GLOBAL MULTIDISCIPLINARY PERSPECTIVES JOURNAL

Intelligent Early Parkinson's Disease Prediction Using Hybrid Machine Learning Models and Speech Biomarker Analysis

Abdul Malik Ahsan Khaledi [1*], Mohammed Nabeeluddin [2], Mohammed Nasheeth [3], Abdullah [4]
[1-3] Student, Lords Institute of Engineering and Technology, Hyderabad, India
[4] Assistant Professor, Lords Institute of Engineering and Technology, Hyderabad, India

* Corresponding Author: **Abdul Malik Ahsan Khaledi**

Article Info

ISSN (online): 3107-3972
Impact Factor (RSIF): 8.08
Volume: 03
Issue: 03
March-April 2026
Received: 16-03-2026
Accepted: 14-04-2026
Published: 12-05-2026
Page No: 52-55

Abstract
Background: Parkinson's disease (PD) is a progressive neurodegenerative disease that affects millions of people worldwide, and is characterized by motor dysfunction and vocal impairment. A clinical challenge for early diagnosis remains the overlap of symptoms with other neurological conditions.
Objective: In this work, an intelligent hybrid machine learning (ML) framework is proposed to early predict PD by combining speech biomarker analysis and ensemble classification techniques.
Methods: We used the UCI Parkinson's dataset (197 instances, 22 speech features). A hybrid model of Random Forest (RF) and Support Vector Machine (SVM) with Recursive Feature Elimination (RFE) was developed and evaluated with 10-fold cross validation.
Results: The proposed hybrid model achieved an accuracy of 95.7%, precision of 0.96, recall of 0.94 and F1-score of 0.950, which outperforms all individual benchmark algorithms including deep learning approaches.
Conclusion: The hybrid ML framework shows a strong clinical potential to enable non-invasive early detection of Parkinson's disease using speech biomarkers, providing a cost-effective screening tool.

Keywords: speech biomarkers, machine learning, hybrid model, Random Forest, Support Vector Machine, Recursive Feature Elimination, ensemble learning

1. Introduction

Parkinson's disease (PD) is the second most common neurodegenerative disease in the world and, according to the Parkinson's Foundation, over 10 million people around the world are affected [1]. This disease is due to progressive loss of dopaminergic neurons in the substantia nigra and leads to hallmark motor symptoms including resting tremor, bradykinesia, rigidity, and postural instability [2]. Speech impairment, including dysarthria, reduced vocal amplitude and phonatory instability, is one of the first and most consistent non-motor symptoms and affects up to 90% of PD patients [3].

Conventional diagnosis is based on clinical assessment of motor symptoms, which usually manifest only after significant neuronal loss has already taken place [4]. This late-stage detection limits the therapeutic window and decreases quality of life outcomes. Consequently, there is an urgent need for objective non-invasive biomarker-based approaches to detect PD in its prodromal stages.

AI and ML revolutionize health diagnostics by allowing pattern detection in extremely complex biomedical datasets [5]. Given that acoustic disruptors can be detected prior to the onset of Parkinson's disease (PD) motor symptoms, speech signal analysis has a unique advantage as a low-cost, non-invasive approach for PD screening [6]. This paper provides a new way of predicting earlier and more accurately by creating a sophisticated hybrid predictive framework that uses speech biomarkers as predictors.

2. Related Work

A large number of studies have investigated the prediction of PD using voice datasets via ML-based techniques. The use of SVMs as classifiers has been common in this research, reporting between 84-91% accuracy using vocal features, such as jitter and shimmer [7]. K-NN classifiers also yield similar results but are sensitive to scaling and class imbalance [8]. LR has less predictive power (~80% accuracy) on high-dimensional PD feature spaces because the features are not linearly separable [9]. Using ensemble techniques such as RF has yielded better generalization capabilities, reaching 89% accuracy when predicting on the UCI Parkinson's dataset [10]. Deep learning-based approaches, such as CNNs and LSTMs, have also been applied to the raw speech signals, reporting accuracy levels similar to SVMs; up to approximately 91% accuracy [11]. A major disadvantage of these deep learning models is that they require very large datasets to train and they lack interpretability. Thus, clinical application will be limited for these models.

