MJI
Media Jurnal Informatika
Vol. 18, No. 1, June 2026
p-issn : 2088-2114, e-issn : 2477-2542

Parkinson's Disease Classification Using Vocal Biomarkers, XGBoost, and SHAP

Wafiq Mariatul Azizah [a,1,*], Irma Amelia Dewi [a,2]

[a] Informatics Program, Institut Teknologi Nasional Bandung, Bandung 40124, Indonesia
[1] wafiqmariatula@gmail.com*; [2] irma_amelia@itenas.ac.id
* corresponding author

ARTICLE INFO

Article history
Received 2026-06-23
Revised 2026-06-24
Accepted 2026-06-27

Keywords
Parkinson's disease
XGBoost
SHAP
SMOTE
Voice biomarkers

ABSTRACT

Parkinson's disease (PD) is a progressive neurodegenerative disorder affecting more than 11.77 million people worldwide. Voice signal analysis has gained attention as a non-invasive screening approach because nearly 90% of PD patients experience measurable speech impairments. However, previous machine learning studies on PD voice datasets commonly face several limitations, including class imbalance that may lead to data leakage, the use of accuracy as the primary evaluation metric, and limited utilization of model interpretability methods. This study proposes a PD classification pipeline integrating SMOTE, XGBoost, and SHAP using the UCI Parkinson dataset, which consists of 195 samples and 22 acoustic features. A quantitative experimental approach was employed using 5-fold stratified cross-validation, where SMOTE was applied only to the training data within each fold to prevent data leakage, while SHAP was used for feature analysis and feature reduction experiments. The results showed that SMOTE improved the F1-Score from 0.9400 to 0.9527 and the Accuracy from 0.9077 to 0.9282. The final model achieved a mean AUC-ROC of 0.9614 and a Recall of 0.9592 across five folds. SHAP analysis showed differences between SHAP feature rankings and XGBoost built-in importance, with MDVP:Shimmer exhibiting the largest ranking change. In addition, the top-8 SHAP-ranked features achieved performance comparable to the full 22-feature model, obtaining an Accuracy of 0.9282 and an AUC of 0.9612. These findings indicate that the proper application of SMOTE and SHAP-based feature selection can improve model evaluation and provide additional information for feature analysis in Parkinson's disease classification.

1. Introduction

Parkinson's Disease (PD) is a progressive neurodegenerative disorder that ranks as the second most common neurological condition after Alzheimer's disease. A study based on the Global Burden of Disease 2021 reported that in 2019 more than 8.5 million individuals suffered from Parkinson's disease, with disability rates attributable to the disease increasing by 81% since 2000, and global prevalence continuing to rise to 11.77 million cases in 2021, with a higher disease burden in men than women amid global population aging [1]. These conditions position PD as one of the public health problems requiring greater attention, particularly in early detection efforts to slow disease progression and improve patients' quality of life.

One of the clinical characteristics that distinguishes Parkinson’s disease from other neurological disorders is the high incidence of speech impairment among patients. Studies show that around 90% of Parkinson’s patients experience speech changes such as pitch variation, frequency changes, amplitude differences, and reduced vocal stability. These symptoms can appear before visible motor symptoms [2]. Therefore, voice signal analysis can be used as a simple, non-invasive, and low-cost screening method, especially in areas with limited access to neurological specialists.

The UCI Parkinson dataset, comprising 195 voice recordings and 22 biomedical features, was first introduced by Little et al. [3] and remains one of the most widely used benchmark datasets in recent studies [2], [4], [5]. Alshammri et al. [4] compared several machine learning algorithms including KNN, SVM, Decision Tree, Random Forest, and MLP on the same dataset. Meral et al. [5] conducted a comparative analysis of six classifiers including SVM, k-NN, Decision Tree, Neural Network, Ensemble, and Stacking with hyperparameter optimization using Bayesian Optimization, Grid Search, and Random Search on a Parkinson's disease vocal dataset. Abu Sayed et al. [6] further examined the effectiveness of vocal biomarkers combined with machine learning algorithms for Parkinson's disease detection, highlighting the potential of acoustic features for disease diagnosis.

In addition, Sheikhi and Kheirabadi [7] proposed a Rotation Forest-based ensemble method combined with feature selection for Parkinson's disease severity prediction. Srinivasan et al. [8] developed a multiclass machine learning approach by integrating multiple classification algorithms to improve detection performance. Furthermore, Rahayu and Sudrajat [9] demonstrated the effectiveness of SMOTE in handling class imbalance in disease classification using XGBoost and Random Forest, with Random Forest achieving the highest accuracy of 98.8%. Sedigh Malekroodi et al. [10] conducted a systematic review of voice-based ML and deep learning approaches for PD detection, similarly noting that methodological consistency in feature extraction and validation strategies remains a key challenge across existing studies.

Nevertheless, a review of the existing literature reveals several recurring methodological problems. First, the UCI Parkinson dataset has an inherent class imbalance, where the proportion of positive PD samples far exceeds that of healthy samples [2]; many studies disregard this issue, causing models to be biased toward the majority class with seemingly high accuracy metrics that do not reflect genuine detection capability. This problem has also been observed in other classification domains, where improper handling of imbalanced data leads to misleading performance estimates [11]. Second, a number of studies suffer from data leakage due to the application of oversampling techniques prior to the separation of training and test data, rendering the claimed performance non-generalizable [12], [13]. Third, the majority of existing approaches focus solely on improving performance metrics without considering model interpretability, whereas in the context of clinical diagnosis, decision transparency is a crucial aspect for gaining the trust of both medical personnel and patients [14]. These methodological gaps constitute the departure point of this research.

