Diagnosis of Parkinson's Disease (PD) in Early Stages Through Voice Analysis Using Machine Learning Algorithm

Submitted By

Sazzad Hossain

Section: A

ID: 211-35-3167

Department of Software Engineering

Daffodil International University

Supervised By

Mr. A.H.M Shahriar Parvez

Associate Professor

Department of Software Engineering

Daffodil International University

Thesis submitted in fulfillment of the requirements for the award of the degree of Bachelor of Science

Fall - 2024

APPROVAL

This thesis titled on " **Diagnosis of Parkinson's Disease (PD) in Early Stages Through Voice Analysis Using Machine Learning Algorithm**", submitted by **Sazzad Hossain (ID: 211-35-3167)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Chairman

Dr. Imran Mahmud
Associate Professor & Head
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

Nuruzzaman Faruqi
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

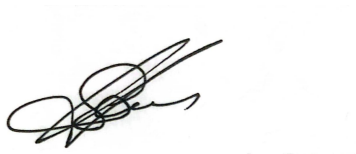
Md. Rajib Mia
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Md. Fazle Munim
Associate Director & Vice President
Government & Public Sector
Ernst & young (EY)

SUPERVISOR’s DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science.



(Supervisor’s Signature)

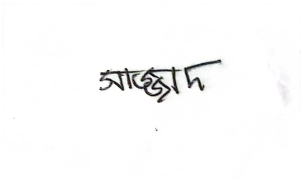
Full Name : Mr. A.H.M Shahriar Parvez

Position : Associate Professor

Date : January 2025

STUDENT'S DECLARATION

I confirm that this thesis is the result of my original work, except where quotations and citations have been appropriately acknowledged. I further declare that it has not been submitted previously or simultaneously for any other degree at Daffodil International University or any other institution.



(Student's Signature)

Full Name : Sazzad Hossain

ID Number : 211-35-3167

Date : January 2025

ACKNOWLEDGEMENT

First and foremost, I express my heartfelt gratitude to Almighty Allah for granting me the strength, wisdom, and perseverance to complete this research. Throughout my academic journey, I have been profoundly grateful for the unwavering love, support, and encouragement of my parents, whose faith in me has always been my greatest source of motivation and inspiration.

I extend my sincere thanks to my supervisor, Associate Professor Mr. A.H.M Shahriar Parvez, for his invaluable advice, support, and guidance throughout this research. His expertise and insights have significantly influenced this work. I am also deeply appreciative of the departmental head, Dr. Imran Mahmud, for his continuous support, thoughtful guidance, and constructive feedback, which were instrumental in helping me successfully complete this journey. Lastly, I extend my gratitude to my friends, colleagues, and everyone who supported and encouraged me throughout this process.

ABSTRACT

Parkinson's Disease, a progressive neurodegenerative disorder, is challenging to diagnose at its early stages due to symptom overlap with other conditions and the subtle onset of clinical features. Leveraging vocal impairments, which affect up to 90% of PD patients even during early stages, this research demonstrates the potential of speech analysis as a non-invasive, cost-effective diagnostic tool. A systematic workflow was employed, involving data preprocessing, exploratory data analysis, feature selection, and the application of five machine learning models: Logistic Regression, Random Forest Classifier, Support Vector Classifier (SVC), Gradient Boosting Classifier, and XGBoost Classifier. These models were trained and evaluated on a dataset containing phonetic features extracted from voice recordings, with performance measured through accuracy, precision, recall, F1 score, and AUC. The findings highlight the efficacy of ensemble learning models, particularly Gradient Boosting and XGBoost, in accurately classifying PD cases. These results validate the use of vocal biomarkers and machine learning in advancing diagnostic precision for neurodegenerative diseases.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Parkinson's disease (PD) is a neurodegenerative disease characterized by primary defects in motor functions, and is revealed through symptoms such as tremors, bradykinesia, muscle rigidity, and postural instability. Degeneration of dopaminergic neurons in the brain, reducing the efficient communication between the brain cells, hence leading to a deficit in motor functions primarily caused this disorder [1]. PD is the second most common neurodegenerative disorder, after Alzheimer's, affecting 1–2% of people over the age of 60, causing major chronic disability worldwide [2]. The diagnosis of this condition must be performed as early as possible because symptoms such as resting tremor and bradykinesia develop only when major neuronal destruction has taken place.

Despite symptomatic treatments continuing to improve, the arrest of neurodegeneration in PD remains an unmet challenge. Early diagnosis is challenging as clinical features of PD closely overlap with other diseases and the insidious onset blurs the early symptoms of PD [3]. Symptomatic treatments aim to restore the brain levels of dopamine without affecting the underlying neurodegenerative process. Novel methods of diagnosis are urgently required, capable of diagnosing PD during their preclinical stages. Recently, vocal biomarkers have emerged as a promising non-invasive diagnostic toolbox. Since almost 90% of patients affected by PD develop some forms of vocal impairment even at its early stages, speech analysis can represent an effective method for the diagnosis at early stages [4, 5]. Advances in the area of CI and ML provide this section with the necessary set of tools for high-accuracy analysis of speech samples for early detecting the disease [6].

The current diagnostic practices rely heavily on clinical evaluations, which might not be able to catch the symptoms at an early stage and delay interventions. Diagnosis by traditional methods involves MRI or motion-based tests, which are resource-intensive and less accessible. Symptoms also often overlap with other disorders,

further complicating diagnostics. This demands the need for innovative, cost-effective, accessible methods for early detection of PD.

This study will explore the integration of vocal biomarkers with deep machine learning models for improved early diagnosis of PD. According to studies, as many as 90% of patients with PD have vocal impairments, such as reduced pitch variability, irregular rhythm, and monotone speech, even during prodromal stages [4,7]. Vocal analysis is a non-invasive, low-cost alternative to conventional MRI or motion-based approaches. This has changed with developments in computational intelligence and machine learning; modern algorithms are now capable of detecting subtle changes in the vocal features associated with PD using techniques such as spectral analysis and acoustic modeling [5, 8].

This paper aims to explore the potential of vocal biomarkers in the diagnosis of PD and the capabilities of machine learning models in improving accuracy. The study addresses several challenges by capitalizing on state-of-the-art developments in non-invasive diagnostic technology for improved diagnostic precision and prognosis in neurodegenerative disorders such as PD.

CHAPTER 2
LITERATURE REVIEW

Govindu et. al. [9] emphasis on the early detection of Parkinson's Disease using voice as biomarker and applied machine learning. This paper differentiates four ML models: Random Forest, Support Vector Machine, Logistic Regression, and K-Nearest Neighbors, each using the MDVP audio dataset. Among all of these, the Random Forest model yielded the best performance with an accuracy of 91.83% and a sensitivity of 0.95; therefore, it can be considered as the best method for detecting Parkinson's disease in telemedicine applications. Feature reduction was also done with PCA, hence SVM reached 91.75%. Of the rest, the KNN classifier performed best in this balanced dataset, as it can handle such data more easily, having equal positive and negative samples. These results underline the robustness and versatility of Random Forest in handling unbalanced data, while SVM was excellent in reduced datasets. dependence on audio data is the only biomarker, small dataset size, and the need for data balancing techniques are limitations of the dataset. Using extra biomarkers, such as REM sleep patterns, in order to improve detection and generalize the application are possible options for future.

Favaro et. al. [10] addresses early detection of PD using speech-based features in a novel dataset, ParkCeleb, containing longitudinal speech recordings data from public figures from hollywood. The investigation covers data points from 10 years before diagnosis up to 20 years after diagnosis, while covering speech attributes related to pitch variability, pause duration, and speech rate. Experimental results on classification achieved high accuracy—Area Under the Curve up to 0.93—for detecting prodromal symptoms. These findings also point out the potential of speech analysis in early disease detection and treatment monitoring.

The key limitations are the potential biases due to dataset variability, which includes differences in recording conditions and demographic mismatches. Further, public recordings may lack representative linguistic subtlety, thereby limiting generalization to diverse populations. Despite these, this study contributes to extending the phenotypic analyses of PD based on speech patterns.

