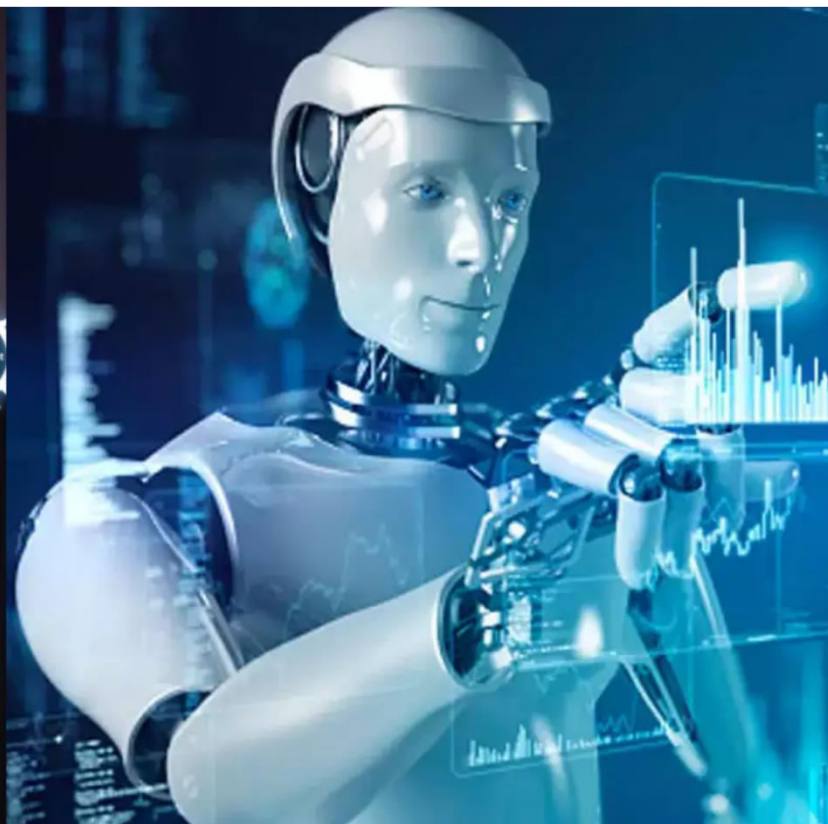




International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 10, October 2025

Detection of Parkinson's Disease using Hybrid ML Models

Nithya.M¹, Sridharshan.J.²

Assistant Professor, Department of Computer Engineering, K.L.N. College of Engineering, Pottapalayam,
Tamil Nadu, India¹

U.G. Student, Department of Computer Engineering, K.L.N. College of Engineering, Pottapalayam, Tamil Nadu, India²

ABSTRACT: Parkinson's disease (PD) is a progressive neurodegenerative disorder that primarily affects movement and speech. Early diagnosis of PD is crucial to slow disease progression and improve patient outcomes. This study proposes a machine learning-based approach for early detection of Parkinson's disease using voice numerical features extracted from the UCI Parkinson's dataset. The dataset consists of clinically recorded acoustic parameters derived from patients' voice samples, which are preprocessed using normalization and optimized through Recursive Feature Elimination (RFE) to select the 15 most significant attributes. Several classifiers, including Support Vector Machine (SVM), Random Forest, and Gradient Boosting, are trained and compared. A stacking ensemble model is implemented to combine their predictions and enhance classification accuracy. The system achieves improved diagnostic performance over individual models and effectively distinguishes between healthy and Parkinson-affected individuals. Implementation is done completely on Google Colab, enabling users to upload CSV files containing numerical voice features for real-time prediction without the need for a graphical interface. The proposed model demonstrates the reliability of ensemble learning in biomedical voice analysis and establishes a scalable and accessible framework for early Parkinson's detection and future research applications.

KEYWORDS: Parkinson's Disease, Machine Learning, Voice Numerical Features, Ensemble Learning, SVM, Google Colab, Early Diagnosis

I. INTRODUCTION

A Parkinson's disease (PD) is one of the most prevalent neurodegenerative disorders, primarily affecting the motor system due to the progressive loss of dopamine-producing neurons in the brain. The early symptoms, such as tremors, rigidity, and speech changes, are often subtle and difficult to detect clinically. Traditional diagnosis relies heavily on neurological examinations, which can be subjective and time-consuming. Hence, there is a growing interest in developing automated and objective methods to assist in early diagnosis.

