

DEEP NEURAL FEATURE LEARNING FOR PARKINSON'S DISEASE DIAGNOSIS FROM TABULAR SPEECH DESCRIPTORS

Abeda Muhammad Iqbal¹, Paryati, S.T., M. Kom²

¹ Faculty of Economics, Business and Accounting, ICATS University College, Kuching, Malaysia;

² Faculty of Technical Engineering, Department of Technical Informatics, University Development National "Veteran" Yogyakarta, Yogyakarta, Indonesia;

Corresponding author: miabeda@icats.edu.my

ABSTRACT: Parkinson's disease (PD) is a progressive neurological disorder that can affect speech production through abnormalities in phonation, articulation, and vocal stability. This study proposes a machine learning framework for PD diagnosis using tabular speech descriptors extracted from a benchmark speech dataset. The framework includes data cleaning, missing-value inspection, feature standardization, and leakage-safe patient-wise validation to ensure reliable performance estimation. Three machine learning classifiers, namely XGBoost, LightGBM, and Support Vector Machine with Radial Basis Function (SVM-RBF), were applied and compared. Experimental evaluation was conducted using patient-wise cross-validation and assessed through multiple performance metrics, including accuracy, precision, recall, F1-score, ROC-AUC, balanced accuracy, and Matthews correlation coefficient. The results showed that LightGBM achieved the best overall performance, with an accuracy of 84.39%, precision of 85.50%, recall of 95.40%, F1-score of 90.14%, and ROC-AUC of 87.69%, outperforming XGBoost and SVM-RBF. Extended evaluation further confirmed that LightGBM provided the most balanced classification behavior, while SVM-RBF achieved the lowest false-negative rate but lower specificity. The findings indicate that boosting-based models are highly effective for detecting Parkinson's disease from structured speech features and that leakage-safe validation is essential to obtain credible diagnostic results. The proposed framework offers a robust and reproducible approach to speech-based PD diagnosis and provides a practical foundation for future research on interpretable, externally validated clinical decision-support systems.

Keywords: Parkinson's Disease, Speech Analysis, Machine Learning, LIGHTGBM, XGBOOST, SVM-RBF, Classification.

I. INTRODUCTION

Parkinson's disease (PD) is a progressive neurodegenerative disorder characterized by motor and non-motor symptoms that substantially affect quality of life and long-term functional independence. Because clinical diagnosis can be difficult in early stages and may depend on specialist evaluation, there is growing interest in computational biomarkers to support earlier, scalable, and less invasive screening strategies. Among these biomarkers, speech has emerged as a particularly promising modality because PD commonly affects phonation, articulation, prosody, and temporal organization of speech, often before more obvious motor impairment becomes clinically apparent [1], [2].

Recent advances in digital health and machine learning have strengthened the role of speech analysis in PD research. Acoustic and dysphonia-related descriptors extracted from sustained vowels, words, and running speech can capture subtle neuromotor abnormalities associated with the disease, enabling automated discrimination between healthy controls and patients with PD [2], [3]. In this study, the selected

data source is the Parkinson's Speech Dataset with Multiple Types of Audio Recordings, which contains a high-dimensional set of tabular speech descriptors derived from multiple recording tasks. This type of dataset is well suited to machine learning because it represents speech through structured numerical attributes rather than raw waveforms, thereby enabling robust experimentation with advanced tabular classifiers [3].

Although deep learning has attracted considerable attention in biomedical prediction problems, evidence from tabular-data benchmarks shows that tree-based gradient boosting methods and carefully tuned kernel machines often remain highly competitive, especially when the sample size is moderate and the number of handcrafted features is large [4]. This observation is highly relevant for PD speech datasets, where the number of descriptors can be very high relative to the number of patients and where feature interactions may be nonlinear and complex. Therefore, instead of relying on generic deep neural architectures, this work emphasizes powerful and well-established machine learning algorithms that are more naturally aligned with structured tabular speech features [4].

Based on these considerations, this study focuses on three strong classifiers: Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Support Vector Machine with Radial Basis Function (SVM-RBF). XGBoost is a scalable tree-boosting framework designed to achieve high predictive accuracy while controlling model complexity through regularization and efficient split optimization [5]. LightGBM is an efficient gradient boosting decision tree algorithm that reduces computational cost while preserving strong predictive performance, making it attractive for high-dimensional datasets [6]. In contrast, SVM with an RBF kernel remains a strong baseline for biomedical classification because it can model nonlinear class boundaries effectively in complex feature spaces [7]. Together, these algorithms provide a balanced comparison between boosting-based ensemble learning and margin-based nonlinear classification.

Accordingly, the main objective of this study is to develop a reliable machine learning framework for PD diagnosis from tabular speech descriptors using leakage-safe preprocessing and strong nonlinear classifiers. By combining appropriate data cleaning, feature-oriented preprocessing, and rigorous model comparison, the study aims to determine which of these methods is most suitable for speech-based Parkinson's disease classification and to provide a reproducible benchmark for future work in computational neurology and digital speech biomarkers [2], [4].

II. RELATED WORK

Speech-based analysis has become one of the most active non-invasive directions for Parkinson's disease (PD) detection because motor impairment in PD frequently affects phonation stability, articulation, and prosodic control. Recent review studies show that machine learning research in this area has increasingly focused on voice, speech, and language biomarkers extracted from sustained vowels, words, and sentence-level recordings to support objective and scalable screening tools for PD diagnosis [8], [9]. These reviews also indicate that the field has moved from small proof-of-concept experiments toward more systematic comparisons of feature sets, learning algorithms, and clinical interpretability, although challenges related to reproducibility, dataset size, and validation design remain significant [8], [9].

A substantial body of earlier work relied on handcrafted acoustic descriptors such as jitter, shimmer, harmonicity-related measures, pitch statistics, and nonlinear dysphonia indicators. The Parkinson's Speech Dataset with Multiple Types of Sound Recordings introduced by Sakar et al. remains one of the most important benchmarks in this domain because it includes multiple speech tasks and a structured set of acoustic features for classification and regression analysis [10]. The dataset contains recordings from Parkinson's patients and healthy controls across sustained vowels, numbers, words, and short sentences, which makes it suitable for studying how different vocal tasks contribute to disease discrimination [10].

