Machine Learning-Based Framework for Early Detection of Distinguishing Different Stages of Parkinson’s Disease

Archana Panda[1]and Dr Prachet Bhuyan[2]
[1]GIFT, Bhubaneswar
[2]School of Computer Science Engineering , KIIT University

Abstract

Parkinson’s disease (PD) is caused by a disruption in the brain cells that produce dopamine, a substance that allows brain cells to communicate with one another. Dopamine-producing cells in the brain are in charge of movement control, adaptation, and fluency. When 60–80% of these cells are lost, there is insufficient dopamine production, and Parkinson's motor symptoms appear. It is believed that the disease begins many years before the movement-related symptoms appear, so researchers are looking for ways to identify the non-movement symptoms that appear early in the disease as early as possible, thereby halting the disease's progression. Accurately detecting Parkinson’s disease during an early stage is unquestionably critical for delaying its steady progress and providing patients with access to disease-modifying therapy. To that end, the premotor stage of Parkinson’s disease should be closely monitored. Based on the results of the test, a technique is introduced to determine whether an individual has Parkinson's disease or not. Premotor characteristics Particularly, several indicators have been considered to detect Parkinson’s disease at an early stage. A comparison of the proposed different Machine Learning models will be used based on relatively small data sets of healthy individuals and early Parkinson’s disease patients reveals. The designed model should achieve the highest accuracy on average. In this study, we provided the feature importance of the Parkinson's disease (PD) detection process based on the Machine Learning Process. Decision Trees, Random Forest, Neural Networks, Deep Learning, Gradient Boosted Tree, and Support Vector Machines, algorithms were used to classify Parkinson’s patients. These algorithm’s feature elimination outperformed all other methods. With the fewest number of voice features, 87.18% accuracy was achieved for Parkinson’s diagnosis. We find that these techniques perform well in classifying early Parkinson’s disease and healthy normal people, with high accuracy. The analysis of non-invasive biological markers for disease detection is critical for accurate clinical diagnosis. As a result, the analysis can be used to detect Parkinson’s disease at an early stage.

Keywords: Parkinson’s disease, Machine Learning, Early detection, Classifier.

1. Introduction

Parkinson’s disease is a progressive neurodegenerative brain disease that progresses slowly. The term “neurodegenerative” refers to the loss of brain cells. Dopamine is normally produced by brain cells in specific areas of the human brain. These cells are focused in an area of the brain known as the ventral striatum. Dopamine is a chemical that communicates between the ventral striatum and other areas of the brain that control movement. Dopamine enables people to move in unison. When 60 to 80% of endorphin cells are lost, not enough endorphins are produced, and Parkinson's disease (PD) motor symptoms occur. The nervous system, lower brain stem, and olfactory tracts are the first to show signs of Parkinson’s disease. The disease is thought to begin several years before movement disorders such as loss or decrease of ability to smell, sleep and constipation, tremor and slowing of movement. As a result, researchers are looking for ways to identify these nonmovement symptoms that appear early in the disease as soon as possible, to halt the disease's progression. 90% of Parkinson’s disease patients have vocal impairments. Because of its ease of implementation and high accuracy, machine learning (ML) is increasingly being used for medical disease diagnosis. In the literature review papers for image classification to be used for ML in image processing, ML has been used for the treatment of Parkinson's disease. This paper focuses on studies conducted after Parkinson’s disease has been diagnosed, using ML methods to estimate the cognitive consequences of PD and predict the tremor level of PD patients using an ML application.

Support vector machine (SVM), neural networks, decision trees, and Naive Bayes are some classification techniques. The study's goal is to analyse and compare the performance of four of the above-mentioned classification techniques upon Parkinson's diagnosis. First, we try comparing the classifier accuracy on the actual and partitioned PD datasets, and then we compare their effectiveness using the attribute values selection technique.

2. Related Work

Several studies have focused on the use of ML methods for the automatic recognition of Parkinson's disease. Islam et al. used comparative analysis to detect Parkinson's disease effectively using Random tree (RT), SVM, and Feedforward Back Propagation Neural Network (FBANN). For each classification, a 10-fold cross-validation analysis was performed. The proposed model was successful. Sharma and Giri assessed the model’s performance with Artificial Neural Networks (ANN), K-nearest neighbour (KNN), and SVM with radial basis function. The models were extremely accurate. Shian Wu and Jiannjong Guo used Factor Analysis, Logistic Regression, Decision Trees, and Artificial Neural Networks to see if voice features could distinguish a Parkinson’s disease patient from a healthy one. They stated that the decision tree has the lowest classification error of the three methods, the logistic regression model has the second-lowest, and ANN has the highest classification error.

