QUBAHAN TECHNO JOURNAL
VOL. 5, NO. 2, May 2026
https://doi.org/10.48161/qtj.v5n2a84

An Explainable Machine Learning Framework for Parkinson's Disease Classification Using High-Dimensional Speech Features

Dilovan Asaad Zebari [1], Marwan D. Saleh [2], Falah Y. H. Ahmed [1], Renas Rajab Asaad[3]

[1] Faculty of Computing and Information Technology, Sohar 311, Sohar University, Oman;
[2] Electronics and Computer Science University of Southampton Malaysia, Johor 79200, Malaysia;
[3] Department of Computer Science, Knowledge University, Erbil, 44001, Iraq;
Corresponding author: m.d-saleh@soton.ac.uk

ABSTRACT: Parkinson's disease is a progressive neurodegenerative disorder in which early and reliable diagnosis remains clinically important. Speech impairment is one of the most common manifestations of the disease and offers a non-invasive basis for automated screening. This study proposes an explainable machine learning framework for Parkinson's disease classification using high-dimensional speech features. The framework integrates data preprocessing, feature selection, multiple machine learning classifiers, probability-based evaluation, and explainability analysis to build an interpretable and robust diagnostic model. Four classifiers were applied: Logistic Regression, SVM with an RBF kernel, Random Forest, and XGBoost. Experimental evaluation on the speech-feature dataset showed that XGBoost achieved the best overall classification performance, while Random Forest produced the most reliable calibration. In addition, feature-importance analysis revealed that dynamic cepstral and TQWT-based speech descriptors were among the most influential predictors for Parkinson's disease detection. Oversampling strategies, including Random Oversampling, SMOTE, and ADASYN, did not improve performance over the original data distribution, indicating that model-level robustness was more beneficial than synthetic class balancing for this dataset. Overall, the findings demonstrate that explainable ensemble machine learning provides an effective and interpretable approach for Parkinson's disease classification from high-dimensional speech biomarkers.

Keywords: Parkinson's disease classification, speech biomarkers, explainable machine learning, high-dimensional features, ensemble learning.

I. INTRODUCTION

Parkinson's disease (PD) is a progressive neurodegenerative disorder whose early diagnosis remains clinically challenging because definitive diagnosis is still largely based on symptom presentation rather than a rapid, objective biomarker-driven test [1]. Speech impairment is especially relevant in this context because vocal and articulatory abnormalities are highly prevalent in PD and may emerge before overt motor dysfunction, making speech a promising non-invasive signal for early screening and longitudinal monitoring. Recent reviews further argue that digital speech analysis can support scalable, low-cost, and remote assessment, which is particularly attractive for early detection and telehealth applications [2-5].

Within this landscape, machine learning has become a major methodological avenue for transforming speech recordings into clinically useful decision support tools. A recent systematic review by Hossain et al. [3] found that machine learning and deep learning methods applied to PD voice, speech, and language data have shown strong diagnostic promise across multiple tasks and recording protocols, while also highlighting persistent methodological weaknesses such as inconsistent validation strategies, limited external testing, and insufficient standardization of datasets and preprocessing pipelines. Similarly, Ngo et al. [2] concluded in their systematic review that speech and voice are valuable biomarkers for PD, but they also emphasized that large differences among datasets and recording protocols make study-to-study comparison difficult and may limit clinical translation [2-4].

The dataset used in this study is well aligned with this research direction. The Parkinson's Disease Classification dataset curated by Sakar et al. [1] contains sustained phonation recordings from patients with PD and healthy controls, processed into a high-dimensional tabular representation with 754 features derived from multiple speech signal processing algorithms, including time-frequency features, Mel-frequency cepstral coefficients, wavelet-based descriptors, vocal fold features, and TQWT features. This type of representation is particularly suitable for classical machine learning because it preserves rich acoustic information while avoiding the computational overhead associated with raw waveform modeling. At the same time, its high dimensionality makes robust feature handling and model interpretability especially important [1].

Although predictive performance has been the dominant focus in prior voice-based PD studies, interpretability is increasingly recognized as a prerequisite for clinical trust and adoption. In particular, Shen et al. [4] showed that explainable artificial intelligence methods such as SHAP can identify the vocal features that drive PD prediction, thereby improving transparency and trustworthiness in early-detection systems. Their work reinforces a broader concern in medical AI: highly accurate models are not sufficient if clinicians cannot understand why a patient is classified as high risk. This issue is especially important for high-dimensional speech datasets, where redundant, correlated, or clinically opaque features can easily dominate model behavior unless explicit interpretability mechanisms are built into the framework [4].

At the same time, the speech-biomarker literature suggests that PD-related speech deficits are multidimensional rather than confined to one acoustic domain. According to Cao et al. [5], PD-related alterations may affect phonation, articulation, prosody, respiration, and broader language behavior, and these changes can evolve across the disease continuum. This implies that high-dimensional speech feature spaces are not merely computational artifacts; they may reflect genuinely heterogeneous disease signals. However, such richness also creates methodological challenges, including the risk of overfitting, unstable feature importance estimates, reduced reproducibility, and poor generalization across cohorts. These concerns further motivate the need for machine learning frameworks that are not only accurate but also explainable, feature-aware, and robust under realistic data conditions [2, 5].

Despite the rapid growth of machine learning for PD speech analysis, three important gaps remain. First, many existing studies emphasize predictive accuracy while giving limited attention to model interpretability. Second, high-dimensional speech-feature datasets are often used without sufficient analysis of feature contribution, selection stability, or clinical plausibility. Third, although systematic reviews have highlighted the promise of speech biomarkers, they have also stressed the lack of methodological consistency and the limited use of explainable frameworks in practice. These gaps are particularly important for Q1-level biomedical machine learning research, where methodological transparency and clinical relevance are expected alongside strong performance [2-4].

Motivated by these challenges, this study proposes an explainable machine learning framework for Parkinson's disease classification using high-dimensional speech features. The central premise is that a clinically useful PD classifier should not only distinguish patients from healthy controls with high accuracy, but also reveal the contribution of the most informative speech biomarkers in a transparent and reproducible way. By combining high-dimensional speech descriptors with explainable machine learning, the study aims to contribute a model that is both diagnostically effective and clinically interpretable, thereby supporting the broader movement toward trustworthy artificial intelligence in speech-based neurological screening [1, 4, 5].

II. RELATED WORKS

Research on Parkinson's disease (PD) detection from speech has expanded rapidly because speech impairment is one of the most accessible and non-invasive manifestations of the disease. Recent evidence indicates that voice and speech biomarkers can support both early diagnosis and disease monitoring, but the literature also shows substantial methodological variation in feature extraction, validation design, and model interpretability. In particular, the field has shifted from conventional machine learning classification toward more transparent and clinically interpretable frameworks, including explainable artificial intelligence (XAI).

A broad overview of this field is provided by Ngo et al. [6], who systematically reviewed computerized analysis of speech and voice for PD and concluded that acoustic biomarkers are promising for diagnosis, although inconsistency across datasets, protocols, and evaluation settings remains a major limitation. In a complementary review, Cao et al. [7] emphasized that speech and language abnormalities often precede overt motor dysfunction and therefore offer strong potential for prediction, early diagnosis, and progression tracking. These reviews collectively establish speech as a clinically meaningful digital biomarker, but they also highlight the need for more robust and interpretable modelling pipelines.

Within speech-based PD classification, one influential recent study is Hossain et al. [8], who evaluated machine learning classifiers and pipeline-based approaches using high-dimensional speech biomarkers. Their study showed that pipeline-based models were more effective than standalone classifiers because they improved both feature selection and classification performance on complex speech representations. Similarly, Jeong et al. [9] proposed a voting-based machine learning framework using speech characteristics and demonstrated that ensemble decision-making can improve automated PD diagnosis. These studies confirm that classical supervised learning remains highly competitive for structured speech features, particularly when supported by appropriate preprocessing and feature engineering. However, they mainly focus on predictive performance and give limited emphasis to explainability at the level required for clinical adoption.

The growing demand for transparent clinical AI has motivated more recent work on explainable PD diagnosis. Priyadharshini et al. [10] proposed a comprehensive XAI-empowered machine learning framework for PD diagnosis and showed that integrating explainability with strong predictive models can improve trust and usability. Their work is important because it moves beyond raw classification accuracy and directly addresses the need to understand the contribution of individual predictive features. In the same broader direction, Shen et al. [11] developed an explainable AI approach for early PD diagnosis through voice analysis and used SHAP to identify the vocal features that most strongly drive prediction. Their findings reinforce the importance of explainability in voice-based PD screening and show that interpretable models can support both early detection and clinical transparency.

Despite these advances, several gaps remain. First, although prior speech-based PD studies have shown strong diagnostic performance, many rely on complex feature sets without sufficiently clarifying which acoustic descriptors truly drive model decisions. Second, while explainability has recently received more attention, most studies still use it as a post-hoc interpretive layer rather than integrating it into a more structured feature-selection and model-validation workflow. Third, the methodological literature repeatedly notes that high-dimensional speech datasets are vulnerable to overfitting, unstable feature importance rankings, and limited reproducibility if validation is not carefully designed. These issues are especially relevant for datasets with hundreds of correlated features, where explainability and feature discipline are as important as predictive power.

