Baghdad Science Journal

Volume 22 | Issue 7 Article 28

7-25-2025

# A Study on Machine Learning Based Parkinson Disease Prediction

Suparna Dasgupta
*Department of Information Technology, JIS College of Engineering, West Bengal, India*,
suparnadasguptait@gmail.com

Soumyabrata Saha
*Department of Information Technology, JIS College of Engineering, West Bengal, India*,
som.brata@gmail.com

Pronay Pal
*Department of Information Technology, JIS College of Engineering, West Bengal, India*,
pronay.pal@jiscollege.ac.in

Shilarchana Maiti
*Department of Computer Application, JIS College of Engineering, West Bengal, India*,
shilarchnamaiti@gmail.com

Sudarshan Nath
*Department of Information Technology, JIS College of Engineering, West Bengal, India*,
angrybirdskiller7@gmail.com



## How to Cite this Article

Dasgupta, Suparna; Saha, Soumyabrata; Pal, Pronay; Maiti, Shilarchana; and Nath, Sudarshan (2025) "A Study on Machine Learning Based Parkinson Disease Prediction," *Baghdad Science Journal*: Vol. 22: Iss. 7, Article 28.
DOI: https://doi.org/10.21123/2411-7986.5008

BAGHDAD SCIENCE JOURNAL 2025;22(7):2438–2457

Scan the QR to view the full-text article on the journal website

RESEARCH ARTICLE

# A Study on Machine Learning Based Parkinson Disease Prediction

**Suparna Dasgupta** [1,*], **Soumyabrata Saha** [1], **Pronay Pal** [1], **Shilarchana Maiti** [2], **Sudarshan Nath** [1]

[1] Department of Information Technology, JIS College of Engineering, West Bengal, India
[2] Department of Computer Application, JIS College of Engineering, West Bengal, India

## ABSTRACT

Parkinson's Disease, a neurodegenerative disorder, is one of the major chronic health issues in the world. It causes a severe disorder that affects majorly muscle control but it can also be the reason of affecting senses, cognitive ability and cognitive health. Approximately, 90% of the Parkinson affected people face speech difficulty. Conventional diagnosis methods may be biased which may result in wrong diagnosis of Parkinson's Disease because symptoms are usually elusive. This study aims to assess how well different machine learning algorithms can predict Parkinson's disease using vowel phonation data, with the goal of early detection and more accurate patient assessments. Different machine learning algorithms, like Random Forest, Logistic Regression, Decision Tree, Support Vector Machine, and Boosting algorithms (Gradient, Extreme Gradient, Light Gradient, and Categorical), are evaluated for their prediction ability. Based on vowel phonation data, Random Forest achieved the highest accuracy of 98.4% among the evaluated classifiers in predicting Parkinson's disease. It highlights the prominence of machine learning application for early detection of Parkinson's disease accurately. This research helps create a better way to assess patients' risk of Parkinson's disease, leading to a clearer understanding and supporting future studies in this area.

**Keywords:** Detection, Efficacy, Machine learning, Parkinson, Prediction

## Introduction

One of the most prevalent neurodegenerative disorders, Parkinson disease[1] affects 1–2 persons per 1,000 people over 50 years. Due to the growing older population and rising incidence rates, the projected number of persons living with Perkinson Disease worldwide has more than quadrupled from 1990 onwards. As of 2020, an estimated 9.4 million individuals were still suffering with this condition globally.[2] Just 4% of occurrences of this condition occur in adults under the age of 50, with those over 60 being the group most affected.[3] Movement preparation, instigation, and completion are all a part of Parkinson disease, a degenerative neurological condition involving motor and non-motor symptoms. The symptoms of this condition will manifest differently in each person. Some of the symptoms include tremor, slowness of movement, rigidity of muscles, loss of spontaneous movements, and even abnormalities in speech and writing. Early Parkinson disease may cause speech to seem bland and the face to slow down or disappear.

Parkinson.[4] is the most prevalent neurological ailment, produces considerable impairment, lowers quality of life. Dopamine is a neurotransmitter that is synthesized by nerve cells in this region of the brain. When dopamine levels drop, neurons in the parts have trouble speaking, writing, walking, and doing other basic tasks. Despite the creation

Received 4 February 2024; revised 7 June 2024; accepted 9 June 2024.
Available online 25 July 2025

* Corresponding author.
E-mail addresses: suparnadasguptait@gmail.com (S. Dasgupta), som.brata@gmail.com (S. Saha), pronay.pal@jiscollege.ac.in (P. Pal), shilarchnamaiti@gmail.com (S. Maiti), angrybirdskiller7@gmail.com (S. Nath).

https://doi.org/10.21123/2411-7986.5008

of cardinal symptoms of Parkinson in clinical evaluations, most disease severity rating systems have not been extensively examined and validated. Many individuals have non-motor symptoms before the beginning of Parkinson, but they lack specificity, are difficult to quantify, and vary from patient to patient.

There have been significant improvements in health care services that use pruning technology as a result of technological developments like Artificial Intelligence and Machine Learning.[5,6] Machine learning may leverage past results to make wise judgements on previously unknown contemporary circumstances. Researchers have made ML based methods for predicting the occurrence of Parkinson. The most crucial issues for machine learning methods are how to choose the right classifier and valid features. This study uses several machine learning classifiers to do classification. After conducting training on the provided data, the model's accuracy is validated using the test dataset. The approach applied to construct the model will be determined by the accuracy of the findings. To better predict the Parkinson, our effort will make use of previously collected data to develop ML approaches. Critical parameters for the prediction technique were selected utilizing a data-driven feature selection strategy based on statistical and feature reduction methods. The beginning of Parkinson disease may often be predicted in many instances.[7,8] These methods are either too costly or insufficient to reliably predict a person's risk of developing heart disease. Prediction based on data from patient reports is not always simple to make.

In the context of existing work, one prevalent issue is the limited focus on feature selection techniques and hyperparameter tuning in optimizing model performance for Parkinson disease prediction. Many previous studies have primarily emphasized accuracy metrics without adequately addressing the importance of fine-tuning model parameters or selecting relevant features. This oversight can lead to suboptimal predictive models that may not effectively capture the complexity of the disease.

The combination of hyper-parameter tuning, model development and feature selection avoided the challenge of the previous work which lacked these necessary considerations. There are dozens of articles that have been published in regard to the Parkinson Disease prediction model, but hyper-parameter tuning techniques and feature selection are infrequently used.

Moreover, this phenomenon also extends to realizable over-fitting which is associated with using a narrow range of data sets, as the instruction set contains numerous examples. To ensure the models are applicable in a wide variety of situations, the approach is based on cross-validation thereby testing the model on several datasets.

Another limitation of the research so far is that many studies only train models on existing datasets and do not check whether the model does well on new datasets or different patient populations. It can lead to overfitting and make the models useless in the real world. We generalize models by doing rigorous cross-validation and applying it across many datasets and cohorts of patients. In this way, our models can be extended to different clinical scenarios and thus better predict Parkinson's disease. Our general method solves more problems — we get more accurate and general models.

These were the main findings of this research:

- This paper proposes a novel method for Parkinson disease prediction using machine learning specially focusing on vowel phonation data for preemptive detection and risk assessment.
- Data imbalance (and hence the importance of IQR) is dealt with to control for its performance of the model and generalizing it.
- To construct a generalized model, hyper parameterize by running many large variations of parameter values.
- By testing different Machine Learning Algorithms such as Random Forest, Logistic Regression, Decision Tree, Support Vector Machine, Gradient Boosting Machine, Extreme Gradient Boosting, Light Gradient Boosting, and Categorical Boosting, the study finds out the best one for the prediction of Parkinson disease with accuracy.
- More regarding the features picked, the original push in this particular research is for wrappers to be associated.
- In its current structure, it seems that research gets to be optimal on model performance, tending to have been ignored in former projects focused only on accuracy. The great aspect remains the most significant one in most cases where prediction models are enabled and the previous research efforts are compared.
- Revealed Strategic Hyperparameter Tuning in the improvement of model efficacy since this has been very neglected within the whole scope of the past research endeavors and brings a predicted power in the model.
- Meaningful Model Selection Combined with Feature Selection and Hyperparameter Tuning Thus, the whole process stands equally in guaranteeing higher accuracy levels in predicting Parkinson Disease hence setting a new trend for future research in the field.

- The findings of this study contribute valuable insights into the potential of machine learning algorithms in improving diagnostic accuracy for Parkinson disease, addressing the limitations of traditional diagnostic methods that are often subjective and prone to misdiagnosis.

The remainder part of the paper is outlined as follows. A literature survey has been pursued in the next section followed by the system model and methodology of the proposed system. In the next section the result analysis along with discussion is presented. The conclusion is offered in the final section.

## Literature survey

Recent research employed a variety of machine learning techniques for symptom-based Parkinson patient identification. MRI scans, gait analysis, and genetic data have all been used in studies to predict Parkinson disease, but there is surprisingly little study on the use of hearing loss as an early indicator.

In[9] authors used SVM model to predict senior patients' Parkinson onset using genetic data. They managed to train an SVM model to an accuracy of 0.889, whereas an enhanced SVM model achieves a precision of 0.9183. This finding supports the superiority of auditory data over genetic data in PD categorization. UCI telemonitoring dataset keystroke data was used to train a Random Forest classifier[10] to predict the severity of PD in elderly individuals. The models used in[11] that utilize audio data to categorize PWP are very dependent on MATLAB. Python-trained open-source models, which are both quick and memory-efficient, are used in this study.

Authors[12] demonstrated how ensemble deep learning models applied to phonation data might predict Parkinson disease progression. Deep learning model performance was poor since they didn't apply feature selection. Authors[13] intended to decrease PD diagnosis dependency on wearable technology by training a classical decision tree on 12 complicated speech parameters of the MDVR-KCL dataset. The ResNet[14] model was not trained on the subtleties of audio frequency, but rather on pictures of audio data. Authors[15] used an objective ML model to predict PD cases, their best findings only got up to an accuracy of 85%, leaving room for clinicians' biases.