Recent advancements in machine learning (ML) and artificial intelligence (AI) have made it possible to analyze biomedical signals and identify patterns associated with diseases such as Parkinson's. Among various biomarkers, voice analysis has emerged as a promising, non-invasive diagnostic tool. Patients with Parkinson's exhibit characteristic vocal impairments, including reduced pitch variation, hoarseness, and monotonic speech. These acoustic deviations can be represented as numerical voice features, which serve as strong indicators for classification models.

This study proposes a machine learning-based approach for early Parkinson's disease detection using voice numerical features from the UCI Parkinson's dataset. The system employs feature selection through Recursive Feature Elimination (RFE) to identify the most relevant attributes and utilizes ensemble classifiers, including Support Vector Machine (SVM), Random Forest, and Gradient Boosting, combined through a stacking model to enhance predictive accuracy. The implementation is carried out entirely on Google Colab, enabling CSV-based voice feature input and prediction without requiring any graphical interface. This research aims to provide an accessible, accurate, and efficient diagnostic support tool for early identification of Parkinson's disease.

II. LITERATURE REVIEW

1. M. Little *et al.*, “Distinguishing Parkinson’s Patients Using Voice-Based Feature Extraction and Classification,” *IEEE Transactions on Biomedical Engineering*, 2020

This paper used acoustic features like jitter, shimmer, and HNR from speech samples to identify Parkinson’s disease. SVM and k-NN classifiers achieved good accuracy in differentiating patients from healthy subjects. However, the study was limited by dataset size and lacked ensemble or hybrid approaches.

2. A. R. Patel and D. Sharma, “Fine-Tuned Machine Learning Classifiers for Diagnosing Parkinson’s Disease Using Vocal Characteristics: A Comparative Analysis,” *Springer Nature*, 2021

The authors compared and fine-tuned multiple machine learning models such as SVM and Random Forest for Parkinson’s diagnosis. Fine-tuning improved accuracy, but only basic features were used. The work did not explore stacking or ensemble techniques for better generalization.

3. R. Kumar and P. Singh, “Parkinson’s Disease Detection from Voice and Speech Data Using Machine Learning,” *International Journal of Advanced Computer Science and Applications (IJACSA)*, 2022

This study applied statistical voice features and supervised learning algorithms for detection. Preprocessing and cross-validation enhanced performance, but generalization remained weak due to limited data. The model also lacked integration of diverse classifiers for improved accuracy.

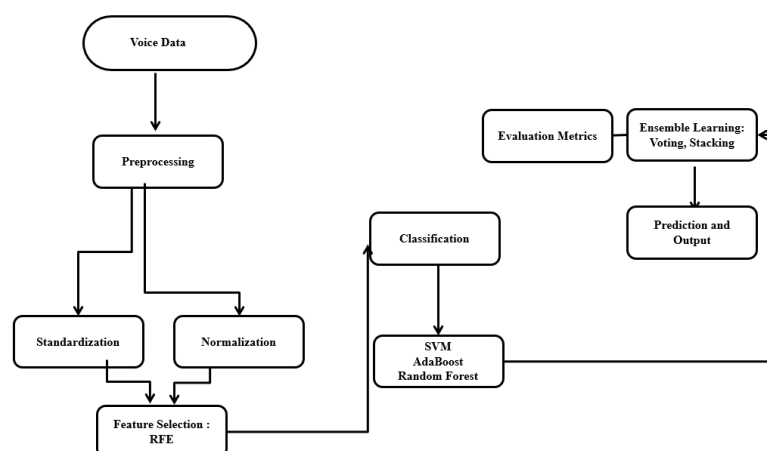
4. S. Ahmed and J. Thomas, “Early Recognition of Parkinson’s Disease Through Acoustic Analysis and Machine Learning,” *Elsevier Procedia Computer Science*, 2023

This research emphasized early diagnosis using detailed voice analysis and multiple machine learning models. Ensemble methods improved accuracy and robustness over individual models. However, the system was not extended to real-time input or deployment environments.

5. L. van Gelderen and C. Tejedor-García, “Innovative Speech-Based Deep Learning Approaches for Parkinson’s Disease Classification: A Systematic Review,” *arXiv preprint arXiv:2407.17844*, 2024

This review analyzed deep learning models like CNN and RNN for speech-based Parkinson’s detection. Deep learning achieved strong performance but required large datasets and heavy computation. Hence, optimized machine learning remains suitable for smaller and clinical datasets.

