IT Journal Research and Development (ITJRD)
Vol.10, No.1, August 2025, E-ISSN : 2528-4053 | P-ISSN : 2528-4061
DOI : 10.25299/itjrd.2025.22310

Improving Parkinson's Diagnosis: A Stacked Ensemble Learning with Vocal Biomarkers and Data Balancing

Oluwasegun Abiodun Abioye[1], Abraham Evwiekpaefe[2], Philip Odion[3] ,Olalekan Joel Awujoola[4]
Directorate of Information and Communication Technology, Nigerian Defence Academy, Kaduna, Nigeria[1,4]
Department of Computer Science, Nigerian Defence Academy, Kaduna, Nigeria[2,3]
segunabioye@nda.edu.ng[1], aeevwiekpaefe@nda.edu.ng[2], poodion@nda.edu.ng [3], ojawujoola@nda.edu.ng[4]

Article Info

Article history:

Received May 7, 2025
Revised May 21, 2025
Accepted Aug 8, 2025

Keyword:

Parkinson's Disease
Ensemble Learning
Vocal Biomarkers
Stacking Classifier
Machine Learning

ABSTRACT

Parkinson's disease (PD) presents diagnostic challenges due to subtle early symptoms and overlap with other movement disorders. This study proposes a stacked ensemble learning approach for early PD detection using vocal biomarkers. A Kaggle dataset with 195 voice recordings from 31 individuals (23 with PD) was used to train a model combining CatBoostClassifier and RandomForestClassifier as base learners, with Logistic Regression as the meta-learner. Class imbalance was addressed using RandomOverSampler, and 20-fold stratified cross-validation ensured robust performance evaluation. Key vocal features such as jitter, shimmer, pitch period entropy (PPE), spread1, and spread2 were extracted to distinguish PD patients from healthy controls. The model achieved 100% classification accuracy, a perfect ROC AUC of 1.00, and a low average Brier Score of 0.0071, reflecting excellent probability calibration. SHAP analysis identified spread2, spread1, and PPE as the most influential features, reinforcing pitch instability as a key PD biomarker. The classifier produced no false positives and few false negatives, indicating high reliability. To evaluate generalizability, the model was tested on the Parkinson's Disease Smartwatch (PADS) dataset, which includes 469 participants. It maintained strong performance, supporting its potential as a non-invasive, voice-based screening tool for early PD diagnosis, particularly in telemedicine.

Corresponding Author:

Oluwasegun Abiodun Abioye
Directorate of Information and Communication Technology
Nigerian Defence Academy
Afaka, Kaduna, Nigeria
Email: segunabioye@nda.edu.ng

1. INTRODUCTION

Parkinson's disease is the second most prevalent neurodegenerative disorder after Alzheimer's, affecting about 1% of individuals over the age of 60 globally. Its incidence rises with age, with the average onset occurring around 60 years. Men are more likely than women to develop Parkinson's, with an estimated male-to-female ratio of 1.5 to 1 [1]. It is a condition marked by the

gradual decline of dopamine-producing neurons in the substantia nigra, leading to a variety of motor and non-motor symptoms, it arises from the death of neurons, leading to reduced dopamine levels in the brain. This dopamine deficiency disrupts communication between synapses, resulting in impaired motor function [2]. In fact, Parkinson's disease (PD) currently affects more than 10 million people globally, highlighting its growing impact on public health. Early and accurate detection of PD plays a crucial role in improving patient outcomes. When diagnosed in its initial stages, timely intervention can enable the prompt initiation of treatments, which may significantly ease symptoms, slow disease progression, and enhance overall quality of life. Early detection also allows for better planning and personalized care, which are essential for managing the complex and evolving nature of the disease [3]. Prodromal indicators of Parkinson's disease (PD) may manifest prior to the emergence of its hallmark motor symptoms. These early manifestations can include hypomimia (reduced facial expressivity), vocal alterations, and gait disturbances characterized by diminished arm swing. Nevertheless, such subtle clinical features are frequently misattributed to normal aging processes, potentially contributing to delayed diagnosis and intervention in PD [4].

Dysphonia and dysarthria are common vocal disturbances in Parkinson's disease (PD), resulting from impaired motor control. Due to the complex nature of PD, no definitive cure currently exists. However, early diagnosis using vocal features, combined with artificial intelligence, offers a reliable, accurate, and non-invasive method for detecting and monitoring the disease [5]. Among the emerging methods for Parkinson's disease detection, vocal feature analysis represents a promising and effective approach [6]. Audio signal analysis has emerged as a critical tool in the early detection of Parkinson's disease (PD). While vocal irregularities in the prodromal stages are often imperceptible to human listeners, they can be effectively captured through detailed acoustic feature analysis. Owing to the non-invasive and low-cost nature of voice assessment, this modality fits well within the framework of telemedicine. It enables remote and scalable screening, where individuals can self-record voice samples using standard mobile devices. This approach holds significant potential for early diagnosis, longitudinal monitoring, and large-scale epidemiological studies without the constraints of traditional clinical settings [7].

Clinicians have traditionally relied on the Unified Parkinson's Disease Rating Scale (UPDRS) to monitor disease progression and assess the efficacy of therapeutic interventions, including pharmacological treatments and surgical procedures. This standardized tool provides a comprehensive evaluation of both motor and non-motor symptoms, enabling objective tracking of patient outcomes over time. Following an early diagnosis of Parkinson's disease (PD), doctors may be able to slow its progression through advanced medical interventions. These include deep brain stimulation, a surgical procedure that delivers electrical impulses to specific areas of the brain to regulate abnormal activity, and pharmacological or therapeutic treatments aimed at stimulating the brain's dopamine-producing neurons. By enhancing dopamine activity, these approaches can help manage symptoms more effectively and potentially delay further neurodegeneration associated with PD [8].