In[16] authors used different machine learning models to categorize patients as having PD based on a dataset of numerous speech biomarkers. Using a unique deep learning model, authors were able to achieve a 96.45% accuracy rate in classification, but at a high cost owing to the model's high memory needs. Authors[17] used a linear classification model with 95% accuracy to classify PD patients' shuffling. Their research focused on patient gait, and subsequent studies recommended using audio and sleep data to enhance findings. Authors[18] analyzed brain MRI images spatially and temporally. To identify MCI in PWP, authors[19] have used a combination of decision trees, random forests, and K-Nearest Neighbors. Authors[20] performed L1-support SVM on vowel phonation dataset for neurological illness patients without feature identification. Authors[21] indicated that ML can identify PD's subtle non-motor symptoms that doctors may overlook. They built a data gain evaluation model to determine PD talents from the dataset. This strategy included many machine learning and data-gathering techniques. Compared to DL-based PD assessment models, this method worked well in PD findings but had insignificant consequences.

Authors[22] provided a technique for identifying Parkinson disease. Weka tools were utilized to construct algorithms for data pre-processing, classification, clustering, and analysis. In[23] authors explored numerous speech signal analysis methods for diagnosing this subjective condition. TQWT excelled in state-of-the-art speech signal computational algorithms used for PD detection feature extraction. Classifier predictions were pooled using ensemble techniques after applying several classifiers to distinct feature groups. The authors tested ML methods for PD patient identification.[24] KNN, SVM, Naive Bayes, and random forest are four machine learning classifiers that were employed to diagnose PD. At 70.26% accuracy and 0.64 precision for test data, the Naive Bayes algorithm detected PD patients. The authors suggested PD diagnosis utilizing feature selection, extraction, and pre-processing classification.[25] Recursive feature elimination and feature significance approaches were employed for feature selection in their study. SVM was shown to have an accuracy of 79.98% before feature selection, however it performed better after selection. The authors suggested a statistical technique to identify subjective illness using vowels and voice parameters. The accuracy rates for SVM and KNN were 91.25% and 91.23%, respectively.[26] The authors proposed comparing performance measures with genetic algorithm-based feature sets and Principal Component Analysis based feature reduction strategies in.[27] They achieved 97.57% accuracy using SVM with RBF and genetic algorithm-based feature sets.

In[28] authors used multi-agent fact analysis to decide reaction. Reinforcement learning, Choice Tree, Naive Bayes classification, and Random Forest approaches were used to construct the multi agent device for speech issue assessment. Authors[29] examined

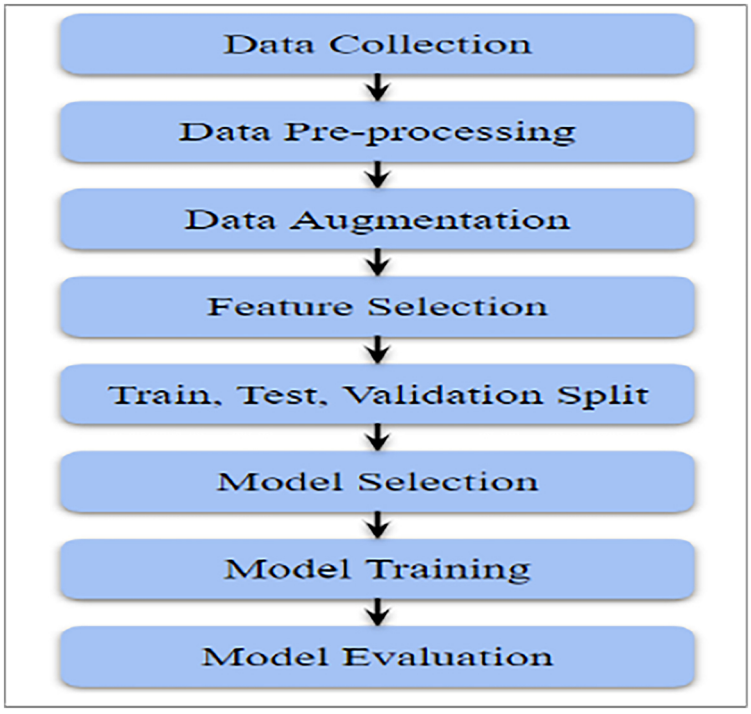


**Fig. 1.** Proposed methodology.

indicators of idiopathic Parkinson disease for several feature groups. Parkinson prediction the usage of artificial neural network research on this machine. After expertly importing the data into the Neural Network environment, the ANN model was 90% correct.

The drawback is that illness prediction solely relies on vowel phonation data, which could not adequately represent the intricacy of biomarkers and symptoms associated with Parkinson disease. Our primary goal of improving the models' performance is commendable, but it risks ignoring how well the models work with other types of patients or in other types of healthcare environments.

## System model and methodology

This section elaborated on the research methods used by the authors of the proposed system model. The proposed methodology involves the acquisition of data from Kaggle, specifically focusing on voice modulations in individuals with Parkinson disease. The dataset encompasses information on jitter, shimmer, and MDVP (Mean Delta Pitch Value) derived from vowel phonation. Following data pre-processing, comprehensive analysis, and visualization, a profound understanding of these attributes is gained. To develop predictive models for classifying audio data into Parkinson disease or healthy categories based on frequency variations, nine distinct models are employed. These models include Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, K-Nearest Neighbours, Gradient Boosting Machine, XGBoost, Light GBM, and Cat Boost. The training phase utilizes 70% of the data, while 20% is reserved for validation. The models are subsequently tested on 10% of the data, and their performance is evaluated using precision, recall, F1-score, confusion matrix, and ROC-AUC score metrics. The procedure of the proposed model is shown in Fig. 1.

### *Data collection*

The dataset for Parkinson disease comprises a diverse range of voice modulation data collected from individuals with Parkinson disease. Variables include jitter, shimmer and MDVP of vowel phonations features. The dataset is obtained from Kaggle repository and carefully curated to ensure data quality. Biomedical voice measurements were collected from 31 individuals, comprising 23 patients diagnosed with Parkinson disease. The data set includes patients within the age range of 46 to 85 years, while normal readings are derived from individuals aged 23 years. Each participant underwent an average of 6 phonation sessions, recorded 195 times per session. The duration of these recordings varied from 1 to 36 seconds. The primary objective of the dataset is

**Table 1.** Feature names and their description.

| Feature Name | Description |
|---|---|
| MDVP:Fo(Hz) | Average vocal fundamental frequency |
| MDVP:Fhi(Hz) | Maximum vocal fundamental frequency |
| MDVP:Flo(Hz) | Minimum vocal fundamental frequency |
| MDVP:Jitter (%), MDVP:Jitter (Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP | Several measures of variation in fundamental frequency |
| MDVP:Shimmer, MDVP:Shimmer (dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA | Several measures of variation in amplitude |
| NHR, HNR | Two measures of the ratio of noise to total components in the voice |
| Status | The health status of the subject (one)-Parkinson, (zero)-healthy |
| RPDE, D2 | Two nonlinear dynamical complexity measures |
| DFA | Signal fractal scaling exponent |
| Spread1, Spread2, PPE | Three nonlinear measures of fundamental frequency variation |

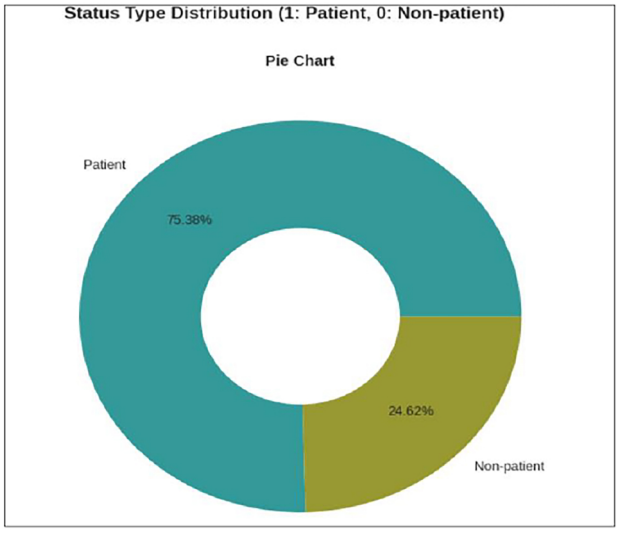


**Fig. 2.** PD Patient and non-patient distribution.

to distinguish between healthy individuals and those with Parkinson disease based on the "status" column, with 0 indicating a healthy status and 1 indicating Parkinson disease. Below table is elaborating on the attributes of 195 records. Feature names and their description are presented in Table 1.

## *Data preprocessing*

### *Data cleaning*

In the initial stage, our focus was on identifying null values and duplicate records within the dataset and employed the Pandas library for this purpose. After a thorough examination, it has been determined that there are no null values or duplicate records present in the dataset. The absence of null values signifies that each field or attribute in the dataset contains valid and complete information, contributing to the overall integrity of the data. Additionally, the absence of duplicate records indicates that each entry in the dataset is unique, preventing redundancy and ensuring the accuracy of our analysis.

### *Data exploration*

To gain insights into the fundamental statistics and characteristics of the data, conducted exploratory data analysis (EDA). This analytical approach facilitated the identification of patterns, trends, and potential outliers within the dataset see Fig. 2.

The pie chart presented above illustrates that within the dataset, approximately 75.38% of the data is associated with individuals diagnosed with Parkinson disease, while the remaining portion represents non-patient instances. This distribution highlights the presence of a class imbalance issue in the data, where one class (PD patients) significantly outweighs the other in terms of representation. Acknowledging this class imbalance is crucial for maintaining a balanced and unbiased perspective in subsequent analyses and model training, as it can impact the performance and reliability of predictive models. Histogram representation of different features is presented in Fig. 3.

Examining the histogram provided, it is evident that a majority of the data exhibits a positive skewness, indicating a concentration of values towards the lower end of the distribution. Notably, features such as spread1, spread2, and D2 demonstrate characteristics of a normal distribution. These particular features display a more symmetrical and balanced pattern in their distribution, suggesting a relatively even spread of values across their respective ranges. Boxplot of different features for outlier detection is presented in Fig. 4.

Outliers in a box plot are identified using the interquartile range (IQR), which measures how spread out the data is. The IQR is the distance between the first quartile (Q1) and the third quartile (Q3). Any data points that fall significantly outside this range are considered outliers. The boxplot's whiskers extend to a predefined multiple of the IQR, and any data


**Fig. 3.** Histogram representation of different features.

points lying beyond these whiskers are considered potential outliers.

### *Dealing with outliers*

To deal with potential outliers in the data, the authors used the Interquartile Range (IQR) method. This involves calculating the IQR, which shows how spread out the middle 50% of the data is. Then, they set upper and lower limits based on a multiple of the IQR. Any data points outside these limits are considered potential outliers.

