

PARKINSON'S DISEASE DETECTION USING MACHINE LEARNING AND SPEECH ANALYSIS: A SUPPORT VECTOR MACHINE APPROACH

Vedant S Chaudhari*¹

*¹UG Student, Department Of Information Technology B.K Birla College Kalyan,
(Empowered Autonomous Status), India.

ABSTRACT

Parkinson's disease (PD) is a widely recognized neurological condition, the early diagnosis of which is essential for timely intervention and effective management. Recent advancements in the intersection of machine learning and speech analysis have paved the way for promising non-invasive diagnostic tools. This research leverages Support Vector Machines (SVM), a powerful machine learning algorithm, to detect PD based on speech-related features. The significance of timely PD diagnosis is underscored by the potential to improve patient outcomes. Drawing inspiration from relevant studies in the field [1-5], this research builds upon a comprehensive dataset and code framework. It employs standardization techniques for feature scaling and the SVM algorithm for predictive modeling. The primary goal is to assess the SVM model's diagnostic capabilities. Performance metrics, such as accuracy and confusion matrices, are employed to gauge the model's predictive prowess. Visualizations provide insights into data distributions, and a detailed classification report offers a holistic view of model performance. This research contributes to ongoing efforts aimed at enhancing PD diagnosis through speech analysis and machine learning. The outcomes have the potential to revolutionize diagnostic procedures, facilitating earlier intervention and more effective disease management. It represents a promising stride toward improving the quality of life for individuals affected by PD.

Keywords: Parkinson's Disease, Machine Learning, SVM, Speech Analysis, Accuracy, Random Forest, Multilayer Perceptron, Gradient Boosting.

I. INTRODUCTION

Parkinsons disease is a widespread neurodegenerative disorder that imposes a substantial burden on individuals and healthcare systems as the second most common neurodegenerative disease following Alzheimers early diagnosis and effective management are crucial areas of medical research although Parkinsons diseases characteristic symptoms including tremors bradykinesia and rigidity serve as diagnostic indicators they often appear only after significant neurodegeneration has occurred making early diagnosis challenging machine learning a branch of artificial intelligence has emerged as a powerful tool for medical diagnostics with its capacity to analyze extensive datasets and detect intricate patterns machine learning offers a non-invasive approach for early Parkinsons disease detection by integrating with speech analysis even before evident motor symptoms emerge the research significance lies in the potential to redefine Parkinsons disease diagnosis and management traditionally diagnosis relied on clinical assessments that were only feasible at advanced disease stages by applying machine learning and speech analysis early Parkinsons disease detection becomes possible enabling timely interventions and personalized treatment strategies this approach enhances patient care and alleviates the financial and social burdens associated with delayed diagnosis this study builds upon a substantial body of literature that emphasizes the promise of machine learning and speech analysis in Parkinsons disease detection previous research has illuminated the role of speech biomarkers and the feasibility of employing machine learning algorithms such as support vector machines SVM to discriminate between healthy and Parkinsons-afflicted individuals based on speech characteristics these studies combined with the growing prevalence of machine learning in medical diagnostics lay the foundation for an in-depth investigation into SVMs application in early Parkinsons detection this paper delves into the realm of machine learning and speech analysis using an SVM approach to diagnose Parkinsons disease it explores the utilization of speech features as potential biomarkers and aims to elucidate the discriminative power of this approach additionally it compares the SVM models performance with other machine learning models including random forest multilayer perceptron MLP neural network and gradient boosting classifier the research contributes to improving early Parkinsons disease diagnosis and represents progress toward more effective patient care the subsequent sections of this paper provide a comprehensive view of the research findings and their implications

by the study end it is anticipated that the research will contribute to the development of innovative diagnostic tools in healthcare furthering the progress of early Parkinsons disease detection.

The related works done are as follows: In [1] research the study centers on the utilization of noninvasive speech assessments for telemonitoring the progression of Parkinsons disease the study harnesses machine learning methodologies to scrutinize speech attributes offering a means to gauge the severity of the disease this investigation highlights the promise of speech analysis as a valuable instrument for monitoring the advancement of Parkinsons disease.