Traditionally, the diagnosis of Parkinson's disease (PD) has primarily depended on a detailed evaluation of a patient's medical history, with particular emphasis on the observation and assessment of characteristic signs and symptoms such as tremors, muscle rigidity, and slowed movement. While this clinical approach remains important, advancements in medical technology have significantly enhanced diagnostic capabilities. Today, physicians can complement traditional methods with modern neuroimaging techniques, including magnetic resonance imaging (MRI) scans, which provide detailed images of brain structures. Additionally, diagnostic tools such as electroencephalograms (EEG) to measure electrical brain activity, speech analysis to detect vocal impairments often associated with PD, and electromyography (EMG) to assess muscle and nerve function, offer more objective and precise insights into the presence and progression of the disease. These innovations contribute to earlier and more accurate diagnoses, enabling more timely and targeted treatment strategies [5].

Machine learning (ML) has emerged as a powerful asset in the field of medical diagnostics. With its ability to analyze intricate patterns and uncover hidden correlations within large and complex datasets, ML offers a promising avenue for improving diagnostic accuracy. In the context of Parkinson's disease (PD), early detection can be significantly enhanced by incorporating subtle behavioral changes and less apparent indicators such as changes in speech patterns or abnormalities in gait into the diagnostic process. By leveraging these nuanced data points, ML models can assist clinicians in identifying PD at earlier stages, potentially leading to more effective intervention and better patient outcomes [9]. Researchers have employed various machine learning techniques to distinguish individuals with Parkinson's disease from healthy controls by analyzing handwriting, voice, and speech samples [10].

Currently, there are no clinically validated biomarkers that can reliably detect PD at its earliest stages, which presents a significant challenge in timely diagnosis and intervention. As a result, there is a growing need to integrate artificial intelligence (AI) techniques into the healthcare system to enhance diagnostic precision and speed. The application of AI in analyzing subtle and non-invasive indicators, such as voice data, could play a transformative role in supporting clinicians and improving outcomes for individuals at risk of developing Parkinson's disease. This study is grounded in the hypothesis that vocal characteristics, when combined with ensemble learning and data balancing techniques, have the potential to serve as powerful predictive biomarkers for the early detection of Parkinson's disease (PD). By systematically analyzing a broad spectrum of voice features such as including jitter, shimmer, fundamental frequency parameters, Recurrence Period Density Entropy (RPDE), Noise-to-Harmonics Ratio (NHR), Pitch Period Entropy (PPE), harmonic-to-noise measures, and Detrended Fluctuation Analysis (DFA); the research aims to construct a robust predictive framework capable of identifying early signs of PD with high accuracy[11].

This study presents a robust ensemble learning pipeline using a stacked model with CatBoostClassifier and RandomForestClassifier as base learners and Logistic Regression as the meta-learner. Trained on a Kaggle Parkinson's dataset, class imbalance is addressed using RandomOverSampler. A 20-fold stratified cross-validation enhances the reliability of performance metrics, including precision, recall, F1-score, ROC AUC, and confusion matrix. SHAP is used for feature-level interpretation, offering transparency and highlighting key vocal biomarkers distinguishing Parkinson's patients from healthy individuals. The approach ensures robustness to high-dimensional data and supports trustworthy AI-driven diagnostics.

The following is a summary of this paper's contribution:

- Investigate the effectiveness of addressing class imbalance using RandomOverSampler to enhance model performance in the classification of Parkinson's disease.
- Enhanced Classification Performance: The ensemble approach improves accuracy and generalization by leveraging diverse model strengths.
- Examine the impact of stratified cross-validation on model robustness and performance, ensuring reliable accuracy estimates in the classification task.
- Explore the role of model interpretability through SHAP (SHapley Additive exPlanations) and feature importance analysis to improve transparency and trust in AI-based medical predictions.
- Highlight the biological relevance of voice biomarkers by investigating the contribution of specific voice features in distinguishing between Parkinson's patients and healthy individuals.

2. STUDY LITERATURE

Recent progress in the application of machine learning (ML) and deep learning (DL) methodologies to non-invasive biomarkers has significantly enhanced efforts toward the early and accurate diagnosis of Parkinson's disease (PD). These technological advancements have addressed

critical challenges inherent in traditional diagnostic approaches, including diagnostic delays arising from symptom overlap with other neurological conditions, the dependence on invasive and often costly clinical procedures, and the pressing need for accessible, efficient, and non-invasive screening modalities capable of supporting early clinical decision-making and improving patient prognoses. Uday Kumar et al., (2022) addressed PD detection using voice impairments, employing an XGBoost Classifier (94.87% accuracy) on the UCI dataset (195 recordings, 22 features: MDVP:Fo, PPE, jitter), with preprocessing (scaling, feature analysis) and multiple-fold validation [12]. Wang et al. (2020) introduced an ensemble feed-forward deep neural network (DEEP_EN) to detect early Parkinson's disease (PD) based on premotor biomarkers, including REM sleep behavior disorder, olfactory loss, cerebrospinal fluid data, and dopaminergic imaging markers. Using the PPMI dataset comprising 183 healthy individuals and 401 early-stage PD patients, they applied log-transformation for preprocessing and optimized training via stochastic gradient descent. DEEP_EN achieved the highest average accuracy of 96.68%, outperforming twelve machine learning and ensemble models, and further identified critical features contributing to PD detection through boosting-based importance analysis [3]. Alshammri et al. (2023) investigated Parkinson's disease (PD) detection using various machine learning (ML) and deep learning (DL) models, including support vector machine (SVM), random forest (RF), decision tree (DT), K-nearest neighbor (KNN), and multi-layer perceptron (MLP), applied to voice signal features. Using a dataset of 195 voice recordings from 31 patients sourced from the UCI Machine Learning Repository, the models were enhanced through techniques such as Synthetic Minority Over-sampling Technique (SMOTE), feature selection, and hyperparameter tuning with GridSearchCV. Among the models, MLP achieved the best performance, with an accuracy of 98.31%, recall of 98%, precision of 100%, and F1-score of 99%. SVM also performed competitively, achieving 95% accuracy, 96% recall, 98% precision, and a 97% F1-score. These findings highlight the potential of ML- and DL-based voice analysis tools for reliable and accessible PD diagnosis [8].