Here's a simple way to calculate the IQR and detect outliers:

- **Step 1: Sort the Data**
  Arrange the data points in ascending order.
- **Step 2: Find the First Quartile (Q1)**
  The first quartile, Q1, is the middle value of the lower half of the data. It marks the point below which 25% of the data lies.
- **Step 3: Find the Third Quartile (Q3)**
  The third quartile, Q3, is the middle value of the upper half of the data. It marks the point below which 75% of the data lies.
- **Step 4: Calculate the Interquartile Range (IQR)**
  IQR = Q3 – Q1
- **Step 5: Define the Lower and Upper Limits**

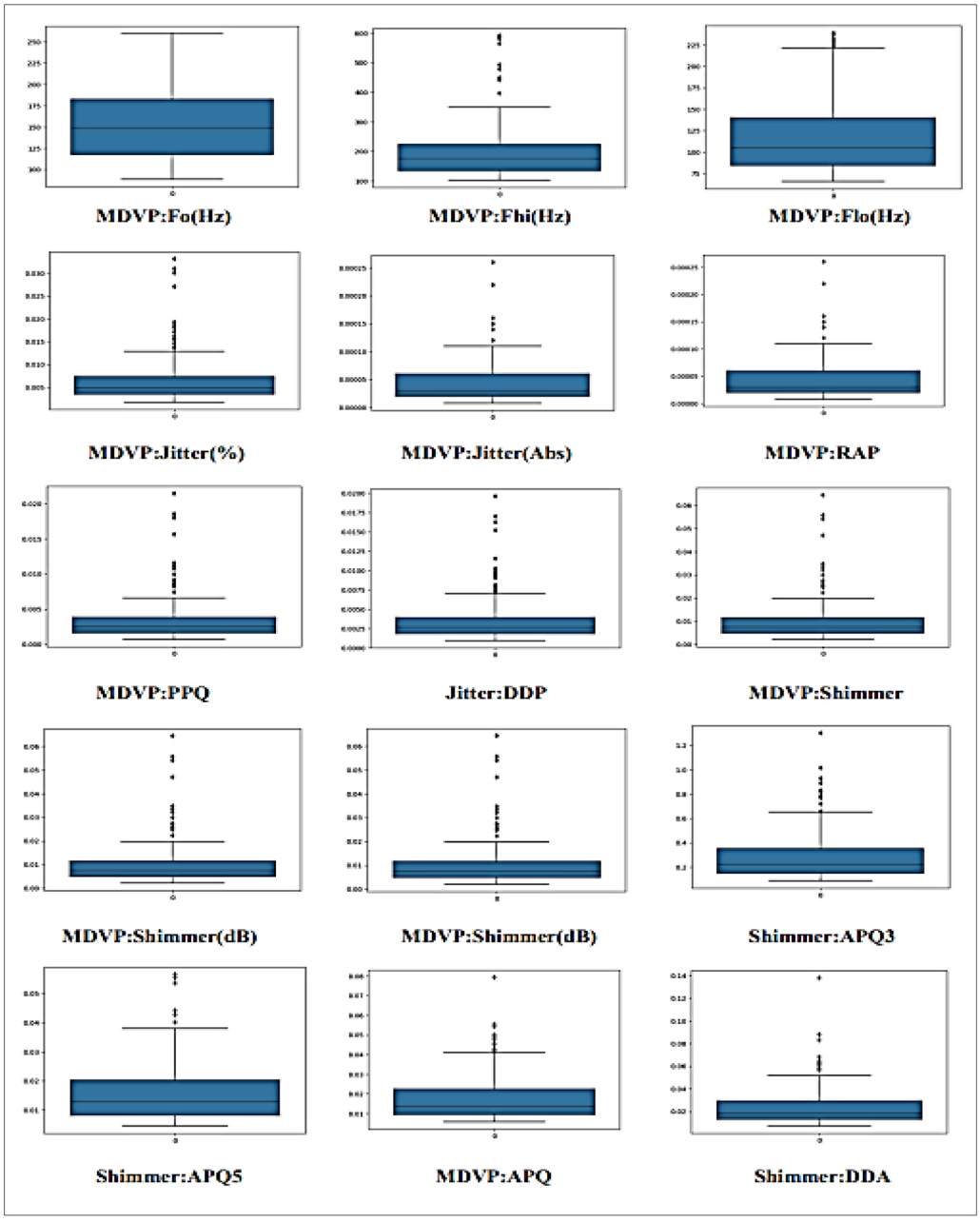


**Fig. 4.** Boxplot representation of different features for outlier detection.

Lower Limit: Q1 − K*IQR
Upper Limit: Q3 + K*IQR
Here, K is a constant that decides how far away from the quartiles a data point can be to be considered a potential outlier. The value of K is usually set to 1.5.

- **Step 6: Identify Outliers**
  Any data points below the lower limit or above the upper limit are considered potential outliers.

To reduce the effect of outliers, replacement with the average value of features was used instead. This includes extreme data that is beyond their set limits but is replaced with the average instead. In turn, this causes the data distribution to be normalized and prevents eventual analysis or model building from its influence. These steps ensure keeping the overall structure while maintaining the features.

### *Feature scaling*

Feature scaling is one of the preprocessing steps applied to the data in order to prepare it for machine learning. This implies making numerical features into the same range, thereby assisting certain models. The

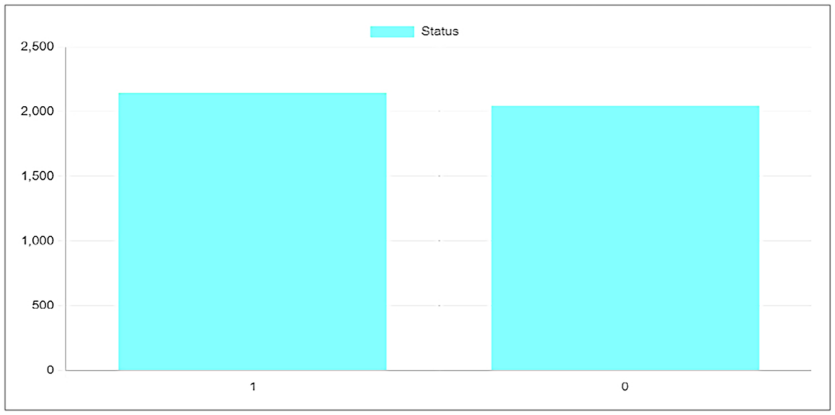


**Fig. 5.** Bar plot of the target variable 'status' representation.

authors applied the Z-Score Normalization method to standardize the features. This transforms each feature's value so that it has a mean of 0 and a standard deviation of 1. In doing this, all features are scaled in a similar fashion so that those features with greater values do not overly dominate the analysis. The Eq. (1) is presented as:

$$\left[ z = \frac{xi - \mu}{\sigma} \right] \tag{1}$$

In this formula, Z is the Z-score, xi is the current value, $\mu$ is the mean of the dataset, and $\sigma$ is the standard deviation. By using this formula, all values are standardized into Z-scores. As a result, the values fall within the range of –1 to +1, which keeps the original data's range but shows how far each value is from the mean, in terms of standard deviations.

### *Data augmentation*

Since the dataset is small, with only 195 entries, the authors used the IQR technique to create more data. This process generated synthetic data points to add to the existing dataset. The IQR technique made sure the new data matched the original dataset's patterns. This data augmentation helped increase the dataset size, giving the machine learning models more examples to learn from, which improves their performance.

Fig. 5 shows a bar plot of the target variable "status".

Authors addressed the issue of limited data by generating 2000 new data points for individuals diagnosed with Parkinson disease and additional 2000 data points for non-PD individuals. Thus, our dataset has now increased to 4,195 patient records. The increase in size also solved the problem of class imbalance to ensure a more balanced mix of PD and non-PD cases. Our goal is to make the training of the machine learning model more reliable and fairer with a more balanced mix of cases.

### *Feature selection*

Feature selection is one of the critical steps that occur during the machine learning process. It helps in the selection of the most appropriate features, which would be best for the model, enhancing performance, avoiding overfitting, and easy interpretation, reducing computation. The authors enable the model to concentrate on meaningful patterns while making it more efficient and interpretable by selecting the most important features.

The wrapper-based approach was implemented in feature selection, where a tree-based model was taken. Wrapper-based approach analyzes the performance of different feature subsets to see which are most significant. This identifies features that give the most impact on the performance of the model. The intention was to refine the accuracy and readability of the model by selecting the best features. The selected features are shown in Table 2.

Our wrapper-based feature selection approach uses four machine learning models: Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbors. Forward feature selection was to iteratively enhance our model's performance from the initial set of 22 features.

Beginning with a null model containing zero features, authors adopted a greedy approach, sequentially adding features one at a time to optimize the model's predictive capabilities. Through this process, the authors identified and selected the best subset of 15 features, leveraging the unique perspectives of each model to collectively enhance the overall feature set for improved model performance.

