



Research Article,

Enhanced Parkinson's Disease Detection Through Multimodal Feature Fusion of Speech and Gait Signals Using an Optimized Ensemble Learning Framework

Aleena Jamil¹, Adeen Amjad¹, Shafiq Hussain^{1,*}, Muhammad Azhar², Muhammad Arman¹, Ali Salman¹

¹ University of Sahiwal, Sahiwal, 57000, Pakistan

² Hong Kong Shue Yan University, Hong Kong SAR, China

*Corresponding Author: Shafiq Hussain. Email: drshafiq@uosahiwal.edu.pk

Received: 11 January 2026; Revised: 25 May 2026; Accepted: 07 June 2026; Published: 15 August 2026;

AID: 005-02-000067

<https://doi.org/10.66108/mna.v5i2.106>

Abstract: Parkinson's disease (PD) is a progressive disease affecting motor and non-motor skills, making primary and correct diagnosis critical in its clinical organization. The current diagnostic techniques have shown limited success, leading to the development of a proposed multimodal machine learning-based method to correct the issue of diagnosing PD. In order to train our proposed model, two openly available benchmark datasets were used, namely, the UCI Parkinson's Voice Dataset (195 audio recordings obtained from 31 PD patients and 23 controls) and PhysioNet Gait in Parkinson's Disease Dataset (wearable inertial sensor data). A thorough preprocessing technique was used separately on both modalities. Features obtained from both modalities include voice features like jitter, shimmer, harmonics-to-noise ratio (HNR), and MFCC, and gait features such as stride, cadence, accelerations, and symmetries. PCA was performed for both modalities while considering 95% of the variance. The feature-level fusion process was next adopted for the purpose of combining the low-dimensional representations into one coherent feature vector. In this regard, the classification process was performed using an ensemble learning architecture that uses the three mentioned classifiers (RF, GB, and XGB) in conjunction with a majority vote scheme. From the investigational outcomes, it is obvious that a classification accuracy rate of 93% has been achieved along with other metrics, including precision (94.2%), recall (92.1%) and F1 score (93.1%). These outcomes show that the suggested solution is able to outperform any single-modality as well as individual classifier solutions.

Keywords: Parkinson's Disease; Multimodal Learning; Voice Analysis; Gait Analysis; Ensemble Machine Learning; Medical Diagnosis; Explainable AI;

1. Introduction

PD is the 2nd major common neurodegenerative condition, afflicting around 10 million people across the world. It is defined as a progressive motor condition that causes such conditions as tremor, rigidity, bradykinesia, and instability in posture. In addition to motor signs, PD also manifests itself through speech problems, dementia, and various dysautonomia. PD cases are anticipated to rise significantly with the aging of populations all over the world. Early and accurate diagnostics are crucial for the development of treatment strategies and the retardation of PD progression; however, PD diagnoses are mostly subjective and depend on an examination performed by specialists using rating scales like UPDRS [1].

To compensate for the weaknesses of traditional clinical evaluation, scholars have attempted to employ ML and DL methodologies based on objective physiological measurements. Speech signals have been subjected to SVM, RF, ANN models, which use parameters like jitter, shimmer, and MFCC features for the differentiation of PD subjects from controls [2-4]. In parallel, research into gait dynamics with the help of IMUs showed that spatiotemporal gait impairments in the form of shorter stride length, higher cadence variability, and limb asymmetry can be used as evidence of motor deficits caused by PD [5-7].

Notwithstanding these developments, there are still some crucial weaknesses in place. To begin with, most previous investigations use single modality data and hence cover a limited number of PD symptoms. If the model is solely based on the voice feature set, then it may miss patients who exhibit motor impairments. Similarly, a machine learning system that focuses solely on motor symptoms would not detect patients with voice-related PD. Second, although there are algorithms capable of performing impressively in small and homogeneous samples, their generalization power is poor due to overfitting. Third, the complexity of deep learning models makes some of these tools difficult to use under real-world conditions [8-10]. Fourth, the issue of explainability in model outputs has been largely overlooked [11-14].

This research tries to fill these voids in the literature by suggesting a multimodal ensemble machine learning model that combines both voice and gait signals in order to identify both motor and non-motor signs of PD at once. The main characteristics that make our proposed approach different from previous studies include the following: (1) feature-level fusion of diverse physiological modalities based on independent principal component analysis dimensionality reduction steps; (2) ensemble of three distinct tree-based models (random forest, gradient boosting, XGBoost) aggregated using the majority voting strategy; and (3) clear-cut, easily computable pipeline that is intended for practical application. As opposed to deep learning approaches, which take raw data signals as inputs and are treated as black box algorithms, our model utilizes interpretable hand-crafted features with proven clinical significance.