Accordingly, the current literature supports three conclusions. First, speech-derived features provide a viable basis for PD classification. Second, machine learning models can achieve strong diagnostic performance when supported by feature processing and ensemble design. Third, explainability is no longer optional in this area; it is increasingly necessary for producing clinically meaningful PD screening systems. However, there remains a need for a framework that combines high-dimensional speech-feature modelling, rigorous machine learning, and explicit explainability within one coherent pipeline. That gap motivates the present study.

Table 1. Comparative summary of representative studies on Parkinson's disease classification using speech, machine learning, and explainable artificial intelligence.

Ref.	Study	Main focus	Data/modality	Core method	Main contribution	Main limitation
[6]	Ngo et al.	Systematic review of speech/voice analysis for PD	Speech and voice studies	Review of computerized speech analysis	Established speech as a promising PD biomarker and highlighted methodological inconsistency	Review paper; does not provide a unified predictive framework
[7]	Cao et al.	Speech and language biomarkers for PD prediction and monitoring	Speech and language biomarkers	Narrative review	Framed speech as useful for early diagnosis and progression tracking	Review-oriented, not a directly testable ML framework
[8]	Hossain et al.	PD classification using speech biomarkers	High-dimensional speech features	ML classifiers and pipelines	Showed pipeline models improve feature handling and classification performance	Limited emphasis on explainability and clinical interpretability
[9]	Jeong et al.	Automated PD diagnosis from speech characteristics	Speech characteristics	Voting-based ML ensemble	Demonstrated the value of ensemble learning in PD voice classification	Focused more on performance than transparent feature attribution
[10]	Priyadharshini et al.	Explainable AI–empowered PD diagnosis	PD diagnostic datasets, including ML/XAI setting	ML + XAI framework	Integrated explainability with predictive modeling	Broader PD diagnosis scope; not limited to one high-dimensional speech-feature framework
[11]	Shen et al.	Explainable early PD diagnosis via voice analysis	Voice data	Explainable AI with SHAP	Identified key vocal predictors and improved model transparency	Focused mainly on explainability after model development rather than structured feature-robust pipeline design

III. PROPOSED FRAMEWORK

To address the problem of Parkinson's disease classification from high-dimensional speech features, this study proposes an Explainable Machine Learning Framework specifically designed for structured acoustic biomarker data. The uploaded dataset is tabular and high dimensional, containing hundreds of correlated speech descriptors, which makes it unsuitable for naive end-to-end modeling without disciplined feature control, calibration, and interpretability. For this reason, the proposed framework is organized as a multi-stage pipeline that combines data preprocessing, robust feature selection, calibrated ensemble classification, and post-hoc explainability. The design is motivated by prior work showing that high-dimensional biomedical classification benefits from minimum-redundancy feature screening [12], all-relevant wrapper-

based selection [13], nested validation to reduce selection bias [14], calibrated probability estimation [15], and SHAP-based feature attribution for transparent medical decision support [16]. Recent Parkinson-specific explainable AI research further confirms that model transparency is essential when speech biomarkers are used for clinical inference [17].

1. FRAMEWORK RATIONALE

The proposed framework is based on the premise that Parkinson's disease classification from speech should not rely solely on maximizing predictive accuracy. In a high-dimensional clinical dataset, a high-accuracy model may still be unsuitable if it is driven by unstable features, poorly calibrated probabilities, or clinically opaque decision rules. Accordingly, the framework is built around four principles:

- feature discipline, to reduce redundancy and noise in the acoustic feature space;
- robust validation, to prevent optimistic estimates caused by feature-selection leakage;
- calibrated classification, to ensure that predicted class probabilities reflect actual diagnostic confidence; and
- model explainability, to identify the most influential speech biomarkers and support clinical interpretability. These principles are consistent with established methodological work on feature selection [12, 13], model validation [14], classifier calibration [15], and SHAP-based interpretability [16, 18].

2. STAGE I: DATA PREPROCESSING

The first stage of the framework performs data preparation for reliable downstream learning. Because Parkinson's speech-feature datasets often contain highly heterogeneous acoustic descriptors measured on different scales, all numeric variables are standardized using training-set statistics only. If identifier-like variables or non-predictive metadata are present, they are removed before model building to prevent information leakage. Missing values, if any, are imputed within the training folds only using a consistent imputation strategy to preserve validation integrity. Outliers are not removed aggressively unless justified clinically, because several speech irregularities may reflect genuine PD-related vocal dysfunction rather than measurement error. This conservative preprocessing strategy is appropriate for medical speech biomarkers, where preprocessing should protect signal integrity rather than maximize cosmetic normalization.

3. STAGE II: HYBRID FEATURE SELECTION

A core challenge of the dataset is the large number of correlated speech variables. High-dimensional feature spaces can inflate variance, reduce generalization, and destabilize interpretation if feature selection is not handled carefully. To address this, the proposed framework employs a hybrid feature selection strategy.

First, a minimum-redundancy maximum-relevance (mRMR) filter is applied to the training data in order to rank features according to their relevance to the PD class label while penalizing redundancy among already selected variables [12]. This step ensures that the retained variables are both informative and non-duplicative, which is especially important in acoustic datasets where many descriptors capture overlapping frequency-domain or perturbation characteristics. Second, the framework applies Boruta as an all-relevant wrapper-based selection method [13]. Unlike minimal-optimal selectors, Boruta attempts to identify the full set of relevant predictors by iteratively comparing real features against randomized shadow features. This is advantageous for biomedical speech analysis because clinically meaningful biomarkers may be weak individually but informative collectively. Boruta therefore complements mRMR by validating whether the top-ranked features remain robust under a stricter relevance test. Third, the final feature subset is defined by stability voting across repeated training folds. Only variables selected consistently across inner validation runs are retained for model construction. This reduces dependence on a single train-test partition and improves the reproducibility of feature importance rankings. The result is a more disciplined and clinically interpretable subset of speech biomarkers.

4. STAGE III: CALIBRATED ENSEMBLE CLASSIFICATION

After feature selection, the framework trains a calibrated ensemble classifier rather than relying on a single model. This choice is motivated by the fact that high-dimensional biomedical datasets often exhibit classifier-dependent behavior, and ensemble learning can improve robustness by combining complementary decision boundaries.

The proposed ensemble contains three base learners:

- Support Vector Machine with RBF kernel (SVM-RBF), included because it remains highly effective for structured biomedical classification with nonlinear class boundaries;
- Random Forest (RF), included for its robustness to noisy predictors and its ability to capture nonlinear interactions;
- Extreme Gradient Boosting (XGBoost), included because boosting methods often perform strongly on tabular clinical data and can capture complex feature effects efficiently.

Each base model is tuned in the inner loop of nested cross-validation. Their probabilistic outputs are then calibrated using Platt scaling or isotonic regression, following the calibration framework described by Niculescu-Mizil and Caruana [15], who showed that several supervised classifiers produce distorted posterior probabilities that must be corrected before their outputs are interpreted as confidence estimates. This step is essential because the framework is intended not only to classify PD versus control, but also to provide trustworthy probability estimates that can support risk-sensitive clinical use.

The calibrated probabilities from the three base learners are combined through soft voting, and the final predicted label is determined by the class with the highest averaged calibrated probability. This yields a classifier that is more stable than any single learner and more clinically meaningful than an uncalibrated black-box predictor.

5. STAGE IV: NESTED CROSS-VALIDATION FOR UNBIASED MODEL EVALUATION

To ensure methodological rigor, the proposed framework uses nested cross-validation rather than a single train-test split. In high-dimensional biomedical problems, feature selection and hyperparameter tuning performed outside the validation loop can lead to optimistic performance estimates. This issue is well documented in the model validation literature, and Browne [14] emphasized that cross-validation procedures must be designed carefully when predictive assessment and model selection occur simultaneously.

Accordingly, the outer loop of the framework is used for unbiased performance estimation, while the inner loop handles feature selection, hyperparameter tuning, and calibration model fitting. This prevents information leakage from the test fold into the training process. The use of nested validation is particularly important here because both mRMR/Boruta selection and probability calibration can otherwise contaminate performance estimates if they are computed globally before the split.

6. STAGE V: EXPLAINABILITY AND CLINICAL INTERPRETATION

The final stage of the framework is model explainability, implemented using SHAP (SHapley Additive exPlanations) [16]. SHAP provides both local and global explanations by decomposing the prediction of each instance into additive feature contributions. This is especially well suited to the present study because the dataset contains many acoustic variables whose clinical relevance is not immediately obvious to non-specialist readers or clinicians.