Our feature selection strategy extended by incorporating a Tree-based model, leveraging Random

**Table 2.** Wrapper based feature selection technique and their description.

| Logistic Regression | SVM | Decision Tree | KNN |
|---|---|---|---|
| PPE | PPE | MDVP:Shimmer(dB) | MDVP:Shimmer(dB) |
| Spread1 | MDVP:Fo(Hz) | PPE | MDVP:Jitter(%) |
| MDVP:Fo(Hz) | MDVP:Shimmer(dB) | Spread1 | PPE |
| MDVP:APQ | Spread1 | Shimmer:APQ5 | Spread1 |
| D2 | D2 | MDVP:Jitter(%) | MDVP:Fo(Hz) |
| Spread2 | MDVP:Jitter(Abs) | MDVP:Fo(Hz) | MDVP:RAP |
| MDVP:Fhi(Hz) | Spread2 | D2 | D2 |
| MDVP:Flo(Hz) | MDVP:Jitter(%) | MDVP:RAP | MDVP:Fhi(Hz) |
| NHR | MDVP:Fhi(Hz) | MDVP:Shimmer | Spread2 |
| Shimmer:APQ5 | Shimmer:DDA | MDVP:Fhi(Hz) | MDVP:Shimmer |
| MDVP:RAP | Shimmer:APQ3 | Spread2 | MDVP:APQ |
| MDVP:Jitter(Abs) | Shimmer:APQ5 | RPDE | DFA |
| Shimmer:APQ3 | RPDE | Shimmer:APQ3 | Shimmer:APQ3 |
| Shimmer:DDA | HNR | DFA | RPDE |
| MDVP:Shimmer | MDVP:APQ | MDVP:APQ | Shimmer:APQ5 |

**Table 3.** Features selected using tree-based model and their description.

| Random Forest | GBM | XGBM | Light GBM | Cat Boost |
|---|---|---|---|---|
| PPE | PPE | MDVP:Shimmer(dB) | MDVP:Shimmer(dB) | D2 |
| Spread1 | MDVP:Fo(Hz) | PPE | MDVP:Jitter(%) | MDVP:Fhi(Hz) |
| MDVP:Fo(Hz) | MDVP:Shimmer(dB) | Spread1 | PPE | MDVP:Fo(Hz) |
| MDVP:APQ | Spread1 | Shimmer:APQ5 | Spread1 | Spread2 |
| D2 | D2 | MDVP:Jitter(%) | MDVP:Fo(Hz) | Spread1 |
| Spread2 | MDVP:Jitter(Abs) | MDVP:Fo(Hz) | MDVP:RAP | RPDE |
| MDVP:Fhi(Hz) | Spread2 | D2 | D2 | Shimmer:APQ5 |
| MDVP:Flo(Hz) | MDVP:Jitter(%) | MDVP:RAP | MDVP:Fhi(Hz) | PPE |
| NHR | MDVP:Fhi(Hz) | MDVP:Shimmer | Spread2 | DFA |
| Shimmer:APQ5 | Shimmer:DDA | MDVP:Fhi(Hz) | MDVP:Shimmer | MDVP:Fo(Hz) |
| MDVP:RAP | Shimmer:APQ3 | Spread2 | MDVP:APQ | MDVP:Shimmer(dB) |
| MDVP:Jitter(Abs) | Shimmer:APQ5 | RPDE | DFA | MDVP:APQ |
| Shimmer:APQ3 | RPDE | Shimmer:APQ3 | Shimmer:APQ3 | HNR |
| Shimmer:DDA | HNR | DFA | RPDE | Jitter:DDP |
| MDVP:Shimmer | MDVP:APQ | MDVP:APQ | Shimmer:APQ5 | Shimmer:APQ3 |

Forest, Gradient Boosting Machine, XGBoost, Light GBM, and Cat Boost. These models facilitated the identification of the most important features based on their calculated feature importance scores. This comprehensive approach aimed to capture and prioritize features that collectively contribute the most to the predictive power of the model, enhancing its robustness and performance across multiple tree-based algorithms. Selected features by tree-based model is presented in Table 3.

### *Final selected features*

Using various approaches, 15 features have identified those consistently emerged as common selections across different methods, and these are presented in Table 4.

**Table 4.** Final selection of important features and their description.

| | |
|---|---|
| Top 15 Important Features | MDVP:Fo(Hz) |
| | MDVP:Fhi(Hz) |
| | MDVP:Flo(Hz) |
| | MDVP:Jitter(%) |
| | MDVP:Jitter(Abs) |
| | MDVP:RAP |
| | MDVP:PPQ |
| | MDVP:Shimmer(dB) |
| | Shimmer:APQ3 |
| | HNR |
| | RPDE |
| | Spread1 |
| | Spread2 |
| | D2 |
| | PPE |

### *Train, test and validation data generation*

The dataset was partitioned into three distinct sets: Training, Testing, and Validation. Initially, 90% of the data was allocated to the Training set, while the remaining 10% was reserved for the Test set. Within the Training set, 20% of the data was further set aside for Validation, serving as an independent subset for fine-tuning and optimizing the model during the training process. This partitioning strategy aimed

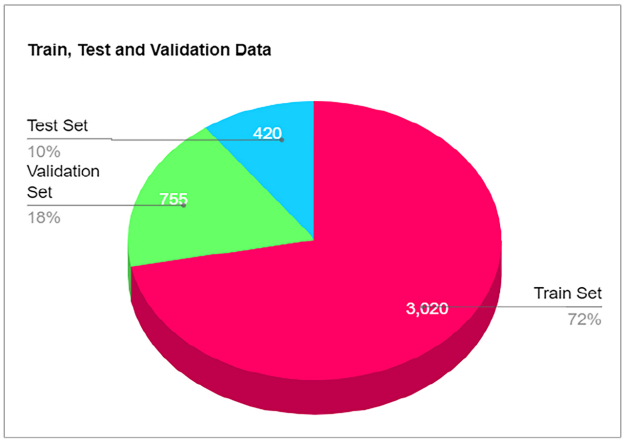


**Fig. 6.** Pie chart representation of train, test and validation data.

to ensure a robust evaluation of the model's performance on unseen data while allowing for effective model validation and parameter tuning. Pie Chart representation of Train, Test and Validation Data are presented in Fig. 6.

### *Model selection*

In this proposal authors have opted for a diverse ensemble of nine models to tackle our classification task. These models encompass a range of algorithms and techniques, each bringing its unique strengths and characteristics to the table. This comprehensive set of models enables us to explore various approaches and select the most suitable one based on the specific nuances of our dataset and the nature of the classification problem at hand. Fig. 7 represented the Logistic Function, while Fig. 8 presented the Parkinson Disease Prediction Using SVM, Fig. 9 represented the Decision Tree, Fig. 10 presented the Random Forest Classifier, and Fig. 11 represented the Parkinson Disease Prediction using KNN.

### *Logistic regression*

Logistic regression serves as a statistical technique primarily employed for binary classification tasks, wherein the objective is to predict the outcome of a categorical dependent variable with two possible values. Widely utilized in machine learning and statistics, logistic regression models the relationship between independent variables and the probability of a specific outcome. Diverging from linear regression, which forecasts continuous outcomes, logistic regression utilizes the logistic function (sigmoid function). This transformation confines the predicted values within the range of 0 to 1, making it particularly suitable for scenarios like audio data analysis. In the context of Parkinson Disease classification, where the

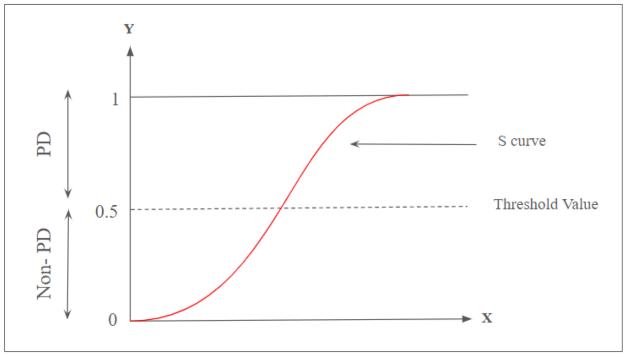


**Fig. 7.** Logistic function (sigmoid function) representation.

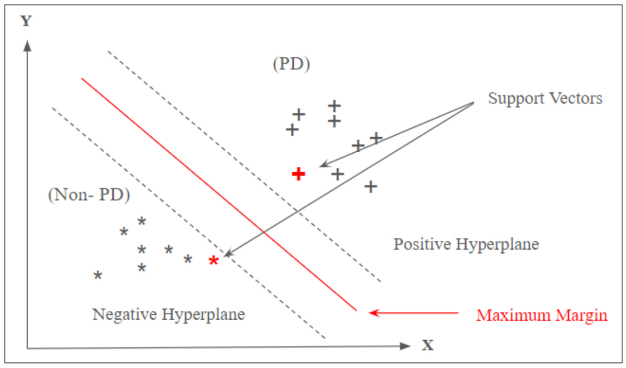


**Fig. 8.** Using SVM parkinson disease prediction representation.

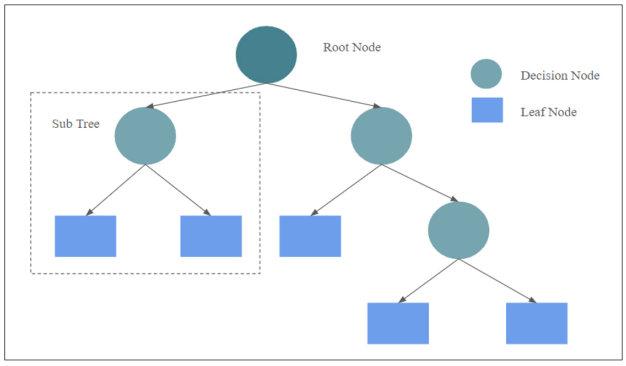


**Fig. 9.** Decision tree representation.

attributes influencing the prediction are not linearly correlated but exhibit an exponential pattern, logistic regression proves to be an apt and effective modelling approach. Below is the Eq. (2) of sigmoid function:

$$\text{Sigmoid}\,(x) = \frac{1}{1+e^{-x}} \tag{2}$$

Here, Sigmoid(x) denotes the output of the sigmoid function for a given input x. $e^{-x}$ represents the base of the natural logarithm where x signifies the input to the function.

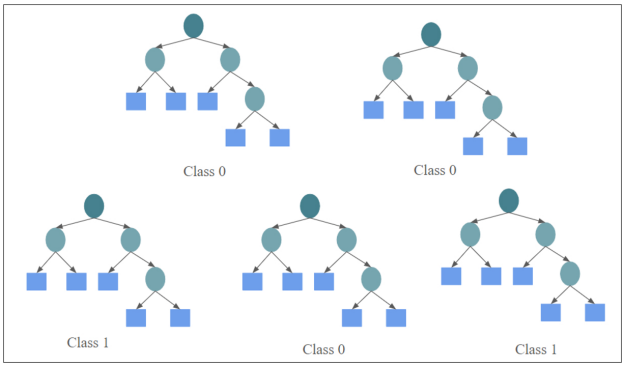


**Fig. 10.** Random forest classifier representation.

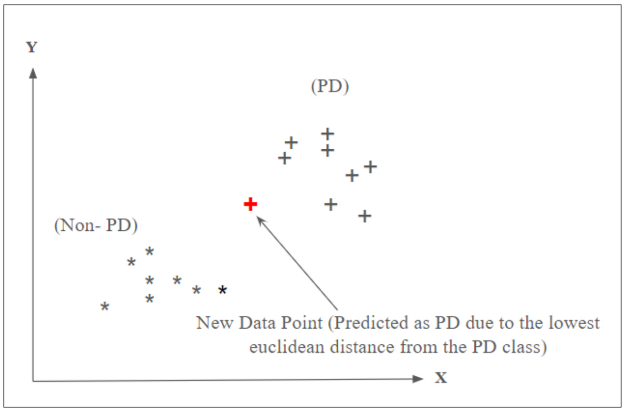


**Fig. 11.** Using knn parkinson disease prediction representation.

In our specific context, a threshold value of 0.5 is considered. Instances where the output value surpasses this threshold are designated as Parkinson Disease (PD), while those falling below the threshold are classified as non-Parkinson Disease (non-PD).