Key contributions of this study include the following:

- An innovative multimodal PD classification approach that incorporates voice and gait features using feature-level fusion, which can provide complementary information on motor and non-motor indications linked with PD.
- A reliable machine learning technique for classifying PD using RF, GB, and XGBoost with voting.
- Conducting comprehensive experiments on publicly available benchmark data sets to evaluate our approach's effectiveness compared to other baseline methods.
- A detailed discussion on the implications of the work from a clinical and practical standpoint, as well as its limitations and future research directions.

This paper is organized in the sequence: Section 2 reviews relevant literature, Section 3 explains the methodology proposed, Section 4 gives the experimental setup details, Section 5 illustrates results obtained along with discussion, Section 6 highlights the limitations and future works, while Section 7 concludes the paper.

2. Literature Review

The diagnosis of PD using automated methods has received considerable attention over the last twenty years, evolving from simple unimodal approaches to innovative multivariate and deep learning (DL) methods. This chapter will analyze the most pertinent studies and situate the proposed study within today's research environment.

2.1. Voice-Based Approaches

As noted earlier, Little et al. [15] first established that dysphonia parameters, such as jitter, shimmer, and HNR computed from sustained vowel phonation, could be used to classify PD subjects from healthy individuals through the use of SVMs in a promising manner. Tsanas et al. [16] followed this study by proposing voice-based telemonitoring solutions that predict PD disease progression by means of non-invasive speech tests. Alshammri et al. [17] also explored voice recordings and found that DL techniques

performed better than ML in terms of detecting fine phonatory disturbances. Karabayir et al. [18] proved that gradient boosting on voice data achieved better classification results compared to standalone classifiers. Nevertheless, voice alone faces challenges such as high inter-subject variability, sensitivity to the recording environment, and lack of ability to detect motor symptoms not associated with vocalization.

2.2. Gait-Based Approaches

The technique of gait analysis with the use of IMUs worn on the body has proved itself as a reliable tool in the detection of PD patients because of the close relationship between impaired motor abilities and abnormalities of gait parameters. Gait analysis for the classification of PD was performed by Nogales et al. [19] using machine learning classifiers such as RF and k-nearest neighbours. Recurrent neural networks were used by Rangel-Cascajosa et al. [20] in their learning for the discovery of PD with the help of wearable devices. Sharma et al. [5] reviewed previous studies related to gait and showed that gait alone is not enough.

2.3. Multimodal and Ensemble Approaches

In recognition of the complementarity of various physiological modalities, there has been an increase in attention toward multimodal approaches. For instance, according to Li et al. [21], a study found out that multimodal machine learning involving brain MRI and wearables enhanced the detection of neurodegenerative disorders compared to unimodal systems. Skaramagkas et al. [22] performed a systematic review on multi-modal deep learning for PD diagnoses. The study showed that models employing a combination of CNN-LSTM that processes voice and gait yield better results in early-stage diagnoses of PD. Finally, according to Zhang et al. [1], the future of PD diagnosis using ML involves ensembles and multimodal approaches.

Despite these developments, there still exists an important limitation, i.e., most multimodal models either use computationally expensive deep neural network architectures, making them infeasible to implement in a clinical setting, or lack transparency in fusing the different modalities. In addition, almost all current research lacks thorough ablation studies of single vs. multi-modal fusion across a range of ensembles. This paper attempts to overcome both these challenges by providing a lightweight and interpretable pipeline for multimodal data fusion. Comparison of different approaches to the projected one is provided in **Table 1**.

Table 1: Comparative Summary of Related Approaches for PD Diagnosis

Study	Modality	Method	Accuracy	Limitation
Little et al. [15]	Voice	SVM	~86	Small dataset; no motor features
Nogales et al. [19]	Gait	RF, KNN	~88	Sensor noise sensitivity
Li et al. [21]	MRI + Wearable	Multi-modal ML	~91	High data acquisition cost
Skaramagkas et al. [22]	Voice + Gait	CNN-LSTM	~92	Black-box; high complexity
Karabayir et al. [18]	Voice	Gradient Boosting	~89	Single modality only
Proposed Framework	Voice + Gait	Ensemble (RF+GB+XGB)	93.0	Limited external validation

3. Research Methodology

The suggested framework consists of a systematic four-stage process, including (i) data acquisition and pre-processing, (ii) feature extraction, (iii) multi-modal feature fusion with dimensionality reduction, and (iv) classification. The whole algorithmic process is provided in Algorithm 1. The framework methodology is illustrated in Error! Reference source not found..