Luna et. al. [11] proposes an improved version of the Smallest Normalized Difference Associative Memory algorithm, called ISNDAM, for PD detection from voice recordings. ISNDAM introduces an enhancement in the classification performance by including a feature selection phase and was tested on two publicly available datasets. The experiments prove that ISNDAM yields a classification accuracy of 99.48% for Dataset 1 and 99.66% for Dataset 2. Results also proved that this performed exceptionally well among the tested 70 different machine learning models from the WEKA platform and further provided results superior to those reported for some previous studies on both Dataset 1 and Dataset 2, hence proving that ISNDAM is better than these in classifying PDs.

Although ISNDAM turned out to be successful, the method still has its shortcomings: it relies on small and very specific datasets, and additional validation with real-world data is required. While the method is tolerant of noise, its scalability and performance on larger, more diverse datasets remain untested. Work could be done in the future on enhancing robustness across broader datasets and the integration of additional biomarkers to enhance its diagnostic utility.

Tracy et. al. [12] focuses on the voice as a biomarker for early detection of PD by analyzing acoustic features from voice recordings of PD patients and healthy controls. It accentuates the power of machine learning models in differentiating early-stage PD. Gradient boosted trees were the best among those tested, achieving an AUC score of 0.95. It also furthers the issue of identity confounding, where leakages between training and test sets inflate performance; when this was addressed, the best AUC stood at 0.88—what more realistic performance has shown.

These would involve limitations in self-reported diagnosis that may or may not be clinically validated and an observed significant deviation in participant demographics. Dataset imbalance and alteration to the PD severity scale may also make a difference. Further research work should be performed on clinically validated datasets with various populations presenting related movement disorders to better enhance model robustness.

Sayed et. al. [13] has presented the possibility of vocal biomarkers in combination with sophisticated machine learning algorithms for early detection of PD. In the present study, the parameters used for analyzing speech signals are jitter, shimmer, and harmonic features. LightGBM, XGBoost, AdaBoost, and SVM were some of the models used in the study. Among all, LightGBM emerged as the most powerful model, identifying 96% accuracy, 100% sensitivity, and 94.43% specificity, hence emerging as a robust candidate for accurate prediction of PD. The XGBoost model produced the highest AUC, at 97%, which is an indication of strong classification capability.

Although these results for the study appear promising, the few limitations include a small dataset with imbalance; hence, methods for oversampling such as SMOTE were considered. Advanced imaging can also be incorporated for further diagnostics improvement. This research has outlined the role of non-invasive methods for improvement in early detection of PDs and underlined the role that machine learning plays in the diagnosis of neurodegenerative diseases.

Chintalapudi et. al. [14] has targeted the early detection of PD using voice biomarkers along with advanced ML models. The used classifiers are three in number: SVM, KNN, and RF. Feature reduction techniques include PCA and SMOTE to improve the model's efficiency. About each given analyzed model, RF showed quite huge accuracy—97.4%; thereafter, the SVM was presenting 85.1 percent, and KNN with an accuracy of 80.7%. This study shows how both enhancements in balancing and simplifications accomplished when reducing a dataset using PCA are reinforced.
This research underlines the problem of small sample sizes and unbalanced datasets that may limit the generalization of the model. Whereas the proposed algorithms significantly improve the classification accuracy, the results may not generalize to diverse datasets. Larger datasets and additional biomarkers should be studied in future research to validate the findings and extend the results to more general applications.

Wang et. al. [15] reviews the performances of various deep learning and machine learning models for the early detection of PD using premotor features such as REM

sleep behavior disorder, olfactory loss, cerebrospinal fluid biomarkers, and dopaminergic imaging. In this study, data from PPMI including 183 healthy individuals and 401 early PD patients were used. The deep learning model achieved the best performance, detecting 96.45% of the images correctly; this algorithm outperformed 12 other traditional machine learning models.

Although the results are very promising, several limitations are important to note in the study, including reliance on a relatively small dataset, which may limit generalizability. Further, although deep learning models are really good at handling complex and nonlinear relationships, interpretability remains a challenge for them because of their "black-box" nature. The future work will focus on validation of these findings with more and diverse datasets to make the results more robust, exploring other biomarkers to increase the accuracy of the prediction.

Karapinar Senturk [16] proposes a concept for early diagnosis of PD with phonetic features extracted from the voice data based on machine learning techniques. The approach combines two techniques, namely Recursive Feature Elimination and Feature Importance, for feature selection while evaluating the classifiers such as Support Vector Machines, Artificial Neural Networks, and Classification and Regression Trees. Among all the others, SVM, when combined with RFE, gave the maximum accuracy in classification, that is, 93.84%, therefore, proving to be an effective method for diagnosing PD with less computational overhead. The dataset used had speech signals from a total of 31 individuals out of which 23 were affected by Parkinson's disease; therefore, this can be said to be pretty noninvasive voice analysis compared with the traditional methods that needed MRI or motion-based technique application.

Suppa et. al. [17] investigates the application of machine learning to voice changes in PD. It recorded voice data from 115 PD patients and 108 healthy controls with a focus on early- and mid-advanced stages of PD. It employed SVM for detecting abnormalities and monitoring disease progression. The results showed that 84% of PD patients demonstrated impairments in voice, with the severity of insult correlating to the progression of the disease. Of note, L-Dopa improved voice symptoms but did not restore normalcy. Machine learning achieved a high

diagnostic accuracy in distinguishing early-stage PD patients from controls, thus offering great potential as an objective biomarker for early detection. However, a number of limitations are present in this study: inability to consider daily vocal fluctuations and the relatively small sample size are the most important. Furthermore, age differences between the various patient groups and healthy controls may have biased the results. Confirmation of these results using larger datasets is needed, together with an investigation into other voice features that could give further improvement in classification.

CHAPTER 3

METHODOLOGY

The workflow of this study is outlined in figure 3.1, illustrating the systematic approach taken for Parkinson's Disease classification using machine learning models. The process begins with the dataset, which undergoes data preprocessing to ensure quality and consistency. This includes cleaning the data and preparing it for analysis. The workflow is divided into two main branches: data splitting and scaling and exploratory data analysis (EDA).

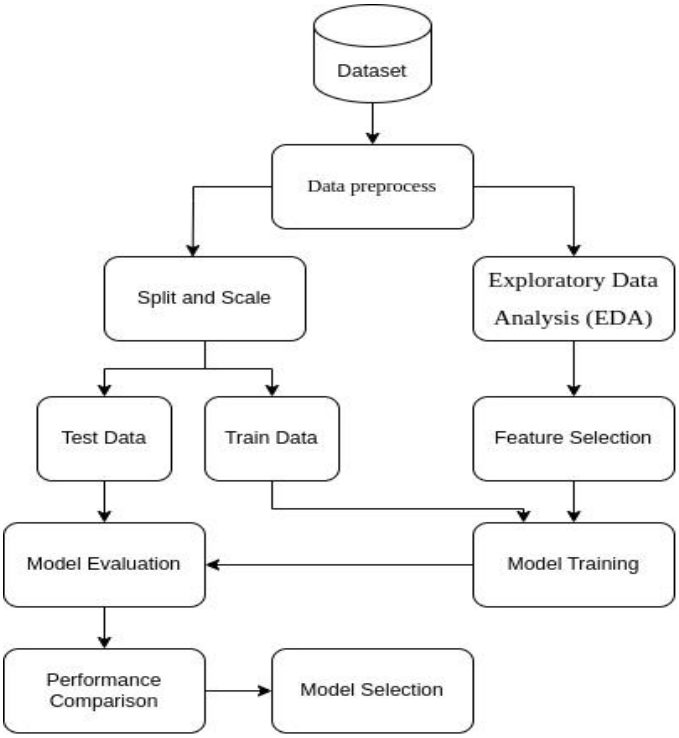


Figure 3.1 : Workflow Diagram

In the first branch, the dataset is split into training and testing subsets, followed by feature scaling to standardize the data. The second branch involves EDA, where key features are analyzed and selected for model training. The selected features are used to train multiple machine learning models. After training, these models are evaluated using the testing data to assess their performance across various metrics. The model evaluation phase compares the performance of all models, leading to model selection based on metrics such as accuracy, precision, recall, F1 score, and AUC. Finally, a performance comparison is conducted to identify the best model for

Parkinson's Disease classification, completing the workflow. This structured approach ensures reliable and interpretable results, highlighting the most effective model for the task.