### *Support vector machine*

SVMs work based on the best-known boundary, known as a hyperplane, that can distinguish data points into different classes in a high-dimensional space. For the prediction of Parkinson's disease, SVMs learn clinical markers or genetic data and project them into a multi-dimensional space. It finds the hyperplane which classifies between the two classes while minimizing the error. A kernel function in SVM actually transforms the input data into the higher dimensional space, allowing this classifier to identify the complex relationships. Therefore, it makes the SVM an effective tool for resolving nonlinear relationships, as well as the complexity due to Parkinson's disease.

### *Decision tree*

Simple yet powerful, decision trees operate on a principle of predicting Parkinson disease. The algorithm partitions the input dataset based on the most informative features to optimize the tree structure, maximally separating the positives and the negatives with respect to Parkinson disease. For each node in the tree, the algorithm picks that feature which can best discriminate between the two classes from the branches leading to the subsequent nodes. The process continues until one of the stopping criteria is satisfied: either a specified depth is reached or homogeneity is achieved in the terminal nodes. Since decision trees inherently discover complex decision boundaries and provide for interpretability by visualizing the path from root nodes to the final leaf node, they are precious in the Parkinson disease prediction task because they find some of the key features and their interactions in the dataset for understanding what causes the disease and making good predictions on new cases.

### *Random forest classifier*

Random Forest operates based on an ensemble learning principle specifically by building up multiple decision trees and averaging their outputs towards robust and accurate Parkinson disease prediction. For the purpose of disease prediction, in the Random Forest, every single tree is developed based on a bootstrap sample of the original dataset and has randomness both in data as well as in feature selection. For the purpose of tree building up, at every node, a random subset of features is split, thereby creating diversity among trees. The ultimate prediction is given by aggregating the predictions coming from individual trees, typically done using a mechanism of majority vote. The two sources of randomness, the initial creation of a tree and then averaging different trees, would help reduce overfitting and enhance the generalization capability. Random Forest's capacity for modeling complex relations in the dataset, handles high-dimensional feature spaces and gives an estimation of feature importance. Therefore, it is a very effective and reliable tool for predicting Parkinson disease.

### *K-nearest neighbours*

This algorithm is proximity-based classification and, therefore, is highly applied in the prediction of Parkinson disease. In terms of disease prediction, every single element in the dataset lies in some feature space, and this algorithm classifies a new instance by determining its k-nearest neighbors based on some chosen distance metric, usually Euclidean distance. The majority class of all these neighbors will then be the one predicted for that new instance. KNN never assumes an underlying data distribution and automatically adapts to patterns in local minima. These

**Table 5.** Parameters selection for hyper parameter tuning and their description.

| Model Name | Parameters | Values |
|---|---|---|
| Logistic Regression | max_iter | 50,100,150 |
| | Warm_start | True, False |
| | Fir-intercept | True, False |
| SVM | C | 1,3,5 |
| | kernel | rbf, sigmod |
| | tol | 0.1, 0.01, 0.001 |
| Decision Tree, Random Forest | criterion | gini, entropy |
| | splitter | best, random |
| | min_samples_split | 2, 4, 6 |
| | Warm_start | True, False |
| KNN | n_neighbors | 3, 5, 7 |
| | P | 1, 2 |
| | metric | euclidean, manhattan, minkowski |
| GBM | loss | log_loss, exponential |
| | learning_rate | 0.1, 0.01, 0.001 |
| | criterion | friedman_mse, squared_error |
| | min_samples_split | 2, 4, 6 |
| | max_depth | 3, 5, 7 |
| XGBM | objective | binary:logistic |
| | n_estimators | 100, 200, 300 |
| | learning_rate | 0.1, 0.01, 0.001 |
| | max_depth | 3, 5, 7 |
| | gamma | 0.1, 0.2, 0.3 |
| Light GBM | boosting_type | gbdt, rf, dart, goss |
| | num_leaves | 31, 37, 41 |
| | learning_rate | 0.1, 0.01, 0.001 |
| | n_estimators | 100, 150, 200 |
| | objective | binarys |
| | min_child_samples | 20, 30, 40 |
| Cat Boost | iterations | 100, 200,300 |
| | learning_rate | 0.1, 0.01, 0.001 |
| | depth | 6, 8, 10 |
| | l2_leaf_reg | 3, 5, 7 |

*Red marked values are selected after hyper parameters tuning using GridSearchCV.

**Table 6.** Precision, recall, F1-score and ROC AUC of different models and their description.

| | Precision | | Recall | | F1-Score | | |
|---|---|---|---|---|---|---|---|
| Model Nam | **0** | **1** | **0** | **1** | **0** | **1** | ROC AUC |
| Logistic Regression | 0.73 | 0.90 | 0.90 | 0.72 | 0.80 | 0.80 | 0.86 |
| SVM | 0.90 | 0.82 | 0.76 | 0.93 | 0.93 | 0.87 | 0.86 |
| Decision Tree | 1.00 | 0.97 | 0.96 | 1.00 | 0.98 | 0.98 | 0.99 |
| Random Forest | 1.00 | 0.97 | 0.96 | 1.00 | 0.98 | 0.98 | 1.00 |
| KNN | 0.85 | 0.84 | 0.80 | 0.88 | 0.82 | 0.86 | 0.86 |
| GBM | 0.96 | 0.98 | 0.98 | 0.97 | 0.97 | 0.97 | 0.99 |
| XGBM | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.99 |
| Light GBM | 1.00 | 0.94 | 0.92 | 1.00 | 0.96 | 0.97 | 0.99 |
| Cat Boost | 1.00 | 0.95 | 0.93 | 0.99 | 0.98 | 0.97 | 1.00 |

characteristics make the algorithm especially strong in situations when decision boundaries cannot be defined and are complex at the same time. In the context of Parkinson disease, it makes use of the similarity in patterns of features to classify the subjects. Thereby, it provides a simple yet highly efficient way to predict

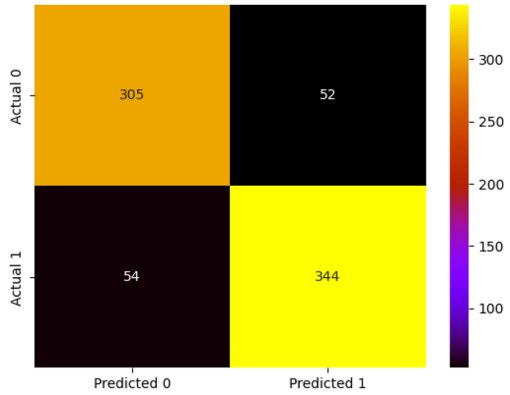


**Fig. 12.** Logistic regression confusion matrix representation.

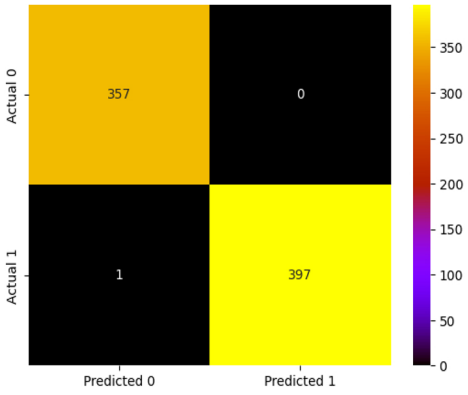


**Fig. 15.** Random forest confusion matrix representation.

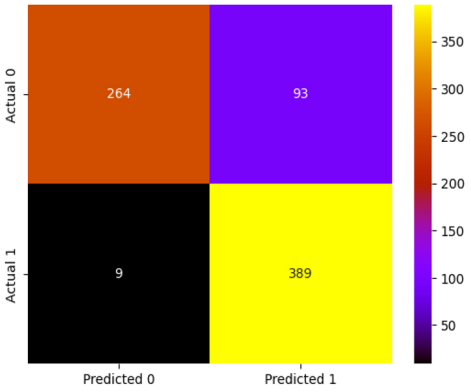


**Fig. 13.** SVM confusion matrix representation.

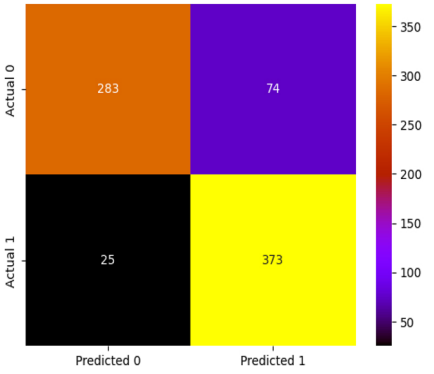


**Fig. 16.** KNN confusion matrix representation.

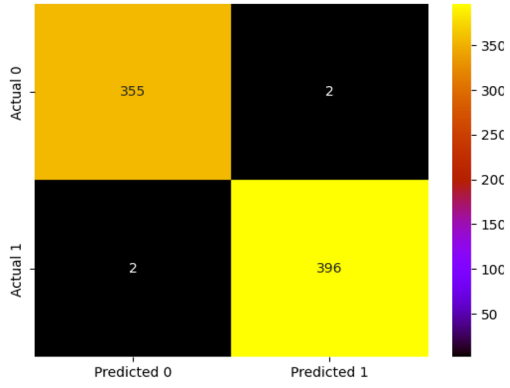


**Fig. 14.** Decision tree confusion matrix representation.

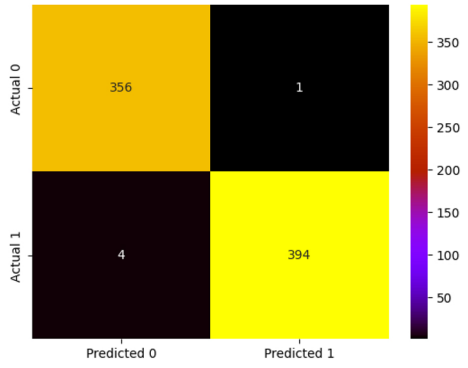


**Fig. 17.** GBM confusion matrix representation.

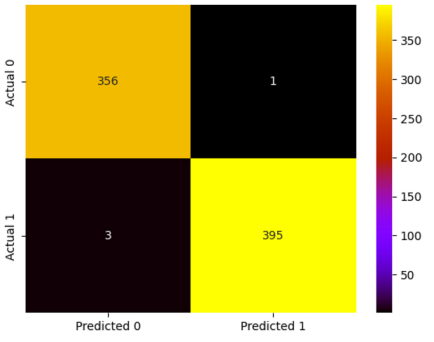


**Fig. 18.** XGBM confusion matrix representation.

the existence or absence of the disease by relying on the features of the nearest instances in the feature space.