This [2] study addresses the challenge of quantifying Parkinson's condition from a patient's voice in the Interspeech ComParE 2015 PC Sub-Challenge. The research treats it as a regression task and employs an ensemble learning approach, Random Forest (RF), combined with various types of features. These features include acoustic traits, Automatic Speech Recognition system (ASR) outputs, and non-intrusive intelligibility measures. The system surpasses baseline results with a remarkable 19% improvement in the development set, indicating its effectiveness. This [4] study introduces a method for early Parkinson's disease detection from free-speech, even in uncontrolled background settings. It combines signal and speech processing with machine learning. By utilizing speech databases from different PD stages, the research demonstrates the effectiveness of Random Forest (RF) and Support Vector Machine (SVM) algorithms. When properly tuned, these algorithms offer highly accurate computational tools for PD presence estimation.

II. METHODOLOGY

1. Data Collection and Exploration Data Source:

Data Description: The dataset comprises biomedical voice measurements from individuals, including features like jitter, shimmer, pitch, and other voice-related parameters.

Data Exploration: We utilized the Pandas library to load and inspect the dataset. The initial review included an examination of the first 10 records, dataset dimensions, data types, and the presence of null values. Basic statistics were computed using the describe() function to analyze data distribution. We explored the class distribution of the 'status' variable to understand the balance between healthy and unhealthy individuals and conducted statistical summaries for both classes to identify potential differences.

2. Data Preprocessing

Feature Selection: Relevant features for the predictive model were selected, while the 'name' and 'status' columns were removed. 'Name' serves as an identifier, and 'status' is the target variable.

Data Splitting: The dataset was divided into training and testing sets using an 80-20 split ratio to enable model training and evaluation on separate data.

Standardization: Standardized the feature values to have a mean of 0 and a standard deviation of 1. Standardization is essential when working with machine learning models.

3. Model Development and Hyperparameter Tuning

A. Model Selection: In our investigation, we explored various models like Random Forest, Support Vector Machine (SVM), Multilayer Perceptron (MLP) Neural Network, Gradient Boosting Classifier. Use of these models in this study provides a holistic perspective on machine learning techniques for Parkinson's disease detection. Each model offers unique strengths, from complex pattern recognition in Random Forest and SVM to non-linear modeling with MLP, and enhanced accuracy through the Gradient Boosting Classifier. Utilizing this diverse set of models ensures a robust analysis and a comprehensive evaluation of their performance, contributing to a deeper understanding of their suitability for Parkinson's disease diagnosis.

B. Hyperparameter Tuning: Hyperparameter optimization was conducted for SVM and the Gradient Boosting Classifier using GridSearchCV. This step aimed to determine the best combination of hyperparameters for each model.

4. Model Training and Evaluation

Model Training: Each model was trained on the training dataset using the best hyperparameters identified through GridSearchCV.

Model Evaluation: The performance of each model was rigorously evaluated on both the training and testing datasets. The primary evaluation metric used was accuracy.

5. Prediction for New Data

The methodology demonstrated how to apply the trained SVM model to predict whether an individual is healthy or unhealthy based on a new set of feature values. These new data points were standardized using the same scaling parameters as the training data.

6. Model Performance Visualization

Confusion Matrix: The confusion matrix was computed and visualized to provide insights into the model's ability to correctly classify healthy and unhealthy individuals.

Class Distribution Visualization: The distribution of the 'status' variable was visualized to understand the balance between the two classes in the dataset.

7. Classification Report:

A comprehensive classification report was generated, encompassing metrics such as precision, recall, F1-score, and support for each classes. This report offers a detailed overview of the model's performance, particularly with imbalanced datasets.

This section details the research approach including data collection preprocessing training and estimation of the SVM model the primary aim is to make sure clarity and enable reproducibility for researchers notably its crucial to mention that this methodology was also employed for various other models like random forest multilayer perceptron MLP neural network and gradient boosting classifier the research paper subsequently provides a thorough comparison of these model performances.

