International Journal of Science and Technology (IJST) | ISSN(O):2049-7318 |2(3):42-46. 2026

Machine Learning-Driven Parkinson's Disease Prediction and Severity Assessment Using Biomedical Voice Features

D. Siva Theja[1], K. Afreen Anjum[2], A. Sathvika[3], J. Bhumika[4], V. Kavya[5], S. Hima Bindu[6]

[1]Assistant Professor, [2,3,4,5,6]UG Students, Aditya College of Engineering, Madanapalle – 517325, Andhra Pradesh, India

Corresponding author: sivatheja.devarinty@gmail.com

Abstract: Parkinson's Disease (PD) is a progressive neurodegenerative disorder that significantly affects speech, movement, and overall quality of life. Early diagnosis of Parkinson's Disease remains a major challenge because traditional clinical methods rely heavily on neurological examinations, expert observations, and expensive imaging techniques. This paper presents an automated Parkinson's Disease detection and recommendation framework using machine learning techniques and biomedical voice features. The proposed system utilizes 22 voice-based attributes such as jitter, shimmer, frequency variation, and harmonic-to-noise ratio to identify patterns associated with Parkinson's Disease. Data preprocessing techniques including normalization and Synthetic Minority Over- sampling Technique (SMOTE) are applied to improve dataset quality and reduce class imbalance. Feature selection is performed using XGBoost feature importance to retain only the most relevant attributes. Multiple machine learning models including K-Nearest Neighbors (KNN), Decision Tree, Naïve Bayes, and XGBoost are trained and evaluated to achieve accurate disease prediction. In addition to detection, the system categorizes the disease into Mild, Moderate, and Severe stages based on confidence scores generated by the model. A recommendation module provides personalized lifestyle, dietary, and wellness suggestions according to the predicted severity level. The complete system is deployed using a Flask-based web application that allows users to receive real-time predictions through an interactive interface. Experimental results demonstrate that the proposed system achieves high accuracy and provides an accessible, low-cost, and non-invasive solution for early Parkinson's Disease diagnosis and management..

Index Terms: Parkinson's Disease, Machine Learning, Voice Signal Analysis, XGBoost, SMOTE, Feature Selection, Flask, Disease Severity Prediction, Recommendation System..

I. INTRODUCTION

Parkinson's Disease (PD) is a chronic neurodegenerative disorder that affects millions of people worldwide. The disease primarily damages dopamine-producing neurons in the brain, leading to symptoms such as tremors, muscle rigidity, slowed movement, and speech impairments. Since Parkinson's Disease progresses gradually over time, early diagnosis is extremely important for improving patient care and slowing disease progression. However, conventional diagnostic methods depend largely on clinical expertise, neurological examinations, and costly imaging technologies such as MRI and PET scans. These methods are often time-consuming, subjective, and inaccessible to individuals living in rural or underdeveloped regions.

One of the earliest indicators of Parkinson's Disease is a change in speech characteristics [3]. Patients often experience reduced pitch variation, breathiness, monotone voice, and articulation difficulties even before visible motor symptoms appear. These speech abnormalities provide an effective non-invasive biomarker for early disease detection. Recent advancements in Artificial Intelligence and Machine Learning have created opportunities to analyze these subtle voice patterns with high precision.

Machine learning algorithms can process large volumes of biomedical voice data and identify hidden correlations that are difficult for humans to detect [4][5]. By leveraging voice-based biomedical features, machine learning systems can provide faster, more accurate, and cost-effective diagnosis compared to traditional approaches.

The proposed system introduces an automated Parkinson's Disease detection and recommendation framework that combines machine learning classification, feature selection, severity prediction, and personalized healthcare recommendations. The framework processes 22 biomedical voice features, applies preprocessing and balancing techniques, and trains multiple machine learning models for accurate prediction. In addition, the system classifies the disease into Mild, Moderate, and Severe stages and generates customized lifestyle recommendations based on the predicted severity level.