3.1 Dataset

The dataset being used in this paper is collected from Kaggle [18]. It incorporates phonetic features extracted from voice recordings of PD patients and healthy individuals. It contains 24 features, which comprise different acoustic properties, such as frequency, jitter, shimmer, and noise-to-harmonic ratios. Here are some important features of the dataset:

Frequency Features: MDVP:Fo(Hz), MDVP:Fhi(Hz), MDVP:Flo(Hz)

Jitter Features: MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP

Shimmer Features: MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA

Noise Features: NHR, HNR

Nonlinear Measures: RPDE, DFA, spread1, spread2, D2, PPE

Furthermore, there is one binary target variable that exists, labeled as 'Status,' which suggested 1 for PD and 0 for healthy individuals.

3.2 Data preprocessing

Data preprocessing was done in different steps to ensure data quality for machine learning analysis. The following steps were performed on the dataset in this process: In order to handle missing values df.isnull().sum() method was used to check and confirm the nonexistence of missing values. If there are any missing values, suitable imputation techniques would be applied to deal with them. This step ensures completeness in the dataset and minimizes errors during model training.

The df.duplicated().sum() method was utilized to find the duplicate rows in the dataset and those rows were removed. This prevents redundancy and avoids introducing bias into the model.

The name column was dropped since it contained non-numeric data and information irrelevant to the classification task. This reduces noise and enhances computational efficiency.

The status column was changed into binary values with 1 corresponding to PD and 0 corresponding to healthy. Such a standardization simplifies the classification problem.

These preprocessing steps ensured that the dataset was clean, consistent, and ready for exploratory analysis and machine learning modeling, which improves model accuracy and efficiency.

3.3 Exploratory Data Analysis (EDA)

For an overall understanding of the structure of the dataset, feature relationships, and their association with PD, exploratory data analysis was done using various visualization techniques, proportion and count of people with and without PD with respect to the status were visualized in Figure 3.2 and in Figure 3.3. These plots helped understand the distribution of classes and checks if any class imbalance exists in the dataset. Figure 3.2 shows the distribution of individuals differentiated as "Healthy" (status = 0) and those diagnosed with Parkinson's Disease (status = 1). It shows an imbalance in the dataset, with a larger proportion of samples belonging to the PD class.

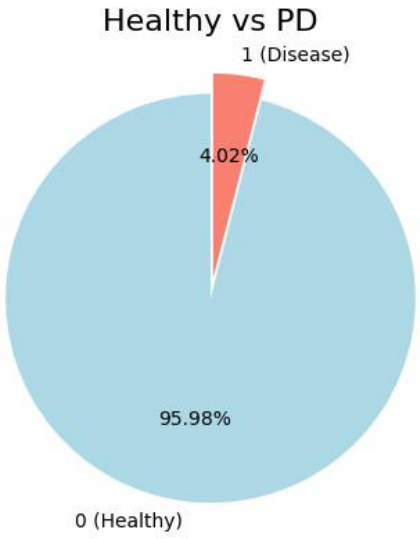


Figure 3.2: Healthy vs PD patient’s demographic

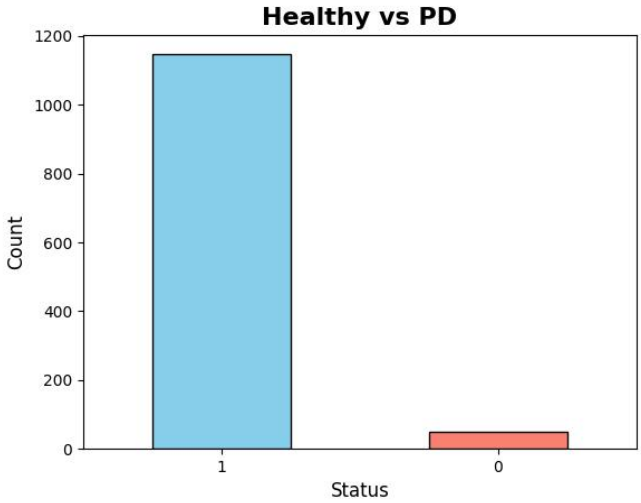


Figure 3.3: Healthy vs PD patient’s numerical difference

Figure 3.3 depicts the count of samples in each class, providing a clearer numerical comparison between the two categories. It confirms the class imbalance, with significantly more samples in the PD category. Figure 3.2 provides an intuitive understanding of proportions, while Figure offers absolute counts, giving complementary perspectives.

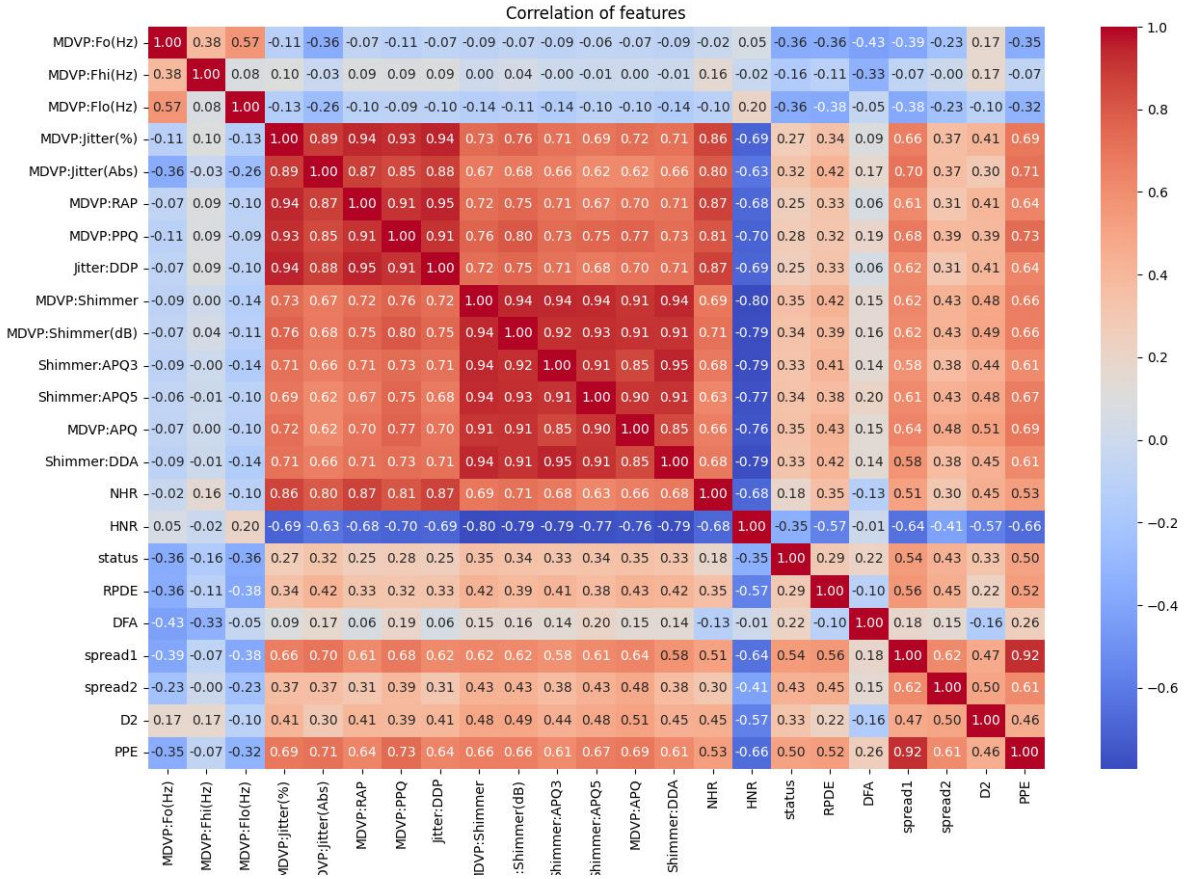


Figure 3.4: Correlation of Features

Figure 3.4 analyzes the correlation between features. It depicts the pairwise correlation between numeric features in the dataset, showcasing the relationships among features. Features like MDVP:Fo(Hz) and MDVP:Fhi(Hz) show strong positive correlations (>0.8), suggesting redundancy. Negative correlations between features such as HNR (Harmonic-to-Noise Ratio) and NHR (Noise-to-Harmonic Ratio) indicate opposing trends in their values. Features with correlations close to 0, such as RPDE and MDVP:Jitter(Abs), indicate independence. Features like MDVP:Fo(Hz) and PPE show prominent correlations with the target variable (status), showing their importance for classification. Figure 3.4 provides a comprehensive overview of feature relationships. It identifies redundancies and highlights features with significant correlations to the target variable.