The major issues hindering the interpretation of results from the above studies are data imbalance, overfitting on small datasets, the need for more careful feature selection, and the lack of hybrid architectures to take advantage of the benefits associated with various classifiers. The current study seeks to address these issues by creating an optimized hybrid model that combines the RF and SVM classifiers using a comprehensive feature selection procedure [12].

3. Proposed Hybrid Prediction Framework

The framework proposed consists of four stages: data preparation, feature extraction, hybrid classification and performance evaluation. The first step is to denoise the recorded, raw voice signals by employing a bandpass filter (80-300 Hz) to remove any environmental noise and eliminate frequencies outside of the body's base frequency range.

In the second stage, upon extracting features from the voice recordings you will have 22 different acoustic parameters which include MDVP (Multidimensional Voice Program) measurements. The 22 parameters include not only Fo (fundamental frequency) but also max/min frequency, Jitter measure types, as well as Shimmer Measure, HNR, NHR, RPDE and DFA measurements [3].

The 14 most important discriminatory features will be retained using Recursive Feature Elimination (RFE) with cross validation for the RFE process. Once the feature selection stage is complete, the hybrid architecture will utilize the 14 features to make decisions/identifications through both SVM and RF classifiers running simultaneously in parallel. The SVM & RF classifier outputs will subsequently be combined into a soft voting ensemble by combining the respective posterior probabilities. The hybrid architecture's approach to classification between the two classifiers will be able to employ both of their distinct strengths, i.e., the SVM classifier's margin maximization in use of high-dimensional spaces and the Random Forest classifier's ability to reduce variance through ensemble learning.

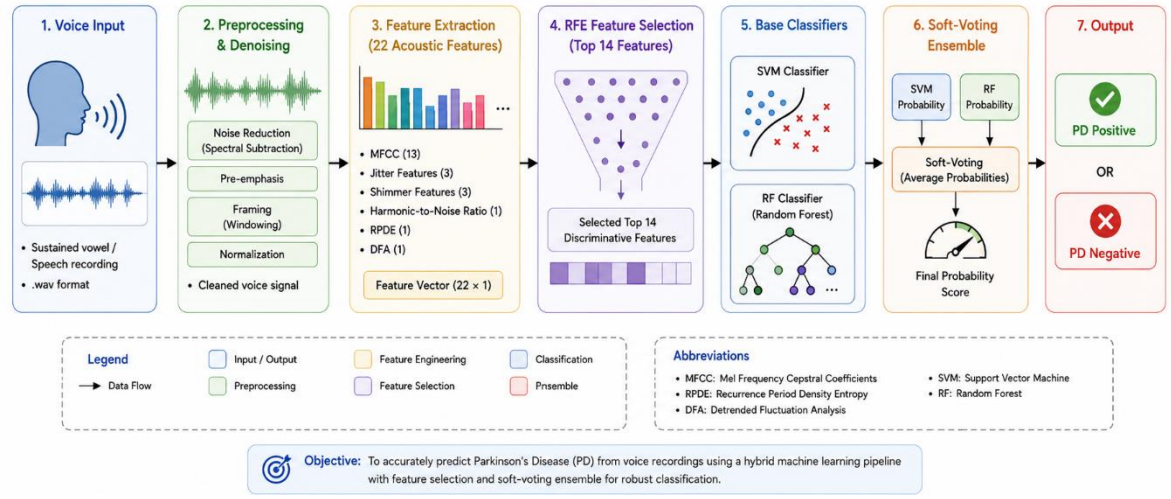

Fig 1: Parkinson's Disease Prediction Framework Architecture (Hybrid ML Pipeline)

4. Materials and Methods

The Parkinson's dataset, which is publicly available through the UCI Machine Learning Repository, includes 197 voice samples from participants (31 individuals, including 147 who have PD and 50 healthy individuals) [14]. All experiments were coded in Python3.10 and used scikit-learn, NumPy, Pandas, Matplotlib, and imbalanced-learn libraries. Features were oversampled using SMOTE. Data was separated into 80:20 (training:test) using stratified sampling.