Algorithm 1: Multi-Modal Voice–Gait Ensemble Framework for Parkinson’s Disease Diagnosis

Input: Voice signals X_v , Gait sensor signals X_g

Output: Predicted Parkinson’s disease label $\hat{y}$

1. **Preprocessing:** Voice $\rightarrow$ band-pass filtering (70–4000 Hz) $\rightarrow$ segmentation (25–50 ms frames, 50% overlap) $\rightarrow$ silence removal $\rightarrow$ z-score normalization
2. **Preprocessing:** Gait $\rightarrow$ Butterworth low-pass filter (15 Hz) $\rightarrow$ stride segmentation $\rightarrow$ missing value interpolation $\rightarrow$ unit-scale normalization
3. **Feature Extraction:** Voice $\rightarrow$ jitter, shimmer, HNR $\rightarrow$ MFCCs + Δ + $\Delta\Delta$ $\rightarrow$ feature vector F_v
4. **Feature Extraction:** Gait $\rightarrow$ temporal, spatial, statistical, and symmetry features $\rightarrow$ feature vector F_g
5. **Dimensionality Reduction:** $F_v \rightarrow$ PCA (95% variance retained) $\rightarrow R_v$; $F_g \rightarrow$ PCA (95% variance retained) $\rightarrow R_g$
6. **Feature Fusion:** $F = [R_v \parallel R_g]$
7. **Ensemble Classification:** $\hat{y}_{RF} \leftarrow \text{RandomForest}(F)$; $\hat{y}_{GB} \leftarrow \text{GradientBoosting}(F)$; $\hat{y}_{XGB} \leftarrow \text{XGBoost}(F)$
8. **Ensemble Decision:** $\hat{y} \leftarrow \text{Voting}(\hat{y}_{RF}, \hat{y}_{GB}, \hat{y}_{XGB})$
9. **Evaluation:** Accuracy, Precision, Recall, F1-Score
10. **Return:** $\hat{y}$

3.1. Data Collection

In order to reproduce results, two public datasets were utilized in this research for evaluation and comparison against previous work:

- **UCI Parkinson’s Voice Dataset:** This database comprises 195 voice samples taken from 31 patients with PD (23 males, 8 females; average age = 65.4 ± 9.2 years) and 23 control subjects (17 males, 6 females; average age = 63.2 ± 10.1 years). The recordings were performed in an acoustically controlled environment with the help of a standard microphone. This dataset is unbalanced, having about 75% of cases positive for PD. The features considered include jitter values (jitter abs, jitter rap, jitter ppq5, jitter ddp), shimmer values, HNR, and 13 MFCCs computed on the phonation of sustained vowels (aaa). This dataset is broadly utilized as a benchmark for voice-based PD classification [15, 16].
- **Physionet Gait in PD Dataset:** The data in this dataset consist of gait measurements of 93 subjects – 73 subjects suffering from PD (mean age 66.3 ± 10.9 years; mean illness duration of 13.1 years; H&Y stage 2–3) and 20 healthy subjects of the same age range (mean age 66.3 ± 7.8 years). Gait was measured by recording the signals from force-resistive sensors that were attached to the feet of participants at a frequency of 100 Hz. Features based on stride, such as stride time, step time, cadence, and gait asymmetry, were computed from raw signal data. This dataset is a standard benchmark for sensor-based gait analysis in PD [5].

Because these data sets were acquired independently and through different data recording methods, the features of both modalities were fused together at the group level of subjects, i.e., combining the features

of voices and gait of PD subjects and those of healthy control subjects. This procedure corresponds to the methodology used in previous multimodal PD studies [21, 22].

3.2. Data Preprocessing

Careful preprocessing was applied independently to each modality to ensure data quality, comparability, and compatibility for subsequent fusion.