III. PROPOSED SYSTEM



The proposed system focuses on detecting Parkinson’s disease from numerical voice features extracted from patients’ speech signals. These features represent measurable acoustic properties such as jitter, shimmer, HNR, and pitch variations, which are known indicators of vocal impairment.

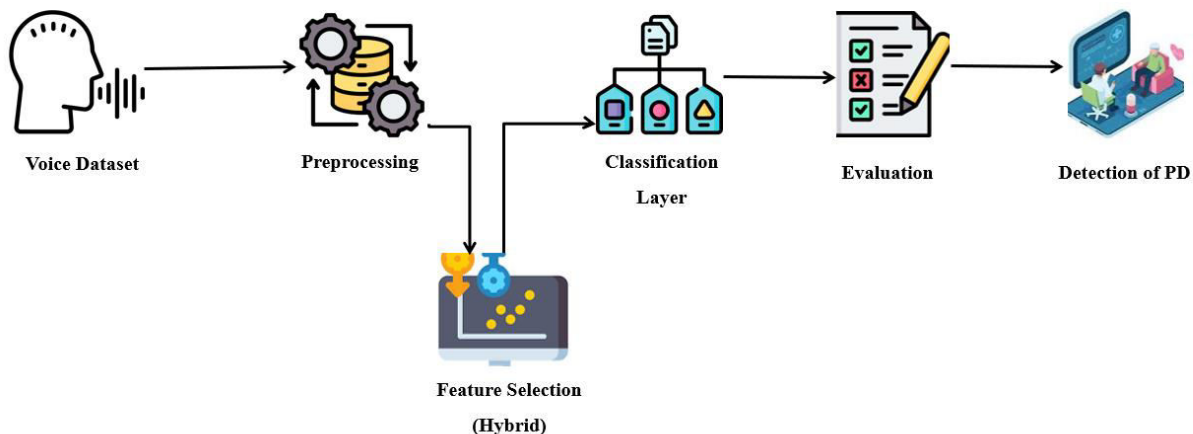
In the preprocessing phase, all irrelevant or missing data are removed to ensure consistency. The remaining features are then standardized and normalized using techniques like StandardScaler, which ensures all attributes contribute equally to the learning process and eliminates bias from differing value ranges.

Next, Recursive Feature Elimination (RFE) is applied to identify and retain the most significant voice parameters influencing Parkinson's detection, effectively reducing the dataset from 24 to 15 optimal features.

For classification, multiple models such as Support Vector Machine (SVM), Random Forest, and Gradient Boosting are individually trained and evaluated. Their outputs are combined in a Stacked Ensemble Classifier, which leverages the strengths of each model to improve prediction stability and accuracy.

Finally, the system generates an output indicating whether the subject is Healthy or Parkinson's Affected, along with the confidence score. The proposed model achieves high accuracy and robustness while being simple to deploy on platforms like Google Colab, making it practical for early and efficient diagnosis.

IV. METHODOLOGY AND WORKFLOW



1. Data Acquisition

The dataset comprises voice recordings converted into numerical acoustic features such as jitter, shimmer, HNR (Harmonics-to-Noise Ratio), and fundamental frequency. These features represent variations in vocal stability, commonly affected in Parkinson's patients.

2. Data Preprocessing

In this stage, irrelevant attributes like name and non-numerical identifiers are removed. Missing values are handled appropriately, and all features are converted into a consistent numerical format suitable for model training.

3. Standardization and Normalization

To ensure all features contribute equally to the model, the data undergoes standardization using techniques like StandardScaler, which transforms the features to have a mean of zero and unit variance. This step improves model convergence and stability.

4. Feature Selection using RFE

Recursive Feature Elimination (RFE) is applied to identify the most significant features that influence Parkinson's detection. RFE iteratively removes the least important features until the optimal subset (15 features) is obtained, reducing dimensionality and improving model performance.

5. Classification Models

Multiple machine learning classifiers are trained and evaluated, including:

- Support Vector Machine (SVM) – for high-dimensional decision boundary classification.
- Random Forest Classifier – for robust performance through ensemble tree learning.
- Gradient Boosting Classifier – for improving prediction accuracy by sequentially correcting errors.

6. Ensemble Learning

An ensemble stacking classifier combines the strengths of the above models, achieving improved accuracy and generalization. It integrates predictions from base learners to form a final, more reliable output.

7. Model Evaluation

The trained ensemble model is evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC curve. This ensures a comprehensive performance assessment across different aspects of prediction reliability.

