





Machine learning based Parkinson's disease Prediction

Anila R Nambiar  *¹ and Akshatha Rani K A  †²

¹Assistant professor, Dept. of MCA, Acharya Institute of Technology, Bangalore

²Dept. of Computer Applications, Acharya Institute of Graduate Studies, Bangalore

Abstract

Early detection of vocal alterations in people with Parkinson's disease (PD) allows for pre-emptive care before the emergence of more severe physical symptoms. This study examines both static and dynamic aspects of communication that are relevant to identifying PD. A comparison between articulation transition features in PD patients and healthy control (HC) speakers shows differences in articulation transitions and trends in the fundamental frequency curve. We suggest collecting time-series data utilizing an unidirectional long-short-term memory (LSTM) model, with an emphasis on the dynamic features of speech signals to identify Parkinson's disease (PD). The study evaluates speech capabilities by analyzing the strength of transitions from voiced to mute segments (offset) and voiced-to-voiced segments. Two assessment techniques are employed, using 10-fold partitioning of the dataset while ensuring no overlap of data from the same individual in the validation process. We recommend using the bidirectional LSTM framework to capture dynamic elements of speech in this investigation, which may provide new insights into PD detection.

Keywords: Parkinson's Disease (PD). Fundamental Frequency Curve. Long-Short Term Memory. Articulation Transitions.

*Email: anila2793@acharya.ac.in Corresponding Author

†Email: rani123chen@gmail.com

1 Introduction

Parkinson's disease is the second most common neurological disorder. It is estimated that over 7 to 10 million people worldwide are affected by Parkinson's disease (PD). Even while the illness is not deadly in and of itself, it has a substantial negative influence on quality of life and frequently results in a shorter life expectancy than in healthy individuals (Office of Communications and Public Liaison, 2023). The inability of people to carry out daily chores, including walking steadily or holding a pen consistently, is an indicator of their declining quality of life. A tremor or shaking of the body that cannot be controlled, bradykinesia, or diminished movement of the limbs, difficulties sitting and standing, loss of balance, muscle stiffness, drooping of the face, difficulty writing and drawing, eventually the loss of control over finger movement, unstable posture, etc (Goubault et al., 2017). are some of the major symptoms of Parkinson's disease. Parkinson's Disease (PD) has been the subject of extensive research for centuries, yet in many cases, the exact cause remains elusive. But in 1961, a strong correlation was found in the brain between dopamine levels and Parkinson's disease. Within the Basal Ganglia, which is a part of the nervous system where neuronal loss and poor regeneration cause dopamine levels to drop, the condition is frequently identified. Dopaminergic therapies are therefore frequently employed to control the illness and limit its course, while full recovery is never certain (Ramesh & Arachchige, 2023)

The outcomes of the experiment demonstrate that the suggested technique significantly outperforms conventional machine learning models using static characteristics in terms of PD detection accuracy. In order to identify the most successful machine learning algorithm for Parkinson's disease prediction, this study assesses the accuracy of six different algorithms. Additionally, it aims to classify patients based on the severity of their condition and assess the stage of the disease. Machine learning is utilized for a range of purposes, including analysis to achieve our objectives. While machine learning methods are widely used, precise prediction of Parkinson's disease is critical, and achieving high accuracy is essential (Alshammri et al., 2023). As a result, various evaluation methods are applied to assess these algorithms, enabling medical professionals and researchers to gain deeper insights into the disease and identify the most effective ways to forecast it. The key contributions of this paper include extracting categorized accuracy relevant to predicting Parkinson's disease, comparing multiple machine learning algorithms, and identifying the best-performing method for Parkinson's prediction.