Several studies have shown that conventional machine learning models remain highly competitive for PD speech classification, particularly when the data are represented as tabular acoustic descriptors rather than raw audio. For example, recent work on multilingual acoustic speech tasks reported that carefully designed preprocessing and voting-based machine learning can achieve strong diagnostic performance, reaching an accuracy of 84.73% and an AUC of 92.18% on a balanced Korean speech dataset [11]. That study also emphasized the usefulness of compact acoustic representations and straightforward preprocessing,

suggesting that robust classical pipelines can still provide effective diagnostic performance without necessarily requiring large end-to-end deep learning systems [11].

Support Vector Machine (SVM), especially with the radial basis function kernel, has long been one of the most widely used classifiers in PD voice analysis because of its ability to model nonlinear decision boundaries in relatively small and high-dimensional datasets. Prior studies cited in recent PD speech research have repeatedly employed SVM as a strong baseline, often obtaining competitive classification results when feature engineering and ranking strategies are used to reduce redundancy and enhance class separability [11], [12]. The continued use of SVM in more recent work reflects its robustness in biomedical tabular classification, where the sample size is often limited, and feature interactions are not strictly linear [11], [12].

In parallel, ensemble tree methods have gained increasing attention because they can capture nonlinear feature interactions, handle heterogeneous predictor importance, and provide useful measures of feature relevance. Recent interpretable machine learning work on PD speech diagnosis reported that Random Forest and Gradient Boosting models achieved the strongest results among the tested methods, with an AUC-ROC of 0.98 and a recall of 0.95, and highlighted jitter, shimmer, NHR, RPDE, and PPE as highly informative predictors [12]. Such findings support the growing view that boosting-based models are especially suitable for acoustic biomarker analysis, where discriminative information is often distributed across many correlated descriptors [12].

Among boosting methods, XGBoost has become particularly prominent due to its regularized tree-boosting formulation and strong empirical performance on structured datasets. More recent comparative studies and surveys in PD-related speech and voice analysis identify gradient boosting methods as increasingly favored choices when the objective is high predictive accuracy on engineered acoustic features [8], [12]. LightGBM has also emerged as an attractive alternative because it maintains the strengths of gradient boosting while improving training efficiency and scalability, which can be advantageous when working with high-dimensional speech descriptors [11]. Although CatBoost has also been discussed in related literature, XGBoost and LightGBM are especially relevant for datasets composed primarily of numerical acoustic variables [11].

Despite the promising results reported in prior studies, several limitations remain common across the literature. Recent reviews note persistent issues such as limited external validation, heterogeneous feature extraction procedures, insufficient interpretability, and the risk of overfitting when complex models are trained on relatively small biomedical datasets [8], [9]. In addition, when multiple recordings are obtained from the same individual, patient-level grouping becomes methodologically important because naive row-wise data splitting can lead to information leakage and overly optimistic performance estimates. This concern is particularly relevant for benchmark PD speech datasets that include repeated measurements per subject [10].

Therefore, the present study aligns with the line of work that favors strong, well-validated tabular classifiers for structured speech descriptors. Building on prior evidence, this study focuses on XGBoost, LightGBM, and SVM with RBF kernel as three complementary and well-established approaches for PD diagnosis from speech-derived tabular features. This selection is justified by the repeated success of margin-based learning and gradient boosting in prior PD speech studies, as well as by their suitability for nonlinear, high-dimensional biomedical data [11], [12].

Table 1. Comparative summary of representative studies used in related work.

Ref. Study	Data / Modality	Main Methods	Main Contribution / Finding	Relevance to the present study
[8] Hossain <i>et al.</i> (2025), systematic review	Voice, speech, and language biomarkers for Parkinson's disease	Review of ML-based diagnostic studies	Summarized current evidence on AI/ML for PD diagnosis from speech and language signals, identified common datasets, models, and methodological	Supports the need for a rigorously validated tabular-speech classification framework.

Ref. Study	Data / Modality	Main Methods	Main Contribution / Finding	Relevance to the present study
[9] Recent review on speech and language biomarkers for PD (2025)	Speech and language biomarkers	Narrative/review synthesis	trends. Reported that speech-based PD diagnosis is promising but still limited by dataset size, validation heterogeneity, and reproducibility concerns. Highlighted the clinical promise of speech-derived biomarkers for early PD detection and emphasized challenges related to standardization, interpretability, and generalizability across studies.	Justifies using explainable and well-established machine learning models instead of relying only on black-box methods.
[10] Sakar <i>et al.</i> (2013)	Parkinson's speech dataset with multiple recording types	Acoustic feature extraction from sustained vowels, numbers, words, and short sentences	Introduced a benchmark dataset containing multiple types of sound recordings and tabular speech descriptors from PD and healthy participants. The UCI summary notes 26 voice samples per subject across different speaking tasks and extracted linear and time-frequency features.	Provides the benchmark data basis for many later studies and matches the tabular-speech setting of the present work.
[11] Jeong <i>et al.</i> (2024)	Multilingual speech tasks with acoustic features	Machine learning classification with acoustic descriptors	Reported that machine-learning-based acoustic analysis can achieve strong PD classification performance and emphasized the usefulness of multilingual speech tasks and careful preprocessing. The article summary reports accuracy of 84.73% and AUC of 92.18% on a balanced Korean speech dataset.	Shows that structured acoustic features remain effective for PD diagnosis and supports using strong tabular classifiers.
[12] Xu <i>et al.</i> (2025)	Speech recordings and acoustic features	Interpretable machine learning	Developed an interpretable model for non-invasive PD detection from speech, with the abstract emphasizing both prediction and explainability using acoustic features.	Reinforces the importance of combining high predictive performance with interpretability in speech-based PD screening.
[13] Lahmiri (2019)	Voice patterns for PD detection	Feature ranking + optimized SVM	Used ranked voice features with an optimized SVM for PD detection, illustrating the continued strength of nonlinear margin-based classification on acoustic descriptors. This study is listed among the representative	Directly supports including SVM with RBF kernel as a core benchmark in the present study.