Figure 3.5 of key features was assessed by observing central tendencies and variability, such as MDVP:Fo(Hz), MDVP:Fhi(Hz), MDVP:Flo(Hz), and HNR. This helped in viewing possible outliers and detecting skewness in the data.

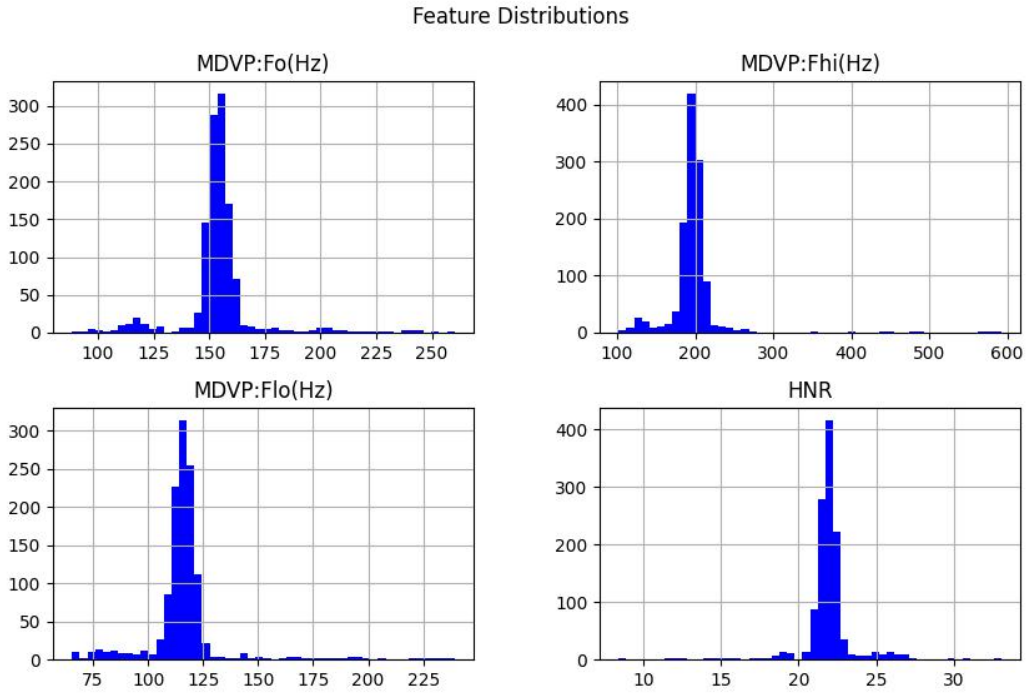


Figure 3.5: Distribution of Features

Figure 3.5 visualizes the distributions of key acoustic features (MDVP:Fo(Hz), MDVP:Fhi(Hz), MDVP:Flo(Hz), and HNR) in the dataset, providing insights into their spread, central tendency, and variability. The distribution of MDVP:Fo(Hz) (Average Fundamental Frequency) is slightly skewed, indicating variability in the vocal frequency of individuals. The right-skewed distribution of MDVP:Fhi(Hz) (Maximum Fundamental Frequency) suggests that most individuals have a maximum frequency within a narrower range, with a few outliers showing significantly higher values. This feature MDVP:Flo(Hz) (Minimum Fundamental Frequency) shows a broader spread compared to MDVP:Fhi(Hz), capturing a wider range of minimum frequencies. The distribution of HNR (Harmonic-to-Noise Ratio) is bell-shaped, indicating that the majority of samples have mid-range HNR values, with fewer samples at extreme values. Figure 3.5 reveals the underlying characteristics of each feature. The skewness in fundamental frequency features (MDVP:Fo(Hz), MDVP:Fhi(Hz), MDVP:Flo(Hz)) suggests variability in vocal behavior, while the normal-like distribution of HNR indicates its potential as a stable predictor.

Figure 3.6 illustrates the relationship between the fundamental frequency (MDVP:Fo(Hz)) and frequency variability (MDVP:Jitter(%)), with points color-coded by the target variable (status). This visualization helps identify whether these features can separate individuals with Parkinson's Disease (PD) from healthy individuals.

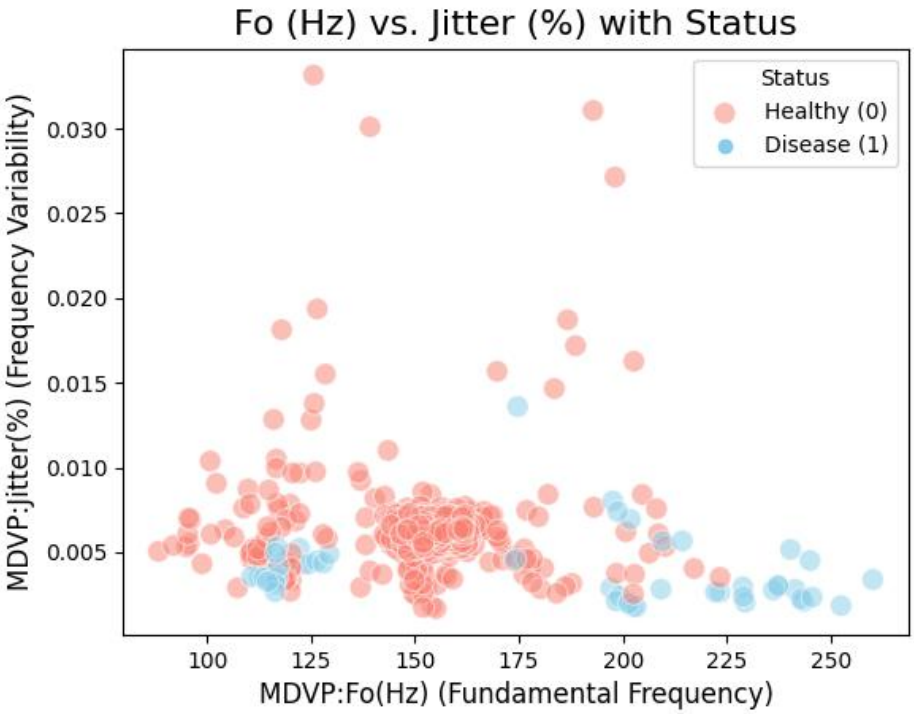


Figure 3.6: MDVP:Fo vs MDVP:jitter(%)

Healthy individuals tend to cluster at higher fundamental frequency values (MDVP:Fo(Hz)) and lower jitter percentages, indicating more stable vocal frequencies. PD individuals exhibit higher variability in jitter percentages, with a broader range of fundamental frequencies. The scatter plot suggests that these two features, particularly MDVP:Jitter(%), are significant predictors for distinguishing between healthy and PD individuals. Some overlap between the two classes is observed, which necessitates the use of additional features for better separability. This scatter plot highlights the distinct patterns in MDVP:Fo(Hz) and MDVP:Jitter(%) across the two classes, underscoring their relevance in PD classification.

Figure 3.6 provides a comprehensive view of pairwise relationships and distributions among key features (MDVP:Fo(Hz), MDVP:Flo(Hz), HNR, NHR, RPDE, PPE) . This visualization helps identify patterns, clustering, and feature separability for Parkinson's Disease (PD) classification.