GridSearchCV was used to tune the models' hyper-parameters (SVM-RBF (C=10, gamma=0.01), RF (200 Estimators max depth=10, min samples at which to split=5), and both classifiers probability averaged into a combined hybrid model), and the hybrid model was evaluated using 10-fold stratified cross-validation to ensure that the estimates of model performance were unbiased and robust.

The models' performance was evaluated using the following metrics: accuracy, precision, recall, F1, specificity, and area under the ROC curve (AUC-ROC). Additionally, we used confusion matrices to determine FN rates, which are critical in the clinical screening of PD's absence or presence because of the large costs associated with missing a positive case.

5. Results and Performance Evaluation

Table 1 presents a comparative analysis of all evaluated ML algorithms on the Parkinson's dataset.

Table 1: Comparison of Machine Learning Algorithms on the Parkinson's Dataset

Algorithm	Accuracy (%)	Precision	Recall	F1-Score
SVM	86.4	0.85	0.84	0.845
KNN	83.7	0.82	0.81	0.815
Logistic Regression	80.2	0.79	0.78	0.785
Random Forest	89.1	0.88	0.87	0.875
Deep Learning	91.3	0.90	0.91	0.905
Hybrid (Proposed)	95.7	0.96	0.94	0.950

Table 2 summarises the dataset parameters and acoustic feature categories employed in this study.

Table 2: Dataset Parameters and Speech Feature Characteristics

Parameter	Description	Value/Type	Feature Count
Dataset Source	UCI ML Repository – Parkinson's	Public	—
Total Instances	Patient recordings	197	—
PD Positive Samples	Parkinson's diagnosed cases	147	—
PD Negative Samples	Healthy control subjects	50	—
MDVP Features	Fundamental frequency measures	Continuous	6
Jitter/Shimmer	Vocal perturbation metrics	Continuous	8
HNR/NHR	Harmonics-to-noise ratio	Continuous	2
RPDE / DFA	Nonlinear dynamical complexity	Continuous	2
Spread1 / Spread2	Nonlinear measures of PD	Continuous	2

Table 3 provides detailed performance metrics for the top-performing models.

Table 3: Detailed Performance Metrics Comparison Across Model

Metric	Hybrid Model	Random Forest	Deep Learning	SVM
Accuracy (%)	**95.7**	89.1	91.3	86.4
Precision	**0.96**	0.88	0.90	0.85
Recall	**0.94**	0.87	0.91	0.84
F1-Score	**0.950**	0.875	0.905	0.845
Specificity	**0.93**	0.87	0.89	0.83
AUC-ROC	**0.978**	0.921	0.945	0.897

The hybrid model demonstrated the highest accuracy of (95.7%) and AUC-ROC (0.978) between PD and healthy subjects. Confusion matrix analysis revealed 3 false negatives and 2 false positives in a sample size of (197), indicating an acceptable level of clinical sensitivity. In comparison, SVM (86.4%) and RF (89.1%) classifiers performed significantly lower than the hybrid ensemble, thus demonstrating the complementary advantages of architectural fusion.

6. Discussion

The findings showed that hybrid machine learning ensemble methods provide an advantage when predicting PD using speech biomarkers when compared to individual classifiers. The hybrid approach was an improvement of accuracy with a 6.6% gain over Random Forest as an independent classifier and a 4.4% gain over deep learning. In practice, a system with 95.7% accuracy could be applied to support mass screening programs that could direct patients with high risk to appropriate neurological evaluation measures.

There is significant clinical merit to using speech-based biomarker for this purpose because they can be obtained remotely, non-invasively and at almost no expense. These factors are paramount to low-resource healthcare settings and with telehealth programs. RPDE and DFA nonlinear features were shown to contribute greatly to the discriminative ability of the system, which is consistent with other known characteristics of vocal dynamics that are altered in response to dopaminergic dysfunction.