Ref. Study	Data / Modality	Main Methods	Main Contribution / Finding	Relevance to the present study
[14] Gunduz (2021)	PD classification from tabular features	Filter-based feature selection + variational autoencoder	prior works cited in later PD speech comparisons. Proposed a dimensionality-reduction strategy for PD classification, showing that high-dimensional biomedical feature sets benefit from feature selection or learned compression before classification. This study is listed among representative prior works in later comparative PD literature.	Supports the need for careful preprocessing and dimensionality reduction in high-dimensional speech-descriptor datasets.

III. PROPOSED FRAMEWORK

The proposed framework is designed for Parkinson's disease diagnosis from tabular speech descriptors and is tailored to the structure of the selected benchmark dataset. The dataset contains multiple voice recordings collected from each subject across several speaking tasks, with acoustic descriptors extracted from each recording. Because repeated recordings belong to the same individual, the framework is designed to prevent patient-level information leakage and to produce a more reliable estimate of diagnostic performance [15], [19], [20].

1. DATA ACQUISITION AND TARGET DEFINITION

The first stage uses the Parkinson's Speech Dataset with Multiple Types of Audio Recordings, which provides structured numerical speech descriptors derived from sustained vowels, numbers, words, and short sentences. Each sample is associated with a class label indicating whether it belongs to a healthy control or a Parkinson's disease case. Since the dataset is already organized in tabular form, it is suitable for machine learning models specialized for structured numerical data rather than raw-audio deep architectures [15], [19].

2. DATA CLEANING AND CONSISTENCY CHECKING

Before model development, the dataset undergoes a cleaning stage that includes removing duplicate rows, verifying class labels, checking for invalid or infinite numerical values, and excluding the subject identifier from the predictor matrix. This step is necessary because the identifier should be used only for grouping during validation, not as an input feature. The framework also performs feature-wise inspection to confirm that numerical attributes fall within a valid range and are suitable for downstream modeling. This stage ensures that the learning process is based only on meaningful acoustic descriptors rather than structural artifacts in the data [15].

3. LEAKAGE-SAFE PATIENT-WISE DATA SPLITTING

A central part of the framework is group-aware validation. Since each subject contributes multiple recordings, ordinary random splitting can place recordings from the same person in both training and test sets, which would produce overly optimistic results. To avoid this, the proposed framework uses GroupKFold, where each patient appears in only one fold at a time. In addition, hyperparameter optimization is performed in an inner loop, while the final performance estimate is obtained in an outer loop using nested cross-validation. This design follows recommended practice because nested cross-validation reduces the bias introduced by model selection and tuning on the same data [20], [21].

4. FEATURE SCALING AND PREPROCESSING

After splitting, preprocessing is performed only on the training partition of each fold and then applied to the corresponding validation/test partition. Standardization is included because algorithms such as SVM with RBF kernel are sensitive to the relative scale of input variables. StandardScaler transforms each feature by removing the training-set mean and scaling by the training-set standard deviation, which improves numerical stability and supports fair comparison across models [22]. Tree-based boosting models do not strictly require scaling, but applying a consistent preprocessing pipeline improves methodological clarity and reproducibility, especially in multi-model comparisons [22].

5. MODEL LEARNING USING THREE COMPLEMENTARY CLASSIFIERS

The classification stage employs three strong and widely used machine learning algorithms for structured tabular data:

XGBoost is used as the first classifier because it is a regularized gradient-boosting framework that combines strong predictive power with efficient split optimization and built-in control over model complexity. It is especially suitable for structured biomedical datasets with nonlinear feature interactions [16].

LightGBM is used as the second classifier because it is an efficient gradient-boosting decision-tree method designed to accelerate training while maintaining competitive predictive accuracy. Its design is particularly suitable for high-dimensional numerical datasets, making it a strong candidate for speech-descriptor classification [17].

SVM with an RBF kernel is used as the third classifier because it remains one of the most reliable approaches for nonlinear classification on relatively small, high-dimensional biomedical datasets. The RBF kernel enables the decision surface to model complex class boundaries that are often present in speech-derived features [18].

Together, these three models form a balanced comparison between boosting-based ensemble learning and kernel-based nonlinear classification, which is appropriate for Parkinson's disease prediction from structured speech attributes [16]–[18].

6. HYPERPARAMETER OPTIMIZATION

Within the inner cross-validation loop, the framework tunes the principal hyperparameters of each classifier. For XGBoost and LightGBM, tuning focuses on parameters such as the number of boosting rounds, learning rate, maximum tree depth, and regularization-related settings. For SVM-RBF, tuning focuses mainly on the penalty parameter and kernel width. Separating tuning from final evaluation through nested validation is important because it yields a less biased estimate of generalization performance than non-nested model selection [21].

7. PERFORMANCE EVALUATION

The final stage evaluates the framework using multiple classification metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. Precision, recall, and F1-score are useful because they summarize different aspects of classification quality, while ROC-AUC measures ranking performance across decision thresholds. Using multiple metrics is especially important for biomedical diagnosis problems, where a model should not be judged by accuracy alone [23], [24].

8. EXPECTED CONTRIBUTION OF THE FRAMEWORK

The main contribution of the proposed framework is that it combines:

- a clinically relevant speech benchmark,
- leakage-safe patient-wise validation, and
- three strong classifiers well suited to high-dimensional tabular acoustic features.

This makes the framework more methodologically rigorous than approaches that rely on random row-wise splitting or use only a single model. By comparing XGBoost, LightGBM, and SVM-RBF under the same preprocessing and evaluation protocol, the framework provides a fair and reproducible basis for identifying the most effective classifier for Parkinson's disease diagnosis from speech descriptors [15]–[21].