### *Gradient boosting machine*

Gradient Boosting Machine is based on an ensemble learning technique that makes a predictive model for Parkinson disease prediction through several weak learners; the common form of weak learner is decision

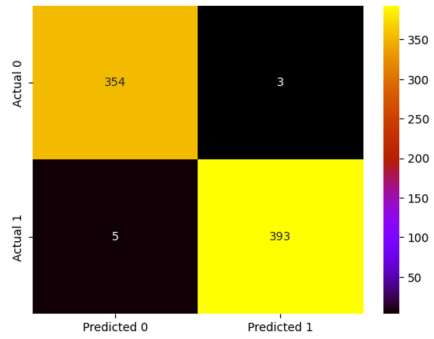


**Fig. 19.** Light GBM confusion matrix representation.

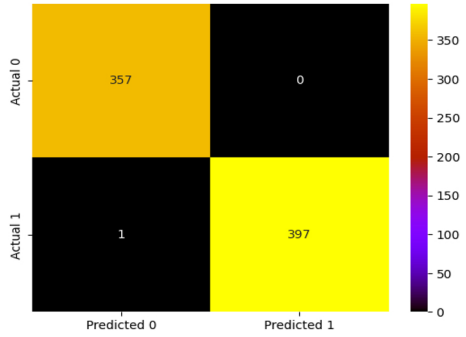


**Fig. 20.** Cat boost confusion matrix representation.

tree. It proceeds iteratively such that each tree will try to improve the errors committed by the combined ensemble so far. In each iteration, it will find the errors of the current ensemble by calculating the gradient of a defined loss function. A weak learner will then be trained to minimize these gradients and added to the ensemble, thus improving the predictive accuracy of the overall model. GBM adjusts by giving more weight to instances that were misclassified in previous iterations. This is continued until a fixed number of trees is reached or satisfactory performance is achieved. GBM strength The strength of GBM is due to its capacity to deal with the complex relationship present in the data, automatic capture of non-linear patterns, and high predictive accuracy for Parkinson disease based on the collective strength of multiple decision trees.

### *Extreme gradient boosting*

XGBoost is an advanced implementation of gradient boosting with focus on accuracy and efficiency. Thus, it is appropriate for Parkinson's disease prediction. It builds a predictive model by combining multiple weak learners that typically are decision trees in an additive manner. It utilizes a regularized objective function with both loss and regularization terms to measure prediction error and control the complexity of the model. In every boosting round XGBoost looks at the existing ensemble's performance and fits in a new tree to correct all the errors. Gradient based optimization techniques and parallel computation are used for the training and the algorithm reduces the time consumption and increases its efficiency. Furthermore, XGBoost also allows features like tree pruning and column subsampling to eliminate overfitting and enhance the generalization power. Its applicability to manage missing data, nonlinear relationships, and high-dimensional feature spaces. This makes XGBoost a very powerful, accurate tool for prediction of Parkinson disease while remaining robust and interpretive.

### *Light gradient boosting machine*

Light GBM is an efficient gradient boosting framework developed to work well for large-scale and computationally efficient machine learning problems, hence being the most suited predictor of Parkinson disease. It works on sequential ensemble decision tree construction wherein every decision tree train to rectify the mistake that was done by the previously built decision trees. Unlike traditional gradient boosting, Light GBM does the growth of the tree leaf-wise, not depth-wise. This strategy efficiently explores the feature space, and at each level, it identifies the most informative features to split upon. Moreover, Light GBM also introduces the histogram-based learning approach where, for faster computations, feature values are binned. The objective function is optimized by the use of a gradient-based approach such that the resultant prediction accuracy may be the best. Its ability to handle large datasets, high-dimensional feature spaces, and to provide fast and accurate predictions makes Light GBM a robust tool in predicting patients with Parkinson disease, especially in scenarios where efficiency in computation is crucial.

### *Categorical boosting*

CatBoost, like all other gradient boosting algorithms, works on an ensemble learning principle to predict Parkinson disease. However, CatBoost has innovations that it introduces to increase its efficiency and predictive performance. It uses a category-specific approach to handle categorical features, making the algorithm avoid manual encoding-increasing model accuracy. This model also uses ordered boosting where the best ordering of the categorical feature happens at tree building, hence there is a minimization in overfitting. The symmetric structure of the tree also helps decrease overfitting and reduces bias in training caused by unbalanced splits. Other features include CatBoost's unique handling of missing data, thereby making it

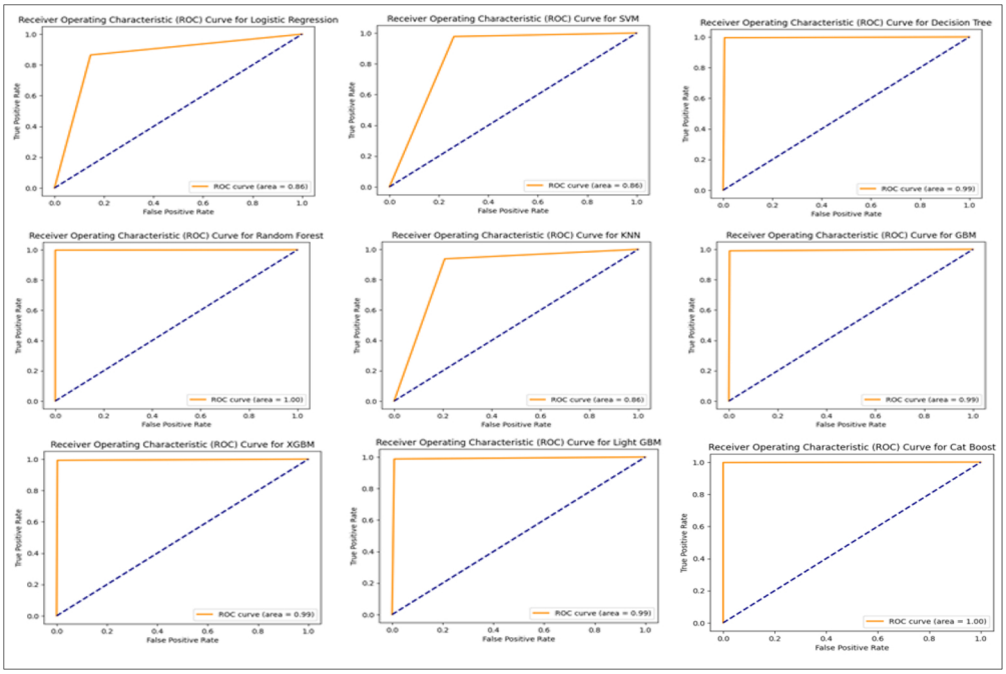


**Fig. 21.** Different models ROC AUC curve representation

much more robust. The algorithm of CatBoost works on gradient optimization with improvement that is iterative at the level of the predictive model. All of these make CatBoost very good for predicting Parkinson disease: It balances these aspects of accuracy, efficiency, and use, especially in scenarios with heterogeneous data types and missing values.

### *Model training*

A total of 5-fold cross-validation is used while setting the value of K for training our model. This entails breaking down the given dataset into subsets, and during training and validation, the procedure goes through five trials, where at each trial a different subset becomes the validation set, and all the remaining data are used in training. All the results will then be averaged to give an overall robust impression of the given model's performance across different sets of data subsets. Model training was conducted in three steps: first, training the model on all possible features; second, the model was only trained on selected features; finally, hyper parameter tuning for achieving the maximum accuracy was carried out using Grid Search Cross-Validation. Such an extensive step was taken for the validation of the performance of the model at different subsets of features as well as hyper parameter optimization of the best model with robust and accurate prediction.

Considered parameters for hyper parameter tuning is presented in Table 5.

### *Model evaluation*

To assess the performance of these models, various authors have chosen a pretty representative set of metrics: ROC-AUC curve, confusion matrix, precision, recall, and F1 score are among them. These metrics collectively give a pretty thorough evaluation of the models' ability to discriminate between classes, detect true positives and negatives and balance precision and recall. The ROC-AUC curve would suggest the trade-off between true positives and false positives, and the confusion matrix basically breaks down further into detailed classification outcomes. Accuracy, precision, recall, and F1 score measure the overall effectiveness and balance in how the models were picking up on the important patterns within the data.