Figure 3.7: Pairplot of key features

On the diagonal, KDEs reveal significant class separability for features like HNR (harmonic-to-noise ratio) and PPE (pitch period entropy), with minimal overlap between healthy and PD samples. MDVP:Fo(Hz) vs. MDVP:Flo(Hz) shows distinct distributions between healthy and PD samples, indicating their potential predictive value. NHR vs. HNR displays an inverse relationship, with PD samples clustering toward higher noise levels. Some features, like RPDE vs. PPE exhibit moderate overlap between classes, suggesting they may be better used in combination with other features for classification. Features like HNR and PPE appear particularly discriminative based on their clustering behavior, making them strong candidates for classification models. The pairplot underscores the importance of these features in distinguishing PD from healthy individuals. While certain features exhibit strong

class separability. This visualization validates the relevance of the selected features and informs their role in feature selection for machine learning models.

3.4 Feature Selection

The feature selection process, guided by EDA and visualizations, ensured that only the most relevant and non-redundant features were included for machine learning modeling. This step reduced computational complexity, enhanced model performance, and aligned with the goal of leveraging vocal biomarkers for effective PD classification. The insights gained from the infographics provided a robust foundation for building accurate and interpretable classification models.

3.5 Model Training and Evaluation

To prepare the dataset [18] for machine learning modeling, the data was split into training and testing subsets and scaled to ensure uniformity across features. These steps are crucial for robust model training and evaluation.

The dataset was divided into input features (X) and the target variable (y), where X included all features except the status column, and y represented the binary target variable indicating the presence of Parkinson's Disease (PD). The data was split into training and testing sets using an 80:20 ratio. The training set was used to train the machine learning models, while the testing set was reserved for evaluating model performance. A random seed (random_state=1) ensured reproducibility of the data split.

A StandardScaler was applied to transform the data. The scaler normalized each feature by removing its mean and scaling it to unit variance. The training data (X_train) was scaled using the fit_transform method, which computed the necessary scaling parameters (mean and standard deviation) and applied them to the training set. The testing data (X_test) was scaled using the transform method, ensuring the same parameters were applied without recalculating, thereby maintaining consistency between the datasets.

Five machine learning models with diverse characteristics are employed to classify PD from healthy individuals. Each model was configured with tuned hyperparameters to strike a balance between bias and variance, ensuring robust and generalizable performance.

Logistic Regression is a linear model that predicts probabilities for binary classification tasks using the logistic function. It is simple, interpretable, and serves as a baseline model [19]. It acts as a benchmark to compare more complex models and evaluate the linear separability of the dataset.

Random Forest Classifier ensemble learning method that builds multiple decision trees and aggregates their predictions to improve accuracy and reduce overfitting [20]. It captures non-linear relationships and reduces overfitting through ensemble averaging.

Support Vector Classifier (SVC) is a robust algorithm that uses kernel methods to find optimal hyperplanes for separating classes in high-dimensional spaces [21]. It handles complex, non-linear boundaries effectively, making it suitable for datasets with overlapping classes.

Gradient Boosting Classifier is a boosting algorithm that builds sequential decision trees, each correcting errors from the previous one, for enhanced accuracy [22]. It balances complexity and performance, capturing subtle patterns while being resistant to overfitting.

XGBoost Classifier is an advanced implementation of gradient boosting that combines speed and performance optimization [23]. It offers powerful predictive performance with regularization, making it highly effective for structured data.

The models are evaluated with accuracy, precision, f1 score, recall and AUC curve. In order to calculate each metrics confusion matrix is drawn. Here is the confusion matrix for each model:

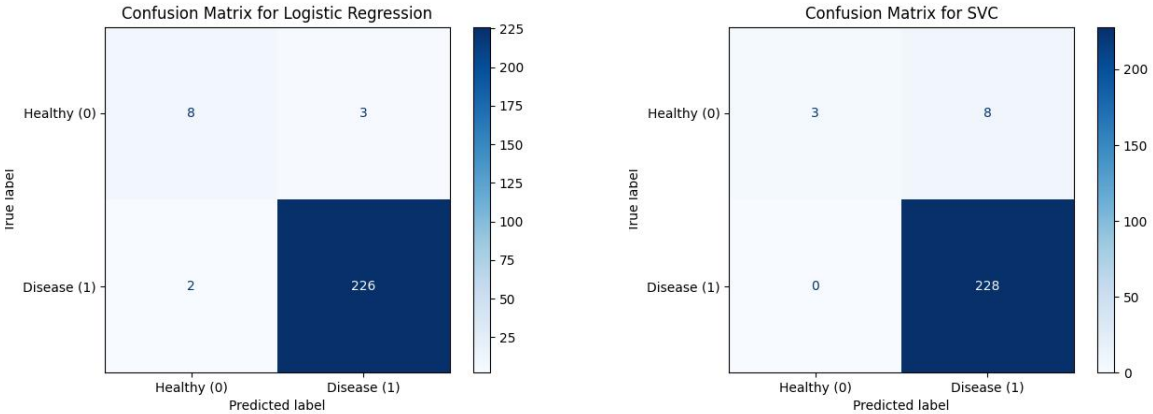
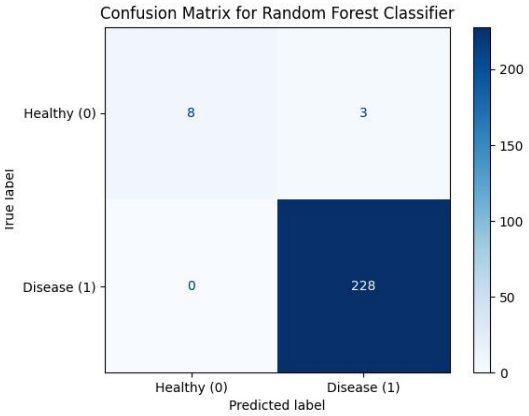


Figure 3.8 (a) : Logistic Regression

Figure 3.8 (b) : SVC



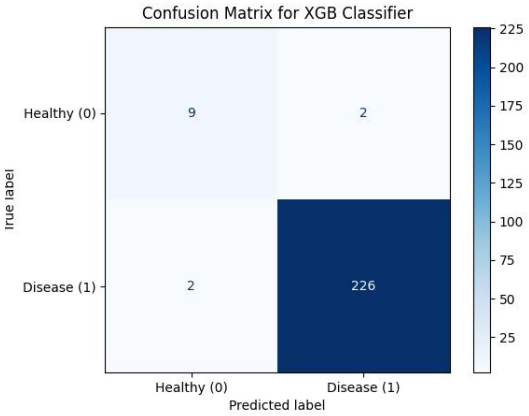


Figure 3.8 (c) : Random Forest Classifier

Figure 3.8 (d) : XGB Classifier

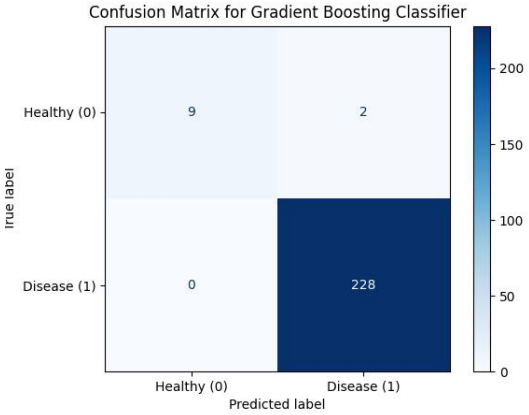


Figure 3.8 (e) : Gradient Boosting Classifier

True Positives (TP): Correctly predicted PD cases.

True Negatives (TN): Correctly predicted healthy cases. Predicted

False Positives (FP): Healthy cases incorrectly classified as PD.

False Negatives (FN): PD cases incorrectly classified as healthy.

From the confusion matrix accuracy, precision, recall and f1 score is calculated.