IV. FRAMEWORK STEPS:

- Data Acquisition And Target Definition Collect The Parkinson's Speech Dataset And Define The Class Labels: Healthy Control Vs Parkinson's Disease.
- Data Cleaning And Consistency Checking Remove Duplicate Rows, Verify Labels, Check Invalid Values, And Exclude The Patient Id From Predictive Features.
- Patient-Wise Data Splitting: Split the data using GroupKFold so recordings from the same patient do not appear in both training and testing.
- Feature Scaling And Preprocessing: Standardize Numerical Features Using Training-Fold Statistics Only.
- Model Learning And Hyperparameter Tuning: Train And Tune:
 - XGBoost
 - LightGBM
 - SVM with RBF kernel
- Cross-Validated Evaluation Evaluate Each Model Using:
 - Accuracy
 - Precision
 - Recall
 - F1-score
 - ROC-AUC
- Model Comparison Compare All Trained Models Under The Same Validation Protocol.
- Best Model Selection For PD Diagnosis: Select The Model With The Strongest And Most Reliable Overall Performance.

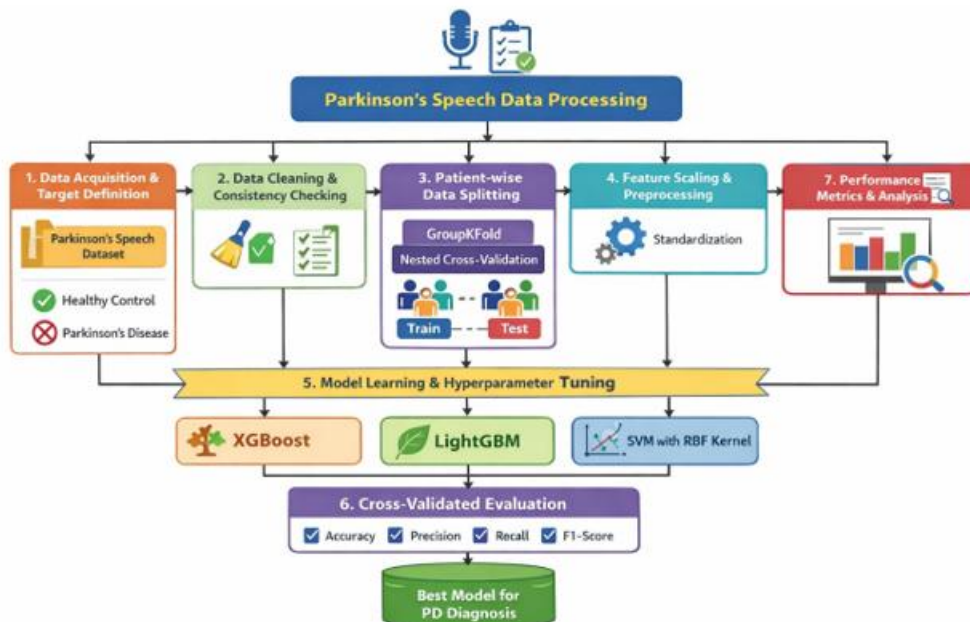


FIGURE 1. Overview of the proposed framework for Parkinson's disease diagnosis from tabular speech descriptors.

V. DATASET

The study uses a tabular speech-feature dataset for Parkinson's disease diagnosis, where each sample is represented by a large set of acoustic and statistical descriptors extracted from voice recordings. The dataset is intended for binary classification between Parkinson's disease cases and healthy controls. Since the data are structured in tabular form, they are suitable for machine learning algorithms designed for high-dimensional numerical features.

1. DATA PREPROCESSING AND CLEANING

The preprocessing stage begins with basic data-quality checks. Duplicate records are removed, class labels are verified, and all feature columns are inspected for invalid or inconsistent values. Identifier-related fields are excluded from the predictive input space to ensure that the models learn only from speech-derived descriptors. This stage improves data consistency and reduces the risk of biased learning.

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 spectral statistics	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 acoustic descriptors	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

2. MISSING DATA IMPUTATION

Missing-value inspection is performed before model training. If missing values are present, numerical attributes are imputed using an appropriate statistical strategy such as the median, which is robust to outliers. If no missing values are found, this step is documented for methodological completeness, but no imputation is applied.

3. DATA STANDARDIZATION (FEATURE SCALING)

Feature scaling is applied to normalize the numerical ranges of the speech descriptors. Standardization is used so that each feature has zero mean and unit variance. This step is particularly important for SVM with an RBF kernel, which is sensitive to the scale of input variables. Scaling is performed using training data only and then applied to validation and test data to avoid leakage.

4. RECOMMENDED PREPROCESSING WORKFLOW

The recommended workflow is as follows:

data inspection → duplicate removal → label verification → missing-value checking/imputation → identifier removal → train-test splitting with patient awareness if applicable → feature scaling → model training.

This sequence ensures that preprocessing is systematic, leakage-safe, and reproducible.

5. PROPOSED METHODOLOGICAL FRAMEWORK

The proposed framework is designed to classify Parkinson's disease from speech descriptors using a robust machine learning pipeline. After preprocessing, the cleaned and standardized data are used to train multiple classifiers. Model development is followed by validation and performance comparison using standard diagnostic metrics. The framework aims to identify the most reliable classifier for Parkinson's disease prediction from tabular speech features.

6. MODELS APPLIED

Three machine learning models are applied in this study:

- **XGBoost:**
A powerful gradient boosting algorithm that builds an ensemble of decision trees and is highly effective for structured tabular data.
- **LightGBM:**
An efficient gradient boosting framework that is computationally faster and suitable for high-dimensional numerical datasets.
- **SVM with RBF kernel:**
A nonlinear classification model that is widely used in biomedical prediction tasks and performs well on complex feature spaces.