This study proposes a PD classification pipeline that combines SMOTE, XGBoost, and SHAP. The contribution of this study does not lie in the novelty of the individual methods, as each method has been used separately in previous studies, but in the way they are integrated within a single classification framework. In particular, SMOTE is applied only to the training data within each cross-validation fold to avoid data leakage, which is a limitation found in several previous studies. In addition, SHAP is used not only to explain the model predictions but also to compare feature rankings with XGBoost built-in importance and to evaluate feature reduction through top-N feature subsets. The XGBoost-SHAP combination has also been applied successfully in other neurodegenerative disease studies, including Alzheimer's disease diagnosis [15]. Therefore, the contribution of this study lies in the development of a classification pipeline that emphasizes proper SMOTE implementation, reliable cross-validation, and the use of SHAP for feature analysis. This approach is intended to address the limitations identified in previous studies and to improve the interpretation of the classification results.

2. Method

2.1 Type and Approach of Research

This study uses a quantitative approach with an experimental research design. The quantitative approach is chosen because the study focuses on measuring and comparing numerical performance metrics, including accuracy, precision, recall, and F1-score, obtained from a structured biomedical dataset. The experimental design is applied by modifying components of the machine learning pipeline, specifically the use of SMOTE within stratified cross-validation folds and the evaluation of SHAP-based feature subsets compared to the full 22-feature set. This is done to observe their effects on classification performance.

The study uses the UCI Parkinson dataset, which consists of 195 voice recordings with 22 acoustic features, and applies a machine learning pipeline integrating SMOTE [15], XGBoost [16], and SHAP [17].

The model is evaluated using 5-fold stratified cross-validation to ensure more stable and generalizable performance results, especially considering the imbalanced class distribution in the dataset.

2.2 Object and Scope of Research

The object of this research is the classification of Parkinson's disease based on acoustic voice features extracted from sustained phonation recordings. The study operates within the healthcare domain, specifically targeting the binary classification task of distinguishing PD-positive individuals from healthy controls. The scope of this research is bounded to the UCI Parkinson dataset, comprising 195 voice recordings and 22 biomedical acoustic features, and does not extend to other PD datasets, imaging modalities, or clinical validation beyond computational experimentation.

2.3 Data Collection Techniques

Data collection in this study was conducted through documentation of a secondary source, namely the UCI Parkinson dataset publicly available in the UCI Machine Learning Repository, originally introduced by Little et al. [3]. The dataset comprises 195 voice recording samples obtained from 31 individuals, of whom 23 were diagnosed with Parkinson's disease and 8 were healthy controls, resulting in an inherently imbalanced class distribution with 147 positive and 48 negative samples. No primary data collection was performed, as the dataset constitutes an established benchmark widely used in Parkinson's disease detection research.

The dataset encompasses 22 acoustic features grouped into five categories as presented in Table 1. These features capture different aspects of vocal signal degradation commonly associated with hypokinetic dysarthria in Parkinson's patients.

Table 1. Acoustic feature groups in the UCI Parkinson dataset

Feature group	Count	Features	Aspect measured
Fundamental frequency	3	MDVP:Fo, Fhi, Flo	Mean, maximum, and minimum vocal fold vibration frequency
Jitter perturbation	5	Jitter(%), Jitter(Abs), RAP, PPQ, DDP	Cycle-to-cycle variation in vocal fold vibration period
Shimmer perturbation	6	Shimmer, Shimmer(dB), APQ3, APQ5, APQ, DDA	Cycle-to-cycle variation in voice signal amplitude
Noise-to-harmonics ratio	2	NHR, HNR	Proportion of noise to harmonic components in the signal
Nonlinear measures	6	RPDE, D2, DFA, spread1, spread2, PPE	Complexity and irregularity of vocal signal dynamics

2.4 Tools and Materials Used

This study was implemented using Python as the primary programming language, supported by several scientific computing libraries: scikit-learn [18], [19] for preprocessing, cross-validation, and evaluation metrics; XGBoost [16] for gradient boosting classification; imbalanced-learn [20] for SMOTE-based oversampling; and SHAP [17] for model interpretability analysis. Data manipulation and exploration were conducted using NumPy and pandas, while matplotlib and seaborn were used for visualization. All experiments were executed in a Jupyter Notebook environment running on a standard personal computer.

2.5 Research Procedures or Stages

The research was conducted in five sequential stages as illustrated in Figure 1.