2 Literature Review

Athanasios Tsanas et al.'s (2012) introduced a novel method for distinguishing between Parkinson's disease (PD) patients and control participants by detecting dysphonia using machine learning algorithms. In their study, they presented PPE, a new and robust dysphonia measure that performs effectively even in challenging and unpredictable environments. Their research utilized data from 195 sustained vowel phonations collected from 31 individuals, 23 of whom were diagnosed with PD. The participants' ages ranged from 46 to 85, and six phonations were recorded for each individual. After applying feature filtering, they identified ten largely uncorrelated measures and explored all possible feature combinations, finding that four provided the most accurate classification. Their proposed model achieved an accuracy rate of 90.4%. The study concluded that the best classification results were obtained by combining traditional frequency-to-noise ratios with unconventional methods. Das's (2010) evaluated the effectiveness of different classification methods for accurately diagnosing Parkinson's disease. Many classifiers used for PD detection rely on Sass-based software in his analysis. Regression, neural network, decision trees, and DMneural were the various classifiers used. The rate of correctness for this was 84.3% for the decision tree, Regression's 88.6%, for the neural network, 84.3%, with the highest level of precision was recorded at 92.9%. The employed dataset was split into training and testing done. Hyper parameter adjustment was done individually for each classifier.

Participants in the study included both presumed healthy persons and patients with Parkinson's disease (PD). To find the best features, lower the overall dimension of the feature vector, and categorize the information using k-nearest neighbors (k-NN), a genetic algorithm (GA) was employed. The dataset was drawn from the Parkinson dataset available in the UCI repository and included 197 samples of speech from 31 individuals. In another significant study, Ramage et al.'s (2024) introduced a method for separating participants into PD and control groups. The data for their research was collected from 40 individuals, 20 of whom had Parkinson's disease, while the other 20 were healthy. Each participant provided 26 speech samples, including sustained vowels, syllables, short sentences, and numbers. For classification, they used Support Vector Machines (SVM) and k-nearest neighbors (k-NN), along with cross-validation techniques known as Partial Leave-One-Out (s-LOO) and Partial Leave-One-Subject-Out (LOSO).

Ramani and Sivagami's (2011) employed data extraction methods to classify a combination of control and Parkinson's disease participants. Their study utilized 197 audio samples from which 22 features were extracted for analysis. Binary logistic regression using an ID3, C4.5, k-NN, Random Tree (RT), and SVM, LDA, and PLS were among the classification models they utilized. To the disorderly tree, accuracy was attained at 100%, while

for the k-NN, C4.5, and LDA, accuracy was at or above 90%. The algorithm C-PLS had the lowest accuracy, which was reported at 69.74%. Bhattacharya and Bhatia's (2010) recommended techniques based on SVM and ANNs (artificial neural networks). The UCI repository was utilized in order to acquire the dataset. The network of multilayer perceptrons, or MLPs was built on an ANN and had two layers. It had been found the support vector machine made higher vital outcomes compared to MLP. The accuracy obtained by using SVM with the linear and puk kernels was 91.79% and 93.33%, respectively. The MLP succeeded in associate precision of ninety-two.31%. To carry out the categorization, Nissar et al.'s (2021) RBF network and multilayer perceptron were utilized. There are 136 continuous vowels in the data they used phonations, during eighty-three the phonations was recorded. The fifty-three phonations were noted from the normal people. The network was trained using 112 phonations, and it was tested using 24 phonations. RBF network performed better than the multilayer perceptron on PD. 86.66% and 83.33% accuracy rates were attained for the test and training sets utilizing MLP, whereas the RBF network yielded 90.12% and 87.5% accuracy for the test and training sets respectively. Review the papers, was to develop a Parkinson's disease prediction system using the inputs as shown. By comparing the accuracy, reliability, memory, and f-measure scores obtained from the different categorization algorithms, such as Logistic Regression, K-Nearest Neighbor, we identified the best algorithm among DT (Decision Tree), RF (Random Forest), SVM (Support Vector Machine), and NB (Naive Bayes) would be the most accurate at predicting the fact of being of Parkinson's disease.