The remainder of this paper is organized as follows. Section II presents the literature review and identifies existing research gaps. Section III explains the proposed methodology and system architecture. Section IV describes implementation details and technologies used. Section V discusses experimental results and evaluation metrics. Finally, Section VI concludes the work and outlines future enhancements.

II. LITERATURE REVIEW

Machine learning-based approaches for Parkinson's Disease detection have gained significant attention in recent years due to their ability to analyze biomedical voice features and identify hidden disease patterns [1][2]. Researchers have explored multiple machine learning models, feature extraction techniques, and ensemble approaches to improve diagnostic accuracy.

Niloofar Fadavi et al. proposed a Parkinson's Disease identification framework using acoustic analysis and machine learning models such as Support Vector Machine (SVM) and Gradient Boosting. Their system achieved an accuracy of 94.87% using biomedical voice measurements. However, the approach faced limitations in handling sequential data and lacked real-time deployment capabilities.

Saravanan Srinivasan et al. introduced a multiclass machine learning method for Parkinson’s Disease identification using Recursive Feature Elimination and feature importance techniques. Models including KNN, KSVM, and Feedforward Neural Networks were evaluated on the UCI Parkinson’s dataset. Although the study achieved good classification performance, it mainly focused on voice-based symptoms and ignored other contextual information.

Mahmoud Essam et al. proposed an adaptive machine learning framework using clustering-based speech feature selection for Parkinson’s diagnosis. Gradient Boosting and AdaBoost models produced high prediction accuracy. The study highlighted the effectiveness of selecting meaningful speech features before classification.

Cristian David Rios-Urrego et al. explored speech- based techniques for distinguishing Parkinson’s Disease from Essential Tremor using Gaussian Mixture Models and SVM. Their system achieved strong classification performance but suffered from limited multilingual datasets.

Syed Nisar Hussain Bukhari et al. applied ensemble machine learning methods combined with preprocessing techniques such as SMOTE and Principal Component Analysis. Their approach achieved 96% accuracy and demonstrated that ensemble learning significantly improves prediction robustness.

Kaushik Sekaran et al. developed a deep learning framework using Self-Operational Convolutional Neural Networks for Parkinson’s Disease classification. The model produced strong accuracy and sensitivity values but required large datasets and high computational resources.

Despite these advancements, many existing systems suffer from several limitations. Most approaches only determine whether Parkinson’s Disease is present and do not provide severity estimation or healthcare recommendations. Additionally, many studies rely on small and imbalanced datasets, reducing generalization capability across diverse populations. Existing systems also lack user-friendly deployment frameworks that support real-time interaction and accessibility.

The proposed system addresses these limitations by integrating feature selection, dataset balancing using SMOTE, ensemble-based machine learning classification, severity-level prediction, and personalized recommendation generation into a single Flask-based web application [6][7].

III. PROPOSED METHODOLOGY

A. System Architecture

The proposed Parkinson’s Disease Detection and Recommendation System follows a modular machine learning architecture consisting of five major stages: data collection, preprocessing, feature selection, machine learning classification, and recommendation generation.

Initially, biomedical voice features are collected from the Parkinson’s dataset. The raw data undergoes preprocessing operations including cleaning, normalization, and balancing. Feature selection techniques are then applied to identify the most relevant attributes contributing to accurate prediction. The refined feature set is passed into multiple machine learning classifiers such as KNN, Decision Tree, Naïve Bayes, and XGBoost. The best-performing model is selected for final deployment.

The prediction module determines whether the patient is affected by Parkinson’s Disease. If the disease is detected, the severity estimation module categorizes the condition into Mild, Moderate, or Severe stages based on confidence scores. Finally, a recommendation engine generates personalized suggestions related to diet, exercise, therapy, and lifestyle management.