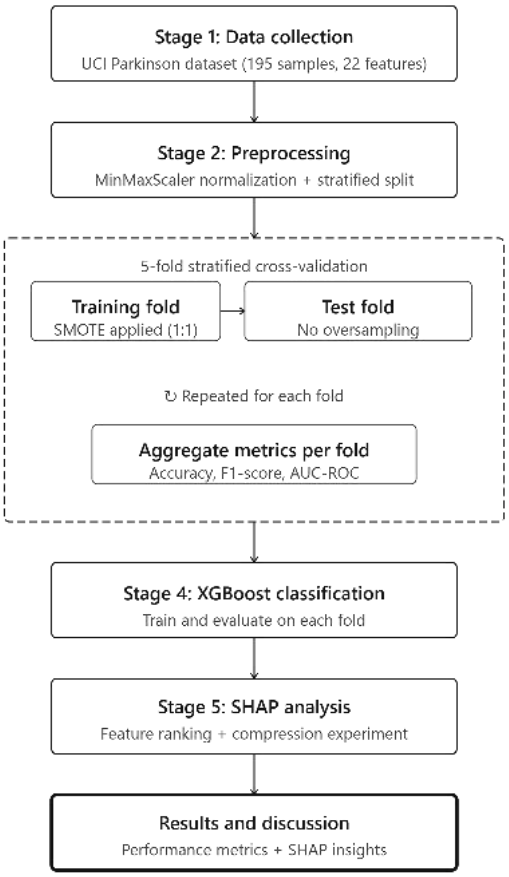

Fig 1. Research pipeline for Parkinson's disease classification using SMOTE, XGBoost, and SHAP

The first stage involved data collection and preliminary exploration of the UCI Parkinson dataset, which consists of 195 voice recordings and 22 acoustic features. The exploration process included analyzing the class distribution to identify the imbalance between PD-positive (147 samples) and healthy (48 samples) classes, as well as calculating descriptive statistics to observe the range and distribution of each feature before preprocessing.

The second stage involved data preprocessing before cross-validation. All 22 features were normalized to the range of 0 to 1 using MinMaxScaler to ensure consistent feature scaling. The dataset was then divided into five stratified folds to maintain the original class distribution in each fold.

The third stage applied 5-fold stratified cross-validation as the evaluation method. In each iteration, the data were separated into training and testing sets. SMOTE [16] was applied only to the training data to balance the minority class until a 1:1 class ratio was obtained, while the testing data remained unchanged to avoid data leakage. The evaluation metrics, including Accuracy, F1-Score, and AUC-ROC, were calculated for each fold and averaged to obtain the final performance results.

The fourth stage involved training and evaluating the XGBoost classifier [17] within the cross-validation process. For each fold, the model was trained using the SMOTE-balanced training data and evaluated on the testing data. The prediction results and probability scores were then used to calculate the evaluation metrics, and the results from all folds were averaged to obtain the final model performance.

The fifth stage applied SHAP analysis [18] after model training. First, the mean absolute SHAP values obtained from the testing data were used to rank the 22 features according to their contribution to the prediction results. The resulting ranking was then compared with the XGBoost built-in feature importance. Second, a feature reduction experiment was conducted by retraining the model using several top-N SHAP feature subsets ($N \in \{3, 5, 8, 10, 15, 22\}$). Each subset was evaluated using the same 5-fold stratified cross-validation and SMOTE procedure to determine the smallest feature subset that could maintain the classification performance of the full model.

2.6 Data Analysis Techniques

Model performance was evaluated using four classification metrics, namely accuracy, precision, recall, and F1-score, which were calculated across the five cross-validation folds. These metrics are derived from the confusion matrix, which consists of four prediction outcomes: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). In this study, the positive class refers to PD samples, while the negative class refers to healthy subjects.

Accuracy measures the proportion of correctly classified samples among all samples, as shown in Equation (1). Although accuracy provides an overall measure of model performance, it is not sufficient when used alone for imbalanced datasets [4], [14]. A model that predicts only the majority class may still achieve a high accuracy value despite performing poorly on the minority class.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (1)$$

Precision measures the proportion of samples predicted as PD that are actually PD-positive, as defined in Equation (2). A high precision value indicates that the model produces fewer false positive predictions, which is important in screening settings because unnecessary referrals may affect both patients and healthcare resources [2], [14].

$$\text{Precision} = \frac{TP}{TP + FP} \qquad (2)$$

Recall measures the proportion of actual PD samples that are correctly identified by the model, as defined in Equation (3). In the context of clinical screening, recall is considered an important evaluation metric alongside F1-score because a low recall value indicates that some PD patients are missed [2], [8]. In early detection applications, missing actual cases is generally more concerning than producing false positive predictions.

$$\text{Recall} = \frac{TP}{TP + FN} \qquad (3)$$

F1-score is the harmonic mean of precision and recall, as defined in Equation (4). Unlike the arithmetic mean, the harmonic mean is affected by large differences between precision and recall, making F1-score an appropriate metric for evaluating classifiers on imbalanced datasets [14], [20]. Given the 3:1 class imbalance in the UCI Parkinson dataset, F1-score and recall were considered the main evaluation metrics in this study.

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (4)$$

In addition to these four metrics, AUC-ROC was computed to assess the model's ability to discriminate between classes across all decision thresholds, independent of class distribution [5], [14]. AUC-ROC is defined as the area under the Receiver Operating Characteristic curve, which plots the True Positive Rate (Recall) against the False Positive Rate across varying classification thresholds, as defined in Equation (5). An AUC value of 1.0 indicates perfect discrimination, while a value of 0.5 indicates performance equivalent to random guessing[8], [10].

