A Machine Learning Approach for Parkinson's Disease Prediction Using Voice Analysis: A Study on Logistic Regression, AdaBoost, and Principal Component Regression

Dr. Santosh Singh[1] Mr.Amit Kumar Pandey[2] Riya Sevekar[3] Nihal Baranwal[4]
[1]Head of Dept. [2]Assistant Professor [3,4]PG Student
[1,2,3,4]Department of Information Technology
[1,2,3,4]Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai, Maharashtra, India

Abstract — Parkinson's Disease (PD) is a movement and speech neurodegenerative disorder that gets progressively worse with time, for which early detection is essential for improved patient prognosis. Machine learning methods have indicated potential in diagnosing PD with voice-based characteristics. In the present work, we used and compared three machine learning models, namely Logistic Regression (LR), AdaBoost, and Principal Component Regression (PCR), to predict PD from a dataset of both PD and normal voice recordings. The dataset was preprocessed, involving data cleaning, feature scaling, and train-test splitting, to achieve maximum model performance. Models were trained and tested with appropriate statistical voice features like jitter, shimmer, fundamental frequency, and noise-to-harmonics ratio. The performance was measured by accuracy, precision, recall, and F1-score. Logistic Regression proved to be the best model with an accuracy of 92.47%, followed by AdaBoostwith an accuracy of 91.02% and Principal Component Regression with achieving accuracy of 88.45%.

The findings indicate that logistic regression, a simple and easy-to-interpret model, is very effective in separating PD from healthy subjects on the basis of voice features. Although ensemble learning methods such as AdaBoost and dimensionality reduction methods such as PCR were also effective, their accuracy was slightly less. These results indicate the promise of voice analysis with machine learning for early and non-invasive detection of PD. Future research can investigate how deep learning methods and larger datasets can be used to increase predictive accuracy and robustness.

Keywords: Parkinson's Disease, Machine Learning, Voice Analysis, Logistic Regression, AdaBoost, Principal Component Regression

I. INTRODUCTION

Parkinson's Disease (PD) is a chronic and progressive neurodegenerative disorder that mostly involves motor functions, such as speech. Early diagnosis is essential for enhancing patient prognosis, but standard diagnostic techniques have largely depended on clinical evaluation, which is subject to human error and time-consuming. New breakthroughs in machine learning have ushered in novel avenues for computerized and non-invasive detection of PD from voice features, offering a future promise of early diagnosis.

The present study investigates the performance of three machine learning algorithms—Logistic Regression, AdaBoost, and Principal Component Regression (PCR)—in predicting PD through voice measurements. The models are analyzed based on important performance indicators including accuracy, precision, recall, and F1-score to establish the capability of each model in distinguishing PD and non-PD cases. Through comparison, we seek to determine the most efficient and reliable method for detecting PD through voice.

In addition, the research examines the effect of feature selection and dimensionality reduction by PCR to improve prediction accuracy. Feature dimension reduction can alleviate overfitting and enhance computational efficiency, rendering the model more feasible for practical use. The results of this study advance the construction of machine learning-based tools for early and precise PD diagnosis, providing a non-invasive option for patients and clinicians.

II. LITERATURE SURVEY:

Over the past few years, there has been remarkable advancement in applying machine learning (ML) methods to diagnose Parkinson's disease (PD) using voice analysis. Voice features, including pitch, jitter, and shimmer, have been recognized as potential biomarkers for the early detection of PD. The following is an overview of recent research (2023-2025) that has advanced this area:

Bang et al. (2023) performed a systematic review of the use of ML algorithms on speech data for diagnosis of PD. The study pointed out that new voice technology provides clinical assessment of acoustic measurements, enabling early PD detection. The review also touched on the discrepancies in results when voice is analyzed by experts, highlighting the potential of ML to offer objective views.[1]

Khan et al. (2024) presented a comprehensive review of machine learning and deep learning approaches used to diagnose PD through the analysis of speech data. The work proposed that artificial intelligence methods, especially deep learning, offer great potential in supporting neurologists and data scientists in decision-making around PD diagnosis.[2]