3 Background

Artificial intelligence (AI) is used in machine learning to enable programs automatically gain knowledge from their errors and get better over time without requiring explicit design. When people educate a computer to do a task considerably more quickly, they are often using machine learning. There are three categories for machine learning:

Machine learning is divided into three types-

1. Supervised Learning

Supervised learning involves training a computer using data that has been labeled. In this approach, the model learns from data where the correct responses or labels are already provided. This means that the data includes examples with known outcomes, allowing the system to learn from these pre-labeled instances.

The following are examples of algorithms for expecting Parkinson's disease detection:

- Logistic Regression
- Decision Tree
- Random Forest
- Naive Bayes
- K Nearest Neighbor
- SVM
- XG Boost

2. Unsupervised Learning

The process of instructing the computer how to utilize unlabeled, unclassified data and letting the algorithm react to the input unsupervised is known as unsupervised learning. The computer won't be trained, in contrast with supervised learning, because no teacher is present. Unsupervised learning is divided into two types-

- Clustering
- Dimensionality Reduction.
- Reinforcement Learning

Reinforcement learning is a feedback-based learning method where a computer learns to operate in different environments by interacting with them and observing the results. The system is rewarded for positive actions and receives penalties or criticism for negative ones, guiding it to improve its performance over time.

4 Methodology

24 important features were chosen from a dataset consisting of 194 items. The research was conducted using stratified 10-fold cross-validation in conjunction with machine learning classifiers. The classifiers were trained, and the outcomes were compared. Decision Tree (DT), Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Naive Bayes (NB), Random Forest (RF), Logistic Regression, and XGBoost were among the classifiers employed in the investigation. PD identification consists of two essential phases: instruction and assessment. Before using the raw data to construct the profound understanding model, there is a preliminary step of pre-processing and standardization. The parameter values of deep learning models are selected to minimize the function that is lost during training. In the testing step, the previously constructed model with the selected parameters is then used to identify PD.

1. Data Collection: Data gathering and choosing the instruction and assessment datasets are the first steps in predicting accuracy. The UCI dataset serves as the source of the dataset. We used 80% training dataset and 20% testing dataset for this project.
2. Attribute Selection: The characteristics of a dataset are called attributes. For Parkinson's disease, numerous attributes are used, including vocal fundamental frequency, gender, age, and age group and tone components. Predicted output is also provided in terms of 0 and 1 (see table 1) .

Table 1. Data Attributes

S.no	Attributes	Description	Type
1	Name	Subject Name and recording number	ASCII
2	MDVP:Fo(Hz)	Average vocal fundamental frequency	Numeric
3	MDVP:Fhi(Hz)	Maximum vocal fundamental frequency	Numeric
4	MDVP:Flo(Hz)	Minimum vocal fundamental frequency	Numeric
5	MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP	Several indicators of variation.	Numeric
6	MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA	Numerous actions of variation in amplitude.	Numeric

3. Data Preprocessing: We need to divide each category field into dummy columns containing 1s and 0s in order to function with categorical data. Among the most crucial tasks to complete in order to obtain precise results is this one.