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(t)\, d(\text{FPR}(t)) \qquad (5)$$

3. Results and Discussion

3.1 Presentation of Research Results

The class imbalance handling experiment was conducted by comparing model performance under two configurations that is with and without SMOTE using identical hyperparameters and 5-fold stratified cross-validation. Figure 2 presents the class distribution of the training data in Fold 1 before and after the application of SMOTE.

Fig 2. Class Distribution of Training Data Before and After SMOTE

As shown in Figure 2, before SMOTE the training data in Fold 1 consisted of 39 healthy samples versus 117 PD samples (ratio 3.0:1), proportional to the overall dataset distribution. After SMOTE, the healthy class was augmented to 117 samples through the generation of 78 synthetic instances via nearest-neighbor interpolation, producing a balanced training set of 234 samples at a 1:1 ratio. The quantitative impact of this balancing on model performance metrics is presented in Table 2.

Table 2. Model Performance Comparison With and Without SMOTE (5-Fold CV)

Metric	Without SMOTE	With SMOTE	Difference
Accuracy	0.9077 ± 0.0466	0.9282 ± 0.0493	+0.0205
F1-Score	0.9400 ± 0.0295	0.9527 ± 0.0319	+0.0127
AUC-ROC	0.9695 ± 0.0217	0.9614 ± 0.0295	−0.0081

Table 2 shows that SMOTE consistently improved Accuracy and F1-Score relative to the baseline configuration without oversampling. The improvement in Accuracy from 0.9077 to 0.9282 and in F1-Score from 0.9400 to 0.9527 suggests that balancing the training class distribution enabled the model to better capture the characteristics of the minority class, which was previously underrepresented at a 3:1 ratio. However, AUC-ROC showed a slight decrease from 0.9695 to 0.9614, accompanied by a marginal increase in standard deviation from 0.0217 to 0.0295. This trade-off can be attributed to the effect of synthetic sample generation on the model's predicted probability distribution: models trained on imbalanced data tend to assign higher confidence scores to the majority class, which can artificially sharpen the ROC curve across thresholds, whereas SMOTE moderates this bias by exposing the model to a balanced distribution, consequently reducing AUC marginally. This phenomenon is consistent with findings by Vandewiele et al. [12], who observed that oversampling can alter probability calibration even when classification performance improves. Given that the AUC value remains within the excellent range (>0.95) and F1-Score improved meaningfully, this trade-off is considered acceptable for the purpose of clinical screening. Based on these results, the SMOTE configuration was adopted as the final pipeline for all subsequent experiments.

Table 3. 5-Fold Stratified Cross-Validation Results (with SMOTE)

Fold	Accuracy	F1-Score	AUC-ROC
1	0.9744	0.9836	0.9741
2	0.9231	0.9492	0.9259
3	0.9487	0.9643	0.9931
4	0.9487	0.9667	0.9793

5	0.8462	0.9000	0.9345
Mean ± Std	0.9282 ± 0.0493	0.9527 ± 0.0319	0.9614 ± 0.0295

Table 3 indicates that the model achieved consistently high performance across all folds, with mean AUC-ROC of 0.9614 ± 0.0295 and mean F1-Score of 0.9527 ± 0.0319. To examine the composition of classification errors in greater detail, predictions from all five folds were aggregated into a single confusion matrix, as presented in Figure 3.

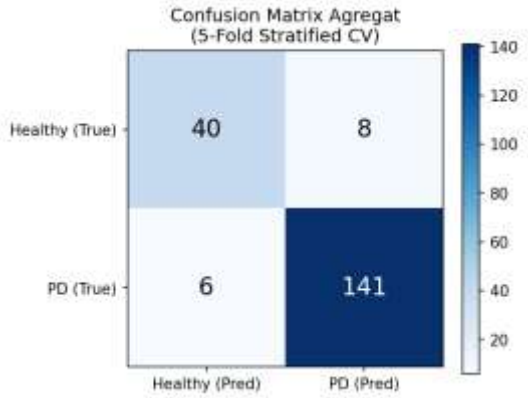

Fig 3. Aggregate Confusion Matrix for 5-Fold Stratified Cross-Validation

Figure 3 shows that, out of 195 samples evaluated across five folds, the model correctly classified 141 of 147 PD samples as True Positive and 40 of 48 healthy samples as True Negative, resulting in a PD Recall of 0.9592 and a Precision of 0.9463. To compare feature attribution between SHAP and the built-in feature importance of XGBoost, the mean absolute SHAP values were calculated on the test data from all folds and then compared with the built-in importance scores. Figure 4 presents the comparison of the top 10 features obtained from both methods.