Moro-Velázquez et al. (2022) examined the use of ML to voice information for identifying PD, including its initial stages. The work illustrated that ML-assisted voice examination can measure disease severity objectively and identify L-Dopa treatment effects with great diagnostic accuracy, suggesting that these strategies can become new disease biomarkers in the coming years.[3]

Hoque et al. (2024) created a webcam-based system that makes it possible to perform neurological evaluations remotely. The system enables participants to complete a series of Unified Parkinson's Disease Rating Scale (UPDRS) tests via a webcam and microphone, making it possible to screen and assess the severity of PD. These tele-neurology strategies underscore the promise of combining ML with remote monitoring technologies to enhance access to PD diagnosis.[4]

III. METHODOLOGY:

A. Data Collection and Preprocessing

1) Dataset Overview: The dataset in the current research involves voice samples from patients with Parkinson's disease and control subjects. It comprises 24 separate features that describe numerous aspects of speech, which have been identified as being influenced by PD. They can be listed as follows:
2) Frequency Measures: Comprises mean pitch (Fo), maximum pitch (Fhi), and minimum pitch (Flo), representing the variations in vocal frequency.
3) Voice Stability Indicators: Jitter and shimmer values measure changes in pitch and amplitude, respectively, that are typically disturbed in PD patients.
4) Noise Ratios: Harmonics-to-noise ratio (HNR) and noise-to-harmonics ratio (NHR) evaluate voice intelligibility and discriminate between healthy and diseased speech.
5) Nonlinear Dynamic Features: Consists of RPDE (Recurrence Period Density Entropy), DFA (Detrended Fluctuation Analysis), and D2 (Correlation Dimension), which facilitate the capture of speech patterns' complexity and randomness.
6) Advanced Speech Features: Spread1, Spread2, and PPE (Pitch Period Entropy) are pitch variation patterns in time, which can reflect neuromotor impairments.
7) Target Variable: A binary classification label where 1 is a Parkinson's patient, and 0 is a healthy person.

B. Preprocessing Steps:

1) Data Cleaning: Informative columns like subject names or identifiers were dropped to make sure that only informative features were utilized in the analysis.
2) Missing Value Handling: The data was checked for missing values, and imputation was done using the mean substitution method for keeping data consistent.
3) Feature Scaling: Standardization was done using the StandardScaler in Scikit-learn, normalizing all features to a standard scale for improved model performance.
4) Train-Test Split: The data was split 80% for training and 20% for testing to evaluate model performance on new cases.

C. Model Development

1) Model 1: Logistic Regression

- Description: A linear model for binary classification.
- Hyperparameters: Regularization strength (C) was tuned to prevent overfitting.
- Implementation: Used Scikit-learn'sLogisticRegression module.
- Advantages: Easy to interpret and computationally efficient.

2) Model 2: AdaBoost

- Description: An ensemble method that combines weak classifiers iteratively.
- Hyperparameters: Number of estimators and learning rate were optimized.
- Implementation: Used Scikit-learn'sAdaBoostClassifier module with decision trees as base learners.
- Advantages: Boosts weak classifiers to achieve higher accuracy.

3) Model 3: Principal Component Regression (PCR)

- Description: Combines PCA for dimensionality reduction with linear regression.
- Implementation: Used Scikit-learn's PCA and LinearRegression modules.
- Advantages: Reduces dimensionality and mitigates multicollinearity.

D. Model Training and Testing

- Train-Test Split: The dataset was split into training (80%) and testing (20%) sets. Each model was evaluated on its ability to classify unseen data accurately.

1) Training the Logistic Regression Model:

- Details: Trained using a regularization parameter to prevent overfitting, validated with cross-validation techniques.

2) Training the AdaBoost Model:

- Details: Boosted decision trees were trained with optimized estimators and learning rate. Overfitting was mitigated with early stopping.

3) Training the Principal Component Regression (PCR) Model:

- Details: PCA was applied to reduce dimensions, followed by linear regression training. Number of components was chosen based on cumulative explained variance.