Another study by Osiris et al. (2023) aimed to differentiate between Parkinson's disease (PD) and Hereditary Ataxias using machine learning techniques, based on gait characteristics collected from smartphone inertial motion sensors (iPhone 5S). The study involved 67 participants, 53 with PD and 14 with Hereditary Ataxias. Feature selection methods were employed for dimensionality reduction, and five classification algorithms were evaluated. The Support Vector Machine (SVM) achieved the highest performance, with an accuracy of 92.7%, precision of 91.1%, sensitivity of 96.2%, and specificity of 89.1%. These results suggest the potential for this approach to inspire further research and offer therapeutic applications [13]. Govindu and Palwe (2023) compared the performance of four machine learning models; Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression on Parkinson's disease detection. The results showed that the Random Forest classifier outperformed the others, achieving a detection accuracy of 91.83% and a sensitivity of 0.95. This study highlights the potential of machine learning in telemedicine, offering new possibilities for Parkinson's disease diagnosis and patient care [14]. Lim et al. (2022) explored the potential of combining acoustic and facial expression analysis using machine learning for early Parkinson's disease (PD) detection. A total of 371 participants were enrolled across training and validation cohorts. Using smartphone-based recordings during a reading task, nine machine learning classifiers were applied to integrated biometric features. The model achieved an AUROC of 0.85 in the training set and 0.90 in the validation cohort. Their findings demonstrate that multimodal voice and facial analysis can effectively aid in the early identification of PD [4]. Iyer et al. (2023) investigated the reliability of telephone-collected voice recordings for Parkinson's disease (PD) detection using machine learning. They gathered sustained vowel /a/ samples from 50 PD patients and 50 healthy controls in naturalistic settings. The study introduced a novel application of a pre-trained Inception V3 convolutional neural network via transfer learning to analyze voice spectrograms, capturing intensity variations across time and frequency. The deep learning model outperformed traditional classifiers, demonstrating the feasibility and accuracy of using mobile voice data for PD classification [6]. Naeem et al. (2025) explored the diagnostic utility of vocal biomarkers for early Parkinson's disease (PD) detection using multiple machine learning

models, including Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), and Decision Tree (DT). The study utilized 195 voice recordings from 31 individuals and addressed class imbalance using the Synthetic Minority Over-sampling Technique (SMOTE), with Principal Component Analysis (PCA) applied for feature reduction. Among the models, RF achieved the highest accuracy of 94% and precision of 94%, followed by SVM with 92% accuracy and 91% precision. The results underscore the relevance of vocal features combined with advanced ML methods for non-invasive and reliable PD diagnosis, particularly in early stages where conventional detection remains challenging [5]. Elshewey et al. (2023) proposed a classification framework for Parkinson's disease (PD) diagnosis using a Bayesian Optimization-Support Vector Machine (BO-SVM) model. The study leveraged Bayesian Optimization to fine-tune hyperparameters across six machine learning classifiers: Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Naive Bayes (NB), Ridge Classifier (RC), and Decision Tree (DT). Utilizing a dataset with 195 instances and 23 features, the models were evaluated based on accuracy, F1-score, recall, and precision. Results showed that the SVM classifier, particularly after hyperparameter optimization with BO, achieved the highest performance with an accuracy of 92.3%, underscoring the effectiveness of BO-SVM in enhancing diagnostic accuracy for PD [15].

Table 1. Related Research

Reference	Dataset	Sample Size	Data Type	Model	Accuracy
[12]	UCI Machine Learning Repository	195 recordings (31: 23 PD, 8 healthy)	Voice recordings	XGBoost	94.87%
[3]	PPMI	584 (401 PD, 183 healthy, aged 30+)	Clinical, imaging, CSF	DEEP_EN (Ensemble DL)	96.68%
[8]	UCI Machine Learning Repository	195 recordings (31: 23 PD, 8 healthy)	Voice recordings	MLP (Multi-Layer Perceptron)	98.31%
[13]	Custom (Manuel Velasco Suarez Institute)	67 (53 PD, 14 HA)	Smartphone accelerometer	SVM	92.7%
[14]	UCI/PPMI	195 recordings (31: 23 PD, 8 healthy, aged 46–85 PD, 23 healthy)	Vowel phonations	Random Forest	91.83%
[4]	Custom (smartphone)	371 (186 PD, 185 controls; train: 112 PD, 111 controls;	Voice, facial recordings	Combined (voice + facial)	AUROC 0.90 (validation)

		validation: 74 PD, 74 controls)			
[6]	Custom (telephone)	100 (50 PD, 50 HC, PD age 66.6, HC age 47.9)	Sustained vowel /a/	Inception V3 (CNN)	AUROC 0.97
[5]	UCI Machine Learning Repository	195 recordings (31 PD, healthy)	Vocal recordings	Random Forest	94%
[15]	UCI Machine Learning Repository	195 recordings (31: 23 PD, 8 healthy)	Voice recordings	BO-SVM (Bayesian Optimizati on-SVM)	92.3%

As summarized in Table 1, several studies have employed the UCI Machine Learning Repository to classify Parkinson's disease using voice recordings. Models such as XGBoost (94.87%) [12], MLP (98.31%) [8], Random Forest (94%) [5], and BO-SVM (92.3%) [15] achieved high accuracy, but they share key limitations: lack of class imbalance correction, minimal cross-validation, and no model explainability. Our study addresses these shortcomings by using a stacked ensemble model comprising CatBoostClassifier and RandomForestClassifier, with Logistic Regression as a meta-learner. Our pipeline integrates RandomOverSampler for class balance, 20-fold stratified cross-validation for robust evaluation, and SHAP for feature-level model interpretation while offering a more reliable.

3. RESEARCH METHOD

This study aims to develop a robust ensemble learning pipeline in improving parkinson's diagnosis. The dataset utilized in this study was sourced from Kaggle and consists of biomedical voice measurements collected from 31 individuals, of whom 23 were diagnosed with Parkinson's disease (PD). It comprises 195 voice recordings, where each row represents a recording and each column a specific vocal feature. The primary objective is to classify individuals as either healthy or affected by PD based on the status variable (0 = healthy, 1 = PD). Key features include fundamental frequency measures (MDVP:Fo(Hz), MDVP:Fhi(Hz), MDVP:Flo(Hz)), jitter and shimmer-related metrics that capture frequency and amplitude variations, respectively, as well as additional nonlinear measures (RPDE, D2, DFA, spread1, spread2, PPE) and noise-related features (NHR, HNR).