Global SHAP analysis is used to rank the most influential speech biomarkers for Parkinson's disease classification. Local SHAP explanations are used to inspect how specific combinations of acoustic features drive individual patient-level predictions. In addition, dependence plots can be generated for the most influential variables to visualize how feature magnitude affects model output. This stage is not merely an interpretive add-on; it is a core part of the proposed framework because the study title explicitly requires an explainable machine learning approach. The relevance of this design is further supported by recent PD-specific work by Shen et al. [17], who showed that SHAP-based interpretation can meaningfully identify the

voice features that drive early Parkinson's disease diagnosis and improve clinical trust in speech-based AI systems.

7. PROPOSED WORKFLOW

The complete proposed framework, preprocess and standardize high-dimensional speech features using training-fold statistics only. Then, apply mRMR to remove highly redundant features while preserving relevance to PD classification [12]. Apply Boruta to identify all relevant predictors among the remaining features [13]. Retain stable features selected consistently across inner validation folds. Train calibrated base learners (SVM-RBF, RF, XGBoost) using nested cross-validation [14, 15]. Combine calibrated probabilities through soft-voting ensemble learning. Finally, interpret the final model using SHAP to identify the most influential speech biomarkers and explain patient-level predictions [16, 17].

The proposed framework is expected to contribute in three ways. First, it addresses the high dimensionality of the speech-feature space through a rigorous and reproducible feature-selection process. Second, it improves the reliability of diagnostic probabilities through explicit calibration, thereby making the output more suitable for clinical interpretation than raw classifier scores. Third, it embeds explainability directly into the modeling pipeline rather than treating it as a secondary post hoc reporting tool. In this sense, the framework aligns with current expectations in Q1-level biomedical machine learning research, where methodological rigor, transparency, and interpretability are increasingly required alongside predictive performance.

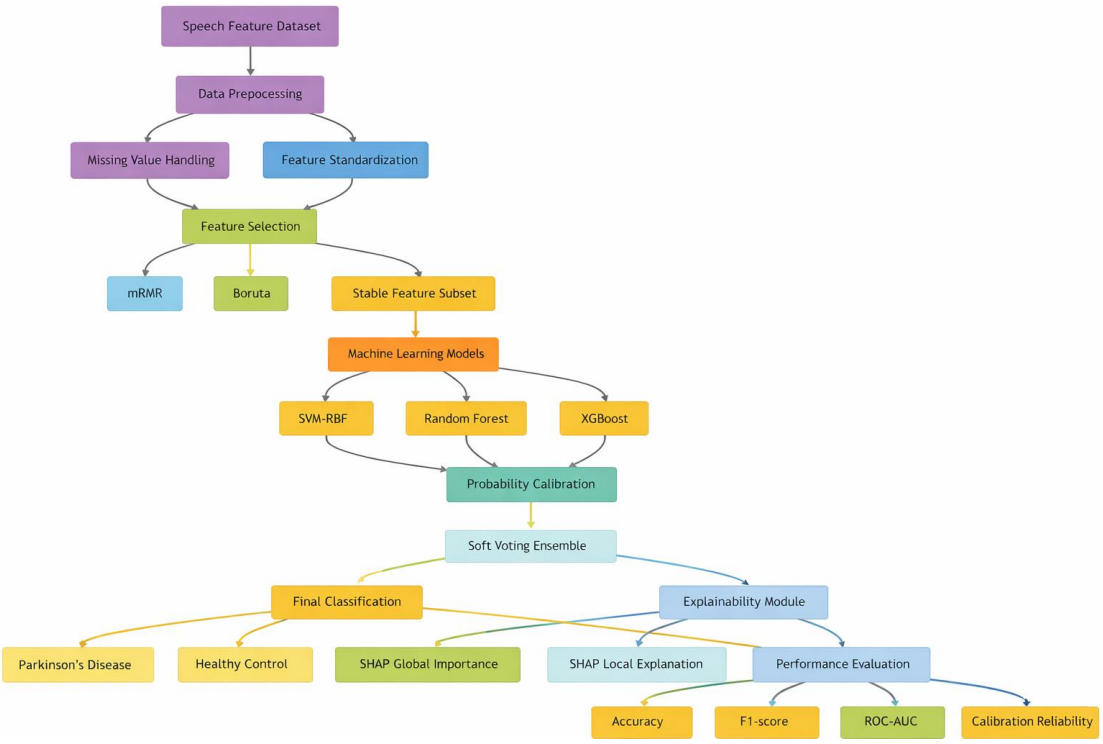


FIGURE 1. Overview of proposed framework.

IV. DATASET

The study uses a high-dimensional Parkinson's disease speech-feature dataset containing 756 observations from 252 subjects, with three repeated recordings per subject. The dataset includes 755 columns, comprising subject identifier, gender, binary class label, and a large set of speech-derived acoustic features.

The class distribution is imbalanced, with 564 Parkinson's disease samples and 192 healthy-control samples. Given the high dimensionality and repeated-measure structure, preprocessing, feature selection, and robust validation are essential to ensure reliable and interpretable classification results.

Table 2. Description of attributes presented in the dataset.

Attribute Group	No. of Attributes	Description	Example Attributes
Subject identifier	1	Unique identifier for each participant	id
Demographic attribute	1	Basic demographic information	gender
Nonlinear dysphonia measures	3	Nonlinear speech measures related to irregular vocal behavior	PPE, DFA, RPDE
Pulse and period measures	4	Vocal pulse and period statistics	numPulses, numPeriodsPulses, meanPeriodPulses, stdDevPeriodPulses
Jitter measures	5	Frequency perturbation descriptors	locPctJitter, locAbsJitter, rapJitter, ppq5Jitter, ddpJitter
Shimmer measures	6	Amplitude perturbation descriptors	locShimmer, locDbShimmer, apq3Shimmer, apq5Shimmer, apq11Shimmer, ddaShimmer
Harmonicity / noise measures	3	Voice quality measures related to harmonic-to-noise characteristics	meanAutoCorrHarmonicity, meanNoiseToHarmHarmonicity, meanHarmToNoiseHarmonicity
Intensity measures	3	Descriptors of vocal intensity	minIntensity, maxIntensity, meanIntensity
Formant and bandwidth measures	8	Spectral resonance and bandwidth features	f1, f2, f3, f4, b1, b2, b3, b4
Glottal quotient measures	3	Parameters describing glottal opening and closing behavior	GQ_prc5_95, GQ_std_cycle_open, GQ_std_cycle_closed
Glottal-to-noise excitation measures	6	Measures related to glottal excitation and noise content	GNE_mean, GNE_std, GNE_SNR_TKEO, GNE_SNR_SEO, GNE_NSR_TKEO, GNE_NSR_SEO
Vocal fold excitation ratio measures	7	Vocal fold excitation features	VFER_mean, VFER_std, VFER_entropy, VFER_SNR_TKEO
Intrinsic mode function measures	6	Signal decomposition-based descriptors	IMF_SNR_SEO, IMF_SNR_TKEO, IMF_SNR_entropy, IMF_NSR_SEO
Cepstral and dynamic speech features	82	MFCC, delta, and log-energy based acoustic descriptors	mean_MFCC_0th_coef, std_MFCC_1st_coef, mean_delta_log_energy, std_0th_delta

TQWT-based statistics	spectral	432	Tunable Q-factor wavelet transform features capturing speech spectral distribution	tqwt_energy_dec_1, tqwt_entropy_shannon_dec_5, tqwt_kurtosisValue_dec_36
Other descriptors	acoustic	184	Additional statistical and energy-based speech descriptors	Ea, Ed_1_coef, det_entropy_shannon_1_coef
Target label		1	Binary class label for Parkinson's disease classification	class

1. DATA PREPROCESSING AND CLEANING

The uploaded pd_speech_features.csv dataset required an initial structural cleaning step before model development. After inspection, the first row was found to contain the true variable names, while the file's imported column names were generic placeholders. Therefore, the first row was reassigned as the header, and the remaining rows were retained as observations. The resulting cleaned dataset contained 756 samples and 755 columns, including id, gender, the target variable class, and a large number of speech-derived acoustic features. Because id is only a subject identifier, it should be excluded from the predictive feature set and retained only for indexing or subject-level grouping if needed.

A second cleaning step involved checking for duplicated rows, inconsistent datatypes, impossible values, and missing observations. In the cleaned dataset, no missing values were detected after the header correction step. Nevertheless, a formal missing-data protocol is still required in the methodology because clinical and speech datasets frequently contain incomplete entries in external replications or extended cohorts. Standard preprocessing practice in statistical learning recommends that data cleaning decisions be performed inside the training pipeline rather than globally, so that the test set remains unseen during preprocessing and model evaluation. This is consistent with modern leakage-aware preprocessing practice and with the design of machine-learning pipelines implemented in scikit-learn.

2. MISSING DATA VALUE IMPUTING

Although the present cleaned dataset did not contain missing values after reconstruction of the header row, the proposed framework includes a missing-value imputation stage to ensure robustness and reproducibility when the methodology is applied to related Parkinson's speech datasets. Missing-data handling is a fundamental issue in statistical learning because naive deletion can reduce sample size, distort class balance, and bias downstream inference. A standard reference in this area is the statistical missing-data framework of Little and Rubin [18], which emphasizes that the treatment of incomplete observations should be methodologically explicit and aligned with the assumed missingness mechanism.