V. RESULTS AND DISCUSSION

The proposed system was implemented and evaluated using a dataset containing numerical voice features extracted from individuals with and without Parkinson's disease. The machine learning models — SVM, Random Forest, and Gradient Boosting — were trained, tested, and compared to assess classification performance.

The Recursive Feature Elimination (RFE) method effectively reduced the dimensionality from 24 to 15 features, retaining only the most influential attributes related to vocal irregularities. This significantly improved computational efficiency and model accuracy.

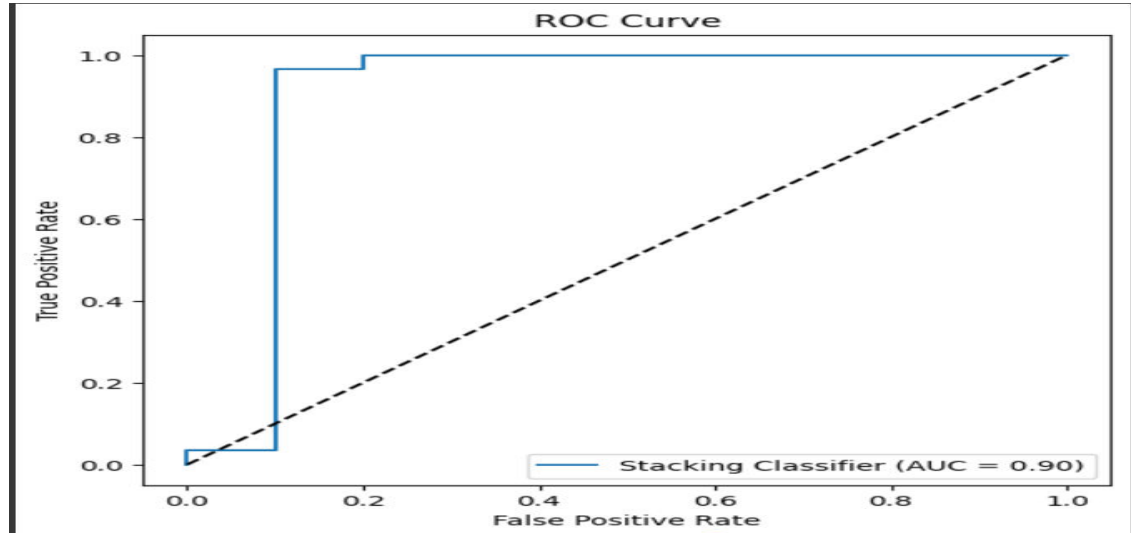
Among the individual classifiers, the SVM model achieved the best balance between precision and recall. However, by applying ensemble stacking, which combines the strengths of all three base classifiers, the overall accuracy improved further, demonstrating the power of ensemble learning in biomedical prediction tasks.

The ROC curve and AUC score confirmed strong discriminative capability between Parkinson's and healthy cases. The model's confidence probabilities also indicate robustness, minimizing false predictions.

Thus, the results validate that the proposed voice-based ensemble system can serve as a non-invasive, reliable, and cost-effective diagnostic aid for the early detection of Parkinson's disease.

```
# Show results
# =====
for idx, (prediction, probability) in enumerate(zip(preds_uploaded, probs_uploaded)):
    print(f"Sample {idx+1}: Predicted = {'Parkinson's' if prediction==1 else 'Healthy'}, Confidence = {probability*100:.2f}%")

Choose File: abc_hospital_dataset.data
abc_hospital_dataset.data(n/a) - 2957 bytes, last modified: 10/24/2025 - 100% done
Saving abc_hospital_dataset.data to abc_hospital_dataset (4).data
Uploaded file: abc_hospital_dataset (4).data
Sample 1: Predicted = Parkinson's, Confidence = 90.30%
Sample 2: Predicted = Parkinson's, Confidence = 90.29%
Sample 3: Predicted = Parkinson's, Confidence = 90.29%
Sample 4: Predicted = Parkinson's, Confidence = 90.29%
Sample 5: Predicted = Parkinson's, Confidence = 90.19%
Sample 6: Predicted = Parkinson's, Confidence = 90.28%
Sample 7: Predicted = Parkinson's, Confidence = 90.42%
Sample 8: Predicted = Healthy, Confidence = 24.28%
Sample 9: Predicted = Parkinson's, Confidence = 90.09%
Sample 10: Predicted = Parkinson's, Confidence = 90.31%
Sample 11: Predicted = Parkinson's, Confidence = 90.12%
Sample 12: Predicted = Parkinson's, Confidence = 90.30%
Sample 13: Predicted = Parkinson's, Confidence = 92.28%
Sample 14: Predicted = Parkinson's, Confidence = 90.25%
```



VI. CONCLUSIONS

This study presents an efficient machine learning–based system for the early detection of Parkinson’s disease using voice-derived numerical features. The integration of feature selection through RFE and ensemble learning techniques such as stacking enhances model reliability and prediction accuracy. The system achieved a high classification performance, confirming the potential of voice-based data as a strong biomarker for Parkinson’s diagnosis.