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

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN} \qquad F1\ Score = 2 \times \frac{precision \times Recall}{Precision + Recall}$$

Model	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	0.979	0.986	0.991	0.989	0.980
Random Forest Classifier	0.983	0.982	1.000	0.991	0.994
SVC	0.966	0.966	1.000	0.982	N/A
Gradient Boosting Classifier	0.987	0.991	0.995	0.993	0.997
XBG classifier	0.983	0.991	0.991	0.991	0.994

Table 3.1: Performance metrics of models

Table 3.1 summarizes the performance metrics of five machine learning models used for Parkinson's Disease (PD) classification: Logistic Regression, Random Forest Classifier, Support Vector Classifier (SVC), Gradient Boosting Classifier, and XGBoost Classifier. The metrics include Accuracy, Precision, Recall, F1 Score, and Area Under the Curve (AUC), providing a comprehensive evaluation of each model's performance.

Figure 3.9 provides a visual evaluation of the classification performance of the machine learning models used for Parkinson's Disease (PD) detection. The ROC curve plots the True Positive Rate (TPR) (sensitivity) against the False Positive Rate (FPR) at various classification thresholds, with the Area Under the Curve (AUC) quantifying overall performance.

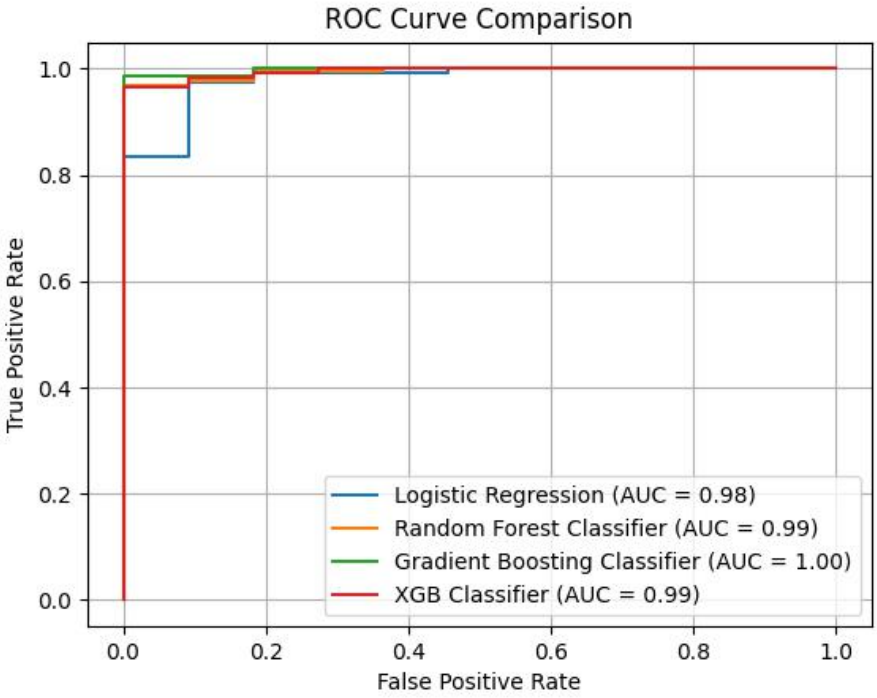


Figure 3.9 : AUC curve of models

The ROC curve comparison highlights the superior performance of ensemble learning models, particularly Gradient Boosting and XGBoost, for PD detection. The steep curves and high AUC values of these models indicate their ability to minimize both false positives and false negatives, making them ideal for applications requiring high precision and recall.

CHAPTER 4
RESULT AND DISCUSSION

The study evaluated the performance of five machine learning models: Logistic Regression, Random Forest Classifier, Support Vector Classifier (SVC), Gradient Boosting Classifier, and XGBoost Classifier for Parkinson's Disease (PD) classification. The models were assessed based on their **c**onfusion matrix, ROC-AUC curve, accuracy, precision, recall, and F1 score to determine their suitability for the task. Table 3.1 and table 4.1 show the best model for PD classification.

Criteria	**Best Model**
Confusion Matrix	Gradient Boosting
ROC-AUC Curve	Gradient Boosting
Accuracy	Gradient Boosting
Recall	Random Forest / SVC
Precision	Gradient Boosting / XGBoost
F1 score	Gradient Boosting

Table 4.1 : Comparisons of models

The Gradient Boosting Classifier achieved the highest overall performance, with an accuracy of 0.987, precision of 0.991, recall of 0.995, and an F1 score of 0.993. It also had the highest AUC (0.997), demonstrating superior class separation and minimal misclassification. The XGBoost Classifier closely followed, achieving an accuracy of 0.983, precision of 0.991, recall of 0.991, and an F1 score of 0.991, with an AUC of 0.994. These results highlight the robustness of ensemble learning models for PD classification tasks.

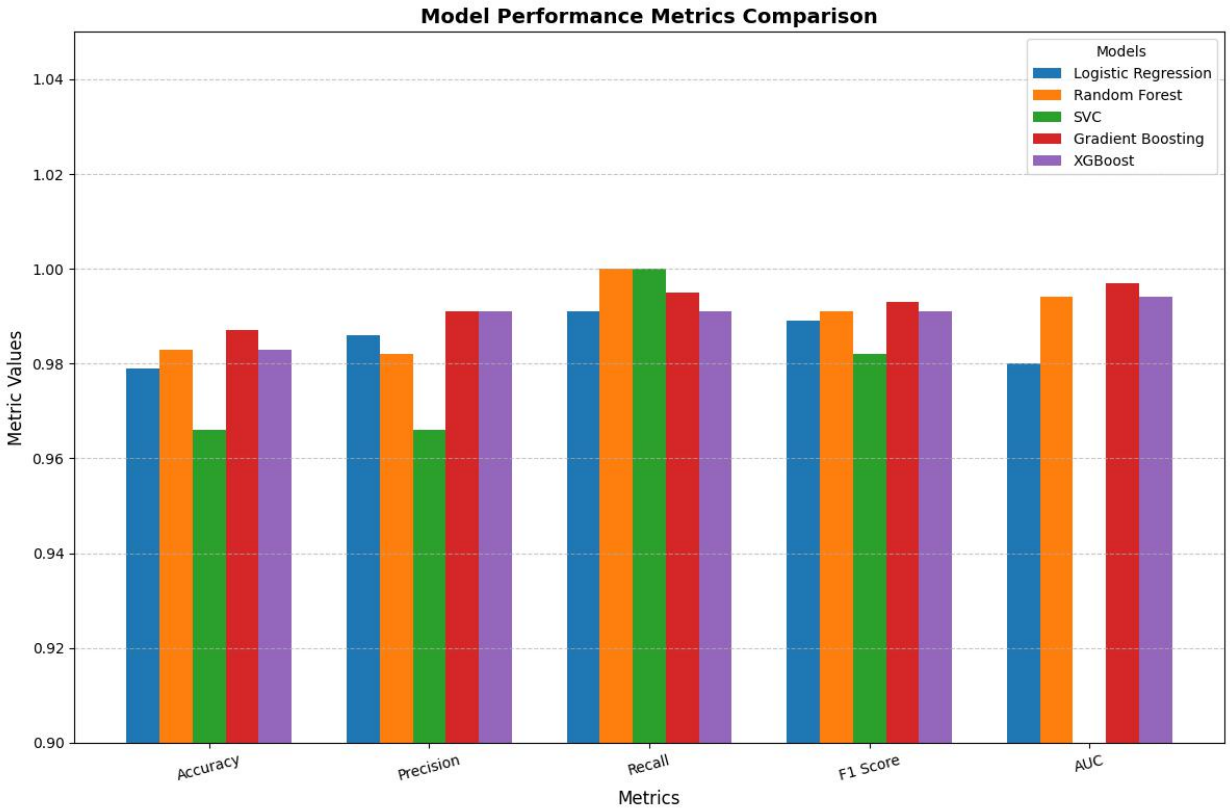


Figure 4.1 : Comparisons of models

The Random Forest Classifier also performed well, achieving an accuracy of 0.983, with a perfect recall of 1.000, ensuring no PD cases were missed. Its precision (0.982) and F1 score (0.991) reflected balanced performance, making it suitable for applications where sensitivity is critical. The SVC demonstrated perfect recall (1.000) but had lower precision (0.966) and accuracy (0.966), indicating a tendency to over-predict PD cases. Despite this, its high sensitivity makes it valuable for screening programs. Logistic Regression served as a robust baseline model, achieving an accuracy of 0.979, precision of 0.986, recall of 0.991, and an F1 score of 0.989. While reliable, it was outperformed by the ensemble methods in handling non-linear relationships.