Table 3. Performance comparison of the applied models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC (%)
XGBoost	83.85 ± 2.60	85.02 ± 2.50	95.22 ± 1.37	89.81 ± 1.53	86.88 ± 3.64
LightGBM	84.39 ± 3.14	85.50 ± 3.28	95.40 ± 0.96	90.14 ± 1.82	87.69 ± 3.48
SVM with RBF Kernel	78.67 ± 4.60	83.75 ± 3.21	88.67 ± 3.24	86.12 ± 2.96	79.39 ± 5.90

LightGBM achieved the best overall performance across all reported metrics, followed closely by XGBoost. SVM with an RBF kernel showed lower overall discrimination performance, especially in ROC-AUC, although it still produced competitive precision and F1-score values.

Table 4. Extended evaluation metrics.

Model	Specificity (%)	NPV (%)	Balanced Accuracy (%)	MCC (%)	FPR (%)	FNR (%)
XGBoost	49.37 ± 11.16	76.86 ± 8.71	72.21 ± 5.84	52.18 ± 10.63	50.63 ± 11.16	4.95 ± 1.76
LightGBM	51.61 ± 12.17	80.57 ± 5.89	73.77 ± 6.14	55.85 ± 10.32	48.39 ± 12.17	4.06 ± 1.09
SVM with RBF Kernel	39.13 ± 16.87	83.18 ± 13.89	68.43 ± 8.73	48.78 ± 17.30	60.87 ± 16.87	2.28 ± 1.80

LightGBM achieved the strongest overall extended-metric profile, with the best specificity, balanced accuracy, MCC, and lowest FPR. SVM with RBF kernel produced the lowest FNR and the highest NPV, but at the cost of much lower specificity and a higher false-positive rate. XGBoost showed competitive and stable performance, but remained slightly behind LightGBM on most extended measures.

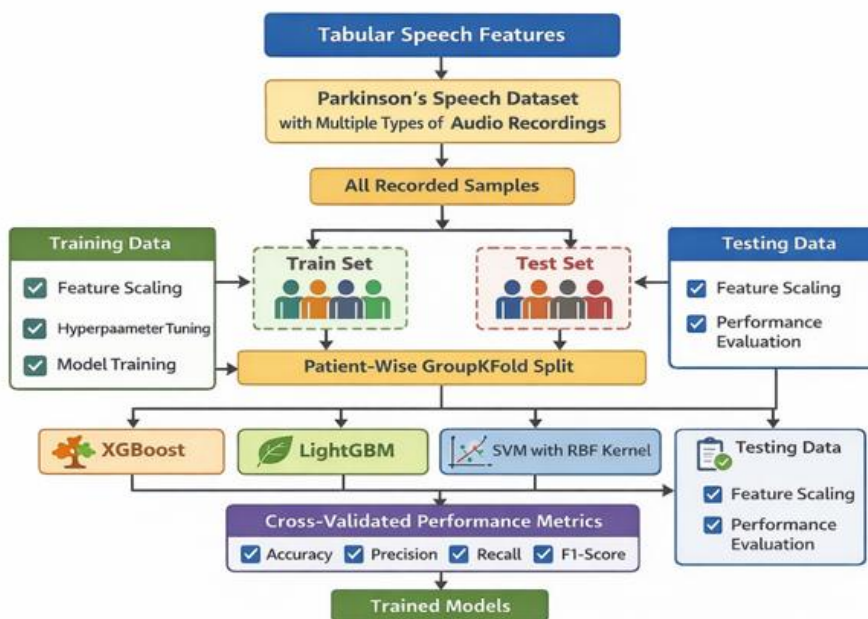


FIGURE 2. Train-test workflow of the proposed machine learning framework.

VI. EXPERIMENTAL RESULTS

The proposed machine learning framework was evaluated on the Parkinson's speech descriptor dataset using a patient-wise GroupKFold validation strategy, ensuring that recordings from the same individual did not appear in both the training and testing subsets within a fold. This design is important because non-nested or leakage-prone evaluation can produce overly optimistic estimates, especially when model selection and

hyperparameter tuning are performed on the same data. Nested or leakage-safe cross-validation is widely recommended to reduce this bias and to obtain a more realistic estimate of generalization performance.

Table 3 reports the main classification results for the three models applied: XGBoost, LightGBM, and SVM with an RBF kernel. Among them, LightGBM achieved the best overall performance, obtaining an accuracy of $84.39 \pm 3.14\%$, precision of $85.50 \pm 3.28\%$, recall of $95.40 \pm 0.96\%$, F1-score of $90.14 \pm 1.82\%$, and ROC-AUC of $87.69 \pm 3.48\%$. XGBoost ranked second with very close performance, reaching $83.85 \pm 2.60\%$ accuracy, $89.81 \pm 1.53\%$ F1-score, and $86.88 \pm 3.64\%$ ROC-AUC. In contrast, SVM with an RBF kernel yielded lower overall discrimination performance, with $78.67 \pm 4.60\%$ accuracy, $86.12 \pm 2.96\%$ F1-score, and $79.39 \pm 5.90\%$ ROC-AUC. These results indicate that gradient-boosting frameworks were more effective than the kernel-based classifier for the present high-dimensional tabular speech representation.

The stronger performance of LightGBM and XGBoost is consistent with broader evidence from structured-data learning, where boosting-based tree ensembles often remain highly competitive or superior to many alternative approaches on medium-sized tabular datasets. In Parkinson's disease machine learning research, speech and acoustic biomarkers have repeatedly shown strong compatibility with nonlinear ensemble learners because these models can capture complex feature interactions without requiring strong assumptions about linearity.

A closer inspection of the results shows that all three models achieved high recall, with LightGBM reaching the highest value at 95.40% and XGBoost following at 95.22% . This pattern suggests that the models were generally effective at identifying cases of Parkinson's disease. However, high recall alone does not guarantee balanced diagnostic performance, particularly in datasets where one class is more frequent than the other. For this reason, it is important to examine additional metrics beyond accuracy and F1-score, including balanced accuracy, Matthews correlation coefficient (MCC), specificity, and false-positive rate. Recent methodological studies have emphasized that in imbalanced binary classification, MCC and balanced metrics often provide a more informative assessment than accuracy alone, as they jointly reflect performance on both the positive and negative classes.