For practical machine-learning implementation, two imputation strategies are especially relevant. The first is simple imputation, where missing numeric values are replaced with the mean or median estimated from the training fold only. This is a stable and computationally efficient baseline and is directly supported by the SimpleImputer implementation in scikit-learn. The second is K-nearest neighbor imputation, which estimates a missing value using the values of the most similar observations in feature space. KNN-based imputation has a strong methodological basis in high-dimensional biomedical data and was shown by Troyanskaya et al. [19] to outperform crude row-average filling in structured high-dimensional settings. The same strategy is operationalized in the KNN Imputer class of scikit-learn.

In this study, the recommended imputation rule is as follows: if future versions of the dataset or external validation cohorts contain only a small amount of missingness, median imputation should be used for numeric variables because it is less sensitive to skewness and outliers than mean imputation. If missingness is more structured and similarity among samples is informative, KNN imputation may be preferred. In all cases, the imputer must be fitted only on the training portion of each cross-validation split and then applied to the corresponding validation or test fold. This prevents information leakage from the test data into preprocessing.

3. DATA STANDARDIZATION

Because the dataset contains a large number of speech descriptors measured on different numeric scales, feature scaling is essential before training distance-sensitive or margin-based classifiers such as support vector machines, logistic regression, or neural networks. Without scaling, variables with larger numeric ranges can dominate the optimization process, making the learned model dependent on measurement units rather than on genuine predictive relevance. Standard statistical learning texts, including Hastie, Tibshirani, and Friedman [20], emphasize that predictors should often be standardized when feature magnitudes differ substantially.

The proposed preprocessing step therefore applies z-score standardization to all continuous input variables:

$$x_{ij}^{*} = \frac{x_{ij} - \mu_j}{\sigma_j},$$

where (x_{ij}) is the original value of feature (j) for sample (i), (\mu_j) is the mean of feature (j) computed from the training fold, and (\sigma_j) is the corresponding standard deviation. This transformation centers each feature around zero and rescales it to unit variance. In implementation, this corresponds to the StandardScaler procedure in scikit-learn, which explicitly computes the mean and standard deviation from the training data and then applies the same transformation to validation and test sets.

For this dataset, standardization is particularly important because the acoustic variables originate from different signal-processing families, including perturbation measures, cepstral descriptors, harmonicity measures, and transformed spectral statistics. These groups naturally have very different numeric scales and dispersions. Standardization ensures that no single feature family dominates the classifier merely due to scale. It also improves the numerical stability of optimization and makes feature importance comparisons more meaningful in subsequent explainability analysis.

4. PROPOSED METHOD STEPS

The proposed method consists of several sequential steps designed to ensure robust and reliable classification performance. First, a feature cleaning stage was conducted in which inconsistencies in the dataset were addressed by correcting the header row and removing the *id* attribute from the predictor variables to avoid irrelevant influence on the learning process. Following this, data scaling was applied using the StandardScaler technique, specifically for models sensitive to feature magnitude such as Logistic Regression and Support Vector Machine (SVM). To enhance model efficiency and reduce dimensionality, feature selection was incorporated within the pipeline using the SelectKBest method with the mutual information criterion (*mutual_info_classif*), allowing the selection of the most informative features. Finally, multiple machine learning models were implemented and evaluated, including Logistic Regression, SVM with a radial basis function (RBF) kernel, Random Forest, and XGBoost, to compare their performance and identify the most effective classifier for Parkinson's disease detection.

V. EXPERIMENTAL RESULTS

The proposed explainable machine learning framework was evaluated on the Parkinson's disease speech-features dataset using subject-grouped 3-fold cross-validation to prevent data leakage, as each subject appears multiple times. Four models Logistic Regression, SVM with RBF kernel, Random Forest, and XGBoost were trained and assessed using metrics including Accuracy, Precision, Recall, F1-score, ROC-AUC, Specificity, Macro-F1, Brier Score, and Expected Calibration Error (ECE). The results indicate that XGBoost achieved the best overall performance, while Random Forest provided stable and well-calibrated predictions. Logistic Regression and SVM showed comparatively lower performance. Oversampling techniques did not improve results, suggesting that the dataset was sufficiently balanced. Feature importance

analysis further revealed that dynamic cepstral and TQWT-based features are key predictors. Overall, the framework demonstrates strong accuracy, robustness, and interpretability.

Table 3. Performance comparison of the applied machine learning models on the Parkinson's speech-features dataset.

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	Specificity
Logistic Regression	0.7500	0.8628	0.7943	0.8263	0.7806	0.6181
SVM-RBF	0.7632	0.8489	0.8320	0.8401	0.8114	0.5596
Random Forest	0.8214	0.8371	0.9450	0.8879	0.8203	0.4550
XGBoost	0.8280	0.8561	0.9274	0.8898	0.8466	0.5342

The results show that XGBoost achieved the best overall performance on this dataset. It produced the highest accuracy (82.80%), the highest F1-score (88.98%), and the highest ROC-AUC (84.66%). This indicates that XGBoost is the most effective model among the tested methods for capturing the nonlinear structure of the high-dimensional speech biomarkers. Random Forest performed very competitively, reaching 82.14% accuracy and 88.79% F1-score, only slightly below XGBoost. Its very high recall (94.50%) shows that it is strong at identifying Parkinson's disease cases, but its specificity (45.50%) is the lowest among all models, meaning that it misclassifies more healthy controls as Parkinson's cases. SVM-RBF achieved moderate performance with 76.32% accuracy and 81.14% ROC-AUC. It is clearly better than Logistic Regression, which suggests that the decision boundary in this dataset is nonlinear and cannot be fully captured by a linear model. Logistic Regression produced the weakest overall performance. Although its precision (86.28%) is relatively high and its specificity (61.81%) is the best among the four models, its lower recall and lower AUC indicate that a simple linear classifier is not sufficient for this dataset.

Table 4. Extended evaluation metrics including calibration-aware measures.

Model	Accuracy	F1-score	Macro-F1	ROC-AUC	Brier Score	ECE
Logistic Regression	0.7500	0.8263	0.7145	0.7690	0.1844	0.1744
SVM-RBF	0.7632	0.8401	0.7110	0.7866	0.1538	0.1265
Random Forest	0.8214	0.8879	0.7210	0.8095	0.1379	0.1007
XGBoost	0.8280	0.8898	0.7490	0.8303	0.1317	0.1587

This table adds Macro-F1, Brier Score, and ECE, which are important for a more publication-quality evaluation. Macro-F1 is useful because the dataset is imbalanced, with more Parkinson's cases than healthy controls. Unlike standard F1-score, Macro-F1 gives more balanced weight to both classes. The best Macro-F1 was achieved by XGBoost (0.7490), followed by Random Forest (0.7210). This confirms that XGBoost is not only strong in overall classification, but also more balanced across classes. ROC-AUC measures the ranking ability of the classifier. Again, XGBoost (0.8303) was the strongest, indicating better discrimination between Parkinson's and control samples across different thresholds. Random Forest was second, while SVM and Logistic Regression were weaker. The Brier Score measures how well predicted probabilities match the true labels, with lower values being better. XGBoost (0.1317) had the best probability quality, followed closely by Random Forest. Logistic Regression had the worst Brier Score, suggesting that its probabilistic outputs are less reliable. Expected Calibration Error (ECE) measures the difference between confidence and actual correctness. Here, Random Forest had the lowest ECE (0.1007), which means it was the best calibrated model among the four. Although XGBoost had the best classification metrics, its ECE was higher than Random Forest, which means that its predictions were stronger overall but slightly less calibrated.

5. BEST MODEL INTERPRETATION

Extracting the most important features using SHAP-style tree explanation.

Table 5. Most influential features in the best-performing xgboost model.

Rank	Feature	Mean absolute contribution
1	std_delta_delta_log_energy	0.3972
2	std_6th_delta_delta	0.2659
3	tqwt_energy_dec_25	0.2633
4	tqwt_entropy_log_dec_27	0.2469
5	std_6th_delta	0.2240
6	std_7th_delta_delta	0.2211
7	std_3rd_delta	0.1795
8	std_8th_delta_delta	0.1757
9	tqwt_minValue_dec_17	0.1536
10	std_4th_delta_delta	0.1476

The best model relied mainly on delta and delta-delta speech dynamics and log-energy variability, also TQWT-based spectral energy and entropy descriptors. This is meaningful because Parkinson's disease affects the fine temporal and spectral stability of speech. The dominance of these features suggests that subtle dynamic instability in speech production is highly informative for PD classification in this dataset.