Shirvan et al. suggested a framework for detecting Parkinson’s disease. The K-NN method was used to classify the data. K-NN is the simplest method for grouping similarity. When the information for the distribution of the data is insufficient, K-NN is used as a classifier. This method is divided into two parts: i) Find K close neighbours, ii) Use these close neighbours to determine the class type. It was demonstrated that a high accuracy per four optimised features and medium accuracy per seven optimised features, and low accuracy for nine optimised features were achieved, which is a remarkable result when compared to other studies. Ramani and Sivagami provided an overview of data mining techniques used for classification.

They concluded by demonstrating that the Random Tree algorithm correctly classified the Parkinson's disease dataset and provided maximum accuracy. The accuracy of the linear discriminant analyses C4.5, Cs-MC4, and K-NN is greater than 90%. Rashidah et al. proposed to use the Multilayer Feedforward Neural Network (MLFNN) with Back-propagation (BP) algorithm to develop a model for early detection and diagnosis of Parkinson’s disease. The K-Means Clustering algorithm is used to classify the output of the network as healthy or PD. The results show that the model can be used in the detection and diagnosis of Parkinson's disease due to its high sensitivity, specificity, and accuracy. David et al. suggested methods based on ANN and SVM to assist specialists in the diagnosis of PD. The SVM outperforms the MLP in terms of performance. The SVM has a high accuracy of around 90%. Other parameters with very high accuracy are “sensitivity” and “negative predictive value,” which have high accuracy values.

3. Methodology

In this study, we have designed an empirical model based on Machine Learning with six classifier algorithms. Multi-Dimensional Voice Program (MDVP) of 17 voice features were subjected to PD considered in this study. By using PD, only the most useful features were used for each classification method, a different FS algorithm was used. These algorithms were used to generate new subsets of features and classifications from the original feature set. The model’s performance was assessed using a variety of criteria. The depicts the study's flow chart.

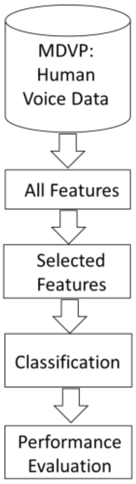


Figure 1: The proposed decision support system's flow chart

The key features of the problems are identified for classification from PD data, which is a data pre-processing process. Satisfactory attribute identification is critical for improving classification accuracy. Dimensionality reduction can improve the performance of the ML method.

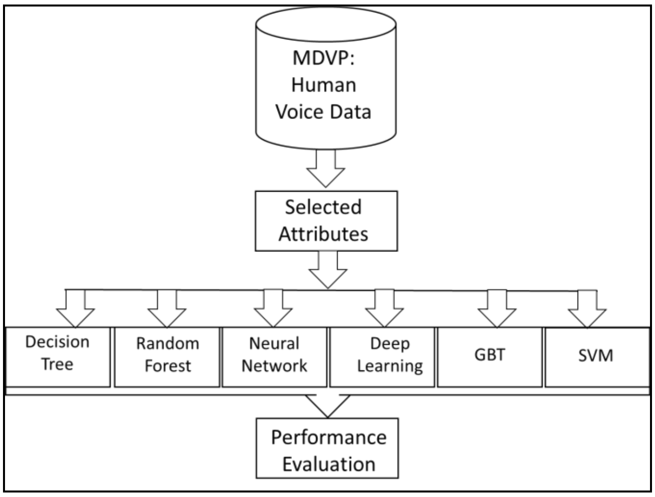


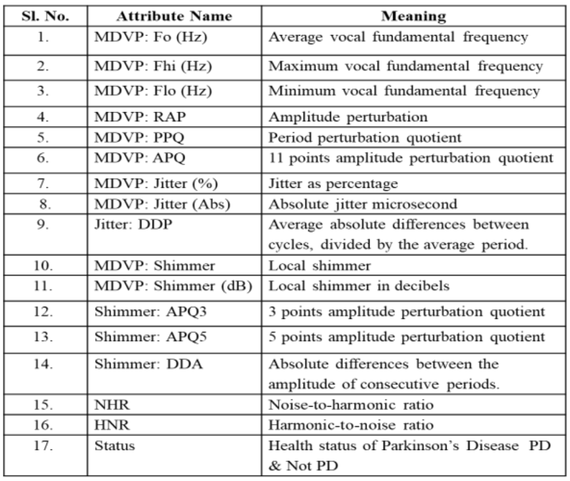
Figure 2: The performance evaluation flow chart

The above flow chart defines that, after preprocessing process it used six different classifier algorithms “Decision Tree, Random Forest, Neural Network, Deep Learning, Gradient Boost Tree (GBT), and Support Vector Machine (SVM)” to measure the performance and the decision has taken for early detection of Parkinson’s Disease through our an empirical model, which has been explained in figure 3.