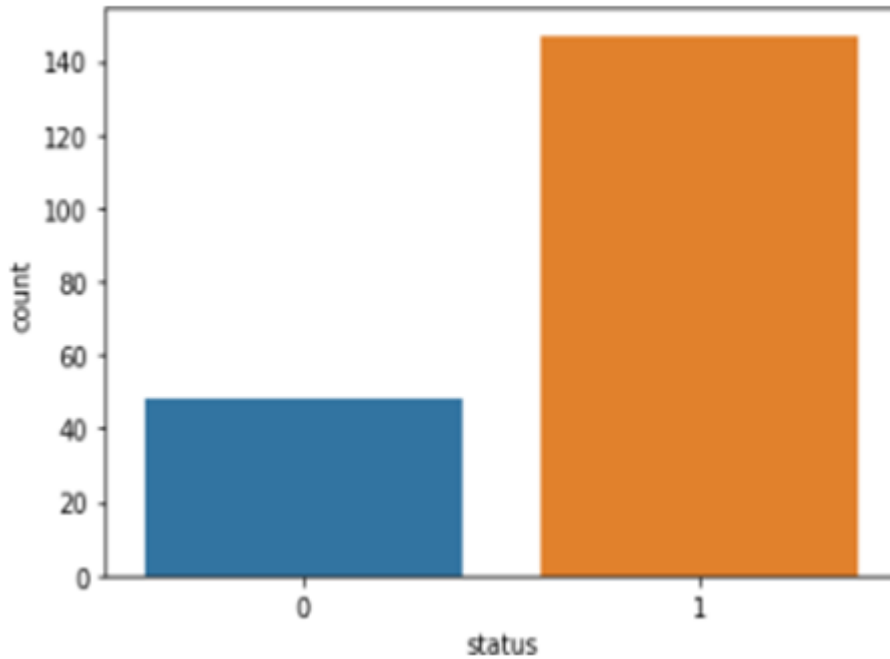


Figure 1. Bar Representation

4. Data Balancing: : Data balancing guarantees that the two output classes are evenly represented in order to go to the next level. The values "0" and "1" in this context denote those who are expected to possess Parkinson's disease and people who do not, respectively (see figure 1).
5. Histogram of attributes: Histograms provide a clear visualization of each data attribute, making it easier to interpret. The main advantage of this type of graphic is its ability to display the distribution of the predicted output (see figure 2).

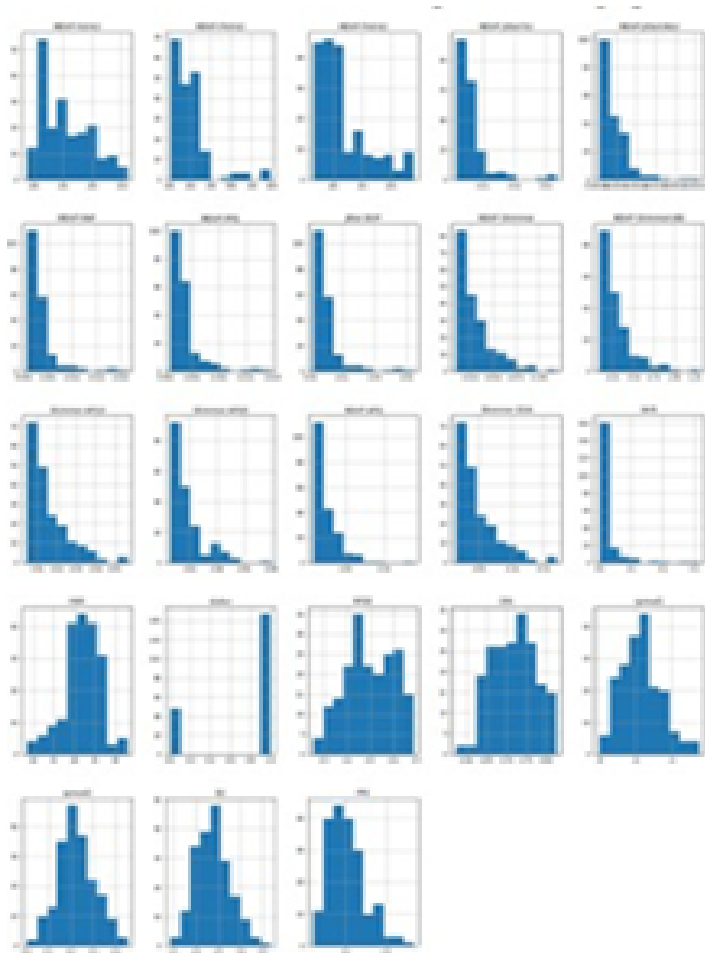


Figure 2. Histogram attributes

5 Algorithms Used

- **Logistic Regression:**

Using the value of one variable as a foundation, one can forecast the value of a different variable using logistic regression. The variable that is independent is used to estimate the value of the dependent variable, which is the one that is being forecasted. The study calculates the elements of the logistic regression calculation in order to ascertain the relevance of the variable that is dependent through the use of several predictor variables.