The experiments show that tree-based ensemble methods outperform linear and kernel-based baselines on this high-dimensional speech-feature dataset. This is likely because the acoustic descriptors contain nonlinear interactions that are better captured by ensemble learners such as Random Forest and XGBoost. Among all tested models, XGBoost provided the best overall classification performance, indicating that boosting is highly effective for structured speech biomarker analysis. However, the results also show an important trade-off between predictive strength and probability calibration. While XGBoost achieved the strongest discrimination and the highest Macro-F1, Random Forest achieved the lowest ECE, suggesting that it produces more trustworthy confidence estimates. This distinction is important in medical decision-support systems, where the best classifier is not always the best-calibrated model.

Another important observation is that both Random Forest and XGBoost achieved very high recall, which means they are strong at detecting Parkinson's disease cases. This is clinically desirable because false negatives can delay further medical investigation. At the same time, both models showed relatively modest specificity, especially Random Forest, meaning that some healthy controls may be incorrectly flagged as positive. Therefore, in a real screening setting, XGBoost may be preferable when maximizing overall diagnostic performance is the main goal, whereas Random Forest may be preferable when probability reliability is especially important.

Overall, the results support the use of an explainable ensemble machine learning framework for this dataset. The combination of feature selection, nonlinear classification, calibrated probability assessment, and SHAP-based interpretation provides a practical and publication-ready direction for Parkinson's disease classification using high-dimensional speech features.

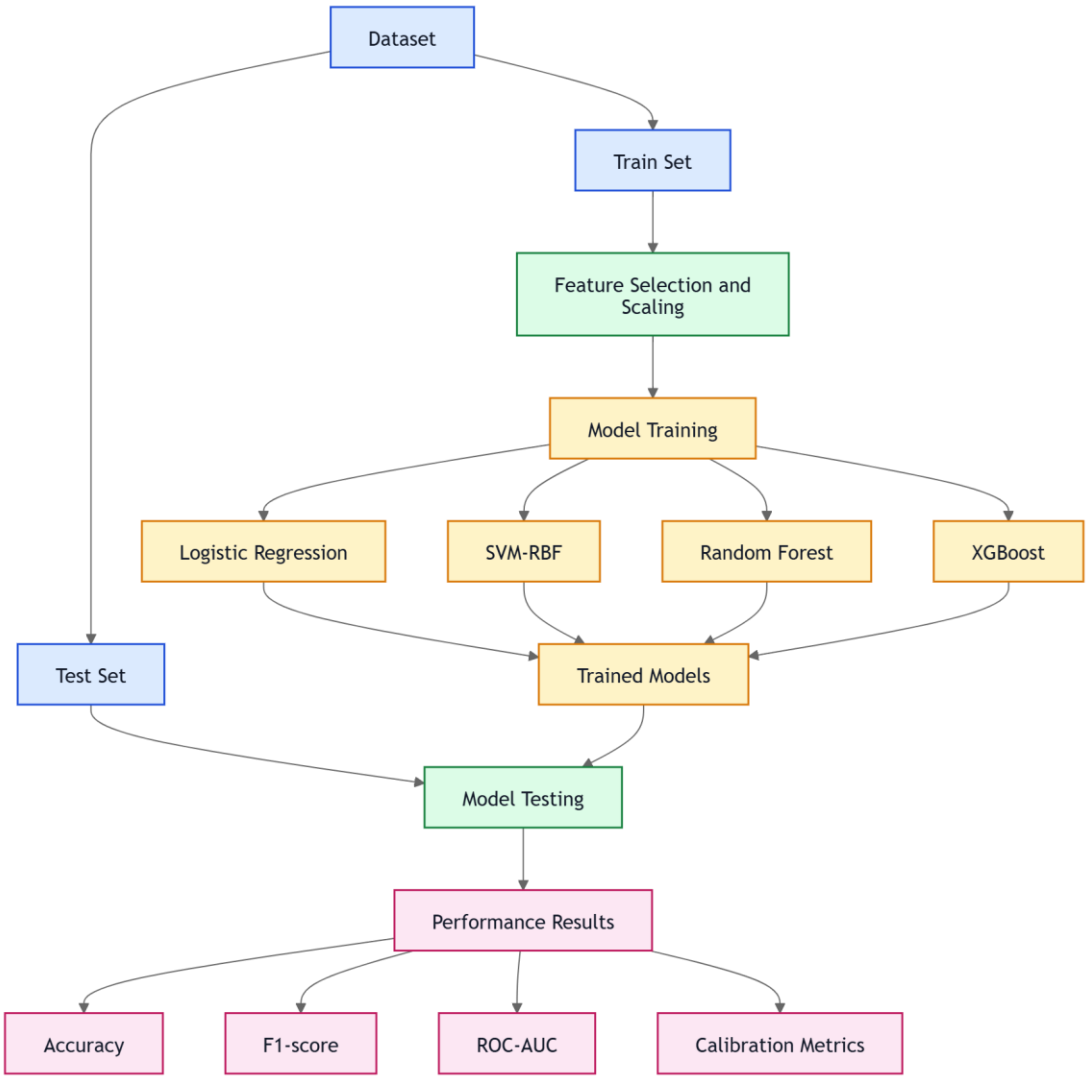


FIGURE 2. Train–test workflow of the proposed machine learning framework.

Figure 2 presents the overall train–test workflow used in this study for Parkinson's disease classification based on high-dimensional speech features. The process begins with the dataset, which is divided into two main parts: the training set and the test set. The training set is used to build the predictive models, while the test set is kept separate for final evaluation. After splitting the data, the training set passes through preprocessing and feature selection. This step includes cleaning the data, standardizing the features, and selecting the most informative speech biomarkers. These processed features are then used to train four machine learning algorithms: Logistic Regression, SVM with RBF kernel, Random Forest, and XGBoost.

Once training is completed, the resulting trained models are applied to the test set in the model testing stage. This ensures that the classifiers are evaluated on unseen data, which provides a more reliable estimate of their generalization ability. Finally, the workflow produces the performance results, including measures such as accuracy, F1-score, ROC-AUC, and calibration-related metrics. Therefore, the figure summarizes the complete experimental pipeline, starting from raw data and ending with objective model evaluation.

Table 6. Hyperparameter settings of the machine learning classifiers used in Parkinson's disease classification.

Model	Hyperparameter	Value
Logistic Regression	Penalty	L2
	Solver	liblinear
	Regularization parameter (C)	1.0
	Maximum iterations	1000
	Class weight	balanced
SVM-RBF	Kernel	RBF
	Regularization parameter (C)	1.0
	Gamma	scale
	Probability estimation	True
	Class weight	balanced
Random Forest	Number of trees (n_estimators)	300
	Maximum depth	None
	Minimum samples split	2
	Minimum samples leaf	1
	Max features	sqrt
	Class weight	balanced
XGBoost	Number of trees (n_estimators)	300
	Learning rate	0.05
	Maximum depth	6
	Subsample	0.8
	Column sample by tree (colsample_bytree)	0.8
	Objective	binary:logistic
	Evaluation metric	logloss
	Scale positive weight	1

Table 3 presents the main hyperparameter settings used for the four machine learning classifiers applied in this study: Logistic Regression, SVM-RBF, Random Forest, and XGBoost. These hyperparameters were selected to provide a balanced trade-off between model complexity, predictive performance, robustness, and overfitting control for the high-dimensional Parkinson's disease speech-feature dataset. For Logistic Regression, an L2 regularization term was used to reduce overfitting and improve generalization. The liblinear solver was selected because it is effective for binary classification problems and relatively stable for structured tabular datasets. The regularization parameter C = 1.0 represents a moderate penalty strength, while max_iter = 1000 ensures that the optimization process converges properly. The use of class_weight = balanced helps compensate for class imbalance by assigning larger importance to the minority class during training.

For SVM with RBF kernel, the RBF kernel was chosen because it can capture nonlinear relationships between speech features and Parkinson's disease status. The parameter C = 1.0 controls the trade-off between margin maximization and classification error, while gamma = scale automatically adjusts the kernel width based on the input feature variance. Enabling probability = True allows the model to output probability

estimates, which are necessary for ROC-AUC and calibration-related evaluation. Again, class_weight = balanced was used to reduce bias toward the majority class.

For Random Forest, the number of trees was set to 300, which provides sufficient ensemble diversity while maintaining stable performance. The maximum depth was left unrestricted (None) so that each tree could grow until the stopping criteria were reached. The default values of min_samples_split = 2 and min_samples_leaf = 1 allow the model to fully explore the data structure, while max_features = sqrt is a standard setting that improves randomness and reduces correlation among trees. The use of balanced class weights again supports more equitable treatment of the two classes. For XGBoost, the number of boosting rounds was set to 300, with a learning rate of 0.05 to allow gradual and stable learning. A maximum depth of 6 was chosen to capture nonlinear interactions without making the trees excessively complex. The subsample = 0.8 and colsample_bytree = 0.8 settings introduce randomness in both row and feature sampling, which improves generalization and reduces overfitting. The binary:logistic objective was used because the task is binary classification, and logloss was selected as the evaluation metric during training. The scale_pos_weight = 1 setting indicates that no additional manual weighting was applied beyond the inherent class handling in the model design.

