IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151182

Parkinson's Disease Detection Through Voice Records Using Machine Learning

Dr. Sudha Katkuri[1], Dr. Azmath Mubeen[2], Dr. D. Hema Latha[3],

Dr. D. Rama Krishna Reddy[4]

Assistant Professor, Department of Business Management, RBVRR Women's College, Hyderabad[1]

Assistant Professor, Department of Computer Science, Veeranari Chakali Ilamma Women's University, Hyderabad[2]

Assistant Professor, Department of Computer Science, Veeranari Chakali Ilamma Women's University, Hyderabad[3]

Assistant Professor, Department of Mathematics, University College of Science, Osmania University, Hyderabad[4]

Abstract: Parkinson's Disease (PD) is the second most prevalent age-related neurological disorder, causing a range of motor and cognitive symptoms. Diagnosis is challenging because PD symptoms closely resemble those of other conditions, including normal aging and essential tremor. Visible symptoms such as difficulty walking and communicating typically emerge around age 50. Although no cure currently exists, certain medications can alleviate some symptoms. This paper presents a machine learning-based approach for PD detection through voice recordings. Several ML algorithms are evaluated, including XGBoost, Random Forest, Decision Tree, Logistic Regression, and Gradient Boosting. Experiments on a biomedical voice measurements dataset show that Random Forest and XGBoost yield the highest accuracy (93.22%), with the XGBoost classifier selected for detailed analysis. The primary aim is to differentiate healthy individuals from those with PD in a non-invasive, cost-effective manner.

Keywords: Parkinson's Disease; XGBoost; Random Forest; Jitter; Shimmer; Machine Learning; Voice Analysis.

I. INTRODUCTION

Parkinson's Disease (PD) is a progressively deteriorating neurological disorder in which symptoms appear gradually over time. While visible symptoms typically manifest in individuals over the age of 50, approximately one in ten patients displays signs before the age of 40 [1]. PD causes the degeneration of specific neurons in the substantia nigra region of the brain, which are responsible for producing dopamine—a chemical messenger that coordinates bodily movements.

Dopamine deficiency leads to progressive motor and non-motor symptoms. Initially, PD commonly presents as tremor or stiffness on one side of the body, such as the hand or arm. As the disease advances, individuals may develop dementia and severe impairments in speech, cognition, and autonomous functions [2]. The broad symptom overlap between early-stage PD and normal aging makes timely diagnosis particularly difficult.

Voice and speech abnormalities are among the earliest and most reliable indicators of PD. Characteristic vocal changes include tremors affecting voice quality, monotone pitch, slurred or slowed speech, and reduced vocal intensity. Acoustic analysis of voice recordings thus offers a promising, non-invasive avenue for early PD detection.

This paper proposes a machine learning framework that analyzes biomedical voice measurements to classify subjects as healthy or PD-positive. The system evaluates five classifiers—Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and XGBoost—and identifies the most effective model for clinical decision support.

A. Symptoms of Parkinson's Disease

PD symptoms vary across individuals and may initially be mild or unilateral. Primary motor symptoms include: (1) Tremor — rhythmic, involuntary shaking, often beginning in a hand or finger (pill-rolling tremor); (2) Bradykinesia — progressive slowing of movement, making routine tasks laborious; (3) Rigidity — muscle stiffness that can cause pain and restrict range of motion; (4) Postural instability — stooped posture and loss of balance; (5) Loss of automatic movements — diminished blinking, facial expression, or arm swing during walking; (6) Speech changes — soft, monotone, or slurred speech; and (7) Micrographia — notably small handwriting.

B. Aetiology and Risk Factors

The exact cause of PD remains unknown, though genetic and environmental factors are implicated. Specific genetic variants (e.g., LRRK2, SNCA mutations) increase susceptibility, and prolonged exposure to herbicides or pesticides is associated with elevated risk [3]. Pathologically, PD is characterised by the accumulation of Lewy bodies—intracellular



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026
DOI: 10.17148/IJARCCE.2026.151182

aggregates of misfolded alpha-synuclein protein—in dopaminergic neurons. Established risk factors include advancing age (peak onset around 60 years), male sex, family history of PD, and chronic toxin exposure.