4. Parkinson’s Dataset

We analyse real-world secondary data on Parkinson’s Disease (PD), where the disease is diagnosed using several extracted features from the human voice’s Multi-Dimensional Voice Program (MDVP). The dataset consists of a human voice with 16 features extracted from 195 people, who had Parkinson’s disease. The following table depicts the mentioned features each column denotes a specific voice feature, and each row corresponds with one of these individual people. The details are given in the following table. These features are extracted from human voices and used to make a diagnosis of Parkinson’s disease to determine who has reached the stages of the disease and who is healthy.

Attribute table



Sl. No.	Attribute Name	Meaning
1.	MDVP: Fo (Hz)	Average vocal fundamental frequency
2.	MDVP: Fhi (Hz)	Maximum 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
12.	Shimmer: 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

Table 1: Attribute table of PD

5. Algorithms

Machine Learning Algorithms are used for the early detection and investigation of Parkinson's Disease. This study has used six algorithms.

- **Decision Tree:** Decision Tree Analysis is a general predictive analysis tool with applications in a variety of fields. In general, decision trees are built using an algorithmic approach that identifies different ways to split a data set based on different conditions. It is among the most commonly used and useful supervised learning methods. Decision Trees are a semi-supervised learning method that can be used for classification as well as regression tasks. The goal is to build a model that can predict the value of the dependent variable using simple decision rules derived from data features. There are a few impurity measures, but for this algorithm, it will be only two steps Entropy and Gini index. Entropy is the volume of information required to accurately describe a sample. The Gini index is a measure of inequality in the sample. It has a value between 0 and 1. Mathematically it is written as,

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

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

- **Random forest:** It is a collection of decision tree algorithms. It is a decision tree extension of bootstrap aggregation (bagging) that can be used for classification and regression problems. Bagging involves the creation of several decision trees, each of which is based on a different bootstrap sample of the training dataset. A bootstrap sample is a sample of the training dataset in which a sample appears more than once, also known as sampling with replacement. Because each decision tree is fit on a slightly different training dataset, and thus has a slightly different performance, bagging is an effective ensemble algorithm. In contrast to traditional decision tree models such as classification and regression trees (CART). Mathematically it is written as,

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

- **Neural Network:** The biological neural networks in the brain, or nervous system, inspired neural networks. It has sparked a lot of interest, and research on this subset of Machine Learning is still ongoing in the industry. A neuron or node is the basic computational unit of a neural network. It receives information from other neurons and computes the result. Weight is assigned to each node/neuron (w). This weight is assigned based on the relative importance of that specific neuron or node. Mathematically it is written as,

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

b = bias, x = input to neuron, w = weights, n = the number of inputs and i = counter

- **Deep Learning:** Deep learning is a machine learning and artificial intelligence method that is aimed to threaten humans and their actions based on certain human brain functions to make active decisions. It is a critical component of data science that channels its modelling based on data-driven techniques through predictive modelling and statistics. There must be some powerful forces driving such a human-like ability to adapt and learn and function accordingly, which we commonly refer to as algorithms. Mathematically it is written as,

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

w_i = weights, x_i = input layer, b = bias

- **Gradient Boost Tree:** Decision Trees with Gradient Boosting The gradient boosting algorithm sequential manner combines weak learners in such a way that each new learner fits the residuals from the previous step, resulting in an improved model. The final model combines the results of each step to produce a strong learner. Decision trees are used as week learners in the gradient boosted decision trees algorithm. The residuals are detected using a loss function.

$$y = A_1 + A_2 + A_3 + (B_1 * x) + (B_2 * x) + (B_1 * x) + e_3 \qquad (6)$$

- **Support Vector Machine:** The Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for classification as well as regression. Though we call it a regression problem, it is best suited for classification. The SVM algorithm's goal is to find a hyperplane in an N-dimensional space that classifies the data points. The size of the hyperplane is determined by the number of features. If there are only two input features, the hyperplane is simply a line. When the number of input features reaches three, the hyperplane transforms into a two-dimensional plane. When the number of features exceeds three, it becomes difficult to imagine. The SVM classifier define in mathematical as,

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

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

6. Data Analysis and Results

We used data from a recognized PD patient to evaluate the relevance of the symptoms because the study sample size was insufficient for independent validation. the severity of motor symptoms, which is a hallmark of Parkinson’s disease. Several ML-based classification models for diagnosis have been reported with higher sensitivity and specificity than the current results. The following model defines the accuracy of six different classifier algorithms and the decision tree is shown for early detection of PD.