Overall, the findings underscore the effectiveness of ensemble models, with Gradient Boosting and XGBoost emerging as the most reliable and accurate classifiers for PD detection. Random Forest demonstrated strengths in recall, while SVC and Logistic Regression provided reasonable performance as complementary models. These results validate the use of advanced machine learning techniques for accurate and reliable disease classification.

CHAPTER 5
CONCLUSION

The findings of this study highlight the superiority of ensemble learning models for the classification of Parkinson’s Disease. Among the evaluated models, Gradient Boosting demonstrated the highest performance across all metrics, including a near-perfect AUC, making it the most effective model for accurately distinguishing PD cases from healthy individuals. XGBoost followed closely, offering robust and consistent performance, while Random Forest excelled in recall, ensuring no PD cases were missed. SVC and Logistic Regression provided reasonable performance, with SVC showing high sensitivity but reduced precision, and Logistic Regression serving as a reliable baseline model.

The limitation is the dataset size, which may limit generalization of the results. Future studies should be carried out on the proposed approach using larger and more diverse datasets to confirm improved applicability.

REFERENCE

1. Braak, H., & Braak, E. (2000). Pathoanatomy of Parkinson's disease. *Journal of Neurology, 247*(Suppl 2), II3–II10. https://doi.org/10.1007/PL00007758
2. Schapira, A. H. (2009). Neurobiology and treatment of Parkinson's disease. *Trends in Pharmacological Sciences, 30*(1), 41–47. https://doi.org/10.1016/j.tips.2008.11.005
3. Razali, R., Ahmad, F., Rahman, F. N. A., Midin, M., & Sidi, H. (2011). Burden of care among caregivers of patients with Parkinson's disease: A cross-sectional study. *Clinical Neurology and Neurosurgery, 113*(8), 639–643. https://doi.org/10.1016/j.clineuro.2011.05.008
4. Ho, A. K., Iansek, R., Marigliani, C., Bradshaw, J. L., & Gates, S. (1998). Speech impairment in a large sample of patients with Parkinson's disease. *Behavioral Neurology, 11*(3), 131–137
5. Amato, F., Rechichi, I., Borzì, L., & Olmo, G. (2022). Sleep quality through vocal analysis: A telemedicine application. *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and Other Affiliated Events (PerCom Workshops)*, 706–711 https://doi.org/10.1109/PerComWorkshops53856.2022.9767372
6. Dash, S. (2023). A systematic review of adaptive machine learning techniques for early detection of Parkinson's disease. In A. Abraham, S. Dash, S. K. Pani, & L. García-Hernández (Eds.), *Artificial Intelligence for Neurological Disorders* (pp. 361–385). Academic Press. https://doi.org/10.1016/B978-0-323-99400-1.00022-3
7. Schalling, E., Johansson, K., & Hartelius, L. (2018). Speech and communication changes reported by people with Parkinson's disease. *Folia Phoniatrica et Logopaedica, 69*(3), 131–141. https://doi.org/10.1159/000479927
8. Rusz, J., Tykalova, T., Ramig, L. O., & Tripoliti, E. (2020). Guidelines for speech recording and acoustic analyses in dysarthrias of movement disorders. *Movement Disorders, 36*(5), 803–814. https://doi.org/10.1002/mds.28465

9. Govindu, A., & Palwe, S. (2023). Early detection of Parkinson's disease using machine learning. Procedia Computer Science, 218, 249–261. https://doi.org/10.1016/j.procs.2023.01.007
10. Favaro, A., Butala, A., Thebaud, T., Villalba, J., Dehak, N., & Moro-Velázquez, L. (2024). Unveiling early signs of Parkinson's disease via a longitudinal analysis of celebrity speech recordings. npj Parkinson's Disease, 10(207). https://doi.org/10.1038/s41531-024-00817-9
11. Luna-Ortiz, I., Aldape-Pérez, M., Uriarte-Arcia, A. V., Rodríguez-Molina, A., Alarcón-Paredes, A., & Ventura-Molina, E. (2023). Parkinson's disease detection from voice recordings using associative memories. *Healthcare, 11*(1601). https://doi.org/10.3390/healthcare11111601
12. Tracy, J. M., Özkanca, Y., Atkins, D. C., & Hosseini Ghomi, R. (2020). Investigating voice as a biomarker: Deep phenotyping methods for early detection of Parkinson's disease. *Journal of Biomedical Informatics, 104*, 103362. https://doi.org/10.1016/j.jbi.2019.103362
13. Sayed, M. A., Tayaba, M., Islam, M. T., Pavel, M. E. U. I., Mia, M. T., Ayon, E. H., Nob, N., & Ghosh, B. P. (2023). Parkinson's Disease Detection through Vocal Biomarkers and Advanced Machine Learning Algorithms. *Journal of Computer Science and Technology Studies, 5*(4), 142-149. https://doi.org/10.32996/jcsts.2023.5.4.14
14. Chintalapudi, N., Dhulipalla, V. R., Battineni, G., Rucco, C., & Amenta, F. (2023). Voice biomarkers for Parkinson's disease prediction using machine learning models with improved feature reduction techniques. *Journal of Data Science and Intelligent Systems*. https://doi.org/10.47852/bonviewJDSIS3202831
15. Wang, W., Lee, J., Harrou, F., & Sun, Y. (2020). Early Detection of Parkinson's Disease Using Deep Learning and Machine Learning. *IEEE Access, 8*, 147635–147646. https://doi.org/10.1109/ACCESS.2020.3016062
16. Karapinar Senturk, Z. (2020). Early diagnosis of Parkinson's disease using machine learning algorithms. *Medical Hypotheses, 138*, 109603. https://doi.org/10.1016/j.mehy.2020.109603
17. Suppa, A., Costantini, G., Asci, F., Di Leo, P., Al-Wardat, M. S., Di Lazzaro, G., Scalise, S., Pisani, A., & Saggio, G. (2022). Voice in Parkinson's Disease: A

Machine Learning Study. *Frontiers in Neurology, 13,* 831428. https://doi.org/10.3389/fneur.2022.831428

18. Shreya Dutta. (2024). Parkinson's Disease [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/9759604
19. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley. https://doi.org/10.1002/9781118548387
20. Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
21. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018
22. Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451
23. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Diagnosis of Parkinson's Disease (PD) in Early Stages Through Voice Analysis Using Machine Learning Algorithm. V_1

ORIGINALITY REPORT

24%	17%	19%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.mdpi.com Internet Source	1%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	umpir.ump.edu.my Internet Source	1%
4	Submitted to Midlands State University Student Paper	1%
5	assets-eu.researchsquare.com Internet Source	1%
6	josevilaseca.github.io Internet Source	1%
7	Submitted to Coventry University Student Paper	1%
8	Submitted to University of Sunderland Student Paper	1%

9	Dinesh Goyal, Bhanu Pratap, Sandeep Gupta, Saurabh Raj, Rekha Rani Agrawal, Indra Kishor. "Recent Advances in Sciences, Engineering, Information Technology & Management - Proceedings of the 6th International Conference “Convergence2024” Recent Advances in Sciences, Engineering, Information Technology & Management, April 24–25, 2024, Jaipur, India", CRC Press, 2025 Publication	1 %
10	www.frontiersin.org Internet Source	<1 %
11	www.ijirset.com Internet Source	<1 %
12	Submitted to University of Essex Student Paper	<1 %
13	Debasis Chaudhuri, Jan Harm C Pretorius, Debashis Das, Sauvik Bal. "International Conference on Security, Surveillance and Artificial Intelligence (ICSSAI-2023) - Proceedings of the International Conference on Security, Surveillance and Artificial Intelligence (ICSSAI-2023), Dec 1–2, 2023, Kolkata, India", CRC Press, 2024 Publication	<1 %
14	Shoogo Ueno. "Bioimaging - Imaging by Light and Electromagnetics in Medicine and	<1 %