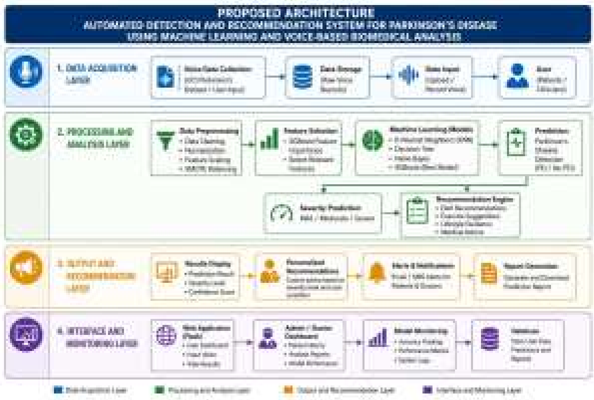

Fig. 1. Proposed Architecture of the Automated Detection and Recommendation System for Parkinson’s Disease Using Machine Learning and Voice-Based Biomedical Analysis

B. Data Collection and Preprocessing

The system uses a Parkinson’s Disease voice dataset containing 22 biomedical voice features. These features include measures related to frequency variation, jitter, shimmer, harmonic-to-noise ratio, and vocal stability.

Data preprocessing is an important stage because machine learning models require clean and normalized input for effective performance. Missing values and inconsistencies are removed from the dataset. Standard Scaler normalization is applied to ensure that all features are transformed into a common scale.

Since the dataset contains class imbalance between healthy individuals and Parkinson’s patients, the Synthetic Minority Over-sampling Technique (SMOTE) is used to generate synthetic minority samples and balance the dataset. This helps improve model generalization and reduces prediction bias.

C. Feature Selection

Feature selection is performed using XGBoost feature importance analysis [8]. The objective of feature selection is to eliminate irrelevant or redundant features while retaining the most influential attributes.

The XGBoost algorithm assigns importance scores to each voice feature based on its contribution to classification performance. Features with higher importance values are selected for training. This process improves model accuracy, reduces overfitting, and decreases computational complexity.

D. Machine Learning Models

Multiple machine learning algorithms are trained and evaluated to determine the best-performing classifier for Parkinson's Disease prediction.

1) K-Nearest Neighbors (KNN)

KNN classifies data samples based on the similarity of neighboring points. It is simple, effective, and suitable for biomedical classification tasks.

2) Decision Tree

Decision Tree models classify data using hierarchical decision rules. They provide interpretability and easy visualization of prediction logic.

3) Naïve Bayes

Naïve Bayes applies probabilistic classification based on Bayes' theorem. It performs efficiently on high-dimensional datasets.

4) XGBoost

XGBoost is an advanced ensemble learning algorithm that combines multiple weak learners to create a highly accurate prediction model [9]. It offers fast execution, feature importance analysis, and strong classification performance.

The models are trained using balanced and normalized datasets. Performance metrics including accuracy, precision, recall, and F1-score are used for evaluation.

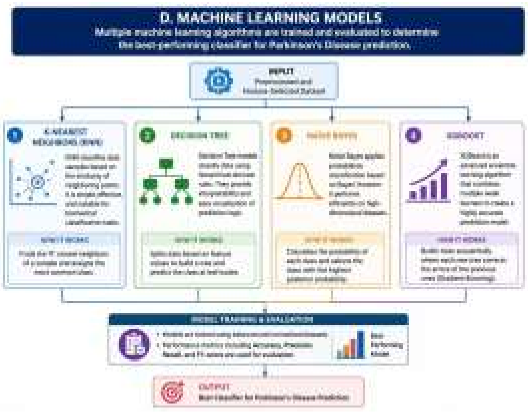

Fig. 2. Evaluation pipeline of ML Algorithms for Parkinson's Disease Detection.

E. Severity Prediction and Recommendation System

The proposed system not only predicts Parkinson's Disease but also estimates its severity level. Severity classification is based on prediction confidence scores generated by the trained model.

Mild Stage: Indicates early symptoms with lower confidence scores.

Moderate Stage: Represents intermediate disease progression.