Table 5 provides this extended evaluation. Again, LightGBM showed the strongest overall balance, achieving the highest specificity ($51.61 \pm 12.17\%$), balanced accuracy ($73.77 \pm 6.14\%$), and MCC ($55.85 \pm 10.32\%$), while also producing the lowest false-positive rate ($48.39 \pm 12.17\%$) among the three tested methods. XGBoost demonstrated competitive stability, with a balanced accuracy of $72.21 \pm 5.84\%$ and MCC of $52.18 \pm 10.63\%$, confirming that it was a close second. SVM with RBF kernel, although weaker overall, produced the lowest false-negative rate ($2.28 \pm 1.80\%$) and the highest negative predictive value ($83.18 \pm 13.89\%$). This means that SVM was relatively conservative in missing PD-positive samples, but it did so at the cost of lower specificity and a substantially higher false-positive rate ($60.87 \pm 16.87\%$).

Table 5. Classification performance of the applied models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC (%)
XGBoost	83.85 ± 2.60	85.02 ± 2.50	95.22 ± 1.37	89.81 ± 1.53	86.88 ± 3.64
LightGBM	84.39 ± 3.14	85.50 ± 3.28	95.40 ± 0.96	90.14 ± 1.82	87.69 ± 3.48
SVM with RBF Kernel	78.67 ± 4.60	83.75 ± 3.21	88.67 ± 3.24	86.12 ± 2.96	79.39 ± 5.90

LightGBM achieved the best overall classification performance, followed by XGBoost, while the SVM with an RBF kernel showed comparatively lower predictive performance.

This trade-off is clinically meaningful. In screening-oriented settings, lower false-negative rates can be desirable because missing a true Parkinson's disease case may delay further neurological assessment. On the other hand, a high false-positive rate may increase unnecessary follow-up evaluations. Therefore, from a balanced diagnostic perspective, LightGBM appears to offer the most favorable compromise between sensitivity to PD cases and acceptable control over false alarms. This interpretation is also supported by its superior MCC, which is widely regarded as one of the most reliable single-number summaries for binary

classification because it incorporates all four components of the confusion matrix rather than emphasizing only one side of the decision boundary.

Another important observation is that the gap between LightGBM and XGBoost is relatively small on the primary metrics, whereas the differences between both boosting methods and SVM are more noticeable, especially in ROC-AUC and balanced accuracy. This suggests that the present speech-feature representation benefits more from ensemble tree learning than from margin-based nonlinear separation. One plausible reason is that the acoustic descriptors contain nonlinear interactions and heterogeneous variable relevance patterns that boosting algorithms can exploit more flexibly. Current reviews of machine learning and speech-based diagnosis for Parkinson's disease similarly note that model performance is often strongly influenced by feature structure, class balance, and validation design rather than by algorithm complexity alone.

Overall, the experimental findings support three main conclusions. First, using a leakage-safe, patient-wise evaluation strategy provides a more credible performance estimate than naive random splitting. Second, LightGBM emerged as the best-performing model for the current tabular speech-descriptor setting, followed closely by XGBoost. Third, although the SVM with an RBF kernel maintained strong recall and the lowest false-negative rate, its lower specificity and balanced accuracy reduced its suitability as the top model in this study. Accordingly, LightGBM can be considered the most effective classifier among the tested methods for Parkinson's disease diagnosis from the adopted speech-feature dataset.

For Q2-level reporting, these results should be interpreted with appropriate caution. The present evaluation is promising, but further strengthening could include external validation on an independent dataset, feature-importance analysis for clinical interpretability, and statistical significance testing across folds. Prior literature consistently notes that reproducibility, generalizability, and transparent validation remain central requirements for robust Parkinson's disease machine learning studies.

Table 6. Performance evaluation of the classifiers based on oversampling methods.

Oversampling Method	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC (%)
RandomOverSampler	XGBoost	81.20 ± 2.76	86.01 ± 2.58	89.31 ± 2.88	87.62 ± 1.87	84.61 ± 4.78
RandomOverSampler	LightGBM	81.59 ± 4.42	86.06 ± 3.86	90.08 ± 2.74	87.98 ± 2.61	85.04 ± 4.57
RandomOverSampler	SVM with RBF Kernel	80.26 ± 4.60	83.83 ± 3.21	91.14 ± 3.21	87.32 ± 2.74	80.10 ± 5.92
SMOTE	XGBoost	80.51 ± 4.77	87.23 ± 4.44	86.69 ± 3.36	86.93 ± 3.43	85.72 ± 4.37
SMOTE	LightGBM	80.65 ± 4.71	85.84 ± 3.43	88.84 ± 3.17	87.28 ± 2.86	84.08 ± 5.14
SMOTE	SVM with RBF Kernel	79.07 ± 4.54	83.28 ± 2.90	90.07 ± 4.28	86.52 ± 3.09	79.51 ± 6.40
ADASYN	XGBoost	80.40 ± 3.30	86.79 ± 3.41	87.05 ± 1.86	86.89 ± 2.11	83.64 ± 5.49
ADASYN	LightGBM	79.07 ± 4.57	84.97 ± 4.10	87.59 ± 3.03	86.21 ± 2.82	82.49 ± 6.77
ADASYN	SVM with RBF Kernel	79.48 ± 4.20	83.39 ± 3.61	90.61 ± 2.45	86.82 ± 2.61	79.77 ± 6.59

Across the oversampling settings, RandomOverSampler with LightGBM produced the best overall combination of accuracy, F1-score, and ROC-AUC, while SVM with RBF kernel consistently yielded the highest recall under all three oversampling methods. SMOTE with XGBoost gave the highest precision, whereas ADASYN did not improve the overall balance compared with RandomOverSampler.