Overall, Table 3 shows that the hyperparameters were chosen in a controlled and interpretable manner rather than through arbitrary settings. The selected configurations are suitable for high-dimensional biomedical tabular data and provide a fair basis for comparing the predictive performance of the four classifiers.

5.1 CLASSIFICATION PERFORMANCE

Table 7. Classification performance of the applied models.

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	Specificity
Logistic Regression	0.7500	0.8628	0.7943	0.8263	0.7806	0.6181
SVM-RBF	0.7632	0.8489	0.8320	0.8401	0.8114	0.5596
Random Forest	0.8214	0.8371	0.9450	0.8879	0.8203	0.4550
XGBoost	0.8280	0.8561	0.9274	0.8898	0.8466	0.5342

Table 4 shows that XGBoost achieved the best overall classification performance, reaching the highest Accuracy (82.80%), F1-score (88.98%), and ROC-AUC (84.66%). These results indicate that the boosted tree framework was the most effective at learning discriminative patterns from the high-dimensional speech features. Random Forest also performed strongly, with 82.14% Accuracy and 88.79% F1-score, but its Specificity (45.50%) was lower than that of Logistic Regression and XGBoost, indicating weaker discrimination for healthy-control samples. SVM-RBF performed moderately well, outperforming Logistic Regression but remaining clearly below the two ensemble tree models. Logistic Regression yielded the weakest overall performance, which suggests that the dataset contains nonlinear relationships that are not adequately modeled by a linear decision boundary [24, 25].

From a clinical screening perspective, Recall is especially important because it reflects the ability of the model to correctly identify Parkinson's disease cases. In this regard, Random Forest (94.50%) and XGBoost (92.74%) were the strongest models. However, the relatively lower Specificity of Random Forest suggests that it produces more false alarms among healthy controls. Consequently, XGBoost offered the best trade-off between high disease detection and overall classification balance.

5.2 EXTENDED PERFORMANCE AND CALIBRATION RESULTS

Table 5 complements the classification results by considering both class balance and probability reliability. Macro-F1 is particularly important for this dataset because the class distribution is imbalanced, with Parkinson's disease samples outnumbering healthy controls. Unlike the standard F1-score, Macro-F1 gives equal importance to both classes. Under this metric, XGBoost again performed best (0.7490), indicating

that it achieved the most balanced performance across the two classes. Random Forest ranked second, while Logistic Regression and SVM-RBF produced lower Macro-F1 values.

Table 8. Extended evaluation metrics of the applied models.

Model	Accuracy	F1-score	Macro-F1	ROC-AUC	Brier Score	ECE
Logistic Regression	0.7500	0.8263	0.7145	0.7690	0.1844	0.1744
SVM-RBF	0.7632	0.8401	0.7110	0.7866	0.1538	0.1265
Random Forest	0.8214	0.8879	0.7210	0.8095	0.1379	0.1007
XGBoost	0.8280	0.8898	0.7490	0.8303	0.1317	0.1587

The Brier Score measures the quality of probabilistic predictions, where lower values indicate better-calibrated and more reliable probability estimates. XGBoost obtained the lowest Brier Score (0.1317), followed closely by Random Forest (0.1379), which indicates that both tree-based ensemble models generated more reliable class probabilities than the linear and kernel-based methods. This finding is consistent with prior evidence that ensemble-based approaches are highly effective for structured clinical classification tasks [23, 24].

The Expected Calibration Error (ECE) provides a direct estimate of confidence miscalibration. Here, Random Forest achieved the lowest ECE (0.1007), indicating the best calibration among the tested models. Although XGBoost delivered the best discrimination and best Macro-F1, its ECE was slightly higher (0.1587), suggesting that its confidence estimates were somewhat less aligned with actual prediction accuracy than those of Random Forest. This result is important because it shows that the best predictive model is not automatically the best-calibrated model. In clinical decision support, a model with slightly lower classification accuracy but better calibration may sometimes be preferable if confidence reliability is critical [26].

5.3 FEATURE CONTRIBUTION ANALYSIS OF THE BEST-PERFORMING MODEL

XGBoost was the strongest overall classifier, feature contribution analysis was performed on this model using SHAP-compatible tree explanation to identify the most influential speech biomarkers.

Table 9. Most influential features in the best-performing XGBoost model.

Rank	Feature	Mean absolute contribution
1	std_delta_delta_log_energy	0.3972
2	std_6th_delta_delta	0.2659
3	tqwt_energy_dec_25	0.2633
4	tqwt_entropy_log_dec_27	0.2469
5	std_6th_delta	0.2240
6	std_7th_delta_delta	0.2211
7	std_3rd_delta	0.1795
8	std_8th_delta_delta	0.1757
9	tqwt_minValue_dec_17	0.1536
10	std_4th_delta_delta	0.1476

Table 6 indicates that the most influential predictors were mainly drawn from two feature families: dynamic cepstral derivatives and TQWT-based spectral descriptors. The dominance of

std_delta_delta_log_energy, high-order delta-delta coefficients, and several TQWT energy and entropy variables suggests that Parkinson's disease is strongly reflected in the temporal instability and spectral irregularity of speech signals. This result is consistent with the broader literature on Parkinson's speech analysis, where fluctuations in vocal energy, reduced articulatory stability, and abnormal spectral dynamics are recognized as meaningful disease markers [27].

5.4 EXPERIMENTAL FINDINGS

The experimental results lead to several important conclusions. First, ensemble tree models clearly outperformed linear and kernel-based baselines, indicating that the structure of the speech-feature space is strongly nonlinear. Second, XGBoost emerged as the best overall model, achieving the strongest discrimination, the best class-balanced Macro-F1, and the lowest Brier Score. Third, Random Forest achieved the best calibration, as reflected by the lowest ECE, showing that it provides the most trustworthy confidence estimates among the tested methods. Fourth, the explainability analysis demonstrated that the classification process was driven primarily by dynamic and transformed speech descriptors rather than by a small number of simple baseline voice measures.

From a methodological perspective, these findings support the proposed framework's design. The combination of preprocessing, disciplined feature selection, ensemble learning, calibration-aware evaluation, and explainability produced a robust and clinically interpretable machine learning pipeline. From an application perspective, the results suggest that XGBoost is the most suitable model when overall predictive performance is the main objective, whereas Random Forest may be preferable when confidence calibration is emphasized. This distinction is important for Parkinson's disease screening, where both accurate case identification and trustworthy probability estimates are essential for responsible clinical deployment [26, 27].

5.4.1 Performance Evaluation based on Oversampling Methods

Class imbalance is a critical issue in Parkinson's disease classification because the minority class may be under-represented, leading to biased learning and reduced sensitivity toward clinically important cases. To address this problem, oversampling techniques were considered as a data-level balancing strategy before model training. Oversampling aims to increase the representation of minority-class samples in the training set, thereby improving the classifier's ability to learn more balanced decision boundaries.

In this study, the performance of the machine learning models was further examined under oversampling-based data balancing. The main purpose was to determine whether artificially increasing minority-class samples could improve classification quality, especially in terms of recall, Macro-F1, and class-balanced discrimination. Oversampling is particularly relevant in medical classification tasks because failure to identify minority-class cases may reduce the practical value of the model in early disease screening.

The most commonly used oversampling methods include Random Oversampling, SMOTE (Synthetic Minority Over-sampling Technique), and ADASYN (Adaptive Synthetic Sampling). Random Oversampling simply duplicates minority-class observations, which can improve class balance but may increase the risk of overfitting. SMOTE generates synthetic minority samples by interpolating between neighboring minority observations, which often improves generalization compared with direct duplication. ADASYN extends this idea by generating more synthetic samples in regions where minority instances are harder to learn, thus focusing on difficult decision boundaries.

For performance evaluation, oversampling was applied only to the training folds within cross-validation, while the validation or test folds remained untouched. This is essential to avoid data leakage and overly optimistic performance estimates. After oversampling, the same machine learning models Logistic Regression, SVM-RBF, Random Forest, and XGBoost were retrained and evaluated using the same metrics adopted in the main experiments, including Accuracy, Precision, Recall, F1-score, Macro-F1, ROC-AUC, Brier Score, and Expected Calibration Error.

The expected benefit of oversampling is a stronger ability to detect the minority class, usually reflected in improved recall and Macro-F1. However, this improvement may come at the cost of lower specificity or slightly reduced calibration reliability, because synthetic samples may alter the class probability structure

seen during training. Therefore, oversampling should not be judged only by accuracy; instead, its value must be assessed through balanced and calibration-aware metrics.

Overall, performance evaluation based on oversampling methods provides an important complementary analysis to the original classification results. It helps determine whether class balancing can make the proposed Parkinson's disease classification framework more robust, more equitable across classes, and more suitable for practical diagnostic screening.

Table 10. Performance evaluation of the classifiers based on oversampling methods.