Severe Stage: Indicates advanced disease conditions with higher confidence scores.

Based on the predicted severity level, the recommendation engine generates personalized healthcare suggestions.

Mild Stage Recommendations

- High-fiber diet and hydration.
- Regular aerobic exercises.
- Yoga and balance training.
- Omega-3 rich foods.

- **Moderate Stage Recommendations**
- Soft and nutritious meals.
- Speech therapy and occupational therapy.
- Calcium and Vitamin D supplementation.
- Strength training exercises.

- **Severe Stage Recommendations**
- Small frequent meals.
- Passive stretching exercises.
- Assisted movement therapy.
- Increased fluid intake and continuous medical monitoring.

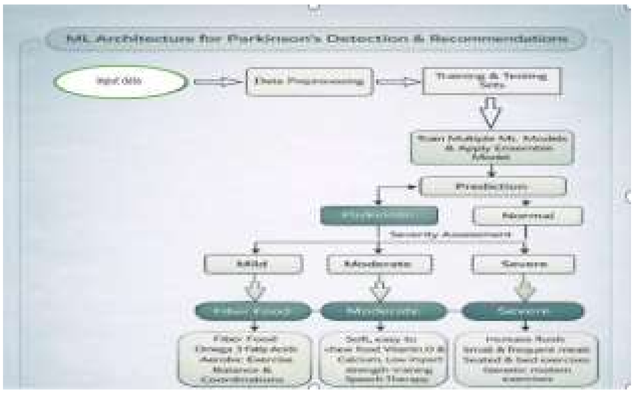

Fig. 3. Hybrid ensemble workflow combining XGBoost (supervised) and Parkinkon's detection and recommendation system.

IV. IMPLEMENTATION

A. Technology Stack

Table II — technology stack of the machine learning driven parkinson's disease prediction and severity assessment model

Component	Technology
Programming Language	Python 3.10
Machine Learning Libraries	Scikit-learn, XGBoost
Data Processing	NumPy, Pandas
Dataset Balancing	SMOTE

Component	Technology
Visualization	Matplotlib
Model Serialization	Pickle
Web Framework	Flask
Development IDE	Visual Studio Code
Operating System	Windows 11

B. Training Dataset

The dataset is divided into training and testing subsets using train-test split techniques. The training set is used to train machine learning models, while the testing set evaluates performance.

After preprocessing and balancing, feature selection is applied using XGBoost. The selected features are passed into multiple classifiers. Hyperparameter tuning and threshold optimization are used to improve prediction performance.

The trained model, scaler, and feature selector are serialized using Pickle and integrated into the Flask web application.

C. ***Web Application Deployment***

The deployment framework is developed using Flask. The web interface allows users to enter biomedical voice features through an HTML form.

The backend processes the input data, performs scaling and feature selection, and generates predictions using the trained model. The output includes:

- Disease Prediction (Positive/Negative)
- Severity Classification
- Confidence Score
- Personalized Recommendations

The web application provides real-time prediction results with low latency and user-friendly interaction.

V. RESULTS AND DISCUSSION

A. Performance Evaluation

The machine learning models are evaluated using standard classification metrics including:

- Accuracy
- Precision
- Recall
- F1-Score
- Confusion Matrix
- ROC Curve

Among all evaluated models, XGBoost achieved the highest classification accuracy due to its ensemble Learning capability and effective feature selection mechanism.

TABLE III — Model Accuracy Results

Model	Accuracy
KNN	89%
Decision Tree	91%
Naive Bayes	88%
XGBoost	96%

The use of SMOTE improved model balance and reduced prediction bias toward majority classes. Feature selection using XGBoost also enhanced classification accuracy by eliminating irrelevant features.

B. Severity Classification Results

The severity prediction module successfully categorized patients into Mild, Moderate, and Severe stages using confidence thresholds.