Empirical Model

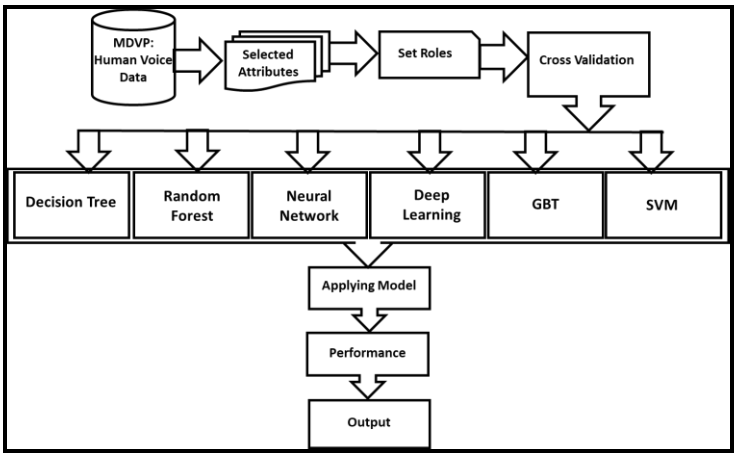


Figure 3: Empirical Model for early detection of PD

We have designed the above model with several steps and operators. In this model first, we have extracted the significant attributes from the said dataset using the selected attribute operator, and then we have selected a categorical variable from the significant attributes using the set-role operator. After that, we used a cross-validation operator for checking accuracy and decision making using 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)” and passing through the result with applying the model and measuring performance operators. Finally, we got the following results, which very nicely compare the accuracy of the different used algorithms 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%

Table 2: Identified PD and Non-PD using all algorithms

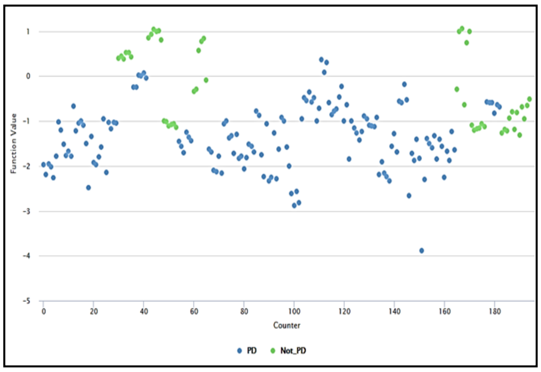


Figure 4: ScaterPlot of PD and Non-PD from total dataset.