C. Research Objectives

This study aims to:

(1) Detect early-stage PD from biomedical voice measurements.

(2) Compare multiple ML classifiers for PD classification accuracy.

(3) Identify the optimal model for non-invasive, cost-effective screening.

(4) Provide a foundation for remote diagnostic and monitoring tools.

II. RELATED WORK

Extensive research has explored automated PD detection using diverse data modalities and computational approaches. Sakar et al. extracted MFCC, Wavelet Transform (WT), and Tunable Q-factor Wavelet Transform (TQWT) features for PD classification. Applying mRMR feature selection with an SVM-RBF classifier on combined TQWT and MFCC features yielded 86% accuracy and an F-measure of 0.84 [4].

Polat combined a Random Forest classifier with the Synthetic Minority Over-Sampling Technique (SMOTE), achieving 94.89% accuracy after minority class oversampling, though the approach risked overfitting [5]. Solana-Lavalle and Rosas-Romero applied Wrapper Feature Selection with an SVM classifier and obtained 94.7% accuracy on a large-scale dataset [6].

Gunduz employed Convolutional Neural Networks (CNN), achieving up to 86.9% accuracy using triple feature sets (WT, TQWT, and baseline) [7]. Nalini et al. compared RNN, LSTM, and MLP architectures for PD voice diagnosis, with LSTM outperforming the others [8]. Hajer et al. applied LDA and PCA with K-means and DBSCAN clustering, with DBSCAN reaching 64% accuracy [9].

Little et al. proposed dysphonia measures incorporating pitch period entropy (PPE) and recurrence/fractal scaling features on a dataset of 23 PD patients and 8 healthy controls, achieving 91.4% accuracy [10]. Moumita et al. used decision-forest and SysFor algorithms with ForestPA features, reaching 94.12% accuracy with minimal decision trees [11]. These studies establish a strong precedent for ML-based voice analysis in PD diagnosis, motivating the comparative evaluation in this work.

III. PROPOSED METHODOLOGY

The proposed system classifies voice recordings as PD-positive or healthy using a pipeline comprising data collection, preprocessing, feature extraction, model training, and evaluation. The XGBoost classifier was selected for final deployment based on its superior performance metrics.

A. Dataset

The study uses the UCI ML Repository Parkinson's Disease dataset, containing 195 biomedical voice recordings from 31 subjects (23 with PD). Each record includes 22 acoustic features capturing fundamental frequency variation, jitter, shimmer, noise ratios, and nonlinear dynamical measures. The target variable (status) indicates PD (1) or healthy (0).

B. Data Preprocessing

Preprocessing steps include: (1) noise reduction to improve audio quality; (2) normalization of feature values to a common scale; and (3) segmentation of recordings into uniform windows. The dataset was split 80:20 for training and testing with stratification to preserve class ratios.

C. Feature Extraction

Key acoustic features used for classification include MDVP:Fo (average fundamental frequency), MDVP:Fhi and MDVP:Flo (maximum and minimum frequencies), MDVP:Jitter and related jitter variants (frequency variability), MDVP:Shimmer variants (amplitude variability), NHR and HNR (noise-to-harmonics and harmonics-to-noise ratios), RPDE and DFA (nonlinear dynamical complexity measures), and PPE (pitch period entropy). These features collectively capture the vocal instability characteristic of PD.