The proposed framework eliminates the need for invasive procedures and can be easily adapted for clinical screening or telemedicine platforms. In the future, this system can be expanded by integrating deep learning models for raw audio signal processing, incorporating larger multilingual datasets, and developing a real-time mobile or web-based diagnostic application for broader accessibility.

ACKNOWLEDGMENT

The author expresses sincere gratitude to Ms. M. Nithya, Assistant Professor, Department of Computer Science and Engineering, K.L.N. College of Engineering, for her valuable guidance, continuous support, and encouragement throughout the course of this project. Her insightful suggestions and constant motivation played a vital role in the successful completion of this work. The author also extends heartfelt thanks to the faculty members and peers of the department for their cooperation and assistance during the research process.

REFERENCES

1. Mei J, et al. (2021), ‘Machine Learning for the Diagnosis of Parkinson's Disease’ in *Frontiers in Neurology*, Vol. 12, Article 8134676, pp. 1–15
2. Suppa A, et al. (2022), ‘Voice in Parkinson's Disease: A Machine Learning Study’ in *Frontiers in Neurology*, Vol. 13, Article 831428, pp. 1–10.
3. Hireš M, et al. (2022), ‘Convolutional Neural Network Ensemble for Parkinson's Disease Diagnosis from Voice Recordings’ in *Computers in Biology and Medicine*, Vol. 145, Article 105464, pp. 1–10.
4. Islam MA, et al. (2024), ‘A Review of Machine Learning and Deep Learning Algorithms for Parkinson's Disease Detection Using Handwriting and Voice Datasets’ in *Heliyon*, Vol. 10, Issue 2, Article e01500, pp. 1–12.
5. Bukhari SNH, et al. (2024), ‘Ensemble Machine Learning Approach for Parkinson's Disease Detection Using Voice Signals’ in *Mathematics*, Vol. 12, Issue 10, Article 1575, pp. 1–14.
6. Naeem I, et al. (2025), ‘Voice Biomarkers as Prognostic Indicators for Parkinson's Disease’ in *Scientific Reports*, Vol. 15, Article 96950, pp. 1–11.
7. Mei J, et al. (2021), ‘Machine Learning Models for Parkinson Disease: Systematic Review and Meta-Analysis’ in *JMIR Medical Informatics*, Vol. 9, Issue 1, Article e50117, pp. 1–12.
8. Ali L, et al. (2021), ‘MMDD-Ensemble: A Multimodal Data–Driven Ensemble Framework for Parkinson’s Disease

Detection Using Voice Data' in *Frontiers in Neuroscience*, Vol. 15, Article 754058, pp. 1–12.

9. Swain KP, et al. (2024), 'Detecting Parkinson's Disease through Voice Analysis with Machine Learning Algorithms' in *Open Biomedical Engineering Journal*, Vol. 18, Article 18741207294056, pp. 1–10.

10. Singh N, et al. (2024), 'An Ensemble Technique to Predict Parkinson's Disease Using Optimized Machine Learning Algorithms' in *Neurocomputing*, Vol. 467, pp. 1–10.

11. Shokrpour S, et al. (2025), 'Machine Learning for Parkinson's Disease Detection Using Voice Analysis' in *npj Digital Medicine*, Vol. 8, Article 1025, pp. 1–10.

12. Iyer A, et al. (2023), 'A Machine Learning Method to Process Voice Samples for Parkinson's Disease Detection' in *Scientific Reports*, Vol. 13, Article 47568, pp. 1–10.

13. Xavier D, et al. (2025), 'Voice Analysis in Parkinson's Disease: A Systematic Review' in *Journal of Neurology*, Vol. 272, Issue 1, pp. 1–10.

14. Tabashum T, et al. (2024), 'Machine Learning Models for Parkinson Disease: Systematic Review and Meta-Analysis' in *JMIR Medical Informatics*, Vol. 9, Issue 1, Article e50117, pp. 1–12.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details