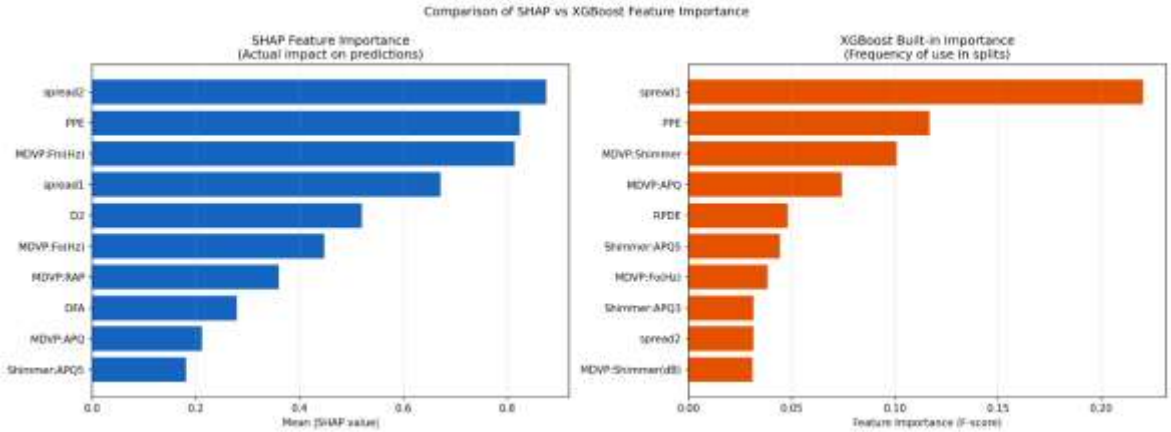

Fig 4. Comparison of SHAP and XGBoost Built-in Feature Importance for the Top 10 Features

Figure 4 reveals that the two methods produce noticeably different rankings, even within the top 10 features, indicating that split frequency and prediction contribution represent different measures of feature importance. The complete ranking of all 22 features is presented in Table 4.

Table 4. SHAP vs XGBoost Built-in Importance Ranking (22 Features*)*

Feature	SHAP Rank	XGB Rank	Difference
spread2	1	9	+8 ▲
PPE	2	2	0

MDVP:Fhi(Hz)	3	11	+8 ▲
spread1	4	1	−3
D2	5	12	+7 ▲
MDVP:Fo(Hz)	6	7	+1
MDVP:RAP	7	17	+10 ▲
DFA	8	19	+11 ▲
MDVP:APQ	9	4	−5 ▼
Shimmer:APQ5	10	6	−4
NHR	11	13	+2
RPDE	12	5	−7 ▼
MDVP:Jitter(%)	13	18	+5 ▲
MDVP:Jitter(Abs)	14	20	+6 ▲
MDVP:Shimmer(dB)	15	10	−5 ▼
Jitter:DDP	16	21	+5 ▲
HNR	17	15	−2
Shimmer:APQ3	18	8	−10 ▼
MDVP:Shimmer	19	3	−16 ▼
MDVP:Flo(Hz)	20	14	−6 ▼
MDVP:PPQ	21	22	+1
Shimmer:DDA	22	16	−6 ▼

Table 4 shows differences in the feature rankings obtained from the two methods. The largest changes were observed in MDVP:Shimmer, which decreased by 16 positions, and DFA, which increased by 11 positions. To further examine the influence of each feature on the prediction results, Figure 5 presents the SHAP beeswarm plot, where each point represents a test sample and is colored based on its original feature value.

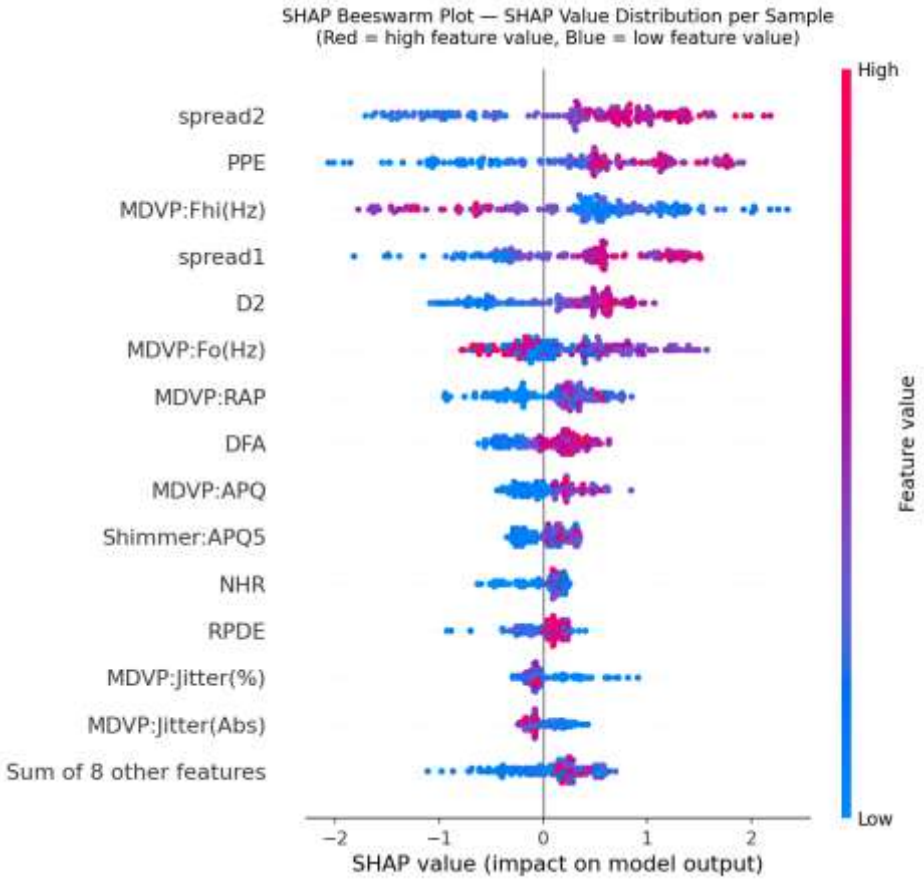

Fig 5. SHAP Beeswarm Plot of Per-Sample SHAP Value Distribution

Figure 5 shows that spread2 has the strongest contribution among all features. Higher feature values generally increase the prediction toward PD, whereas lower values are associated with healthy subjects. For the feature compression experiment, the model was retrained using top-N SHAP-ranked feature subsets, with SMOTE and 5-fold cross-validation applied for each subset size, where N ∈ {3, 5, 8, 10, 15, 22}. Figure 6 presents the AUC-ROC and Accuracy values obtained for each feature subset.