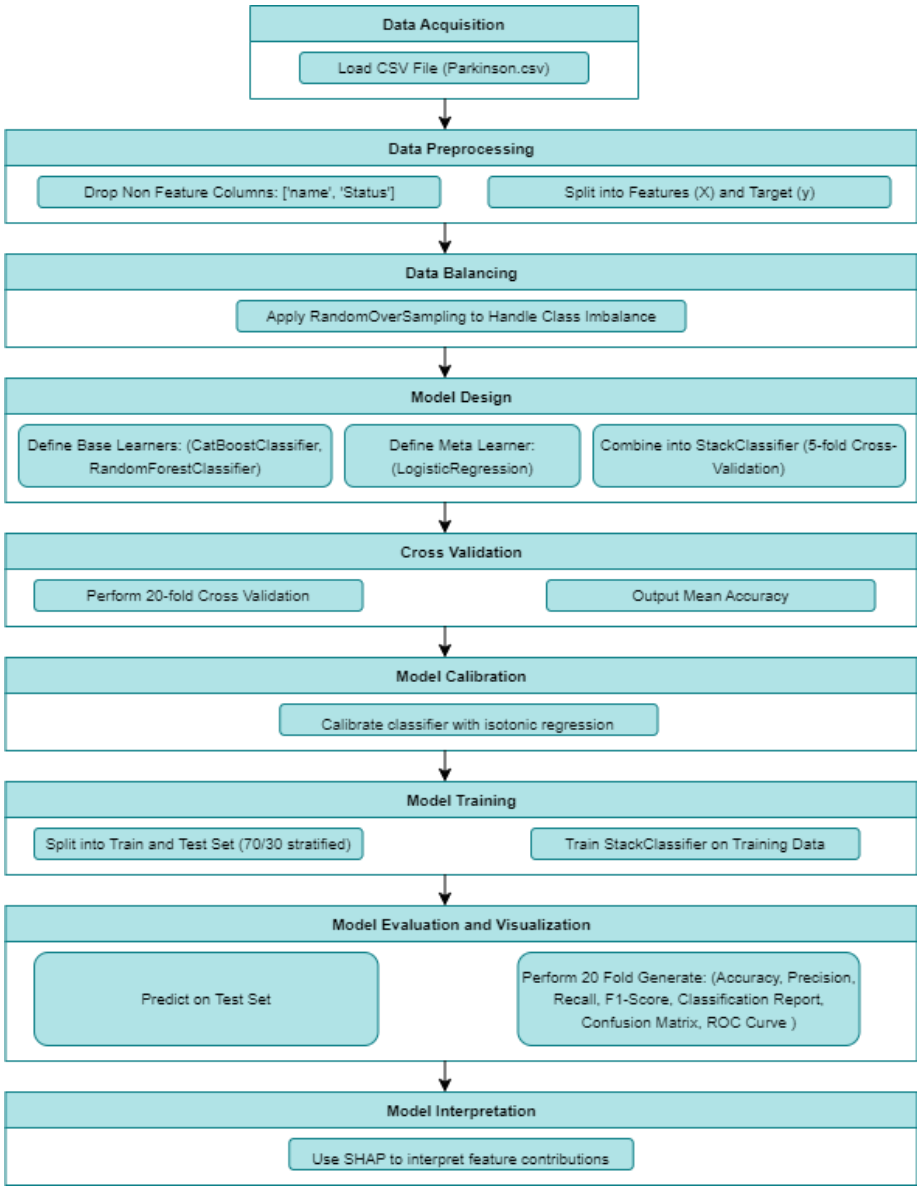

Figure 1. Research Methodology

3.1. Data Gathering

The dataset used in this research is the "Parkinson Disease Detection" available on Kaggle. This dataset consists of 24 features, including various vocal, nonlinear, and signal processing measures derived from voice recordings of individuals, as well as one target column indicating whether the subject has Parkinson's disease. These features reflect fundamental frequency, jitter, shimmer, noise ratios, and nonlinear dynamical complexity, all of which are relevant to the diagnosis of Parkinson's disease. The dataset comprises 24 features related to vocal biomarkers used for Parkinson's disease classification. The Name feature serves as an identifier, consisting of an ASCII subject name and recording number. Features MDVP:Fo(Hz), MDVP:Fhi(Hz), and MDVP:Flo(Hz) represent the average, maximum, and minimum vocal fundamental frequencies, respectively, categorized under frequency features. The jitter-related features include MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, and Jitter:DDP, all capturing variations or perturbations in frequency. Shimmer-related features, reflecting amplitude perturbations, include MDVP:Shimmer,

MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, and Shimmer:DDA. Noise characteristics are captured by NHR (Noise-to-Harmonics Ratio) and HNR (Harmonics-to-Noise Ratio). Non-linear dynamic features include RPDE, D2, DFA, spread1, spread2, and PPE, which measure signal complexity, variation, and entropy. The final column, Status, is an integer target variable indicating health status: 1 for Parkinson's disease and 0 for healthy individuals. All features, except for the identifier and the target variable, are of type float.

3.2 Data Acquisition and Processing

1. **Load CSV File**: This initial step involves importing the "Parkinsons.csv" dataset, likely containing voice or clinical data. The dataset presumably includes features (e.g., 22 acoustic attributes such as MDVP:Fo, jitter, shimmer, PPE) and a target variable (e.g., 'status' indicating PD or healthy). The file is read into a data structure (e.g., a pandas DataFrame in Python), making the raw data available for manipulation. Initial data quality checks, such as handling missing values or outliers begin here, though detailed cleaning is integrated into subsequent steps.
2. **Drop non-feature columns**: The 'name' column, likely a patient identifier, is removed as it offers no predictive value for PD detection. The 'status' column, serving as the target variable (e.g., 1 for PD, 0 for healthy), is excluded from the feature set to prevent data leakage. This step aligns with preprocessing techniques in 2022-Implementing, where irrelevant data was filtered to focus on voice features (e.g., MDVP:Fo, PPE).
3. **Split into Features (X) and Target (y)**: The dataset is divided into X (a matrix of independent variables, e.g., jitter, shimmer, NHR) and y (a vector of the dependent variable, 'status'). This separation is essential for supervised learning, where X trains the model to predict y. The feature set mirrors those in voice-based studies, preparing the data by isolating predictive attributes from the outcome. This combined section ensures the dataset is loaded, cleaned of non-predictive elements, and structured for modeling, setting a solid foundation for PD biomarker analysis.
4. **Data Balancing:** Class imbalance is prevalent in PD datasets, where PD cases often outnumber healthy controls. RandomOverSampler, from the imbalanced-learn library in Python, addresses this by oversampling the minority class (e.g., healthy samples) through random duplication or synthetic sample generation. This balances the dataset, preventing model bias toward the majority class (PD) and improving generalization. Balancing is crucial for fair evaluation, especially given the potential underrepresentation of early-stage PD cases.