Mathematical Formula of Eqs. (3) to (5) Precision, Recall and F1-Score:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{F1-Score} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5}$$

**Table 7.** Accuracy of different machine learning models and their description.

| Model Nam | Accuracy before Hyper-parameter tuning and Feature Selection (%) | Accuracy after Hyper-parameter tuning and Feature Selection (%) | Avg. Model accuracy (%) |
|---|---|---|---|
| Logistic Regression | 84.2 | 86.4 | 87.0 |
| SVM | 89.7 | 92.1 | 92.2 |
| Decision Tree | 89.1 | 89.2 | 95.4 |
| Random Forest | 89.3 | 90.7 | 98.4 |
| KNN | 84.0 | 84.5 | 84.9 |
| GBM | 89.7 | 90.0 | 97.8 |
| XGBM | 89.8 | 90.7 | 97.8 |
| Light GBM | 89.0 | 93.5 | 96.8 |
| Cat Boost | 89.1 | 92.1 | 98. |

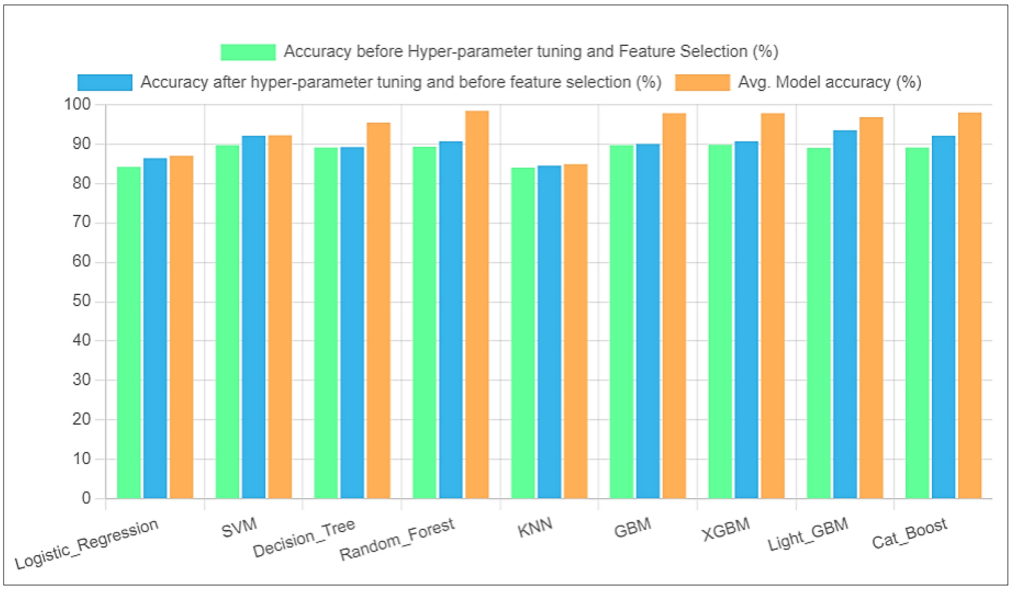


**Fig. 22.** Different model's accuracy bar chart representation.

Here, TP = True Positive, FP = False Positive and FN = False Negative

Precision, Recall, F1-Score and ROC AUC of Different Models are presented in Table 6.

The confusion matrix is used in the evaluation of our classification models that provide an elaborative account of the performance of a classification algorithm by summing up results in a tabular format. This matrix compares predicted labels of the model with actual labels and groups instances into four different outcomes.

- True Positive (TP): Instances correctly predicted as positive.
- True Negative (TN): Instances correctly predicted as negative.
- False Positive (FP): Instances incorrectly predicted as positive (Type I error).
- False Negative (FN): Instances incorrectly predicted as negative (Type II error).

Confusion Matrix of different models is presented below from Figs. 12 to 20.

The Receiver Operating Characteristic curve along with the area under the Curve is basic measuring performance when validating binary classification models. A graphic that displays, at any fixed value of decision threshold, which proportion of truly positive classifications that are ranked "above" is a measure by the false alarm rate at every such value. The AUC measures the overall performance of the model by quantifying the area under the ROC curve, ranging from 0 to 1.

ROC AUC Curve of different models is presented below in Fig. 21.

The AUC scores for Random Forest and CatBoost stand at a perfect 1, signifying their status as perfect classifiers. Similarly, Decision Tree, GBM, XGBoost (XGB), and Light GBM exhibit AUC scores of 0.99, indicating an almost impeccable classification performance. In contrast, Logistic Regression, SVM, and KNN display scores indicating better-than-random

**Table 8.** Comparative analysis of previous study models-our models and description.

| Authors | Logistic Regression | SVM | Decision Tree | Random Forest | KNN | GBM | XGBM | Light GBM | Cat Boost |
|---|---|---|---|---|---|---|---|---|---|
| Mathur et al.[22] | – | – | – | – | 91.2% | – | – | – | – |
| Aich et al.[27] | – | 97.5% | – | 95% | – | – | – | – | – |
| Govindu et al.[28] | 83.6% | 91.7% | – | 83.6% | 83.7% | – | – | – | – |
| Raya et al.[29] | – | 83% | 85% | 88% | 88% | – | – | – | – |
| Patra et al.[30] | – | – | – | 82% | – | 81% | – | – | – |
| Sandhiya et al.[31] | – | – | – | – | – | – | 95% | – | – |
| Celic. et al.[32] | 76% | 75.5% | – | 72.7% | – | 72.3% | – | – | – |
| Jain et al.[33] | – | – | – | 91.5% | – | 90% | – | – | – |
| Proposed | 87% | 92.2% | 95.4% | 98.4% | 84.9% | 97.8% | 97.8% | 96.8% | 98% |

performance in their classification capabilities. Accuracy of different Machine Learning Models is presented in Table 7. The bar chart representation of different model's accuracy is presented in Fig. 22.

It can be observed that following the feature selection process, there is a noticeable improvement in accuracies. Subsequently, after fine-tuning hyper parameters, further enhancements in accuracies are observed. Notably, all reported accuracies represent averages. Among the models, Random Forest stands out with the highest accuracy, reaching 98%. Additionally, the tree-based models, including Decision Tree, Random Forest, GBM, XGBoost, Light GBM, and CatBoost, consistently outperform Logistic Regression, SVM, and KNN in terms of predictive accuracy.

## Results and discussion

Among all the classifiers used in the task of classification of Parkinson disease using vowel phonation data, Random Forest stands out with a high accuracy of 98.4%. The reason behind such excellent performance is that the model fairly considers all attributes within the MDVP dataset. Both Logistic Regression and SVM also report commendable accuracies, whereas KNN shows a less satisfactory accuracy of 84.5% for the prediction of the disease, which is critical to be higher in the said task. The models, ranging from Decision Tree and Random Forest up to GBM, XGBoost, Light GBM, CatBoost, had consistently high precision over 95%. Such models shine when the relationship becomes intricate and nonlinear; they would successfully identify meaningful features, could manage complex interaction effects, and could be very beneficial for feature importances. Indeed, ensemble technique application, mostly with Random Forest and Gradient Boosting, tends to perform better in a prediction model if combined with the results of numerous weak learners. Their immunity to irrelevant features, anti-overfitting capacity, and ability to handle imbalanced datasets make it a good choice for the task of forecasting the outcome of Parkinson disease, which is really challenging. Proper feature selection through wrapper-based and tree-based techniques, indeed, is a critical optimization task in model performance, which usually gets neglected by earlier studies only concentrating on the accuracy aspect. These apart, the incorporation of strategic tuning of hyperparameter not covered in past studies enhances model efficacy. This selection of model or selecting features together along with optimal tuning of a hyperparameter combines to build strong approach ensuring peak accuracy to Parkinson disease.

The comparative analysis reveals that our models is presented in Table 8 and it outperform those utilized in a prior study, primarily owing to our proposed technique. Notably, even in the case of SVM, where Aich et al. achieved a commendable accuracy of 97.5%, our hyper-tuned Random Forest model surpasses this result. It is worth noting that the authors of the previous study employed only a limited number of models for Parkinson disease prediction, whereas our approach explores a more comprehensive range of machine learning algorithms. This broader exploration enables us to identify and leverage the best classification model, contributing to superior predictive performance in the context of Parkinson disease classification.

The limitation of this work is that it only depends on vowel phonation data for the prediction of Parkinson disease, and it may not capture the entire spectrum of symptoms and biomarkers associated with the disease. In fact, though vowel phonation data may give valuable insights, it may fail to account for other clinical features and imaging modalities that would further enhance prediction accuracy. Another limitation of this study is the main focus on optimization of model performance with feature selection and hyperparameter fine-tuning, which pays no regard to generalizability of the models across different patients or environments. Some of these gaps will be filled up in future works through the incorporation of a wider array of data sources

and validation strategies to optimize robustness and applicability across different contexts.

## Conclusion

This study emphasizes the theoretical and practical implications of our research findings. This research indicates the need to integrate wrapper-based and tree-based feature selection with smart hyperparameter adjustment for optimization of performance of the Parkinson disease prediction model. These effective methods, usually neglected in previous research, give the highest accuracy in illness prediction. Our results are of value to healthcare providers and researchers. Our research improves identification and risk assessment of Parkinson disease by improving the prediction models, improving patient outcomes and treatment methods.

However, it is worth noting that some limitations are intrinsic to our proposal. The sole use of vowel phonation data for disease prediction may not totally capture Parkinson symptoms and biomarkers. Optimizing model performance is important, but it may overlook model generalizability across patient groups or clinical situations. Future research efforts would then be possible in addressing some of these issues through the introduction of a broader scope of sources of data as well as diverse validation methods to prove the robustness and applicability of predictive models in different environments.

Several intriguing research directions, in addition to the presented work, open up from this study. Indeed, the application of multimodal sources of clinical, imaging, and genetic data might further raise the precision and accuracy of predictive models for Parkinson disease. The development of interpretable machine learning models could perhaps shed light into the underlying biology and disease progression pathways, which are crucial to advancing the understanding of Parkinson disease. Longitudinal analyses of predictive model utility in predicting disease progression or treatment response at different points of time can add valuable clinical perspectives and help make personalized therapeutic intervention decisions. Such pathways can lead to further research investments that push our knowledge of Parkinson disease forward and support the best care for patients.

## Authors' declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images, that are not ours, have been included with the necessary permission for republication, which is attached to the manuscript.
- Authors sign on ethical consideration's approval.
- No animal studies are present in the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee at JIS College of Engineering, West Bengal, India.

## Authors' contributions statement

S.D.G., S.S., P.P., S.M., and S.N. contributed to the research's design and implementation, the results analysis, and the manuscript's writing.