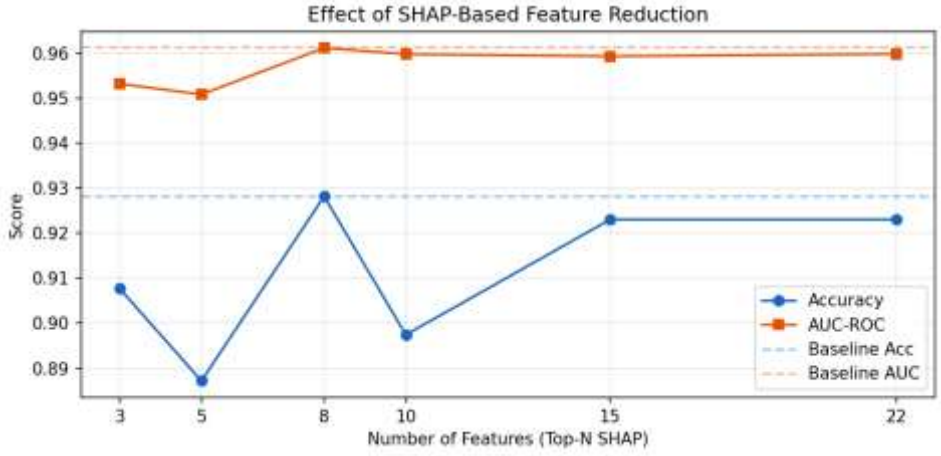

Fig 6. Feature Reduction Effect Curve Based on SHAP Feature Subset

Figure 6 shows that the AUC-ROC values become relatively stable from the top-8 feature subset onward, whereas the Accuracy values vary across the different feature subset sizes. The corresponding numerical results are presented in Table 5.

Table 5. Model Performance Across SHAP-Based Feature Subsets

No. of Features	Features Added	Accuracy	AUC-ROC
Top-3	spread2, PPE, MDVP:Fhi(Hz)	0.9077	0.9533
Top-5	+ spread1, D2	0.8872	0.9509
Top-8	+ MDVP:Fo(Hz), MDVP:RAP, DFA	0.9282	0.9612
Top-10	+ MDVP:APQ, Shimmer:APQ5	0.8974	0.9598
Top-15	+ 5 next features	0.9231	0.9594
Top-22 (all)	+ remaining 7 features	0.9231	0.9599

Table 5 shows that the top-8 SHAP feature subset achieved an Accuracy of 0.9282 and an AUC of 0.9612, which are comparable to the results obtained using all 22 features. Increasing the number of features beyond the top-8 did not improve the Accuracy or AUC values.

3.2 Analysis of Findings

The application of SMOTE improved Accuracy by 2.05 percentage points (0.9077 to 0.9282) and F1-Score by 1.27 percentage points (0.9400 to 0.9527). Without SMOTE, the model was trained on data with a 3:1 PD-to-healthy ratio, which creates an incentive for the classifier to favor the majority class — in this case, PD-positive samples. As a result, the model without SMOTE tends to produce more False Negatives for the healthy class, which directly suppresses both Accuracy and F1-Score. By generating synthetic minority-class samples within each training fold, SMOTE reduced this bias and allowed the classifier to learn a more balanced decision boundary, leading to improved detection of healthy controls without substantially sacrificing PD-class performance. The improvement in F1-Score is particularly meaningful in this context, as F1-Score captures the balance between precision and recall and is more sensitive to minority-class performance than accuracy alone.

The confusion matrix shows that 6 out of 147 PD samples were misclassified as healthy (False Negatives, miss rate 4.08%), and 8 out of 48 healthy samples were misclassified as PD (False Positives). In a clinical screening context, False Negatives are generally more concerning because they represent patients who may not be identified at an early stage. A Recall of 95.92% indicates that the model successfully detected nearly 96 out of every 100 actual Parkinson's cases, showing good performance for a voice-based screening approach.

SHAP analysis showed that the ranking of features based on prediction contribution differs from the ranking produced by XGBoost built-in importance. The largest difference was found in MDVP:Shimmer, which ranked 3rd by XGBoost but 19th by SHAP, indicating that the feature was frequently used during tree construction but had a relatively small contribution to the final predictions. In contrast, DFA and MDVP:RAP increased by 11 and 10 positions, respectively, under SHAP ranking. These findings are consistent with previous studies by Egbo et al. [2] and Nayan et al. [14], which reported that tree-based importance methods may overestimate features involved in repeated splits. The beeswarm plot also showed that spread2 has the strongest influence on the prediction results, where higher spread2 values tend to increase predictions toward PD. Meanwhile, MDVP:Fhi(Hz) and MDVP:Fo(Hz) showed an inverse relationship with PD prediction, which is consistent with the reduced vocal frequency commonly observed in Parkinson's patients.