TABLE IV — Confidence Score vs Severity Level

Confidence Score	Severity Level
0.40-0.60	Mild
0.60-0.80	Moderate
0.80 Above	Severe

The recommendation system generated healthcare suggestions tailored to each severity level, helping patients and caregivers understand appropriate management strategies.

C. Discussion

The proposed framework demonstrates that voice-based biomedical analysis combined with machine learning can provide an effective solution for Parkinson's Disease detection [10][11]. Compared to traditional diagnostic approaches, the system offers several advantages:

- Non-invasive diagnosis using voice analysis.
- Faster and automated prediction.
- Real-time web-based accessibility.
- Severity estimation for disease progression monitoring.
- Personalized healthcare recommendations.

However, certain limitations remain. The dataset size is relatively small and mainly focused on voice biomarkers. Real-world deployment would require larger and more diverse datasets across different languages and demographic groups. Future systems may integrate deep learning, multimodal datasets, and Explainable AI techniques to improve robustness and interpretability.

VI. CONCLUSION AND FUTURE WORK

The proposed approach successfully identifies Parkinson's Disease using voice features and categorizes the disease into Mild, Moderate, and Severe stages. By integrating recommendation generation, the framework extends beyond diagnosis and provides practical healthcare guidance for patients and caregivers.

The deployment of the system through a Flask-based web application ensures accessibility, low cost, and real-time interaction. Experimental evaluation demonstrates that XGBoost-based classification combined with SMOTE balancing achieves high prediction accuracy and strong generalization performance.

Future work will focus on the following enhancements:

- Integration of deep learning techniques such as CNNs and LSTMs.
- Use of larger multilingual and multi-regional datasets.
- Integration of Explainable AI methods such as SHAP and LIME.
- Real-time voice recording support using microphones.
- Cloud deployment for scalable healthcare access.
- Integration with wearable devices and telemedicine platforms.
- Inclusion of motor symptom analysis and multimodal healthcare data.

REFERENCES

1. N. Fadavi et al., “Early Parkinson’s Disease Identification Using Machine Learning and Acoustic Analysis,” 2024.
2. S. Srinivasan et al., “Parkinson Disease Identification Using a Multiclass Machine Learning Method,” 2024.
3. M. Essam et al., “Adaptive Machine Learning Method for Diagnosing Parkinson’s Disease Using Speech Feature Selection,” 2024.
4. Altham et al., “Machine Learning for Detection and Diagnosis of Cognitive Impairment in Parkinson’s Disease,” 2024.
5. T. Tabashum et al., “Comprehensive Analysis of Machine Learning Models for Parkinson’s Disease,” 2024.
6. Q. Du et al., “Machine Learning-Based Local Classification on Class Boundaries for Parkinson’s Disease Detection,” 2024.
7. C. D. Rios-Urrego et al., “Speech-Based Evaluation for Distinguishing Parkinson’s Disease and Essential Tremor,” 2024.
8. M. G. Di Cesare et al., “Machine Learning-Assisted Speech Analysis for Early Parkinson’s Disease Detection,” 2024.
9. V. Ramya et al., “Healthy Food and Exercise Recommendation Application Using Machine Learning,” 2024.
10. H. Al-Nefaie et al., “Machine Learning Algorithms for Parkinson’s Disease Identification Using Voice Features,” 2024.
11. H. Weresh et al., “Variations and Patterns in Parkinson’s Disease Mortality Rates Using Machine Learning Predictions,” 2025.
12. Majda-Zdancewicz et al., “Analysis of Speech Signals in Parkinson’s Disease Patients,” 2024.
13. S. N. H. Bukhari et al., “Speech Signals and Ensemble Machine Learning for Parkinson’s Disease Detection,” 2024.
14. K. Sekaran et al., “Self-Operational Convolutional Neural Networks for Parkinson’s Disease Classification,” 2024.
15. K. Shyamala et al., “Efficient Prediction Model for Early Parkinson’s Disease Diagnosis Using IFRX Model,” 2024.