3.3 Model Design and Training

1. **Define Base Learners**: Two base models are selected to capture diverse data patterns: CatBoostClassifier and RandomForestClassifier. CatBoostClassifier, a gradient boosting framework developed by Yandex, designed to optimize categorical features and missing values, excels in handling complex, non-linear relationships [16]. CatBoost follows the gradient boosting paradigm, where an ensemble of decision trees is built iteratively to minimize a loss function [17]. The objective is to minimize a loss function $L\big(y, F(x)\big)$, where y is thetrue label (e.g., 1 for PD, 0 for healthy), $F(x)$ is the prediction, and x is the feature vector (e.g., jitter, shimmer from "Parkinsons.csv"). The additive model constructs the prediction as a sum of decision trees:

$$F(x) = \sum_{t=1}^{T} f_t \ (x) \tag{1}$$

Where f_t (x) is the $t - th$ decision tree, and T is the total number of tree (iterations) [16].

While RandomForestClassifier, an ensemble of decision trees, reduces overfitting by averaging predictions across trees. These two models provide complementary strengths; CatBoost's boosting for iterative error correction and Random Forest's bagging for robustness and making them ideal for the stacking ensemble.

2. **Define Meta Learner**: LogisticRegression is chosen as the meta-learner in a stacking ensemble. Stacking combines base learner predictions (level-0) into a higher-level model (level-1) to improve accuracy. LogisticRegression's linear nature and interpretability make it ideal for weighting base learner outputs optimally, as it models the relationship between base predictions and the target (e.g., PD status) using a logistic function [18].
3. **Combine into StackingClassifier**: The base learners' predictions are integrated using a StackingClassifier from scikit-learn, employing 5-fold cross-validation. The dataset is split into 5 subsets; the model trains on 4 folds and validates on the 5th, cycling through all combinations.
4. **Split into Train/Test Sets (70/30, stratified)**: Post-design, the balanced dataset is divided into a 70% training set for model fitting and a 30% test set for evaluation. Stratified splitting preserves the PD/healthy ratio, preventing skewed representation.
5. **Train StackingClassifier on Training Data**: The StackingClassifier is trained on the 70% training set. The base learners (CatBoostClassifier, RandomForestClassifier) generate predictions, which the meta-learner (LogisticRegression) uses to learn the final classification rule. The combination of CatBoost and Random Forest ensures the ensemble leverages boosting and bagging, enhancing performance on PD detection.

3.4 Model Evaluation

1. **Perform 20-fold Cross-Validation on the Entire Dataset**: After training the StackingClassifier, its performance is initially assessed using 20-fold cross-validation on the entire balanced dataset (prior to the train/test split). The dataset is divided into 20 subsets; the model trains on 19 folds and validates on the remaining fold, repeating this process 20 times. This extensive validation minimizes variance and bias, providing a robust estimate of the model's generalization ability.
2. **Model calibration:** was performed using isotonic regression via CalibratedClassifierCV. After training the stacked ensemble model, isotonic calibration was applied to the predicted probabilities to improve their alignment with actual outcomes.
3. **Output Mean Accuracy**: The average accuracy across all 20 folds is calculated:

$$Mean\ Accuracy = \frac{1}{20}\sum_{k-1}^{20} Accuracy_k \quad (2)$$

$$Accuracy_k = \frac{Number\ of\ prediction\ in\ fold\ k}{Total\ samples\ in\ fold\ k} \quad (3)$$

This metric represents the proportion of correct PD status predictions.

4. **Predict on Test Set**: Following the initial evaluation, the trained StackingClassifier is used to predict PD status on the unseen 30% test set (from the 70/30 split). This step simulates real-world application, testing the model's ability to generalize to new data.

5. **Perform 20-fold Generalization on Test Set**: The model's performance on the test set is further validated using 20-fold cross-validation, reporting a comprehensive set of metrics:

i. **Accuracy:** The proportion of correct predictions on the test set:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ test\ samples} \quad (4)$$

ii. **Precision**: The proportion of predicted PD cases that are true PD cases:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (5)$$

iii. **Recall:** The proportion of actual PD cases correctly identified:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (6)$$

iv. **F1-Score**: The harmonic mean of precision and recall, balancing the trade-off:

$$F1 - Score = 2\ .\ \frac{Precision\ .Recall}{Precision + Recall} \quad (7)$$

v. **Classification Report**: Summarizes precision, recall, and F1-score for each class (PD, healthy).

vi. **Brier Score**: It evaluates the accuracy of probabilistic predictions. It is defined as:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2 \quad (8)$$

vii. **Confusion Matrix:** A 2x2 matrix showing:

$$\begin{bmatrix} True\ Negatives(TN) & False\ Positives \\ False\ Negatives(FN) & True\ Positives \end{bmatrix} \quad (9)$$

viii. **ROC Curve:** Plots TPR vs. FPR across thresholds, visualizing performance. This multi-metric approach surpasses single-metric evaluations, ensuring a comprehensive assessment of the model's performance on both the entire dataset and the test set.

4. RESULT AND ANALYSIS

This section presents and analyzes the results of the experiment conducted on the 'Parkinson Disease Detection' dataset using a stacked model, which employs CatBoostClassifier and RandomForestClassifier as base learners and Logistic Regression as the meta-learner.