	Biology", CRC Press, 2020 Publication	
15	H.L. Gururaj, Francesco Flammini, S. Srividhya, M.L. Chayadevi, Sheba Selvam. "Computer Science Engineering", CRC Press, 2024 Publication	<1 %
16	econpapers.repec.org Internet Source	<1 %
17	ijrpr.com Internet Source	<1 %
18	Azam, Sikandar. "Characterization of Physico-Chemical Properties of Nano-Sized Particulates and Their Implications on Transport Behavior.", The Pennsylvania State University, 2024 Publication	<1 %
19	www.nature.com Internet Source	<1 %
20	"Machine Intelligence, Tools, and Applications", Springer Science and Business Media LLC, 2024 Publication	<1 %
21	Submitted to Sheffield Hallam University Student Paper	<1 %
22	www.medrxiv.org Internet Source	<1 %

23	Submitted to Liverpool John Moores University Student Paper	<1 %
24	Submitted to University of Warwick Student Paper	<1 %
25	jcreview.com Internet Source	<1 %
26	openbiomedicalengineeringjournal.com Internet Source	<1 %
27	www.researchsquare.com Internet Source	<1 %
28	Zehra Karapinar Senturk. "Early diagnosis of Parkinson's disease using machine learning algorithms", Medical Hypotheses, 2020 Publication	<1 %
29	link.springer.com Internet Source	<1 %
30	Submitted to University of Cincinnati Student Paper	<1 %
31	archives.njit.edu Internet Source	<1 %
32	Submitted to University of East London Student Paper	<1 %

33	Amit Singh Rajawat, Anshika Srivastava. "Recognition of Parkinson's ailment by using various machine learning procedures", Current Psychology, 2024 Publication	<1 %
34	robots.net Internet Source	<1 %
35	Submitted to Daffodil International University Student Paper	<1 %
36	Sangeeta Singh, Sachchida Nand Rai, Santosh Kumar Singh. "Synaptic Plasticity in Neurodegenerative Disorders", CRC Press, 2024 Publication	<1 %
37	Submitted to University of Southampton Student Paper	<1 %
38	Submitted to BB9.1 PROD Student Paper	<1 %
39	cronfa.swan.ac.uk Internet Source	<1 %
40	"Data Science", Springer Science and Business Media LLC, 2022 Publication	<1 %
41	Gunjan Pahuja, Bhanu Prasad. "Deep learning architectures for Parkinson's disease	<1 %

detection by using multi-modal features", Computers in Biology and Medicine, 2022
Publication

42	iris.polito.it Internet Source	<1 %
43	tudr.thapar.edu:8080 Internet Source	<1 %
44	www.hindawi.com Internet Source	<1 %
45	arxiv.org Internet Source	<1 %
46	publications.polymtl.ca Internet Source	<1 %
47	www.efmaefm.org Internet Source	<1 %
48	Cuihua Lv, Lizhou Fan, Haiyun Li, Jun Ma, Wenjing Jiang, Xin Ma. "Leveraging multimodal deep learning framework and a comprehensive audio-visual dataset to advance Parkinson's detection", Biomedical Signal Processing and Control, 2024 Publication	<1 %
49	Giovanni Costantini, Valerio Cesarini, Pietro Di Leo, Federica Amato et al. "Artificial Intelligence-Based Voice Assessment of Patients with Parkinson's Disease Off and On	<1 %

Treatment: Machine vs. Deep-Learning Comparison", Sensors, 2023
Publication

50	Hanguang Xiao. "Diagnosis of Parkinson's disease using genetic algorithm and support vector machine with acoustic characteristics", 2012 5th International Conference on BioMedical Engineering and Informatics, 2012 Publication	<1 %
51	Shahbakhti, Mohammad, Danial Taherifar, and Zahra Zareei. "Combination of PCA and SVM for diagnosis of Parkinson's disease", 2013 2nd International Conference on Advances in Biomedical Engineering, 2013. Publication	<1 %
52	Utkarsh Lal, Arjun Vinayak Chikkankod, Luca Longo. "Fractal dimensions and machine learning for detection of Parkinson’s disease in resting-state electroencephalography", Neural Computing and Applications, 2024 Publication	<1 %
53	core.ac.uk Internet Source	<1 %
54	mdpi-res.com Internet Source	<1 %
55	oa.las.ac.cn Internet Source	<1 %

56	ojs.bonviewpress.com Internet Source	<1 %
57	repository.up.ac.za Internet Source	<1 %
58	Javier Alberto Pérez-Castán, Luis Pérez-Sanz, Lidia Serrano-Mira, Francisco Javier Saéz-Hernando et al. "Design of an ATC Tool for Conflict Detection Based on Machine Learning Techniques", Aerospace, 2022 Publication	<1 %
59	Sadeghifar, Amir. "Reduced Cell Spreading and Force on Talin with Arp2/3 Inhibition Does Not Prevent YAP Activation", The Florida State University, 2023 Publication	<1 %
60	Thangavel Murugan, W. Jai Singh. "Cybersecurity and Data Science Innovations for Sustainable Development of HEICC - Healthcare, Education, Industry, Cities, and Communities", CRC Press, 2025 Publication	<1 %
61	ebin.pub Internet Source	<1 %
62	en.istanbulkongresi.org Internet Source	<1 %
63	www.scirp.org Internet Source	

<1%

64 Chun Wang, Xiaojia Tan, Bokang Zhu, Zehao Zhao, Qian Wang, Ying Yang, Jianqiao Liu, Ce Fu, Junsheng Wang, Yongzhong Lin. "Deep learning-assisted non-invasive pediatric tic disorder diagnosis using EEG features extracted by residual neural networks", Journal of Radiation Research and Applied Sciences, 2024 <1%
Publication

65 Irving Luna-Ortiz, Mario Aldape-Pérez, Abril Valeria Uriarte-Arcia, Alejandro Rodríguez-Molina et al. "Parkinson's Disease Detection from Voice Recordings Using Associative Memories", Healthcare, 2023 <1%
Publication

66 Joselyn Zapata-Paulini, Michael Cabanillas-Carbonell. "Evaluation of machine learning algorithms in the early detection of Parkinson's disease: a comparative study", Indonesian Journal of Electrical Engineering and Computer Science, 2024 <1%
Publication

67 Mohammed Muzaffar Hussain, D.Weslin, S. Kumari, S. Umamaheswari, S. Kamalakannan. "Enhancing Parkinson's Disease Identification using Ensemble Classifier and Data <1%

Augmentation Techniques in Machine Learning", Clinical eHealth, 2023
Publication

68	Osmar Pinto Neto. "Harnessing Voice Analysis and Machine Learning for Early Diagnosis of Parkinson's Disease: A Comprehensive Study Across Diverse Datasets", Research Square Platform LLC, 2023 Publication	<1 %
69	S. I. M. M. Raton Mondol, Ryul Kim, Sangmin Lee. "Hybrid Machine Learning Framework for Multistage Parkinson's Disease Classification Using Acoustic Features of Sustained Korean Vowels", Bioengineering, 2023 Publication	<1 %
70	"IoT and ML for Information Management: A Smart Healthcare Perspective", Springer Science and Business Media LLC, 2024 Publication	<1 %
71	Anna Favaro, Ankur Butala, Thomas Thebaud, Jesús Villalba, Najim Dehak, Laureano Moro-Velázquez. "Unveiling early signs of Parkinson's disease via a longitudinal analysis of celebrity speech recordings", npj Parkinson's Disease, 2024 Publication	<1 %

72	Mehdi Ghayoumi. "Generative Adversarial Networks in Practice", CRC Press, 2023 Publication	<1%
73	N. Shamli, B. Sathiyabhama. "Parkinson's Brain Disease Prediction Using Big Data Analytics", International Journal of Information Technology and Computer Science, 2016 Publication	<1%
74	V. Sharmila, S. Kannadhasan, A. Rajiv Kannan, P. Sivakumar, V. Vennila. "Challenges in Information, Communication and Computing Technology", CRC Press, 2024 Publication	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		