III. MODELING AND ANALYSIS

Models which are used is presented as follows:

Random Forest: Random Forest is a robust ensemble learning technique used extensively in classification tasks. It is designed to overcome the limitations of single decision trees, primarily by reducing overfitting and improving predictive accuracy. The model constructs a "forest" of decision trees, each trained on different subsets of the data. The key concept in Random Forest is "bagging," which entails random resampling of the training data, creating diversity among the trees. Random Forest employs decision trees as its base learners. These trees partition data into subsets based on the values of input features. Bagging reduces the risk of overfitting by training each decision tree on a bootstrapped subset of the data, introducing variation. Random Forest quantifies the importance of features by evaluating their impact on reducing the Gini impurity or Mean Squared Error. Highly important features play a more substantial role in classification.

Support Vector Machine (SVM): Support Vector Machine is a versatile machine learning algorithm known for its efficacy in binary classification. It operates by finding the optimal hyperplane that best separates two classes while maximizing the margin between them. This hyperplane is the decision boundary. SVM can also handle non-linear data effectively through kernel functions. SVM searches for the hyperplane that provides the maximum margin between classes. In 2D, this is a line, while in higher dimensions, it becomes a surface. To deal with non-linear data, SVM uses kernel functions to map data into a higher-dimensional space. Common kernels include the Radial Basis Function (RBF) and polynomial kernels. These are the data points closest to the decision boundary. SVM considers these vectors in determining the optimal hyperplane.

$$w \cdot x - b = 0$$

Where: w is the weight vector. This vector is perpendicular to the hyperplane and dictates its orientation.

x represents the input data point. It's also called a feature vector and is composed of various attributes.

b is the bias term, or the intercept of the hyperplane.

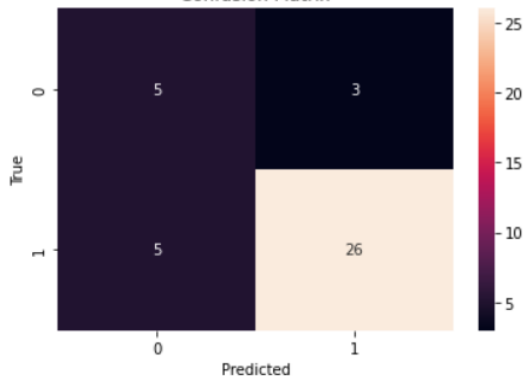
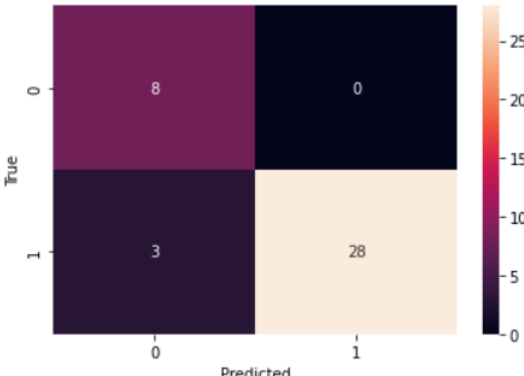
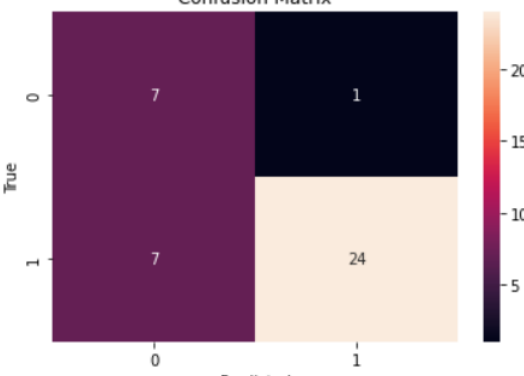
Multilayer Perceptron (MLP) Neural Network: The Multilayer Perceptron (MLP) Neural Network is a fundamental deep learning model. It comprises layers of interconnected neurons, each applying non-linear transformations to input data. MLPs are widely used for tasks like image and speech recognition due to their capacity to model complex relationships in data. Neurons are the fundamental units of MLPs. They process input data by applying weighted sums and biases and then passing the results through activation functions. Activation functions introduce non-linearity into the model, enabling it to learn complex patterns. Common choices include ReLU, Sigmoid, and Tanh. Training an MLP involves backpropagation, an iterative process that

adjusts weights and biases to minimize prediction errors using gradient descent. This process helps the model converge to the optimal state.