4.1. Interpretability of Feature Contributions Using SHAP Analysis

Figure 2 presents the SHAP bar plot, which ranks features based on their mean absolute SHAP values, indicating their average contribution to the model's predictions. The features spread2, spread1, and PPE emerge as the most influential, reflecting their strong impact on model output. Other features such as MDVP:Fhi(Hz), MDVP:Fo(Hz), and DFA show moderate influence, while features like HNR and Shimmer:DDA contribute less significantly. In contrast, Figure 3, the SHAP summary plot, provides a detailed view of individual prediction impacts and highlights how the value of each feature (colored from blue to red) affects the direction and magnitude of model output. High

values of spread2, PPE, and MDVP:Fhi(Hz) tend to push predictions toward the disease class, reinforcing their importance.

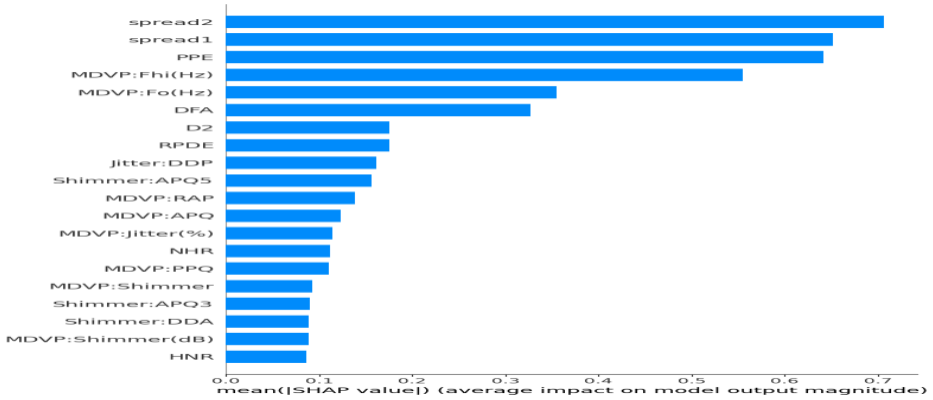

Figure 2. SHAP bar plot

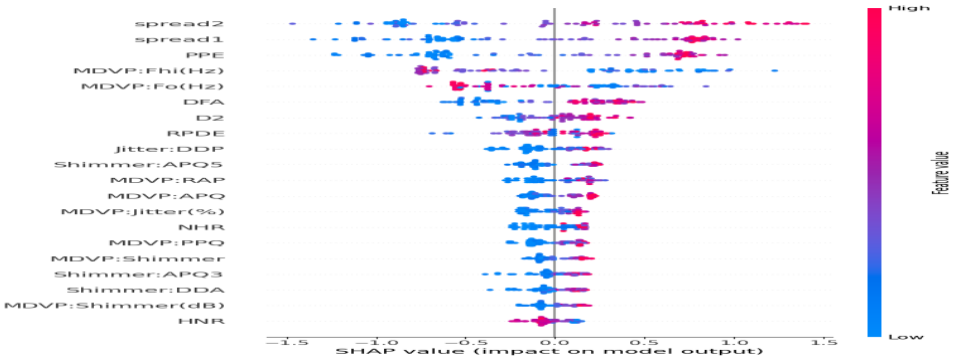

Figure 3. SHAP (SHapley Additive exPlanations) summary plot

4.2. Stacking Ensemble Confusion Matrix

The confusion matrix in Figure 4 illustrates the classification performance of a calibrated stacking ensemble model in distinguishing between Healthy Control and Parkinson's Disease cases. The model achieved perfect classification, correctly identifying all 44 Parkinson's Disease cases (True Positives) and all 45 Healthy Control cases (True Negatives), with zero False Positives and zero False Negatives. This results in 100% accuracy, sensitivity, and specificity, indicating an exceptionally high-performing model on the evaluated dataset. In practical terms, this means the model not only makes correct classifications but also outputs reliable confidence levels, making it a robust and trustworthy diagnostic tool for Parkinson's Disease detection though further validation on external datasets is recommended to ensure its generalizability.

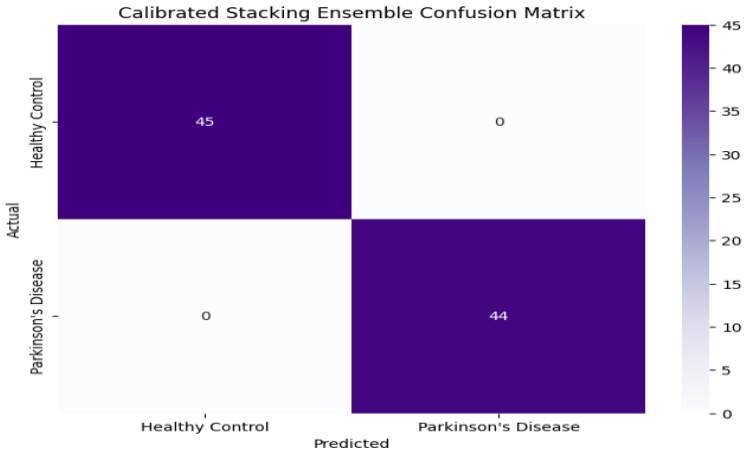

Figure 4. Confusion Matrix of Stacking Ensemble Model

4.3. Stacking Ensemble ROC Curve

The Stacking Ensemble ROC Curve in Figure 5 illustrates the performance of a stacking ensemble model, likely applied to lung cancer detection, by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) with an Area Under the Curve (AUC) of 1.00, indicating perfect classification. The purple ROC curve, which sharply rises along the y-axis from (0,0) to (0,1) and then moves horizontally to (1,1), significantly deviates from the gray dashed diagonal line (representing a random classifier with AUC = 0.5), demonstrating exceptional discriminative ability. This suggests the model, possibly combining multiple deep learning models as seen in studies like Razmjouei et al.'s ensemble approach (achieving 99.85% accuracy), perfectly identifies all positive cases without false positives. However, an AUC of 1.00 raises concerns about potential overfitting, data leakage, or an overly curated dataset, necessitating further validation on external datasets to confirm its real-world applicability for lung cancer diagnosis.