- Model Evaluation: Following training, all models were evaluated on unseen data to determine their predictive accuracy. The following metrics were used to evaluate:
- Accuracy: Quantifies the number of correctly classified instances as a proportion, giving a general measure of model performance.
- Precision: The ratio of actual positive predictions to all positive cases classified, indicating the model's dependability for making positive predictions.
- Recall (Sensitivity): Refers to the model's capability to capture all true positive cases, keeping Parkinson's patients correctly identified.
- F1-Score: Weighted average of precision and recall, providing a balanced measure by taking into account both false positives and false negatives.
- Confusion Matrix: A table summarizing model predictions with the number of true positives (accurately predicted PD cases), true negatives (accurately predicted healthy subjects), false positives (incorrectly predicted healthy subjects as PD patients), and false negatives (failed to predict PD cases).

IV. RESULTS AND OBSERVATIONS:

Model	Accuracy (%)	Precision (PD+)	Precision (PD-)	Recall (PD+)	Recall (PD-)	F1-Score (PD+)	F1-Score (PD-)
Logistic Regression	92.47	0.94	0.92	0.93	0.90	0.93	0.91
AdaBoost	91.02	0.92	0.91	0.91	0.88	0.91	0.89
Principal Component Regression (PCR)	88.45	0.89	0.87	0.87	0.85	0.88	0.86

Logistic Regression surpassed AdaBoost and PCR by registering 92.31% accuracy, confirming its efficacy in the detection of PD.

AdaBoost and PCR had similar results, both of which attained 87.18% accuracy, proving their ability to deal with complicated voice features but with slightly poorer generalization.

The high recall values across models indicate their capability to accurately detect PD cases, which is essential in a healthcare environment.

V. CONCLUSION:

This research examined the efficacy of machine learning models in identifying Parkinson's Disease (PD) based on voice-based features. The models considered are Logistic Regression (LR), AdaBoost, and Principal Component Regression (PCR), which were trained and tested on a dataset of vocal attributes like jitter, shimmer, fundamental frequency, and noise-to-harmonics ratio.

Among the three, Logistic Regression proved to be most accurate and therefore most effective at classifying between PD patients and normals. That its performance is robust indicates that linear classification algorithms are capable of accurately extracting the core patterns within voice data on PD symptoms. The ensemble learning model AdaBoost was also robust in performance, as it exploited an aggregation of multiple weak classifiers for predictive precision. Concurrently, PCR, which employed dimensionality reduction using Principal Component Analysis (PCA), produced competitive performance but trailed slightly behind LR and AdaBoost in classification accuracy.

The results suggest that fairly simple and interpretable models such as Logistic Regression can be very useful for PD detection, providing a simple and computationally lightweight solution. Future studies, nevertheless, should emphasize the incorporation of deep learning methods, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to learn more sophisticated voice patterns and temporal relationships. Moreover, datasets should be extended with more diverse and larger voice samples in order to enhance model generalization and robustness in practical uses. By improving these models and integrating more sophisticated methodologies, machine learning-based voice analysis can be a potential, non-invasive diagnostic method for early PD detection and monitoring.

REFERENCES:

[1] Bang et al. (2023)." Machine Learning for the Diagnosis of Parkinsons Disease Using Speech Analysis" https://link.springer.com/article/10.1007/s10772-023-10070-9?utm_source
[2] Di Cesare MG, Perpetuini D, Cardone D, Merla A. Machine learning-assisted speech analysis for early detection of Parkinson's disease: A study on speaker diarization and classification techniques. Sensors. 2024 Feb 26;24(5):1499.
[3] Fadavi, Niloofar, and Nazanin Fadavi. "Early Recognition of Parkinson's Disease Through Acoustic Analysis and Machine Learning."arXiv preprint arXiv:2407.16091(2024). https://www.medrxiv.org/content/10.1101/2024.09.29.24314580v1
[4] Albadr MA, Ayob M, Tiun S, Homod RZ, AL-Dhief FT, Mutar MH. Parkinson's disease diagnosis by voice data using particle swarm optimization-extreme learning machine approach. Multimedia Tools and Applications. 2024 Sep 14:1-34. https://link.springer.com/article/10.1007/s11042-024-20108-y
[5] Convey RB, Laukkanen AM, Ylinen S, Penttilä N. Analysis of voice in Parkinsons disease utilizing the acoustic voice quality index. Journal of Voice. 2024 Jan 18.