The feature compression experiment showed that the top-8 SHAP-ranked features achieved an Accuracy of 0.9282 and an AUC of 0.9612, which are comparable to the full 22-feature model. Adding more features beyond the top-8 did not improve performance and, in some cases, slightly reduced Accuracy (for example, top-10: 0.8974 and top-15: 0.9231). This suggests that some additional features contribute only limited information to the classification process. Furthermore, the top-8 SHAP subset differs from the feature ranking produced by XGBoost built-in importance. A top-8 selection based on built-in importance would include MDVP:Shimmer, MDVP:APQ, RPDE, and Shimmer:APQ5, while excluding DFA and MDVP:RAP, which were identified by SHAP as important features. These results indicate that SHAP can provide additional information for feature selection beyond the built-in importance scores.

3.3 Implications of the Results

The findings of this study have both practical and methodological implications. From a practical perspective, the result that only 8 acoustic features are needed to achieve performance comparable to the full model suggests that a voice-based PD screening system can be developed with fewer feature extraction steps. This may reduce computational requirements and improve the feasibility of implementation on low-resource devices or in healthcare settings with limited access to neurology specialists. The selected features, which include fundamental frequency measures and nonlinear dynamic features, are also related to speech characteristics commonly observed in Parkinson's patients, providing additional support for the model's predictions.

From a methodological perspective, this study demonstrates the importance of applying SMOTE within each cross-validation fold rather than before data splitting in order to avoid data leakage and obtain more reliable performance estimates. In addition, SHAP provides additional information about feature contributions that may not be captured by XGBoost built-in importance, helping to improve the interpretation of model behavior.

3.4 Limitations of the Study

This study has several limitations that should be noted. First, the UCI Parkinson dataset contains only 195 samples from 31 individuals, which is relatively small and may limit how well the results generalize to larger and more diverse patient populations. Second, all recordings were collected under controlled laboratory conditions, so model performance may differ when applied to voice recordings from real-world clinical environments with background noise or varying recording quality. Third, this study was conducted on a single dataset without external validation on independent PD cohorts, which would be needed to confirm the practical reliability of the identified 8-feature subset. Fourth, the UCI Parkinson dataset contains multiple recordings per individual, which introduces a subject-level leakage risk in sample-based cross-validation: recordings from the same subject may appear in both training and test folds, potentially inflating reported performance relative to true generalization across unseen individuals. Subject-independent cross-validation, where fold splits are performed at the subject level rather than the sample level, would provide a more conservative and realistic performance estimate. Fifth, the SHAP-based feature compression experiment carries an inherent feature selection bias, as the SHAP rankings used to select top-N subsets were derived from a model trained on all 22 features; the selected subsets and the full-feature model therefore share overlapping training information, meaning the compression results should be interpreted as indicative rather than as a fully independent feature selection evaluation. Sixth, XGBoost hyperparameters were not systematically optimized in this study, meaning further tuning could potentially improve performance beyond what is reported here. Future research should address these limitations by testing the pipeline on multiple datasets, incorporating recordings from diverse clinical settings, applying subject-independent cross-validation to obtain unbiased performance estimates, and exploring hyperparameter optimization combined with SHAP-guided feature selection..

4. Conclusion

This study proposed a Parkinson's disease classification pipeline that addresses three limitations commonly found in previous studies, namely class imbalance, data leakage, and limited model interpretability. SMOTE was applied exclusively within each cross-validation fold to prevent data leakage, while XGBoost was used as the classifier and SHAP was employed as the main feature analysis method. The results showed that SMOTE improved F1-Score from 0.9400 to 0.9527 and Accuracy from 0.9077 to 0.9282, while the AUC-ROC value remained within the excellent category. The final model achieved a mean AUC-ROC of 0.9614 and a Recall of 0.9592 across five stratified folds, indicating good and consistent classification performance. SHAP analysis also showed differences between XGBoost built-in importance and feature contributions to prediction, with MDVP:Shimmer ranking 3rd by built-in importance but only 19th by SHAP.

Furthermore, the feature compression experiment showed that the top-8 SHAP-ranked features, namely spread2, PPE, MDVP:Fhi(Hz), spread1, D2, MDVP:Fo(Hz), MDVP:RAP, and DFA, were sufficient to achieve performance comparable to the full 22-feature model in terms of both Accuracy and AUC-ROC. These findings indicate that SHAP can provide additional information for feature selection beyond XGBoost built-in importance. Future studies may validate the proposed pipeline using larger and more diverse datasets,

incorporate recordings from real clinical environments, and investigate hyperparameter optimization combined with SHAP-based feature selection to further improve model generalizability.

Acknowledgment

The authors would like to thank all parties who supported this research.

Declarations

Author contribution. The authors responsible for the study conception, data processing, model development, experimental evaluation, analysis of the results, and manuscript preparation.
Funding statement. This research received no external funding.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

Data and Software Availability Statements

The dataset used in this study is the UCI Parkinson Dataset, publicly available through the UCI Machine Learning Repository at https://archive.ics.uci.edu/dataset/174/parkinsons. All experiments were implemented in Python using the following open-source libraries: scikit-learn, XGBoost, imbalanced-learn, SHAP, NumPy, pandas, matplotlib, and seaborn. No proprietary software or custom datasets were generated during this study.