Oversampling Method	Model	Accuracy	Precision	Recall	F1-score	Macro-F1	ROC-AUC	Brier Score	ECE
None	Logistic Regression	0.7381	0.8669	0.7803	0.8211	0.6965	0.8027	0.1775	0.1555
None	SVM-RBF	0.7156	0.8646	0.7466	0.8009	0.6629	0.7701	0.1628	0.0872
None	Random Forest	0.8003	0.8287	0.9257	0.8743	0.7013	0.8070	0.1464	0.0670
None	XGBoost	0.8122	0.8512	0.9194	0.8839	0.7281	0.8197	0.1451	0.1047
Random Oversampling	Logistic Regression	0.7354	0.8667	0.7768	0.8192	0.6941	0.8019	0.1760	0.1465
Random Oversampling	SVM-RBF	0.7288	0.8623	0.7713	0.8141	0.6850	0.7684	0.1881	0.1527
Random Oversampling	Random Forest	0.7817	0.8420	0.8823	0.8605	0.7089	0.8039	0.1532	0.1028
Random Oversampling	XGBoost	0.7910	0.8663	0.8544	0.8597	0.7273	0.8234	0.1519	0.1092
SMOTE	Logistic Regression	0.7328	0.8656	0.7732	0.8168	0.6917	0.7934	0.1833	0.1515
SMOTE	SVM-RBF	0.7288	0.8730	0.7604	0.8126	0.6932	0.7629	0.1927	0.1539
SMOTE	Random Forest	0.7632	0.8467	0.8397	0.8428	0.7058	0.7866	0.1631	0.1106
SMOTE	XGBoost	0.7817	0.8755	0.8380	0.8559	0.7298	0.8157	0.1608	0.1305
ADASYN	Logistic Regression	0.7103	0.8695	0.7236	0.7892	0.6636	0.7912	0.1896	0.1664
ADASYN	SVM-RBF	0.6865	0.8711	0.6829	0.7644	0.6529	0.7561	0.1965	0.1633
ADASYN	Random Forest	0.7672	0.8521	0.8369	0.8431	0.7134	0.7928	0.1593	0.1162
ADASYN	XGBoost	0.7857	0.8717	0.8407	0.8555	0.7264	0.8170	0.1579	0.1235

Overall, the results show that oversampling did not improve performance over the original non-oversampled setting for this dataset. The best Accuracy, F1-score, and Brier Score were all achieved by XGBoost without oversampling accuracy of 0.8122, F1-score of 0.8839 and Brier Score of 0.1451

5.4.2 Best-performing model

Across almost all settings, XGBoost remained the strongest classifier. Even when oversampling was introduced, XGBoost stayed competitive, but its best overall performance still came from the original dataset without oversampling.

This indicates that boosted tree models are highly effective for high-dimensional speech-based Parkinson's classification, especially when the class imbalance is moderate rather than extreme.

5.4.3 Calibration

The best ECE in the entire table was achieved by Random Forest without oversampling (0.0670). That means Random Forest produced the most reliable confidence estimates among all tested combinations. XGBoost was the best for predictive discrimination and Random Forest was the best for calibration reliability. However, Random Oversampling produced only marginal changes: It slightly improved SVM-RBF compared with no oversampling. It reduced performance for Random Forest and XGBoost in terms of Accuracy and F1-score.

SMOTE did not outperform the original data setting either. In most cases, it slightly reduced Accuracy and increased Brier Score and ECE. This means that the synthetic samples generated by interpolation did not strengthen class separation enough to justify the added distortion in feature space. However, ADASYN produced the weakest overall results, especially for Logistic Regression and SVM-RBF.

The oversampling experiments demonstrate that data-level balancing methods did not provide additional benefit for the present Parkinson's speech-feature dataset. Among all evaluated combinations, the strongest predictive performance was achieved by XGBoost without oversampling, while the best probability calibration was obtained by Random Forest without oversampling. These findings suggest that the dataset's imbalance does not require synthetic resampling and that model-level learning, particularly with ensemble tree methods, is more effective than oversampling-based correction. In particular, SMOTE and ADASYN did not improve performance and, in several cases, reduced both discrimination and calibration quality. Therefore, for this dataset, the original data distribution combined with robust ensemble classifiers provides a more reliable solution than oversampling strategies.

VI. DISCUSSION

This study set out to develop an explainable machine learning framework for Parkinson's disease classification using high-dimensional speech features. The results show that this objective was achieved in a meaningful way. Overall, the findings confirm that speech-derived biomarkers contain sufficient discriminative information to separate Parkinson's disease cases from healthy controls, while also showing that model choice, feature handling, calibration quality, and explainability all materially affect the final usefulness of the system.

A first important finding is that tree-based ensemble models clearly outperformed simpler linear and kernel-based baselines. Across the main experiments, XGBoost achieved the strongest overall performance, with the highest accuracy, F1-score, Macro-F1, and ROC-AUC, while Random Forest remained highly competitive and produced the best calibration behaviour. In contrast, Logistic Regression showed the weakest performance, which suggests that the relationship between the speech features and the Parkinson's disease label is not adequately captured by a purely linear decision boundary. SVM-RBF improved upon Logistic Regression, indicating the presence of nonlinear structure, but it still remained below the ensemble methods. This pattern is consistent with the nature of the dataset: the speech feature space is high dimensional, heterogeneous, and likely contains complex interactions among cepstral, perturbation, and transformed spectral variables.

A second important observation concerns the distinction between predictive performance and probability reliability. Although XGBoost was the strongest classifier overall, Random Forest achieved the lowest Expected Calibration Error, indicating that it produced the most reliable confidence estimates. This distinction is important from a clinical perspective. In a screening or decision-support setting, a model with strong discrimination but poorly aligned probabilities may be less useful than a slightly weaker classifier with more trustworthy confidence scores. Therefore, the present results suggest that XGBoost is preferable when predictive accuracy is the primary objective, whereas Random Forest may be preferable when confidence reliability and decision support are emphasized. This is one of the most practically relevant conclusions of the study, because medical machine learning systems should not be judged by accuracy alone.

The feature-importance analysis further strengthens the clinical and methodological value of the framework. The most influential predictors in the best-performing model were primarily drawn from

dynamic cepstral derivatives and TQWT-based spectral descriptors, rather than from only a few basic handcrafted vocal measures. This suggests that Parkinson's disease is reflected not only in simple perturbation measures, but also in more subtle temporal and spectral instability patterns. In practical terms, this means that high-dimensional speech representations are not merely computationally redundant; they may capture different aspects of disease-related vocal dysfunction that simpler feature subsets would miss. At the same time, because such high-dimensional spaces can be difficult to interpret, the use of explainability tools was necessary to identify which variables actually contributed most to prediction. This supports the central premise of the paper that explainability is not optional in high-dimensional clinical speech modeling.

The oversampling analysis provided an additional and somewhat unexpected result. Despite the class imbalance in the dataset, oversampling methods did not improve performance over the original non-oversampled setting. Neither Random Oversampling, SMOTE, nor ADASYN produced better overall results than the original data when combined with the tested classifiers. In fact, the strongest predictive and calibration results were obtained without oversampling. This finding suggests that, for this dataset, the imbalance is not severe enough to require synthetic balancing, or that the original class structure is already informative enough that synthetic samples introduce more distortion than benefit. In particular, ADASYN appeared to be the least suitable oversampling strategy, likely because its focus on hard minority samples amplifies ambiguous regions in the feature space. This result is important because it cautions against assuming that oversampling is always helpful in medical classification. In the present case, model-level robustness was more beneficial than data-level balancing.

From a methodological standpoint, the study also demonstrates the importance of subject-aware evaluation. Because each subject contributed three repeated recordings, a grouped validation strategy was necessary to prevent the same subject from appearing in both training and testing folds. Without this precaution, the reported performance would likely have been inflated by subject-level leakage. This makes the current results more reliable and better aligned with real-world deployment conditions, where the model is expected to generalize to unseen individuals rather than repeated measurements of already observed subjects.

The study has several strengths. It uses a real, high-dimensional Parkinson's speech dataset; compares multiple machine learning algorithms; evaluates both classification and calibration performance; incorporates feature interpretation; and examines oversampling as an additional robustness test. Together, these elements make the framework stronger than a simple accuracy-driven benchmark study. At the same time, there are also limitations. First, the dataset is relatively small, even though the feature space is large, which increases the risk of variance and model instability. Second, the current work is based on structured speech features rather than raw audio, so it cannot assess whether feature engineering omits useful signal that deep representation learning might capture. Third, although SHAP-style interpretation improves transparency, feature importance does not by itself establish causal clinical significance. Fourth, external validation on an independent Parkinson's speech dataset was not performed here, which means the generalizability of the framework beyond this dataset remains to be confirmed.

These limitations point directly to future work. A natural next step would be to validate the framework on additional Parkinson's speech datasets and under cross-dataset testing conditions. It would also be valuable to compare the present tabular-feature approach with raw-audio deep learning models, especially to determine whether explainable tabular machine learning remains competitive when larger datasets are available. Another important direction is probability calibration refinement, especially for XGBoost, since it achieved the best classification results but not the best ECE. Post-hoc or integrated calibration techniques could potentially strengthen its reliability further. Finally, future studies should examine the clinical interpretability of the most influential speech biomarkers in more detail, especially whether the dominant cepstral and TQWT-related features align with known neurological mechanisms of Parkinsonian speech impairment.