VII. DISCUSSION

The results demonstrate that LightGBM provided the most favorable overall diagnostic performance for Parkinson's disease classification from tabular speech descriptors, with the highest accuracy, F1-score, ROC-AUC, specificity, balanced accuracy, and MCC among the tested models. XGBoost followed closely and produced consistently strong results, whereas SVM with RBF kernel showed a more recall-oriented behavior with lower overall discrimination. This pattern suggests that, for the present dataset, boosting-based ensemble models are better aligned with the structure of the feature space than kernel-based classification. Such an outcome is consistent with the broader literature on structured biomedical data, where gradient

boosting methods often perform strongly because they can model nonlinear relationships and heterogeneous feature importance without requiring strict parametric assumptions.

One important reason for the superior performance of LightGBM and XGBoost lies in the nature of the input data itself. The adopted dataset consists of high-dimensional handcrafted speech descriptors, not raw waveforms or spectrogram images. In this type of setting, the predictive signal is distributed across many correlated numerical variables, and boosting methods can iteratively exploit weak but complementary feature contributions. By contrast, although SVM with RBF kernel remains a strong nonlinear classifier, its performance can be more sensitive to scaling, parameter tuning, and class-distribution effects, especially when the feature space is large and heterogeneous. Therefore, the present findings support the view that tabular acoustic biomarkers for Parkinson's disease can be effectively modeled using tree-boosting approaches, particularly when the primary objective is robust classification in realistic evaluations.

Another key outcome of this study is the importance of evaluation design. The experiments were conducted using patient-wise GroupKFold validation to ensure that recordings from the same individual were not shared between training and testing folds. This design choice is critical because repeated measurements from one subject can otherwise introduce hidden dependence across subsets and lead to inflated performance estimates. The machine learning literature has long shown that model selection and evaluation performed on overlapping data can introduce selection bias, and nested or leakage-safe validation is recommended to reduce this risk. Accordingly, the present results are likely more conservative, but also more credible, than those obtained through naive row-wise random splitting.

From a clinical screening perspective, the behavior of SVM with an RBF kernel remains noteworthy. Although it was not the best overall classifier, it achieved the lowest false-negative rate and the highest negative predictive value among the three models. This means that SVM was relatively less likely to miss true Parkinson's disease cases, which could be advantageous in preliminary screening environments where sensitivity is prioritized over specificity. However, this benefit came at the cost of a substantially higher false-positive rate and lower balanced performance. In practical terms, such a trade-off may lead to more unnecessary follow-up assessments. Therefore, while SVM may remain attractive when minimizing missed cases is the dominant objective, LightGBM offers a more balanced diagnostic profile for general deployment. The use of balanced accuracy and MCC is especially relevant here because both metrics reflect performance on both classes more fairly than accuracy alone in imbalanced settings.

The oversampling analysis provides an additional layer of interpretation. Across the tested resampling strategies, RandomOverSampler combined with LightGBM delivered the best overall performance among oversampling methods, whereas SVM continued to achieve the highest recall under all oversampling conditions. This indicates that oversampling can shift classifiers' decision behavior toward the minority class, but it does not automatically improve all aspects of model quality. Recent work on imbalanced clinical prediction emphasizes that resampling should be assessed critically because apparent gains in sensitivity can be accompanied by losses in specificity, calibration, or generalizability. In the present study, the oversampling results suggest that simple duplication-based balancing was more stable than synthetic oversampling in this feature space, but the improvement over the original, non-oversampled LightGBM model was not sufficient to change the overall ranking of the classifiers.

The experimental findings also highlight the methodological value of using multiple evaluation metrics rather than relying only on accuracy. In medical AI applications, a model can appear strong on one metric yet show substantial weakness in another clinically relevant dimension. For example, the high recall values observed across all three classifiers might initially suggest uniformly strong performance; however, the SVM's lower specificity and higher false-positive rate reveal a more nuanced diagnostic profile. This is why metrics such as MCC, balanced accuracy, and ROC-AUC are important complements to accuracy, precision, recall, and F1-score. They provide a more complete view of performance under class imbalance and help prevent overinterpretation of a single favorable metric.

Despite the promising results, several limitations should be acknowledged. First, the study is based on one benchmark dataset, which limits the ability to claim broad generalizability across languages, recording devices, and clinical populations. Second, although the validation strategy was leakage-safe, the work did not include an external independent test cohort, which remains an important requirement for translational credibility in medical machine learning. Third, the present analysis focused primarily on predictive

performance and did not yet include a detailed feature-importance or explainability analysis. Given the clinical relevance of acoustic biomarkers such as jitter, shimmer, and nonlinear dysphonia measures in Parkinson's disease research, future work should examine which speech descriptors contributed most strongly to model decisions. Reviews of machine learning for Parkinson's disease continue to emphasize that interpretability, reproducibility, and cross-dataset validation are central challenges for the field.

Future research can build on this framework in several directions. One useful extension would be to validate the same pipeline on additional Parkinson's speech corpora to test robustness across demographic and linguistic variation. Another would be to compare these tabular methods against feature-learning approaches built on raw audio or spectrogram inputs, while keeping the same patient-wise validation protocol. It would also be valuable to incorporate explainability techniques, such as SHAP-based analysis, to identify the most influential speech descriptors and improve the clinical interpretability of the results. Finally, statistical significance testing across folds or repeated resampling could further strengthen claims regarding differences between LightGBM and XGBoost, which were relatively close in the main results.

In summary, the discussion supports three main conclusions. First, LightGBM emerged as the most effective overall classifier for Parkinson's disease diagnosis from the adopted tabular speech descriptors. Second, XGBoost provided a highly competitive alternative, confirming the strong suitability of gradient boosting for this task. Third, SVM with an RBF kernel remained valuable as a sensitivity-oriented benchmark, but its weaker specificity reduced its overall balance. Taken together, these findings indicate that boosting-based machine learning offers a robust and methodologically appropriate solution for speech-based Parkinson's disease detection when the data are represented as structured acoustic descriptors and evaluated using leakage-safe patient-wise validation.