References

[1] Y. Luo, L. Qiao, M. Li, X. Wen, W. Zhang, and X. Li, "Global, regional, national epidemiology and trends of Parkinson's disease from 1990 to 2021: findings from the Global Burden of Disease Study 2021," *Front. Aging Neurosci.*, vol. 16, no. January, pp. 1–12, 2024, doi: 10.3389/fnagi.2024.1498756.

[2] B. Egbo, Z. Nigmetolla, N. A. Khan, and P. K. Jamwal, "Explainable machine learning for early detection of Parkinson's disease in aging populations using vocal biomarkers," *Front. Aging Neurosci.*, vol. 17, no. September, 2025, doi: 10.3389/fnagi.2025.1672971.

[3] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1–7, 2009, doi: 10.1109/TBME.2008.2005954.Suitability.

[4] R. Alshammri, G. Alharbi, E. Alharbi, and I. Almubark, "Machine learning approaches to identify Parkinson's disease using voice signal features," *Front. Artif. Intell.*, vol. 6, 2023, doi: 10.3389/frai.2023.1084001.

[5] M. Meral, F. Ozbilgin, and F. Durmus, "Fine-Tuned Machine Learning Classifiers for Diagnosing Parkinson's Disease Using Vocal Characteristics: A Comparative Analysis," *Diagnostics*, vol. 15, no. 5, 2025, doi: 10.3390/diagnostics15050645.

[6] Md Abu Sayed *et al.*, "Parkinson's Disease Detection through Vocal Biomarkers and Advanced Machine Learning Algorithms," *Journal of Computer Science and Technology Studies*, vol. 5, no. 4, pp. 142–149, 2023, doi: 10.32996/jcsts.2023.5.4.14.

[7] S. Sheikhi and M. T. Kheirabadi, "An Efficient Rotation Forest-Based Ensemble Approach for Predicting Severity of Parkinson's Disease," *J. Healthc. Eng.*, vol. 2022, 2022, doi: 10.1155/2022/5524852.

[8] S. Srinivasan, P. Ramadass, S. K. Mathivanan, K. Panneer Selvam, B. D. Shivahare, and M. A. Shah, "Detection of Parkinson disease using multiclass machine learning approach," *Sci. Rep.*, vol. 14, no. 1, pp. 1–17, 2024, doi: 10.1038/s41598-024-64004-9.

[9] A. H. Rahayu and A. Sudrajat, "Improving Heart Disease Severity Prediction Using SMOTE for Imbalanced Data," *Journal of Applied Intelligent System*, vol. 9, no. 2, pp. 250–259, 2024, doi: 10.62411/jais.v9i2.11180.

[10] H. Sedigh Malekroodi, B. Il Lee, and M. Yi, "Voice-Based Detection of Parkinson's Disease Using Machine and Deep Learning Approaches: A Systematic Review," *Bioengineering*, vol. 12, no. 11, pp. 1–29, 2025, doi: 10.3390/bioengineering12111279.

[11] D. Y. Saputra, L. K. Wardhani, and H. Nanang, "Multi-Class Fault Detection under Class-Imbalance in Wireless Sensor Network Using Random Undersampling and Extra Trees," *Media Jurnal Informatika*, vol. 17, no. 2, pp. 367–378, 2025.

[12] G. Vandewiele *et al.*, "Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling," *Artif. Intell. Med.*, vol. 111, 2021, doi: 10.1016/j.artmed.2020.101987.

[13] S. S. Putra and D. Riminarsih, "Evaluating Machine Learning Models Across Feature Extraction and Data Balancing Scenarios for Coretax Sentiment Analysis," *Media Jurnal Informatika*, vol. 17, no. 2, pp. 379–396, 2025.

[14] N. M. Nayan, A. M. Rana, M. M. Islam, J. Uddin, T. Yasmin, and J. Uddin, "An interpretable and balanced machine learning framework for Parkinson's disease prediction using feature engineering and explainable AI," *PLoS One*, vol. 20, no. 10 October, pp. 1–34, 2025, doi: 10.1371/journal.pone.0333418.

[15] N. A. Azhar, M. S. Mohd Pozi, A. M. Din, and A. Jatowt, "An investigation of SMOTE based methods for imbalanced datasets with data complexity analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 7, pp. 6651–6672, 2023, doi: 10.1109/TKDE.2022.3179381.

[16] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. KDD '16, no. 10, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[17] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Section 2, pp. 4766–4775, 2017.

[18] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," in *Journal ofMachine Learning Research 12*, vol. 12, 2011, pp. 633–642. [Online]. Available: https://link.springer.com/10.1007/978-3-031-40336-1_56

[19] S. Prusty, S. Patnaik, and S. K. Dash, "SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer," *Frontiers in Nanotechnology*, vol. 4, no. August, pp. 1–12, 2022, doi: 10.3389/fnano.2022.972421.

[20] H. Hairani, T. Widiyaningtyas, and D. D. Prasetya, "Addressing Class Imbalance of Health Data: a Systematic Literature Review on Modified Synthetic Minority Oversampling Technique (SMOTE) Strategies," *International Journal on Informatics Visualization*, vol. 8, no. 3, pp. 1310–1318, 2024, doi: 10.62527/joiv.8.3.2283.