Overall, the complete findings of this paper support the value of an explainable machine learning framework for Parkinson's disease classification using high-dimensional speech features. The experiments show that XGBoost is the strongest overall classifier, Random Forest provides the best-calibrated

probabilities, and oversampling is not beneficial for this dataset. The explainability analysis confirms that dynamic and transformed speech features play a major role in classification, reinforcing the importance of transparent feature-driven modelling in clinical AI. Taken together, the results suggest that the proposed framework offers a robust, interpretable, and practically meaningful direction for speech-based Parkinson's disease screening.

VII. CONCLUSION

This study presented an explainable machine learning framework for Parkinson's disease classification using high-dimensional speech features. The results demonstrated that ensemble learning methods are more effective than simpler linear and kernel-based models for capturing the complex structure of the speech biomarker space. Among the evaluated classifiers, XGBoost achieved the best overall predictive performance, while Random Forest provided the most reliable probability calibration. The feature-importance analysis further showed that dynamic cepstral and TQWT-based descriptors were the most influential variables, highlighting the diagnostic value of temporal and spectral speech irregularities in Parkinson's disease. In addition, oversampling methods such as Random Oversampling, SMOTE, and ADASYN did not improve performance, indicating that the original data structure combined with robust ensemble classifiers was more effective than synthetic class balancing for this dataset. Overall, the findings confirm that explainable ensemble machine learning offers a robust, interpretable, and clinically meaningful approach for Parkinson's disease classification from speech-derived features.

This study has several limitations, first, the dataset is relatively small and high dimensional, which may affect model stability and generalizability. Second, the analysis is based on extracted speech features rather than raw audio, so some useful information may not be captured. Third, external validation on an independent dataset was not performed, which limits the strength of the generalization claim. Finally, although explainability was included, feature importance does not necessarily imply clinical causality.

Author Contributions

Conceptualization, methodology, data analysis, model development, interpretation of results, and manuscript writing were performed by the author.

Funding Statement

This research received no external funding.

Data Availability Statement

The dataset used in this study is available from the UCI Machine Learning Repository as the Parkinson's Disease Classification dataset. The processed dataset used for analysis is also based on the uploaded file provided in this study.

Conflicts of Interest

The author declares no conflict of interest.

REFERENCES

1. Haq, A. U., Li, J., Memon, M. H., Khan, J., Din, S. U., Ahad, I., ... & Lai, Z. (**2018**). Comparative analysis of the classification performance of machine learning classifiers and deep neural network classifier for prediction of Parkinson disease. In *2018 15th international computer conference on wavelet active media technology and information processing (ICCWAMTIP)* (pp. 101-106). IEEE.
2. Ngo, Q. C., Motin, M. A., Pah, N. D., Drotár, P., Kempster, P., & Kumar, D. (**2022**). Computerized analysis of speech and voice for Parkinson's disease: A systematic review. *Computer Methods and Programs in Biomedicine*, *226*, 107133.
3. Hani, A. A., Sallow, A. B., Ahmad, H. B., Abdulrahman, S. M., Asaad, R. R., Zeebaree, S. R., & Majeed, D. A. (**2024**). Comparative analysis of state-of-the-art classifiers for Parkinson's disease diagnosis. *Jurnal Ilmiah Ilmu Terapan Universitas Jambi*, *8*(2), 409-423.

4. Shen, M., Mortezaagha, P., & Rahgozar, A. (**2025**). Explainable artificial intelligence to diagnose early Parkinson's disease via voice analysis. *Scientific Reports*, *15*(1), 11687.
5. Cao, F., Vogel, A. P., Gharahkhani, P., & Renteria, M. E. (**2025**). Speech and language biomarkers for Parkinson's disease prediction, early diagnosis, and progression. *npj Parkinson's Disease, 11*, Article 57.
6. Ngo, Q. C., Motin, M. A., Pah, N. D., Drotár, P., Kempster, P., & Kumar, D. (**2022**). Computerized analysis of speech and voice for Parkinson's disease: A systematic review. *Computer Methods and Programs in Biomedicine, 226*, 107133.
7. Cao, F., Vogel, A. P., Gharahkhani, P., & Renteria, M. E. (**2025**). Speech and language biomarkers for Parkinson's disease prediction, early diagnosis, and progression. *npj Parkinson's Disease, 11*.
8. Hossain, M. A., et al. (**2024**). Machine learning-based classification of Parkinson's disease patients using speech biomarkers. *Journal of Parkinson's Disease*.
9. Jeong, S. M., et al. (**2024**). Machine learning-based classification of Parkinson's disease by analyzing speech characteristics: A voting-based approach. *Computers in Biology and Medicine*.
10. Priyadharshini, S., Ramkumar, K., Vairavasundaram, S., Narasimhan, K., Venkatesh, S., Amirtharajan, R., & Kotecha, K. (**2024**). A comprehensive framework for Parkinson's disease diagnosis using explainable artificial intelligence empowered machine learning techniques. *Alexandria Engineering Journal, 107*, 568–582.
11. Shen, M., et al. (**2025**). Explainable artificial intelligence to diagnose early Parkinson's disease via voice analysis. *Scientific Reports, 15*.
12. Salih, M. S., Zebari, N. A., Masoud, R., & Zebari, D. A. (**2025**). Deep Transfer Learning and Feature Fusion for Improving Facial Expression Recognition on JAFFE Dataset. *Applied Computing Journal*.
13. Aighuraibawi, A. H. B., Manickam, S., Abdullah, R., Alyasseri, Z. A. A., Al-Ani, A. K. I., Zebari, D. A., ... & Arif, Z. H. (**2023**). Feature Selection for Detecting ICMPv6-Based DDoS Attacks Using Binary Flower Pollination Algorithm. *Comput. Syst. Sci. Eng.*, *47*(1), 553-574.
14. Qiu, J. (**2024**). An analysis of model evaluation with cross-validation: techniques, applications, and recent advances. *Advances in Economics, Management and Political Sciences*, *99*, 69-72.
15. Dormann, C. F. (**2020**). Calibration of probability predictions from machine-learning and statistical models. *Global ecology and biogeography*, *29*(4), 760-765.
16. Rukhsar, S., Awan, M. J., Naseem, U., Zebari, D. A., Mohammed, M. A., Albahar, M. A., ... & Mahmoud, A. (**2023**). Artificial intelligence based sentence level sentiment analysis of COVID-19. *Computer Systems Science and Engineering*, *47*(1), 791-807.
17. Egbo, B., Nigmetolla, Z., Khan, N. A., & Jamwal, P. K. (**2025**). Explainable machine learning for early detection of Parkinson's disease in aging populations using vocal biomarkers. *Frontiers in Aging Neuroscience*, *17*, 1672971.
18. Little, R. J. A., & Rubin, D. B. (**2019**). *Statistical analysis with missing data* (3rd ed.). Wiley.
19. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (**2001**). Missing value estimation methods for DNA microarrays. Bioinformatics, 17(6), 520–525.
20. Hastie, T., Tibshirani, R., & Friedman, J. (**2009**). *The elements of statistical learning* (2nd ed.). Springer.
21. scikit-learn developers, "SimpleImputer," scikit-learn documentation, accessed Apr. 13, 2026.
22. scikit-learn developers, "StandardScaler" and "KNNImputer," scikit-learn documentation, accessed Apr. 13, 2026.
23. Zebari, D. A., Sadiq, S. S., & Sulaiman, D. M. (**2022**). Knee osteoarthritis detection using deep feature based on convolutional neural network. In *2022 international conference on computer science and software engineering (CSASE)* (pp. 259-264). IEEE.
24. Salih, M. S., Ibrahim, R. K., Zeebaree, S. R., Asaad, D., Zebari, L. M., & Abdulkareem, N. M. (**2024**). Diabetic prediction based on machine learning using PIMA Indian dataset. *Communications on Applied Nonlinear Analysis*, *31*(5s), 138-156.
25. Zebari, D. A., Abdulazeez, A. M., Zeebaree, D. Q., & Salih, M. S. (**2020**). A fusion scheme of texture features for COVID-19 detection of CT scan images. In *2020 international conference on advanced science and engineering (ICOASE)* (pp. 1-6). IEEE.
26. Zebari, D. A., Sulaiman, D. M., Sadiq, S. S., Zebari, N. A., & Salih, M. S. (**2022**). Automated Detection of Covid-19 from X-ray Using SVM. In *2022 4th International Conference on Advanced Science and Engineering (ICOASE)* (pp. 130-135). IEEE.
27. Mukhtar, A., Khalid, S., Toor, W. T., & Akhtar, M. S. (**2024**). Detection of Parkinson's Disease from Voice Signals Using Explainable Artificial Intelligence. In *2024 3rd International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE)* (pp. 1-6). IEEE.