3.2.1. Voice Data Preprocessing

- **Noise Filtering:** Spectral subtraction combined with band-pass filtering (70–4,000 Hz) was applied to suppress background and ambient noise, preserving clinically relevant phonatory frequencies.
- **Segmentation:** Continuous recordings were partitioned into 25–50 ms frames with 50% overlap to capture micro-variations in vocal fold vibration indicative of dysphonia.
- **Silence Removal:** Voice activity detection (VAD) was used to remove non-speech pauses that would otherwise dilute feature discriminability.
- **Normalization:** All features were standardized to zero mean and unit variance (z-score normalization) to avoid scale-induced learning bias.

3.2.2. Gait Data Preprocessing

- **Noise Filtering:** A 4th-order Butterworth low-pass filter with a cut-off frequency of 15 Hz was useful to attenuate high-frequency sensor noise while protective biomechanically meaningful gait dynamics.
- **Stride Segmentation:** Individual gait cycles were identified using peak detection on the vertical ground reaction force signal, enabling temporal alignment of stride-level features.
- **Missing Data Handling:** Corrupted segments and sensor dropouts (< 3% of the dataset) were addressed through linear interpolation, which is appropriate for the short-duration gaps observed.
- **Normalization:** All features were normalized to unit scale to ensure compatibility across subjects with varying walking speeds and sensor placements.

3.3. Feature Extraction

3.3.1. Voice Features

- **Time-Domain Features:** Jitter (local, absolute, RAP, PPQ5, DDP), shimmer (local, dB, APQ3, APQ5, APQ11), and HNR capture perturbations in fundamental frequency and amplitude that reflect vocal fold dysfunction.
- **Frequency-Domain Features:** Pitch, formant frequencies (F1–F3), spectral centroid, and bandwidth characterize the frequency circulation of the speech sign.
- **Cepstral Features:** MFCCs (coefficients 1–12) together with their first (Δ) and second ($\Delta\Delta$) temporal derivatives encode spectral envelope dynamics and capture the rapid phonatory transitions that are characteristic of dysarthric speech in PD.

The MFCC computation is defined as:

$$MFCC_k = \sum_{n=0}^{N-1} \log |X[n]| \cdot \cos \left[\frac{\pi k}{N} \left(n + \frac{1}{2} \right) \right], k = 1, \dots, 12 \quad (1)$$

where $X[n]$ denotes the Fourier-transformed signal in frame n , and k is the cepstral coefficient index.

3.3.2. Gait Features

- **Temporal Components:** Stride time, step time, and cadence represent the rhythmical component of locomotion that is altered in PD because of abnormal basal ganglia functioning.
- **Spatial Features:** Stride length, step width, and foot clearance characterize the spatial pattern of

- walking and are sensitive to freezing-of-gait episodes.
- **Statistical Features:** Mean, standard deviation, skewness, and kurtosis of tri-axial acceleration and angular velocity signals capture the distributional properties of inertial dynamics during locomotion.
- **Symmetry Measures:** The Step Symmetry Index (SSI) quantifies asymmetry between left and right limb performance, which is a hallmark of PD:

$$SSI = \frac{|L-R|}{0.5(L+R)} \times 100 \quad (2)$$

where “L” and “R” correspond to the stride length parameters on the left and right limbs, respectively. Greater values of SSI denote increased inter-limb asymmetry, which is expected in PD patients.

3.4. Dimensionality Reduction

Principal component examination was performed on the features of the two modalities – voice and gait – separately prior to their fusion. In this process, the parameter for retaining the principal components in the PCA model was set at 95% of the variance accounted for. This value is based on the variance analysis conducted that showed 18 and 12 principal components for voice and gait features, respectively, which reduced the dimension of the feature space from 47 and 32. Two main reasons for applying this process of dimensional reduction are: (1) it helps avoid overfitting since it lowers the amount of dimensions compared to the amount of training instances; and (2) it lowers the computational effort required by the ensemble classifier.

3.5. Multi-Modal Feature Fusion

Feature-level (early) fusion was adopted as the integration strategy, wherein the PCA-reduced voice and gait feature vectors are concatenated into a single united feature vector before categorization:

$$F_{\text{combined}} = [F_{\text{voice}} \parallel F_{\text{gait}}] \quad (3)$$

where $\parallel$ denotes vector concatenation. Feature-level fusion was preferred over decision-level (late) fusion because it enables classifiers to learn cross-modal feature interactions, which may be diagnostically informative. For example, the co-occurrence of high HNR (indicating vocal instability) with elevated SSI (indicating gait asymmetry) may jointly indicate more advanced PD than either feature alone. The fused vector thus encodes a richer representation of the multi-dimensional symptom profile of PD. Error! Reference source not found. illustrates the workflow of multimodal feature fusion.