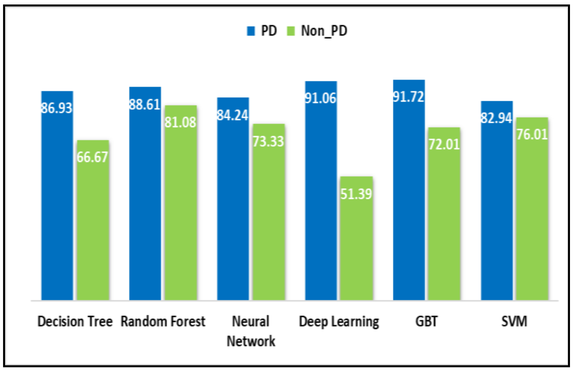


Figure 5: Bar plot of PD and Non-PD using all algorithms

Model Accuracy and Statistics

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

Table 3: Model accuracy in percentage using all algorithms

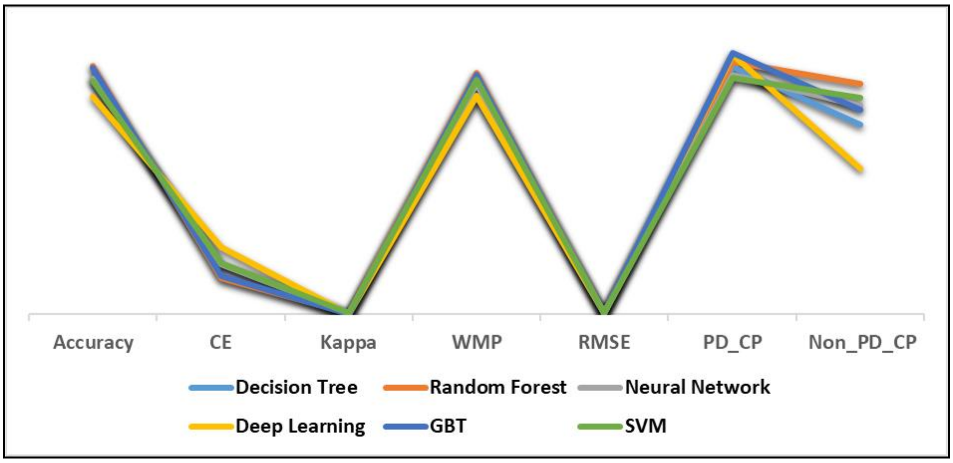


Figure 6: Line graph of Model accuracy using all algorithms

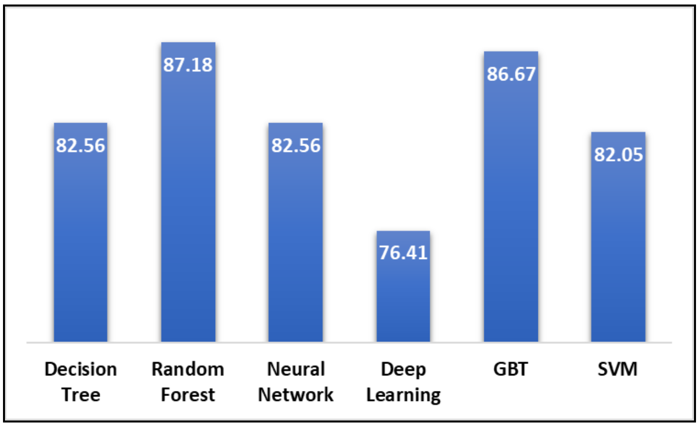


Figure 7: Bar Plot of Model accuracy using all algorithms

As per the above results and comparing the all used six algorithms, we can conclude the Random forest algorithm got the highest accuracy at 87.18%, Kappa is 0.62, RMSE is 0.289, and class precision of true PD and true Non-PD is 88.61%, 81.08%. These values are quite well and highest accuracy, whereas other algorithms are getting less accuracy than Random Forest Classifier algorithm as per the Model Accuracy table 3. With the said result, we can say our empirical model is a robust model and early detected PD. The following trees are shown as per conditions detected for PD and Non-PD.