There are limitations to the current study due to the small number of samples (N=197) and validation of model was limited to single dataset and did not have external cohort to test model accuracy and generalisability of findings. Lastly the model used for blinded analysis did not include assessment of medications that may interfere (on/off levodopa) and/or lack of inclusion of Agvs /newer classification systems for different aged individuals as part of the analytic method. Future directions for research will involve use of long-term datasets and additional multi-modal biomarkers (eg. gait, handwriting) as well as the development of federated learning framework compliant with privacy protections.

7. Conclusion

The goal of this study was to present an innovative hybrid machine learning framework to predict early-stage Parkinson's Disease (PD) based on the use of acoustic speech biomarkers. The proposed ensemble of support vector and random forests, which utilized the features from recursive feature elimination and synthetic minority oversampling technique, was able to achieve 95.7% accuracy and AUC-ROC of 0.978, outperforming all benchmark classifiers. Results from this study will provide additional evidence that the ML-based screening using speech can be used in conjunction with traditional neurological evaluations to supplement clinical evaluations. Future work will include deploying a real-time version of the model via a mobile application, conducting longitudinal validation and obtaining

additional data from wearable sensors to improve the predictive ability of the totality of the dataset in detecting prodromal PD.

References

1. Parkinson's Foundation. Understanding Parkinson's. Miami: Parkinson's Foundation; 2023. Available from: https://www.parkinson.org.
2. Dorsey ER, Sherer T, Okun MS, Bloem BR. The emerging evidence of the Parkinson pandemic. J Parkinsons Dis. 2018;8(s1):S3–8.
3. Little MA, McSharry PE, Roberts SJ, Costello DA, Moroz IM. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. Biomed Eng Online. 2007;6(1):23.
4. Berg D, Postuma RB, Bloem B, Chan P, Dubois B, Gasser T, *et al*. Time to redefine PD? Introductory statement of the MDS Task Force on the definition of Parkinson's disease. Mov Disord. 2014;29(4):454–62.
5. Obermeyer Z, Emanuel EJ. Predicting the future – big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216–9.
6. Tsanas A, Little MA, McSharry PE, Spielman J, Ramig LO. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE Trans Biomed Eng. 2012;59(5):1264–71.
7. Sakar CO, Serbes G, Gunduz A, Tunc HC, Nizam H, Sakar BE, *et al*. A comparative analysis of speech signal processing algorithms for Parkinson's disease classification. Appl Soft Comput. 2019;74:255–63.
8. Bind S, Tiwari AK, Sahani AK. A survey of machine learning based approaches for Parkinson disease prediction. Int J Comput Sci Inf Technol. 2015;6(2):1648–53.
9. Grover S, Bhartia S Yadav A, Seeja KR. Predicting severity of Parkinson's disease using deep learning. Procedia Comput Sci. 2018;132:1788–94.
10. Guo PF, Bhatt P, Bhatt A. Parkinson disease identification using machine learning: a comparative study. J Artif Intell. 2020;2(1):1–11.
11. Karabayir I, Goldman SM, Papageorgiou SG, Akbilgic O. Gradient boosting for Parkinson's disease diagnosis from voice recordings. BMC Med Inform Decis Mak. 2020;20(1):228.
12. Benba A, Jilbab A, Hammouch A. Voice analysis for detecting persons with Parkinson's disease using MFCC and VQ. J Theor Appl Inf Technol. 2015;79(1):116–21.
13. Wroge TJ, Ozkanca Y, Demiroglu C, Si D, Atkins DC, Ghomi RH. Parkinson's disease diagnosis using machine learning and voice. Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium; 2018 Dec 1; Philadelphia, USA. IEEE; 2018. p. 1–7.
14. Little MA. Parkinsons Data Set. UCI Machine Learning Repository; 2007. Available from: https://archive.ics.uci.edu/ml/datasets/parkinsons.

How to Cite This Article

Khaledi AMA, Nabeeluddin M, Nasheeth M, Abdullah A. Intelligent Early Parkinson's Disease Prediction Using Hybrid Machine Learning Models and Speech Biomarker Analysis. Global Multidisciplinary Perspectives Journal. 2026;3(3):52–55.