- **Decision Tree:**

Decision trees are an example of a supervised classification method. In this scenario, non-terminal nodes stand in for tests of one or more qualities, whereas terminal nodes reflect the results of decisions. J48 has been modified. C4.5. The C4.5 rule uses an algorithmic knowledge partitioning process to produce an ordered call tree for the provided dataset. The decision is made by using a depth-first search approach.

- **K Nearest Neighbor:**

The K-Nearest Neighbors algorithm predicts output values by comparing each element to its nearest neighbors using various input values. It is one of the most commonly used machine learning models. KNN classifies new data points based on the classification of their nearest neighbors and measures similarity based on previously stored data points.

- **Random Forest:**

Widely utilized in Issues with regression as well as classification. They creates call trees on entirely distinct datasets and obtains their majority decision about categorization and mean just in the event of a regression. One amongst the foremost vital options within the random forest model program is that the scenario at hand will handle the information a collection of constants as when using regression with variable categories such as within the classification situation. It operates higher outcomes for categorization issues.

- **Support Vector Machine:**

One is the Support Vector Machine also known as or SVM amongst the foremost in style methods for supervised learning, that is employed for categorization also Regression issues. However, primarily, it's used in ML to solve classification challenges. The SVM's objective rule is to create the most straightforward boundary or line that will divide n-dimensional house into categories in order that we'll just put the new data point among the right category among the longer term.

- **Naïve Bayes:**

The Naïve Bayes algorithm is a probabilistic model used for classification. Given an instance to be classified, represented by a vector $x=(x_1, \dots, x_n)$ $x = (x_1, \dots, x_n)$ $x=(x_1, \dots, x_n)$

with n features (independent variables), it calculates the probability of this instance belonging to each possible class.

- XG Boost:

XGBoost is a scalable and distributed machine learning library for gradient-boosted decision trees (GBDT). It is a leading tool for tasks such as regression, classification, and ranking, and provides parallel tree boosting capabilities. To fully understand XGBoost, it's essential to first grasp the underlying concepts of machine learning that it builds upon, including supervised learning, decision trees, ensemble learning, and gradient boosting.

6 Result

Python programming is a good fit for Jupyter Notebook, the simulation tool used. Jupyter Notebooks enable a variety of elements in addition to code, including equations, graphics, references, and additional rich text components. Because these documents allow real-time data analysis and mix code and rich text features, they are the perfect place to put together a statistical description and its findings. Jupyter Notebook is an online interactive application for interactive charts, maps, visualizations, and narrative prose. The accuracy of the algorithms depends on four values, namely True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

The numerical values of TP, FP, TN, and FN are defined as:

- TP = Number of persons with Parkinson's disease
- TN = Number of persons without Parkinson's disease
- FP = Number of persons with no Parkinson's disease but identified as having it
- FN = Number of persons with Parkinson's disease but identified as not having it

Comparing the six distinct approaches employed in this study work, we find that the RF delivers the maximum efficiency after the evaluation and instructional phases relative to the ML approach. The most accurate algorithm is the SVM, which has 97% accuracy (see table 2).