Detection of Parkinson's Disease

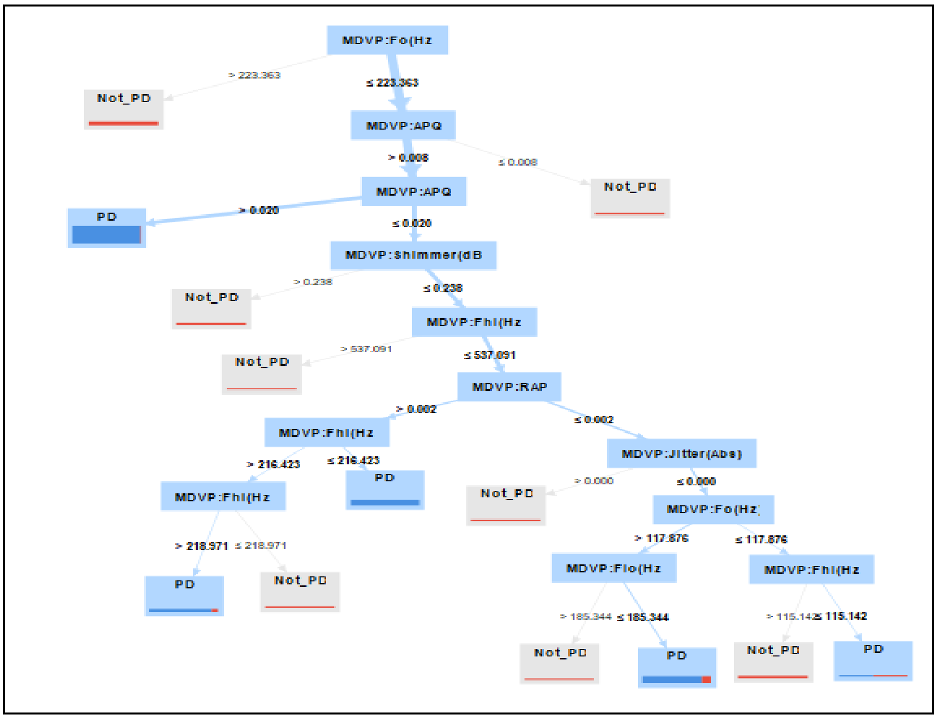


Figure 8: Tree Plot of detecting PD

Detection of Non-Parkinson's Disease

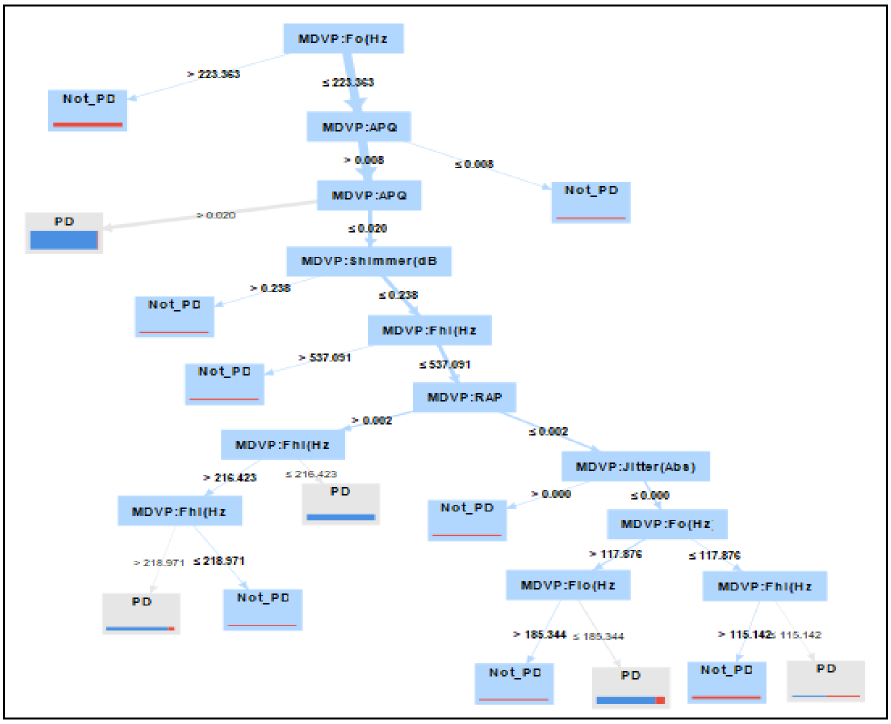
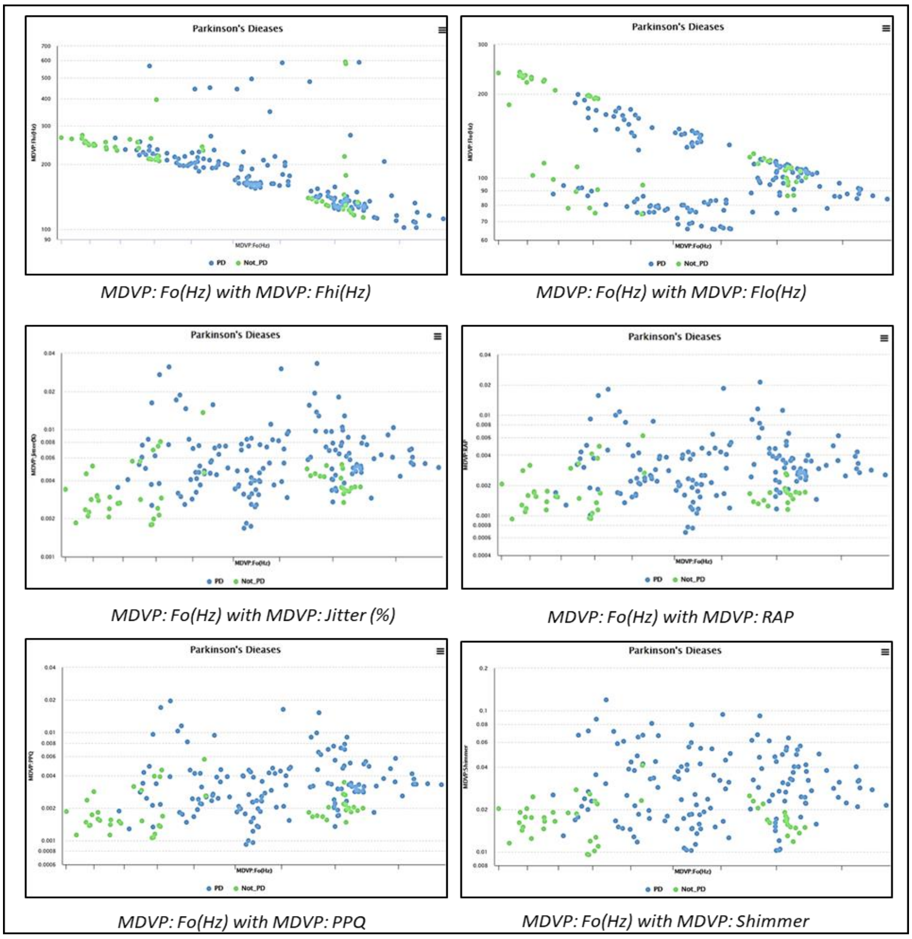


Figure 9: Tree Plot of detecting Non-PD

Scatter plots of PD and Non-PD as per MDVP



MDVP: Fo(Hz) with MDVP: Fhi(Hz)

MDVP: Fo(Hz) with MDVP: Flo(Hz)

MDVP: Fo(Hz) with MDVP: Jitter (%)

MDVP: Fo(Hz) with MDVP: RAP

MDVP: Fo(Hz) with MDVP: PPQ

MDVP: Fo(Hz) with MDVP: Shimmer

Figure 10: Scatter plots of detected PD and Non-PD as per MDVP

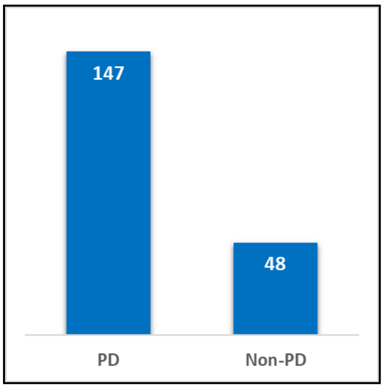


Figure 11: Detected PD and Non-PD as per MDVP

We can observe in the above figure 10, the scatter plot of PD and Non-PD, which is detected as per the features of MDVP: the human voice dataset using the Machine Learning Based Framework. The blue colour dotted denoted PD and the green colour dotted denoted Non-PD. Here we can see the PD percentage is more than the Non-PD. The early detection of Parkinson's Disease has proved to be very efficient and effective. The major features of the comparison line graph are shown in figure 12.