Gradient Boosting Classifier: Gradient Boosting is an ensemble learning technique used for classification and regression. It focuses on creating a robust predictive model by sequentially adding weak learners, often decision trees, and correcting the errors made by the preceding model. Gradient Boosting employs a series of weak learners, usually shallow decision trees. Each new model concentrates on correcting the errors of the previous ones. The core concept of boosting is to give more weight to instances that the previous models struggled to predict correctly. This adaptive learning technique progressively improves accuracy. Gradient Boosting minimizes a loss function by iteratively adjusting predictions using gradient descent. The goal is to find the best-fitting model for the data.

IV. RESULTS AND DISCUSSION

Table 1: Confusion matrix of different models

Models	Confusion Matrix									
Random Forest	 <p>Confusion Matrix for Random Forest:</p> <table border="1"> <tr> <td>True \ Predicted</td> <td>0</td> <td>1</td> </tr> <tr> <td>0</td> <td>5</td> <td>3</td> </tr> <tr> <td>1</td> <td>5</td> <td>26</td> </tr> </table>	True \ Predicted	0	1	0	5	3	1	5	26
True \ Predicted	0	1								
0	5	3								
1	5	26								
Support Vector Machine (SVM)	 <p>Confusion Matrix for Support Vector Machine (SVM):</p> <table border="1"> <tr> <td>True \ Predicted</td> <td>0</td> <td>1</td> </tr> <tr> <td>0</td> <td>8</td> <td>0</td> </tr> <tr> <td>1</td> <td>3</td> <td>28</td> </tr> </table>	True \ Predicted	0	1	0	8	0	1	3	28
True \ Predicted	0	1								
0	8	0								
1	3	28								
Multilayer Perceptron (MLP) Neural Network	 <p>Confusion Matrix for Multilayer Perceptron (MLP) Neural Network:</p> <table border="1"> <tr> <td>True \ Predicted</td> <td>0</td> <td>1</td> </tr> <tr> <td>0</td> <td>7</td> <td>1</td> </tr> <tr> <td>1</td> <td>7</td> <td>24</td> </tr> </table>	True \ Predicted	0	1	0	7	1	1	7	24
True \ Predicted	0	1								
0	7	1								
1	7	24								