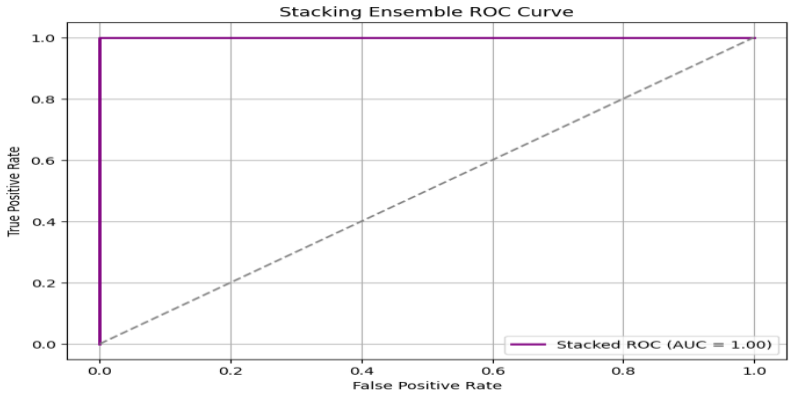

Figure 5. ROC Curve of Stacking Ensemble Model

4.4. Classification Report

The classification report as shown in Table 2 indicates that the stacking ensemble classifier achieved perfect performance in classifying Parkinson's Disease versus Healthy individuals. The model reached an overall accuracy of 100%, correctly predicting all 89 samples with no misclassifications. Both classes Healthy Control and Parkinson's Disease recorded precision, recall, and F1-scores of 1.00, reflecting a flawless ability to identify true positives without introducing false

positives or false negatives. Specifically, the precision of 1.00 indicates that every prediction made by the model for each class was correct, while the recall of 1.00 confirms that the model successfully identified all actual cases in both classes. The F1-score of 1.00 further demonstrates a perfect balance between precision and recall. The macro average and weighted average metrics are also 1.00 across all metrics, suggesting consistent and unbiased performance regardless of class distribution. Extended evaluation results reinforce these findings: the stacking classifier achieved an average cross-validation (CV) accuracy of 98.64% with a standard deviation of ±0.0272, indicating strong consistency across folds. Additionally, the Brier Score a measure of the accuracy of probabilistic predictions was remarkably low at 0.0071 for both the Healthy Control and Parkinson's Disease classes, yielding an average Brier Score of 0.0071 for the binary classification task. This low value highlights not only the model's classification accuracy but also the excellent calibration of its predicted probabilities. Overall, the results affirm that the stacking ensemble classifier is highly effective and reliable for Parkinson's Disease classification on the evaluated dataset. However, external validation on independent datasets is still necessary to ensure robustness and generalizability in real-world clinical settings.

Table 2: Calibrated Stacking Ensemble Classification Report

	Precision	**Recall**	**F1-score**	**Support**
Healthy Control	1.00	1.00	1.00	**45**
Parkinson's Disease	1.00	1.00	1.00	**44**
Accuracy			**1.00**	**89**
Macro Avg	**1.00**	**1.00**	**1.00**	**89**
Weighted Avg	**1.00**	**1.00**	**1.00**	**89**

4.5. Generalizability of the Proposed Model

The generalizability and robustness of the model were evaluated using the Parkinson's Disease Smartwatch (PADS) dataset, which includes clinical assessments of PD patients, individuals with related movement disorders, and healthy controls. Data were collected via a smart-device system comprising two wrist-worn smartwatches and a smartphone, recording 11 neurologist-designed movement tasks. The dataset contains 5,159 measurement steps from 469 individuals, with raw acceleration and rotation data, along with rich metadata on tasks, demographics, medical history, and PD-specific non-motor symptoms. It also reflects demographic diversity in age, gender, handedness, and lifestyle factors, making it a strong foundation for developing and validating sensor-based systems for movement disorder detection in varied populations. [19]. To assess the generalizability and robustness of the proposed model, the proposed model's methodology was applied to the Parkinson's Disease Smartwatch (PADS) dataset with it's classification report shown in Table 3 containing the Brier Score for each Class and the Average Brier Score.

Table 3: Calibrated Stacking Ensemble Classification Report for PADS Dataset

	Precision	**Recall**	**F1-score**	**Support**
Healthy	1.00	1.00	1.00	**83**
Parkinson's	1.00	1.00	1.00	**83**

Other Movement Disorders	1.00	1.00	1.00	**83**
Accuracy			**1.00**	**249**
Macro Avg	**1.00**	**1.00**	**1.00**	**249**
Weighted Avg	**1.00**	**1.00**	**1.00**	**249**

The classification report in Table 4 shows that the proposed model achieved perfect performance in distinguishing Healthy individuals, Parkinson's disease patients, and those with other movement disorders using data from the Parkinson's Disease Smartwatch (PADS) dataset. Each class recorded precision, recall, and F1-score of 1.00 across 83 samples per class, with an overall accuracy of 100%. Both macro and weighted averages were also 1.00, indicating consistent performance. The Brier Scores were extremely low 0.0001 for Healthy, 0.0000 for Parkinson's, and 0.0001 for Other Movement Disorders yielding an average of 0.0000. This confirms that the model made accurate and well-calibrated predictions. In the context of Parkinson's disease detection, these results highlight the model's strong potential for clinical decision support, providing both accurate classification and reliable confidence estimates using wearable sensor data.

Furthermore, the Confusion Matrix and ROC Curve of the above interpretability is shown in Figure 6 and 7 respectively.