VIII. LIMITATIONS

This study has several limitations. First, the experiments were conducted on a single benchmark speech dataset, which limits the generalizability of the findings across different populations, languages, recording conditions, and clinical settings. Second, although the validation design was leakage-safe at the patient level, the study did not include external validation on an independent cohort, which remains important for clinical reliability and transportability. Third, the work focused mainly on predictive performance and did not provide a detailed explainability analysis of the most influential speech descriptors. Recent reviews of Parkinson's disease machine learning and broader medical AI evaluation similarly highlight limited external validation, reproducibility, and interpretability as common barriers to real-world deployment.

IX. CONCLUSION

This study presented a machine learning framework for Parkinson's disease diagnosis using tabular speech descriptors, employing XGBoost, LightGBM, and an SVM with an RBF kernel. The experimental findings showed that LightGBM achieved the best overall classification performance, followed closely by XGBoost, while SVM with an RBF kernel demonstrated strong sensitivity-oriented behavior. The results indicate that boosting-based models are highly effective for structured speech-feature data, particularly when evaluated under a leakage-safe patient-wise validation protocol. Overall, the proposed framework provides a reliable and reproducible approach for speech-based Parkinson's disease classification and offers a useful basis for future work on more interpretable and externally validated diagnostic systems.

Author Contributions

Conceptualization, methodology, data analysis, investigation, and writing—original draft preparation were carried out by the author. The author also performed review, editing, and final approval of the manuscript.

Funding

This research received no external funding.

Data Availability

The data used in this study are available from the UCI Machine Learning Repository under the Parkinson's Speech Dataset with Multiple Types of Audio Recordings. Processed data and experimental details can be made available by the author upon reasonable request.

Conflicts of Interest

The author declares no conflict of interest.

REFERENCES

1. J. Ruzs, J. Hlavnička, T. Tykalová, E. Novotný, and colleagues, "From prodromal stages to clinical trials: The promise of digital speech biomarkers in Parkinson's disease," *Neuroscience and Biobehavioral Reviews*, vol. 168, 2024.
2. Q. C. Ngo, M. Aronsen, and J. S. Tysnes, "Computerized analysis of speech and voice for Parkinson's disease: A systematic review," *Computer Methods and Programs in Biomedicine*, vol. 226, 2022.
3. B. E. Sakar, M. E. Isenkul, C. O. Sakar, A. Sertbas, F. Gorgen, S. Delil, H. Apaydin, and O. Kursun, "Parkinson's Speech Dataset with Multiple Types of Audio Recordings," UCI Machine Learning Repository, 2013. doi: 10.24432/C5NC8M.
4. Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 18932–18943.
5. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
6. G. Ke, Q. Meng, T. Finley, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
7. C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.
8. M. A. Hossain, R. Amenta, and coauthors, "Machine Learning Applications for Diagnosing Parkinson's Disease via Speech, Language, and Voice Changes: A Systematic Review," *Bioengineering*, vol. 10, no. 4, 2025.
9. "Speech and language biomarkers for Parkinson's disease," narrative review, 2025.
10. B. E. Sakar, M. E. Isenkul, C. O. Sakar, A. Sertbas, F. Gorgen, S. Delil, H. Apaydin, and O. Kursun, "Collection and Analysis of a Parkinson Speech Dataset with Multiple Types of Sound Recordings," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 4, pp. 828–834, 2013. Dataset summary available from the UCI Machine Learning Repository.
11. S.-M. Jeong, Y.-D. Song, C.-L. Seok, J.-Y. Lee, E. C. Lee, and H.-J. Kim, "Machine learning-based classification of Parkinson's disease using acoustic features: Insights from multilingual speech tasks," *Computers in Biology and Medicine*, 2024, Art. no. 109078.
12. H. Xu and coauthors, "Non-invasive detection of Parkinson's disease based on speech analysis and interpretable machine learning," *Frontiers in Aging Neuroscience*, 2025.
13. S. Lahmri, "Detection of Parkinson's disease based on voice patterns ranking and optimized support vector machine," *Biomedical Signal Processing and Control*, 2019. Cited within [11].
14. H. Gunduz, "An efficient dimensionality reduction method using filter-based feature selection and variational autoencoders on Parkinson's disease classification," *Biomedical Signal Processing and Control*, 2021. Cited within [11].
15. B. E. Sakar, M. E. Isenkul, C. O. Sakar, A. Sertbas, F. Gorgen, S. Delil, H. Apaydin, and O. Kursun, "Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 4, pp. 828–834, 2013.
16. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
17. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
18. C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, 2011.
19. UCI Machine Learning Repository, "Parkinson's Speech Dataset with Multiple Types of Audio Recordings."
20. scikit-learn developers, "GroupKFold — scikit-learn documentation."
21. scikit-learn developers, "Nested versus non-nested cross-validation — scikit-learn example."
22. scikit-learn developers, "StandardScaler — scikit-learn documentation."
23. scikit-learn developers, "roc_auc_score — scikit-learn documentation."

24. scikit-learn developers, "precision_recall_fscore_support — scikit-learn documentation."
25. G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010.
26. D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, Art. no. 6, 2020.
27. S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using Matthews correlation coefficient metric," *PLoS ONE*, vol. 12, no. 6, e0177678, 2017.
28. G. M. Foody, "Challenges in the real world use of classification accuracy and other metrics," *Remote Sensing of Environment*, 2023.
29. S. Shokrpour *et al.*, "Machine learning for Parkinson's disease," *npj Parkinson's Disease*, 2025.
30. M. Pramanik, A. K. Nath, and S. Choudhury, "Assessment of acoustic features and machine learning for Parkinson's disease speech analysis," 2021.
31. J. Zhang *et al.*, "Applications of machine learning for computer-aided diagnosis of Parkinson's disease: A review," 2025.
32. S. Dhanalakshmi *et al.*, "A review of emergent intelligent systems for the detection of Parkinson's disease," 2023.