Table 2. Comparison table

Algorithm	Accuracy
Logistic Regression	84%
Decision Tree	82%
K-Nearest Neighbor	92%
Random Forest	94%
Naïve Bayes	64%
XG Boost	92%
SVM	97%

7 Conclusion

Data analysis enables the quick identification of patterns and relationships across various classifications, playing a growing role in EEG analysis. This cost-effective clinical test is increasingly applied to detect neurological disorders. While research on classifying Parkinson’s Disease (PD) using resting-state EEG is common, there is a notable gap in studies utilizing motor activation tests or examining disease progression. These studies often lack consistency, with clinical factors such as medication use and disease stage frequently omitted. Furthermore, the datasets are generally small compared to those used in machine learning literature. Despite this, many studies achieved strong classification performance, with accuracy rates exceeding 90% in some cases. A deeper analysis revealed that both model architecture and feature selection were critical for accurate classification, while the EEG preprocessing methods, which varied across studies, had minimal impact. This suggests that future prediction models could bypass manual preprocessing, streamlining the process. As machine learning advances, more sophisticated models are emerging, positioning this review as an early step in applying machine learning to PD research. Future studies could explore alternative approaches to predicting Parkinson’s Disease, using diverse datasets and shifting beyond binary classification (diseased vs. non-diseased) to include different disease stages. There is potential for developing mobile applications for disease forecasting, as well as incorporating deep learning methods and new feature selection techniques to improve prediction accuracy.

References

- Alshammri, R., Alharbi, G., Alharbi, E., & Almubark, I. (2023). Machine learning approaches to identify Parkinson's disease using voice signal features. *Frontiers in Artificial Intelligence*, 6. <https://doi.org/10.3389/frai.2023.1084001>
- Athanasios Tsanas, Max A. Little, Patrick E. McSharry, Jennifer Spielman, & Lorraine O. Ramig. (2012). Novel speech signal processing algorithms for high-accuracy classification of Parkinson s disease. *IEEE Transactions on Biomedical Engineering*, 59, 1264–1271. <http://www.maxlittle.net/publications/TBME-00887-2011.pdf>
- Bhattacharya, I., & Bhatia, M. P. (2010). SVM classification to distinguish Parkinson disease patients. *Proceedings of the 1st Amrita ACM-W Celebration of Women in Computing in India, A2CWIC'10*. <https://doi.org/10.1145/1858378.1858392>
- Das, R. (2010). A comparison of multiple classification methods for diagnosis of Parkinson disease. *Expert Systems with Applications*, 37(2), 1568–1572. <https://doi.org/10.1016/j.eswa.2009.06.040>
- Goubault, E., Nguyen, H. P., Ayachi, F. S., Bogard, S., & Duval, C. (2017). Do bradykinesia and tremor interfere in voluntary movement of essential tremor patients? Preliminary findings. *Tremor and Other Hyperkinetic Movements*, 7. <https://doi.org/10.7916/D822319X>
- Nissar, I., Mir, W. A., Izharuddin, & Shaikh, T. A. (2021). Machine Learning Approaches for Detection and Diagnosis of Parkinson's Disease - A Review. *2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021*, 898–905. <https://doi.org/10.1109/ICACCS51430.2021.9441885>
- Office of Communications and Public Liaison. (2023). Parkinson's Disease: Challenges, Progress, and Promise. National Institute of Neurological Disorders and Stroke. <https://www.ninds.nih.gov/current-research/focus-disorders/parkinsons-disease-research/parkinsons-disease-challenges-progress-and-promise#toc-resources>
- Ramage, A. E., Greenslade, K. J., Cote, K., Lee, J. N., Fox, C. M., Halpern, A., & Ramig, L. O. (2024). Narrative analysis in individuals with Parkinson's disease following intensive voice treatment: secondary outcome variables from a randomized controlled trial. *Frontiers in Human Neuroscience*, 18. <https://doi.org/10.3389/fnhum.2024.1394948>
- Ramani, R. G., & Sivagami, G. (2011). Parkinson Disease classification using data mining algorithms. *International Journal of Computer Applications*, 32(9), 17–22.
- Ramesh, S., & Arachchige, A. S. M. (2023). Depletion of dopamine in Parkinson's disease and relevant therapeutic options: A review of the literature. *AIMS Neuroscience*, 10(3), 200–231. <https://doi.org/10.3934/NEUROSCIENCE.2023017>