D. System Architecture

The processing pipeline follows six stages: (1) Data Collection — voice samples from PD and healthy subjects with demographic metadata; (2) Data Preprocessing — noise removal, segmentation, and normalization; (3) Feature Extraction — acoustic feature computation from raw audio; (4) Model Training — training five ML classifiers with hyperparameter optimization; (5) Model Evaluation — k-fold cross-validation and performance metric analysis; and (6) Prediction — deployment of the best-performing model for new sample classification.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151182

IV. CLASSIFICATION ALGORITHMS

A. XGBoost (Extreme Gradient Boosting)

XGBoost is an optimized, scalable gradient boosting library that constructs an ensemble of weak learners (typically decision trees) sequentially, where each tree corrects the residual errors of the previous one. It incorporates L1/L2 regularization to prevent overfitting, supports parallel processing for efficiency on large datasets, and handles missing values natively. Key hyperparameters include learning rate, max tree depth, and number of estimators.

B. Random Forest

Random Forest is a bagging-based ensemble method that trains a large number of decision trees on random feature subsets of the training data. Predictions are made by majority vote across all trees, which reduces variance and mitigates overfitting. Its robustness to noise and ability to handle high-dimensional feature spaces make it particularly effective for biomedical datasets.

C. Decision Tree

Decision Trees recursively partition the feature space based on information gain or Gini impurity, producing interpretable tree-structured models where internal nodes represent feature tests and leaves represent class labels. The CART (Classification and Regression Tree) algorithm is used. While prone to overfitting when deep, Decision Trees serve as the base learner in ensemble methods.

D. Logistic Regression

Logistic Regression models the probability of class membership using a sigmoid function applied to a linear combination of features. It provides probabilistic outputs and is well-suited to binary classification tasks. Despite its simplicity, it establishes a strong baseline for comparison and offers high interpretability.

E. Gradient Boosting

Gradient Boosting iteratively fits new weak learners to the negative gradient (pseudo-residuals) of a differentiable loss function. The ensemble is built sequentially, with each model correcting errors from the previous iteration via gradient descent in function space. The base learner is CART, and the method effectively reduces both bias and variance.

V. EXPERIMENTAL RESULTS

All five classifiers were trained and evaluated on the PD voice dataset. Table I summarizes the performance of each classifier across key metrics.

TABLE I. Classifier Performance Comparison

Classifier	Accuracy	Precision (PD)	Recall (PD)	F1 (PD)
Logistic Reg.	84.75%	0.87	0.93	0.90
Decision Tree	86.44%	0.93	0.89	0.91
Random Forest	93.22%	0.93	0.98	0.96
Gradient Boost	91.53%	0.91	0.98	0.95
XGBoost	**93.22%**	**0.92**	**1.00**	**0.96**

Both Random Forest and XGBoost achieved the highest accuracy of 93.22%. The XGBoost classifier demonstrated perfect recall (1.00) for PD-positive cases, meaning no PD patient was misclassified as healthy—a critical property for a medical screening tool. The overall classification report for XGBoost on the test set is: precision 0.94, recall 0.93, weighted F1-score 0.93.

Among the healthy class, XGBoost achieved precision of 1.00 and recall of 0.73, indicating occasional misclassification of healthy subjects as PD-positive (false positives). This trade-off is acceptable in a screening context where false negatives (missed PD diagnoses) are clinically more harmful than false positives.

VI. DISCUSSION

The results confirm that voice-based ML classification is a viable non-invasive approach for PD detection. The superiority of ensemble methods (XGBoost, Random Forest) over linear models (Logistic Regression) reflects the non-



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026
DOI: 10.17148/IJARCCE.2026.151182

linear interactions among acoustic features. The XGBoost classifier's perfect PD recall is particularly valuable for clinical screening, where minimizing false negatives is paramount.

The proposed approach offers several advantages over existing diagnostic methods: it is non-invasive (no injections or radioactive tracers), cost-effective (requires only a voice recording device), remote-compatible (can be administered via smartphone), and rapid (results in seconds vs. multiple clinical visits). Compared to prior work achieving 86–94% accuracy using SVM or CNN architectures, the present XGBoost model achieves competitive performance with substantially lower computational overhead.