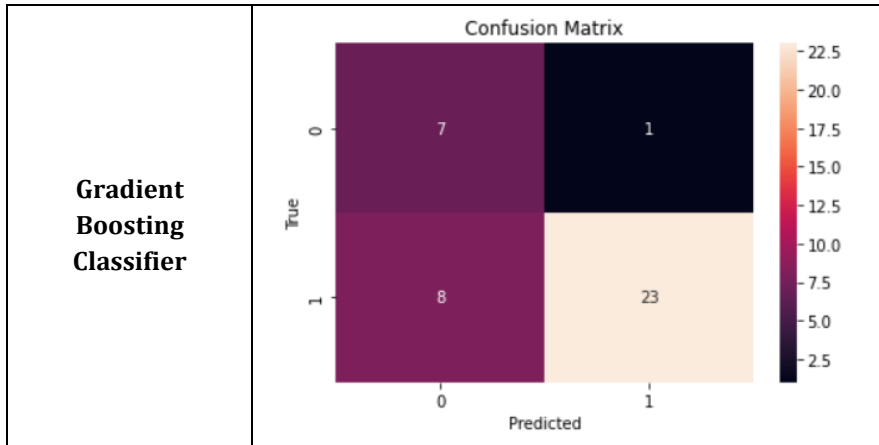
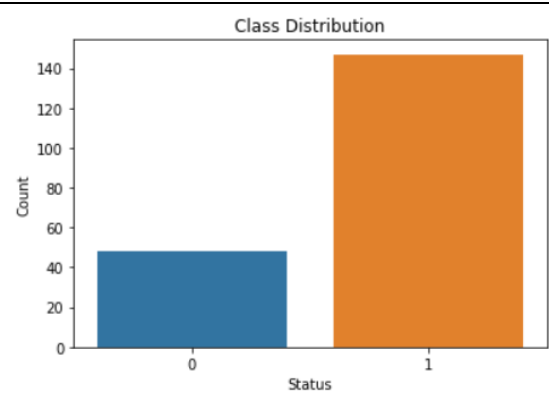
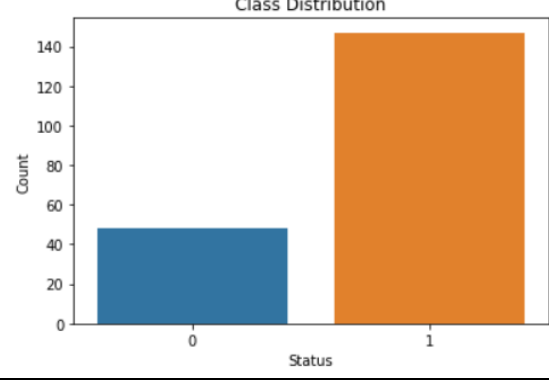
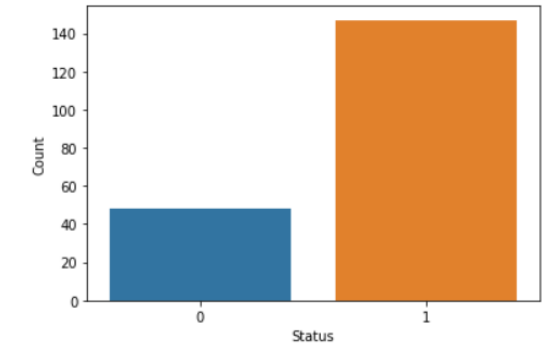


Table 2: Class Distribution of different models

Models	Class Distribution
Random Forest	
Support Vector Machine (SVM)	
Multilayer Perceptron (MLP) Neural Network	