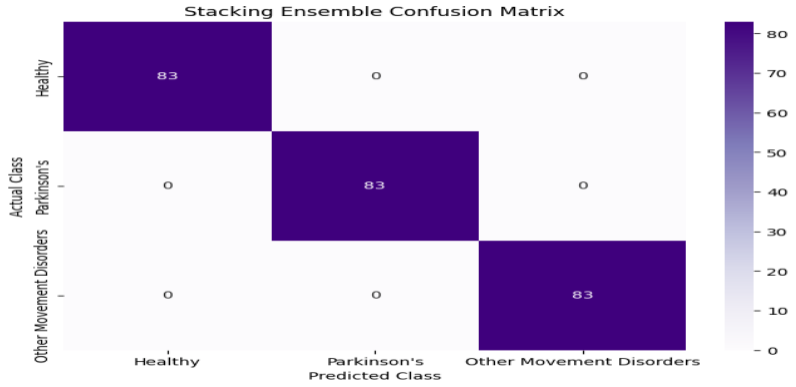

Figure 6. Confusion Matrix of Stacking Ensemble Model on PADS Dataset

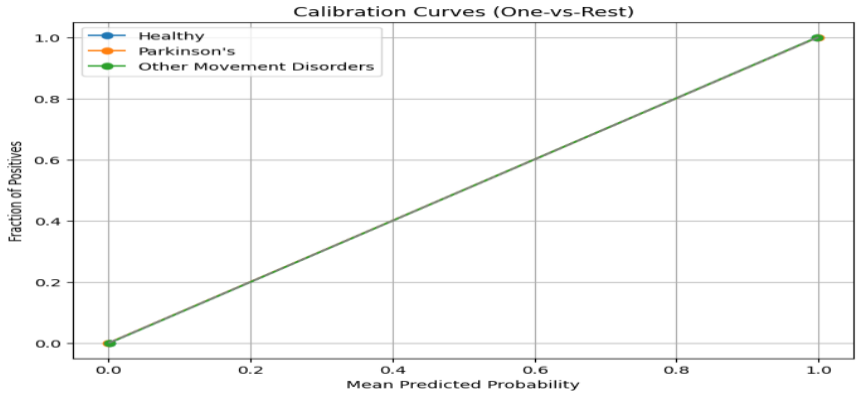

Figure 7. ROC Curve of Stacking Ensemble Model on PADS Dataset

5. CONCLUSION

In conclusion, the stacking ensemble classifier demonstrated exceptional performance in classifying Parkinson's disease versus healthy individuals, achieving high accuracy, precision, recall, and F1-scores. With an overall accuracy of 100%, perfect precision, and a perfect ROC AUC score of 1.00, the model is highly effective in distinguishing between the two classes. These results indicate that the model is both reliable and well-generalized, making it a strong candidate for Parkinson's disease classification tasks, though further validation with external datasets would help confirm its robustness.

REFERENCES

[1] S. Ramesh and A. S. P. M. Arachchige, "Depletion of dopamine in Parkinson's disease and relevant therapeutic options: A review of the literature," 2023, *AIMS Press*. doi: 10.3934/NEUROSCIENCE.2023017.

[2] J.-H. Lee, "Understanding Parkinson's Disorders: Classification and Evaluation Methods, Movement Disorders, and Treatment Methods," *International Journal of Advanced Culture Technology*, vol. 11, no. 3, pp. 9–17, 2023, doi: 10.17703/IJACT.2023.11.3.9.

[3] W. Wang, J. Lee, F. Harrou, and Y. Sun, "Early Detection of Parkinson's Disease Using Deep Learning and Machine Learning," *IEEE Access*, vol. 8, pp. 147635–147646, 2020, doi: 10.1109/ACCESS.2020.3016062.

[4] W. S. Lim *et al.*, "An integrated biometric voice and facial features for early detection of Parkinson's disease," *NPJ Parkinsons Dis*, vol. 8, no. 1, Dec. 2022, doi: 10.1038/s41531-022-00414-8.

[5] I. Naeem, A. Ditta, T. Mazhar, M. Anwar, M. M. Saeed, and H. Hamam, "Voice biomarkers as prognostic indicators for Parkinson's disease using machine learning techniques," *Sci Rep*, vol. 15, no. 1, p. 12129, Apr. 2025, doi: 10.1038/s41598-025-96950-3.

[6] A. Iyer *et al.*, "A machine learning method to process voice samples for identification of Parkinson's disease," *Sci Rep*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-47568-w.

[7] rania khaskhoussy and Y. Ben Ayed, "An I-vector-based approach for discriminating between patients with Parkinson's disease and healthy people," SPIE-Intl Soc Optical Eng, Mar. 2022, p. 34. doi: 10.1117/12.2623240.

[8] L.-C. Chang *et al.*, "Machine learning approaches to identify Parkinson's disease using voice signal features."

[9] M. S. Mirian, R. B. Z Q Zhang, S. Chen, X. Chen, G.-Z. Yang, and Y. G-z, "Detection and assessment of Parkinson's disease based on gait analysis: A survey."

[10] M. A. Islam, M. Z. Hasan Majumder, M. A. Hussein, K. M. Hossain, and M. S. Miah, "A review of machine learning and deep learning algorithms for Parkinson's disease detection using handwriting and voice datasets," *Heliyon*, vol. 10, no. 3, Feb. 2024, doi: 10.1016/j.heliyon.2024.e25469.

[11] L. Yuan, Yao Liu, and Hsuan-Ming Feng, "Parkinson disease prediction using machine learning-based features from speech signal," *Service Oriented Computing and Applications*, vol. 18, no. 1, pp. 101–107, 2024.

[12] U. Kumar, D. S. Baskaran, D. D. Sumathi, and P. G. Scholar, "Implementing a Model to Detect Parkinson Disease using Machine Learning Classifiers," *J Algebr Stat*, vol. 13, no. 1, pp. 99–110, 2022, [Online]. Available: https://publishoa.com

[13] O. Escamilla-Luna, M. A. Wister, and J. Hernandez-Torruco, "Machine Learning Algorithms for Classification Patients with Parkinson's Disease and Hereditary Ataxias," *Journal of Communications Software and Systems*, vol. 19, no. 1, pp. 9–18, 2023, doi: 10.24138/jcomss-2022-0157.

[14] A. Govindu and S. Palwe, "Early detection of Parkinson's disease using machine learning," in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 249–261. doi: 10.1016/j.procs.2023.01.007.

[15] A. M. Elshewey, M. Y. Shams, N. El-Rashidy, A. M. Elhady, S. M. Shohieb, and Z. Tarek, "Bayesian Optimization with Support Vector Machine Model for Parkinson Disease Classification," *Sensors*, vol. 23, no. 4, Feb. 2023, doi: 10.3390/s23042085.
[16] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features." [Online]. Available: https://github.com/catboost/catboost
[17] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: gradient boosting with categorical features support," Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.11363
[18] K. M. Ting and I. H. Witten, "Issues in Stacked Generalization," 1999.
[19] J. Varghese, A. Brenner, M. Fujarski, C. M. van Alen, L. Plagwitz, and T. Warnecke, "Machine Learning in the Parkinson's disease smartwatch (PADS) dataset," *NPJ Parkinsons Dis*, vol. 10, no. 1, Dec. 2024, doi: 10.1038/s41531-023-00625-7.