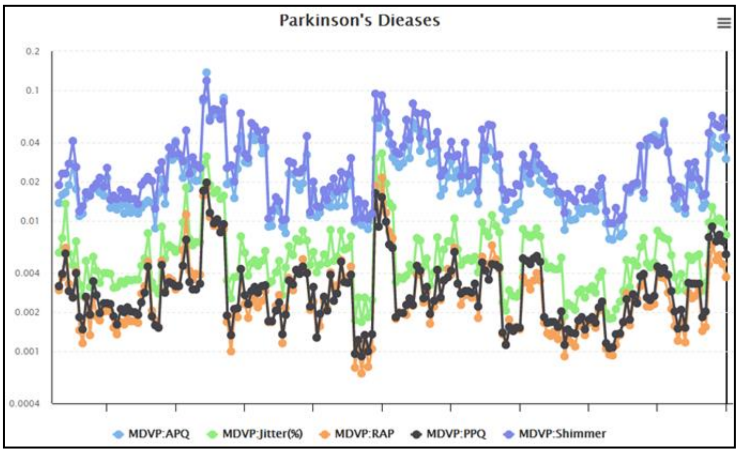


Figure 12: The comparison line graph of major features of MDVP

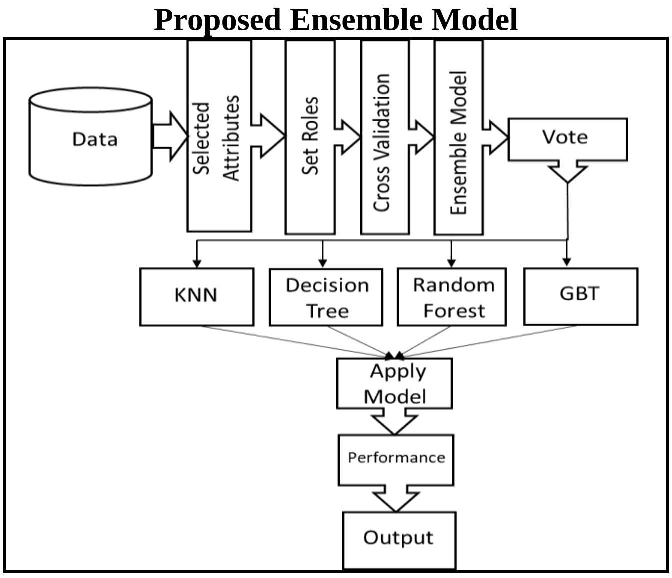


Figure 13: Proposed Ensemble Model

We have designed the above-proposed model with several steps and operators for the ensemble method, it is a well-designed system to produce enriched ML results. In this model first, we have extracted the significant attributes from the said dataset using the selected attribute operator, and then we have selected a categorical variable from the significant attributes using the set-role operator. After that, we used a cross-validation operator to execute the ensemble method for better results. Ensemble methods are techniques for developing multiple models and then combining them to produce better results. Ensemble methods typically yield more accurate results than a single model. This has been demonstrated in several machine learning competitions where the expected-to-win solutions. We used here Voting ensemble methods are employed for classification while averaging is employed for regression. The first step in this method is to create multiple classification/regression models using the training dataset. Each base model can be built using different splits of the same training dataset and the same algorithm, or it can be built using the same dataset but different algorithms. Here we used in voting method with four several ML algorithms i) k-NN, ii) Decision Tree, iii) Random Forest, and iv) GBT together passed through the result by applying the model and measuring performance operators. Finally, we got the results of 92.67% accuracy, which very nicely compares the accuracy of the different used algorithms and detected PD that is explained in model accuracy statistics (table 3). In said accuracy table we can observe Random forests given the highest accuracy of 87.18%. Whereas our proposed model shows 92.67% accuracy. So we can say our proposed model is robust and it shows 5.49% more accuracy. The following accuracy result and graphs are shown here.

ML	Accuracy	AUC	AUC (optimistics)	AUC (pessimistics)
Vote	92.67%	0.831	0.947	0.721