Limitations include dataset size (195 recordings) and demographic homogeneity. The model's generalizability to diverse populations, accents, and recording conditions requires further validation. Additionally, the current framework does not incorporate longitudinal monitoring for disease progression tracking.

VII. CONCLUSION

This paper presented a machine learning framework for Parkinson's Disease detection from biomedical voice recordings. Comparative evaluation of five classifiers demonstrated that XGBoost and Random Forest achieve the highest accuracy (93.22%), with XGBoost attaining perfect recall for PD cases. The system is non-invasive, cost-effective, and suitable for remote administration, making it a promising adjunct to conventional PD diagnosis.

Future work will explore deep learning architectures (CNN, LSTM) on larger multi-lingual datasets, integration of multi-modal signals (gait, handwriting), and deployment of a real-time mobile screening application. Combining ensemble ML methods with deep feature extraction may further improve diagnostic accuracy and enable earlier PD detection.

VIII. FUTURE SCOPE

Several avenues exist for extending this work: (1) incorporating additional classifiers such as K-Nearest Neighbor and Support Vector Machine within an ensemble voting scheme; (2) using larger, demographically diverse datasets to improve model generalizability; (3) combining ML with deep learning (e.g., hybrid CNN-XGBoost) for automated feature learning from raw audio; (4) extending the system to monitor disease progression over time; and (5) integrating additional biomarkers such as gait patterns, handwriting, and EEG signals for multi-modal PD assessment.

REFERENCES

[1] National Institute of Neurological Disorders and Stroke, "Parkinson's Disease Information Page," NINDS, 2023.
[2] Mayo Clinic Staff, "Parkinson's disease," Mayo Clinic, 2023.
[3] W. Dauer and S. Przedborski, "Parkinson's disease: mechanisms and models," Neuron, vol. 39, no. 6, pp. 889–909, 2003.
[4] B. E. Sakar et al., "A comparative analysis of speech signal processing algorithms for Parkinson's disease classification," Appl. Soft Comput., vol. 56, pp. 201–208, 2017.
[5] K. Polat, "A hybrid approach to Parkinson's disease classification using speech signals," Expert Syst. Appl., vol. 38, no. 8, pp. 10026–10031, 2011.
[6] G. Solana-Lavalle and R. Rosas-Romero, "Classification of Parkinson's disease from speech recordings with SVM," Biomed. Signal Process. Control, 2021.
[7] H. Gunduz, "Deep learning based Parkinson's disease classification using vocal feature sets," IEEE Access, vol. 7, pp. 115540–115551, 2019.
[8] N. J. Nalini and S. Palanivel, "Music emotion recognition: the combined evidence of MFCC and residual phase," Egyptian Informatics J., vol. 17, no. 1, pp. 1–10, 2016.
[9] H. Aldhyani et al., "Clustering algorithms for Parkinson's disease voice dataset," IEEE Access, 2020.
[10] M. A. Little et al., "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," IEEE Trans. Biomed. Eng., vol. 56, no. 4, pp. 1015–1022, 2009.
[11] M. Moumita et al., "ForestPA: A decision-forest approach for PD diagnosis," IEEE Access, 2020.
[12] R. H. Abiyev and S. Abizade, "Diagnosing Parkinson's diseases using fuzzy neural system," Comput. Math. Methods Med., vol. 2016, 2016.
[13] A. Abos et al., "Differentiation of multiple system atrophy from Parkinson's disease by structural connectivity," Sci. Rep., vol. 9, p. 16488, 2019.
[14] A. Agarwal, S. Chandrayan, and S. S. Sahu, "Prediction of Parkinson's disease using speech signal with Extreme Learning Machine," in Proc. ICEEOT, 2016, pp. 3776–3779.
[15] C. Ahlrichs and M. Lawo, "Parkinson's disease motor symptoms in machine learning: a review," arXiv:1312.3825, 2013.