results, we applied the ensemble approach using a cross-validation operator. Ensemble approaches are ways to build several models, and then combine them to get superior outcomes. Results using ensemble approaches are frequently more precise than those from a single model. This has been shown in a number of machine learning competitions where the winning solutions were predicted. Here, we used for classification, voting ensemble methods are used, but for regression, averaging is used. Using the training dataset, numerous classification and regression models are first built. The same training dataset and the same algorithm can be used to build each base model, or the same dataset and several techniques can be used. Here, we employed a voting approach with four different machine learning (ML) algorithms: 1) k-NN, 2) Decision Tree, 3) Random Forest, and 4) GBT. The results were then passed through by using the model and performance measurement operators. The final findings were 92.67% accuracy, which compares the accuracy of the various employed algorithms and the detected PD that is discussed in model accuracy statistics (table 3). We can see that Random forests have the highest accuracy of 87.18% in the accuracy chart. While the accuracy of our suggested model is 92.67%.

So, we can conclude that the suggested model is reliable and exhibits 5.49% greater accuracy. Here, the accuracy results and graphs are displayed.

ML	Accuracy	AUC	AUC(optimistic)	AUC(pessimistic)
Vote	92.67%	0.831	0.947	0.721

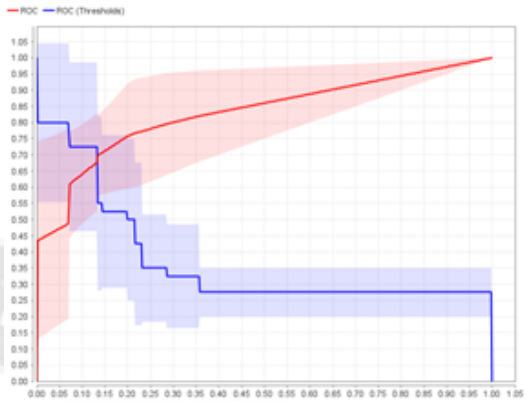

Figure 14: Area under the Curve (AUC)

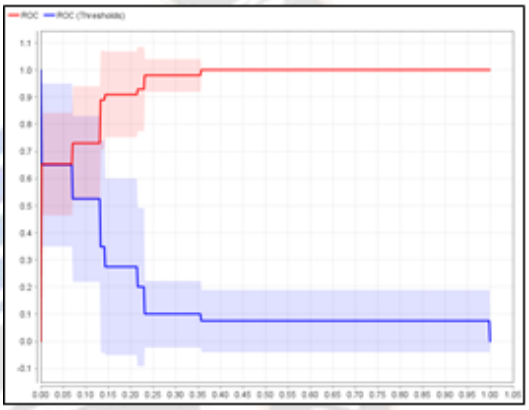

Figure 15: AUC (Optimistic)

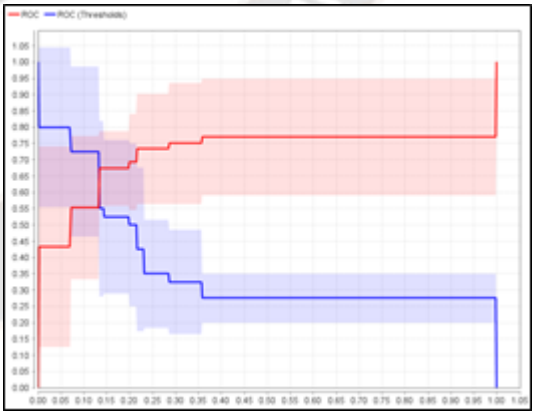

Figure 16: AUC (pessimistic)

The AUCs of 14, 15, and 16 are all close to 1 in the aforementioned figures (ROC red color line). This supports our model's superiority and great separability. Parkinson's disease (PD) is influenced by the optimistic class represented by the red distribution curve, but not by the pessimistic class represented by the blue distribution curve.

VII. CONCLUSION

In this study, we investigated the association between the Parkinson's illnesses and developed an early Parkinson's illness simulation model identification using machining learning techniques. In our investigation, the classifiers "Decision Tree, Random Forest, Neural Network, Deep Learning, Gradient Boost Tree (GBT), and Support Vector Machine (SVM)" are utilized. The training data are designated as (X1, y1)…, (Xn, yn).For a better understanding of the origins of the disease, to start therapies, and to develop effective medications, early detection of Parkinson's disease is essential. This study provided an automated empirical model differentiating between healthy individuals and Parkinson's disease patients based on the attributes.