## References


1. Tysnes O, Storstein A. Epidemiology of parkinson's disease. J Neural Transm. 2017 Aug;124(8):901–915. https://doi.org/10.1007/s00702-017-1686-y.
2. Post B, van den Heuvel L, van Prooije T, van Ruissen X, van de Warrenburg B, Nonnekes J. Young onset parkinson's disease: A modern and tailored approach. J Parkinsons Dis. 2020 Sep 1;10(s1):S29–S36. https://doi.org/10.3233/JPD-202135.
3. Wang W, Lee J, Harrou F, Sun Y. Early detection of parkinson's disease using deep learning and machine learning. IEEE Access. 2020;8:147635–147646. https://doi.org/10.1109/access.2020.3016062.
4. Bajaj R, Sharma V. Education with artificial intelligence based determination of learning styles. Procedia Comput. Sci. 2018;132:834–842. https://doi.org/10.1016/j.procs.2018.05.095.
5. Hassan FH, Omar MA. Recurrent stroke prediction using machine learning algorithms with clinical public datasets: An empirical performance evaluation. Baghdad Sci J. 2021 Dec 20;18(4(Suppl.)):1406. https://doi.org/10.21123/bsj.2021.18.4(Suppl.).1406.
6. Saha S, Dasgupta S, Anam A, Saha R, Nath S, Dutta S. An investigation of suicidal ideation from social media using machine learning approach. Baghdad Sci J. 2023 Jun 20;20(3(Suppl.):1164. https://doi.org/10.21123/bsj.2023.8515.
7. Islam M, Hasan Majumder M, Hussein M, Hossain KM, Miah M. A review of machine learning and seep learning algorithms for parkinson's disease detection using handwriting and voice datasets. Heliyon. 2024 Feb;10(3). https://doi.org/10.1016/j.heliyon.2024.e25469.
8. Rana A, Dumka A, Singh R, Panda MK, Priyadarshi N, Twala B. Imperative role of machine learning algorithm for detection of parkinson's disease: review, challenges and recommendations. Diagnostics (Basel). 2022 Aug 19;12(8):2003. https://doi.org/10.3390/diagnostics12082003.
9. Moradi S, Tapak L, Afshar S. Identification of novel noninvasive diagnostics biomarkers in the parkinson's diseases and improving the disease classification using support vector machine. Biomed Res Int. 2022 Mar 15;2022:1–8. https://doi.org/10.1155/2022/5009892.
10. Keserwani P K, Das S, Sarkar N. A comparative study: prediction of parkinson disease using machine learning, deep learning and nature inspired algorithm. Multimed Tools Appl. 2024. https://doi.org/10.1007/s11042-024-18186-z.

11. Cordella F, Paffi A, Pallotti A. Classification-based screening of parkinson disease patients through voice signal. In IEEE International Symposium on Medical Measurements and Applications, Lausanne, Switzerland, 2021;1–6. https://doi.org/10.1109/MeMeA52024.2021.9478683.
12. Ali L, Chakraborty C, He Z, Cao W, Imrana Y, Rodrigues JJPC. A novel sample and feature dependent ensemble approach for parkinson's disease detection. Neural Comput Appl. 2023 Aug;35(22):15997–6010. https://doi.org/10.1007/s00521-022-07046-2.
13. Huang F, Xu H, Shen T, Jin L. Recognition of parkinson disease based on residual neural network and voice diagnosis. In IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference, Xi'an, China, 2021; 381–386. https://doi.org/10.1109/ITNEC52019.2021.9586915.
14. Wodzinski A, Skalski D, Hemmerling J R, Orozco-Arroyave, Nöth E. Deep learning approach to parkinson disease detection using voice recordings and convolutional neural network dedicated to image classification. In 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2019;717–720. https://doi.org/10.1016/j.procs.2023.01.007.
15. Wroge TJ, Ozkanca Y, Demiroglu C, Si D, Atkins DC, Ghomi RH. Parkinson's disease diagnosis using machine learning and voice. In IEEE Signal Processing in Medicine and Biology Symposium, 2018;1–7. https://doi.org/10.1109/SPMB.2018.8615607.
16. Wang W, Lee J, Harrou F, Sun Y. Early detection of parkinson disease using deep learning and machine learning. IEEE Access. 2020;8:147635–147646. https://doi.org/10.1109/ACCESS.2020.3016062.
17. Alkhatib R, Diab MO, Corbier C, Badaoui ME. Machine learning algorithm for gait analysis and classification on early detection of parkinson. IEEE Sens Lett. 2020 Jun;4(6):1–4. http://dx.doi.org/10.1109/LSENS.2020.2994938.
18. Ricciardi C, Amboni M, Santis C, Ricciardelli G, Improta G, Addio G, *et.al.* Machine learning can detect the presence of mild cognitive impairment in patients affected by parkinson disease. In IEEE International Symposium on Medical Measurements and Applications, Italy. 2020;1–6. https://doi.org/10.1109/MeMeA49120.2020.9137301.
19. Yang X, Ye Q, Cai G, Wang Y, Cai G. PD-ResNet for classification of parkinson's disease from gait. IEEE J. Transl Eng Health Med. 2022;10:1–11. https://doi.org/10.1109/JTEHM.2022.3180933.
20. Haq AU, Li JP, Memon MH, Khan J, Malik A, Ahmad T, *et al.* Feature selection based on l1-norm support vector machine and effective recognition system for parkinson's disease using voice recordings. IEEE Access. 2019;7:37718–34. http://dx.doi.org/10.1109/ACCESS.2019.290635.
21. Mei J, Desrosiers C, Frasnelli J. Machine learning for the diagnosis of parkinson's disease: A review of literature. Front Aging Neurosci. 2021 May 6;13. https://doi.org/10.3389/fnagi.2021.633752.
22. Mathur R, Pathak V, Bandil D. Parkinson disease prediction using machine learning algorithm. Adv. Intell. Syst. Comput. Springer. 2018;357–363. http://dx.doi.org/10.1007/978-981-13-2285-3_42.
23. Sakar CO, Serbes G, Gunduz A, Tunc HC, Nizam H, Sakar BE, *et. al.* A comparative analysis of speech signal processing algorithms for parkinson's disease classification and the use of the tunable q-factor wavelet transform. Appl Soft Comput. 2019 Jan;74:255–63. https://doi.org/10.1016/j.asoc.2018.10.022.
24. Avuçlu E, Elen A. Evaluation of train and test performance of machine learning algorithms and parkinson diagnosis with statistical measurements. Med Biol Eng Comput. 2020 Nov;58(11):2775–2788. Epub 2020 Sep 13. PMID: 32920727. https://doi.org/10.1007/s11517-020-02260-3.
25. Senturk Z. Early diagnosis of parkinson's disease using machine learning algorithms. Med Hypotheses. 2020 May;138:109603. https://doi.org/10.1016/j.mehy.2020.109603.
26. Yaman O, Ertam F, Tuncer T. Automated parkinson's disease recognition based on statistical pooling method using acoustic features. Med Hypotheses. 2020 Feb;135:109483. https://doi.org/10.1016/j.mehy.2019.109483.
27. Aich S, Kim HC, Younga K, Hui KL, Al-Absi AA, Sain M. A supervised machine learning approach using different feature selection techniques on voice datasets for prediction of parkinson's disease. In 21st International Conference on Advanced Communication Technology. 2019;1116–1121. http://dx.doi.org/10.23919/ICACT.2019.8701961.
28. Govindu A, Palwe S. Early detection of parkinson disease using machine learning. Procedia Comput Sci. 2023;218:249–261. https://doi.org/10.1016/j.procs.2023.01.007.
29. Alshammri R, Alharbi G, Alharbi E, Almubark I. Machine learning approaches to identify parkinson's disease using voice signal features. Front Artif Intell. 2023 Mar 28;(6):1–8. https://doi.org/10.3389/frai.2023.1084001.
30. Patra AK, Ray R, Abdullah AA, Dash SR. Prediction of parkinson's disease using ensemble machine learning classification from acoustic analysis. J Phys: Conf Ser. 2019 Nov 1;1372(1):012041. https://doi.org/10.1088/1742-6596/1372/1/012041.
31. Sandhiya S, Ashok. S, Rao G, Prabhu V, Mohanraj K, Azhagumurugan R. Parkinson's disease prediction using machine learning algorithm. In International Conference on Power, Energy, Control and Transmission Systems. 2022;1–5. https://doi.org/10.1109/ICPECTS56089.2022.10047447.
32. Celik E, Omurca SI. Improving Parkinson's disease diagnosis with machine learning methods. Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science. 2019;1–4. https://doi.org/10.1109/EBBT.2019.8742057.
33. Jain D, Mishra AK, Das SK. Machine learning based automatic prediction of parkinson's disease using speech features. In International Conference on Artificial Intelligence and Applications. 2020;351–62. https://doi.org/10.1007/978-981-15-4992-2_33.

# دراسة حول التنبؤ بمرض باركنسون القائم على التعلم الآلي

**سوبارنا داسغوبتا[1]، سوميابراتا ساها[1]، بروناي بال[1]، شيلارشانا مايتي[2]، سودارشان ناث[1]**

[1]قسم تكنولوجيا المعلومات، كلية الهندسةJIS ، ولاية البنغال الغربية، الهند .
[2]قسم تطبيقات الحاسوب، كلية الهندسةJIS ، ولاية البنغال الغربية، الهند.

## الخلاصة

يعتبر تطور مرض باركنسون مزمنًا، ويتميز بتفاقم الأعراض بمرور الوقت، بما في ذلك الرعشات وضعف الحركة والتدهور المعرفي والتغيرات السلوكية. طرق التشخيص التقليدية عرضة للذاتية، مما يؤدي إلى التشخيص الخاطئ بسبب الطبيعة الخفية لتقييم الأعراض. تهدف هذه الدراسة إلى تقييم مدى فعالية خوارزميات التعلم الآلي المختلفة في التنبؤ بمرض باركنسون باستخدام بيانات نطق حروف العلة، بهدف الكشف المبكر وتحسين الدقة في تقييم قابلية المرضى للإصابة بالمرض. العديد من خوارزميات التعلم الآلي، بما في ذلك الغابة العشوائية والانحدار اللوجستي وشجرة القرار وآلة ناقل الدعم ( (SVMوآلة تعزيز التدرج ( (GBMوتعزيز التدرج الفائق ( (XGB وتعزيز التدرج الخفيف ( (Light GBMوالتعزيز الفئوي ( ( Cat Boost، يتم تقييمها لقدراتها التنبؤية. من بين المصنفات التي تم تقييمها، أظهرت Random Forest أعلى دقة تبلغ 98.4% في تصنيف مرض باركنسون باستخدام بيانات نطق الحروف المتحركة. تستكشف هذه الدراسة طرقًا مختلفة لتعزيز الكشف المبكر وتقييم المخاطر لمرض باركنسون. تؤكد النتائج على أهمية الاستفادة من خوارزميات التعلم الآلي للكشف المبكر والتنبؤ الدقيق لمرض باركنسون. يساهم هذا البحث في وضع استراتيجية لتقييم أكثر دقة لمدى تعرض المرضى للإصابة بمرض باركنسون، مما يسهل تحسين الفهم والبحث المستقبلي في هذا المجال.

**الكلمات المفتاحية:** الخوارزمية، المرض، التعلم الآلي، باركنسون، التنبؤ.