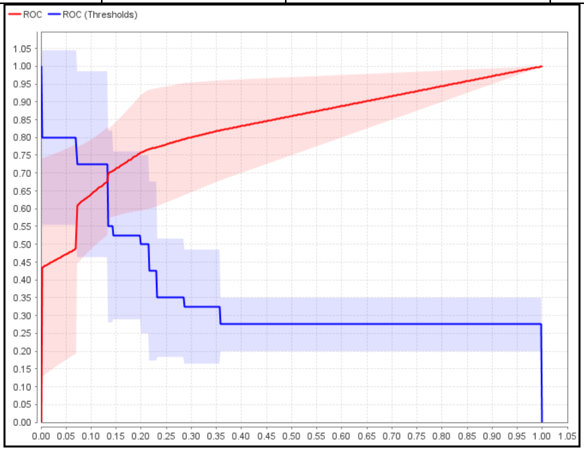
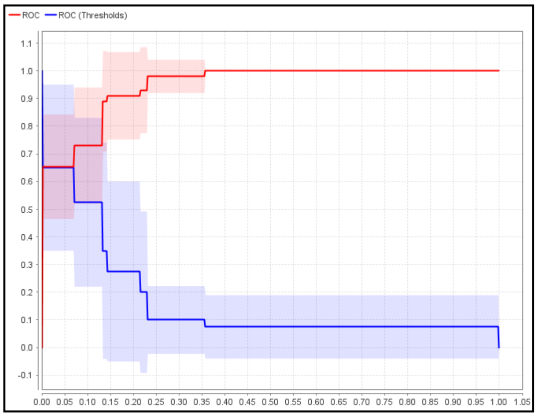


Figure 14: Area under the Curve (AUC)



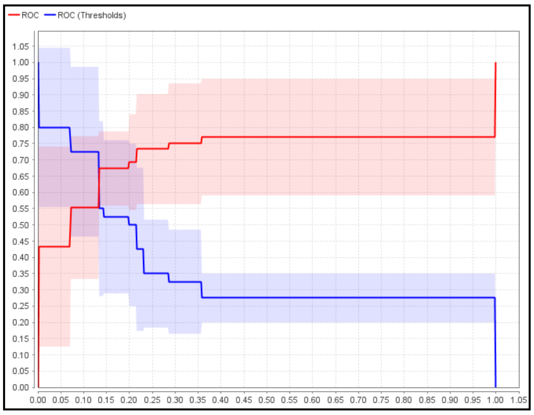


Figure 15: AUC (Optimistic)

Figure 16: AUC (pessimistic)

According to the above figures 14, 15, and 16 all are AUC near 1, (ROC red colour line). It means our model is excellent and it has a good measure of separability. The red distribution curve is an optimistic class which means affected Parkinson's Disease (PD) and the blue curve distribution is pessimistic class means not affected PD.

7. Conclusion

In this study, we used machining learning methods to investigate the relationship between Parkinson’s disease and to develop a model for early Parkinson’s disease detection. The training data is denoted by {(X_1, y_1), ……, (X_n, y_n), and the classifiers “Decision Tree, Random Forest, Neural Network, Deep Learning, Gradient Boost Tree (GBT), and Support Vector Machine (SVM)” are used in our study. Early detection of Parkinson’s disease is critical for gaining a better understanding of the disease’s causes, initiating treatment methods, and developing appropriate treatments. Based on the features, this study proposed an empirical model for automatically distinguishing between normal individuals and patients with Parkinson’s disease.

The empirical model demonstrated good detection capability, with an accuracy of 87.18%. This is primarily due to the Random forest algorithm’s desirable characteristics in learning features from PD data without the need for hand-crafted feature extraction. The results show that the designed model outperforms the six considered machine learning-based frameworks in distinguishing normal people from Parkinson’s disease patients. The performance of the boosting methods, SVM, Neural Network, and deep learning are also comparable. Even though Random Forest outperforms machine learning models, it is difficult to say that the Random Forest classifier is superior to the others.

Our proposed ensemble model is superior to compare with the existing empirical model. Because we can observe the proposed model result and accuracy better than the empirical model. The empirical model's highest accuracy is 87.18%, whereas our ensemble model got 92.67%. So, it has proved that our proposed ensemble model is excellent and robust.

Reference

A. 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,

A. 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.

A. 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.

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.

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.

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 by using classifiers and a small set of vocal features. *Biocybernetics and Biomedical Engineering*, 40(1), 505–516.

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.

I. El Maachi, G.-A. Bilodeau, and W. Bouachir. (2020). Deep 1d-convnet for accurate Parkinson's disease detection and severity prediction from gait. *Expert Systems with Applications*, 143, 113075.

R. Prashanth, S. D. Roy, P. K. Mandal, and S. Ghosh. (2016). High-accuracy detection of early Parkinson's disease through multimodal features and machine learning. *International journal of medical informatics*, 90, 13–21.

R. Prashanth and S. D. Roy. (2018). Early detection of Parkinson's disease through patient questionnaire and predictive modelling. *International journal of medical informatics*, 119, 75–87.

R. Prashanth, S. D. Roy, P. K. Mandal, and S. Ghosh. (2016). High-accuracy detection of early Parkinson's disease through multimodal features and machine learning. *International Journal of medical informatics*, 90, 13–21.

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

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.

Wu, Jiannjong Guo. (2011). A Data Mining Analysis of The Parkinson's Disease. *iBusiness*, 3(1), 2011.

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