With an accuracy of 87.18%, the empirical model showed good detecting capabilities. This is mainly due to the advantages of the Random Forest method in learning features from PD data without the need for manually developed feature extraction. The results show that the proposed model outperforms the six machine learning-based frameworks under consideration in distinguishing between Parkinson's disease patients and healthy persons. SVM, neural networks, deep learning, and boosting approaches all produce similar results in terms of performance. Even if Random Forest surpasses machine learning models, it is challenging to declare that it is the best classifier.

Comparing our suggested ensemble model to the current empirical model, it is superior. Because we are better able to observe the correctness and results of the suggested model than the empirical model. Our ensemble model achieved 92.67% accuracy compared to the empirical model's best accuracy of 87.18%. Thus, it has been shown that our ensemble model is great and trustworthy.

VIII. REFERENCE

[1] S. Ashour, A. El-Attar, N. Dey, H. A. El-Kader, and M. M. A. El-Naby. (2020). Long short-term memory-based patient-dependent model for fog detection in Parkinson’s disease.Pattern Recognition Letters, 131, 23–29,

[2] Wagner, N. Fixler, and Y. S. Resheff. (2017). A wavelet-based approach to monitoring Parkinson’s disease symptoms. *IEEE International Conference on Acoustics*, Speech and Signal Processing (ICASSP). IEEE, 5980–5984.

[3] Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig. (2009). Accurate telemonitoring of Parkinson’s disease progression by noninvasive speech tests.IEEE Transactions on Biomedical Engineering, 57(4), 884–893.

[4] A. Zhao, L. Qi, J. Li, J. Dong, and H. Yu. (2018). A hybrid Spatio-temporal model for detection and severity rating of Parkinson’s disease from gait data. Neurocomputing, 315, 1–8.

[5] D. Braga, A. M. Madureira, L. Coelho, and R. Ajith. (2019). Automatic detection of Parkinson’s disease based on acoustic analysis of speech. Engineering Applications of Artificial Intelligence, 77, 148–158.

[6] G. Solana-Lavalle, J.-C. Galán-Hernández, and R. Rosas-Romero. (2020). Automatic Parkinson’s disease detection at early stages as a pre-diagnosis tool using classifiers and a small set of vocal features. Biocybernetics and Biomedical Engineering, 40(1), 505–516.

[7] Hadjahamadi, A.H. and Askari, T.J. (2012). A Detection Support System for Parkinson’s Disease Diagnosis Using Classification and Regression Tree. Journal of Mathematics and Computer Science, 4, 257-263.

[8] I. El Maachi, G.-A. Bilodeau, and W. Bouachir. (2020). Deep 1d-convnet foraccurate Parkinson’s disease detection and severity prediction from gait. Expert Systems with Applications, 143, 113075.

[9] R. Prashanth, S. D. Roy, P. K. Mandal, and S. Ghosh. (2016). High-accuracydetection of early Parkinson’s disease through multimodal features and machine learning. International journal of medical informatics, 90,13–21.

[10] R. Prashanth and S. D. Roy. (2018). Early detection of Parkinson’s diseasethrough patient questionnaire and predictive modelling. International journal of medical informatics, 119, 75–87.

[11] R. Prashanth, S. D. Roy, P. K. Mandal, and S. Ghosh. (2016). High-accuracydetection of early Parkinson’s disease through multimodal features and machine learning. International Journal of medical informatics, 90, 13–21.

[12] Sharma, A. and Giri, R.N. (2014) Automatic Recognition of Parkinson Disease via Artificial Neural Network and Support Vector Machine. IJITEE, 4, 35-41.

[13] T. Arroyo-Gallego, R. Trincado et al. (2017). Detection of motor impairment in Parkinson’s disease via mobile touchscreen typing. IEEE Transactions on Biomedical Engineering, 64(90),1994–2002.

[14] Wu, Jiannjong Guo. (2011). A Data Mining Analysis of the Parkinson’s Disease. iBusiness, 3(1), 2011.

[15] Zhao, Y.H. and Zhang, Y.X. (2008). Comparison of Decision Tree Methods for Finding Active Objects. Advances in Space Research, 41, 1955-1959.