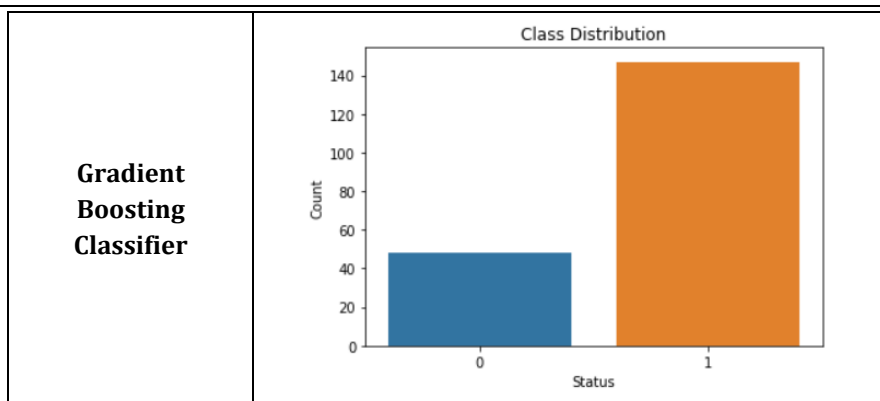


Table 3: Classification Report of different models

Models	Classification Report				
Random Forest		precision	recall	f1-score	support
	0	0.50	0.62	0.56	8
	1	0.90	0.84	0.87	31
	accuracy			0.79	39
	macro avg	0.70	0.73	0.71	39
	weighted avg	0.82	0.79	0.80	39
Support Vector Machine (SVM)		precision	recall	f1-score	support
	0	0.73	1.00	0.84	8
	1	1.00	0.90	0.95	31
	accuracy			0.92	39
	macro avg	0.86	0.95	0.90	39
	weighted avg	0.94	0.92	0.93	39
Multilayer Perceptron (MLP) Neural Network		precision	recall	f1-score	support
	0	0.50	0.88	0.64	8
	1	0.96	0.77	0.86	31
	accuracy			0.79	39
	macro avg	0.73	0.82	0.75	39
	weighted avg	0.87	0.79	0.81	39
Gradient Boosting Classifier		precision	recall	f1-score	support
	0	0.47	0.88	0.61	8
	1	0.96	0.74	0.84	31
	accuracy			0.77	39
	macro avg	0.71	0.81	0.72	39
	weighted avg	0.86	0.77	0.79	39

In Table[3] The precision, recall, F1-score, and support for each classes are summarised in detail. All four models classification report is presented and accuracy of all models present in Table[3].And the confusion matrix is provided in Table[1]. The class distribution of the 'status' variable presented in Table[2]. Findings reveal significant variations in model accuracy with SVM standing out as the most accurate classifier achieving an spectacular accuracy rate of 92 this outcome is consistent with our initial hypothesis which suggested that SVM known for its ability to handle high-dimensional data effectively would perform exceptionally well in the detection of Parkinsons disease the high accuracy achieved by SVM underlines its potential for clinical applications where early detection can remarkably enhance patient outcomes and quality of life random forest with an accuracy rate of 79 also show its potential in the realm of accurate Parkinsons disease detection this altogether learning technique is recognized for its robustness and adaptability to various data types making it a

competitive alternative for diagnostic purposes conversely the MLP neural network and gradient boosting classifier exhibited accuracy rates of 79 and 77 respectively while these models offer promise for disease detection our results suggest the need for further fine-tuning and feature engineering to enhance their classification accuracy potentially to the level attained by SVM and Random forest.

V. CONCLUSION

In summary this study embarked on an extensive exploration in the realm of Parkinsons disease detection utilizing a range of machine learning models such as random forest, SVM, MLP and gradient boosting classifier we achieved varying levels of diagnostic accuracy notably the SVM model emerged as the most accurate achieving an impressive 92 accuracy rate highlighting its potential as a robust diagnostic tool our research underscores the significant role that machine learning plays in the healthcare domain particularly in the early detection of neurodegenerative diseases such as Parkinsons the integration of these models alongside speech analysis showcases a non-invasive and promising approach to diagnosis this work contributes to the existing knowledge in the field by emphasizing the effectiveness of SVMs in detecting Parkinsons disease setting a precedent for further research in this domain however our study is not without limitations particularly concerning the datasets size and diversity future research should explore more extensive and diverse datasets and delve into advanced deep learning methodologies to further enhance diagnostic accuracy in essence the outcomes of this research signify that the synergy between machine learning and speech analysis has the potential to redefine the approach to Parkinsons disease diagnosis by continually pushing the boundaries of machine learning applications we have the opportunity to make early detection more precise and accessible ultimately enhancing the well-being of individuals impacted by this condition.

VI. REFERENCES

- [1] Tsanas, Athanasios, et al. "Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests." *IEEE Transactions on Biomedical Engineering* 63.1 (2016): 164-175
- [2] Random forest-based prediction of Parkinson's disease progression using acoustic, ASR and intelligibility features Author(s): Zlotnik, Alexander; Montero Martínez, Juan Manuel; San Segundo Hernández, Rubén; Gallardo Antolín, Ascensión
- [3] Machine learning based approaches for prediction of Parkinson's disease AK Tiwari - Mach Learn Appl, 2016 - academia.edu
- [4] Automatic detection of Parkinson's disease based on acoustic analysis of speech D Braga, AM Madureira, L Coelho, R Ajith - *Engineering Applications of ...*, 2019 – Elsevier
- [5] Comprehensive Analysis of Machine Learning Techniques for Parkinson's disease Diagnosis D Kumari, R Bhandari - 2023 IEEE 3rd International ..., 2023 - ieeexplore.ieee.org