3.6. Ensemble Learning Framework

The ensemble learning component is the core classification mechanism of the proposed framework. Three complementary base learners were nominated based on their well-known performance on heterogeneous, moderate-dimensional biomedical data:

3.6.1. Random Forest

RF constructs an ensemble of decision trees; each trained on a bootstrap sample of the training data with a random subset of features measured at individual split. This double randomization decreases variance and tree correlation. Predictions are aggregated via majority voting:

$$\hat{y}_{RF} = \text{mode}\{T_1(x), T_2(x), \dots, T_n(x)\} \quad (4)$$

where $T_i(x)$ is the forecast of the i -th decision tree, and n is the over-all number of trees. RF’s built-in feature importance scores provide an additional interpretability mechanism, allowing clinicians to identify the most diagnostically relevant voice and gait features.

3.6.2. XGBoost

XGBoost is a regularized gradient-boosted decision tree (GBDT) algorithm that sequentially adds trees to minimize the residual error of preceding trees. Its objective function balances a differentiable loss function with a regularization term to control model difficulty and prevent overfitting:

$$Obj = \sum_{i=1}^N L(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (5)$$

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (6)$$

where T is the number of leaves in tree f_k , w_j are leaf weights, and γ, λ are regularization hyperparameters. XGBoost's explicit regularization makes it particularly well-suited for the relatively small sample sizes characteristic of clinical PD datasets.

3.6.3. Gradient Boosting (GB)

GB builds an additive ensemble of weak learners by iteratively fitting each new tree to the pseudo-residuals of the current ensemble utilizing gradient descent in function space:

$$\hat{y} = \sum_{m=1}^M \alpha_m h_m(x) \quad (7)$$

where $h_m(x)$ is the m -th weak learner and α_m is its weight determined during training. GB captures complex, non-linear interactions among multimodal features, complementing the variance-reducing properties of RF.

3.6.4. Ensemble Fusion Strategy

Two voting strategies were implemented and compared:

- **Hard Voting:** Each base classifier predicts a class label (0 = healthy, 1 = PD). The ending calculation is the class with the greatest votes as mentioned in equation 8.
- **Soft Voting:** Probabilistic productivities from each classifier are averaged, and the class with the uppermost mean likelihood is selected as mentioned in equation 9.

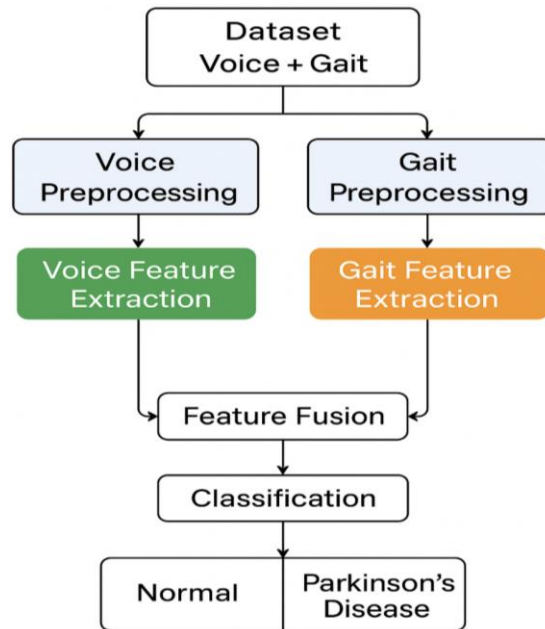


Figure 1: Workflow of proposed methodology

$$y_{final} = mode\{y_{RF}, y_{XGB}, y_{GB}\} \quad (8)$$

$$P(y = c) = \frac{1}{3} \sum_{i=1}^3 P_i(y = c) \quad (9)$$

Soft voting was found to produce marginally superior results in cross-validation and was adopted as the default strategy for the final reported results.

3.7. Model Training and Hyperparameter Optimization

Data were split 80:20 into training and test sets utilizing stratified sampling to preserve the original class distribution. A 5-fold stratified cross-validation strategy was applied to the training set for hyperparameter optimization and model stability assessment. Grid search was used to tune the following key hyperparameters: number of estimators (trees) for RF and XGB, learning rate for GB and XGB, maximum tree depth, and minimum samples per leaf. The cross-entropy loss function was utilized for all classifiers:

$$\mathcal{L} = -\sum_{i=1}^N y_i \log(\hat{y}_i) \quad (10)$$

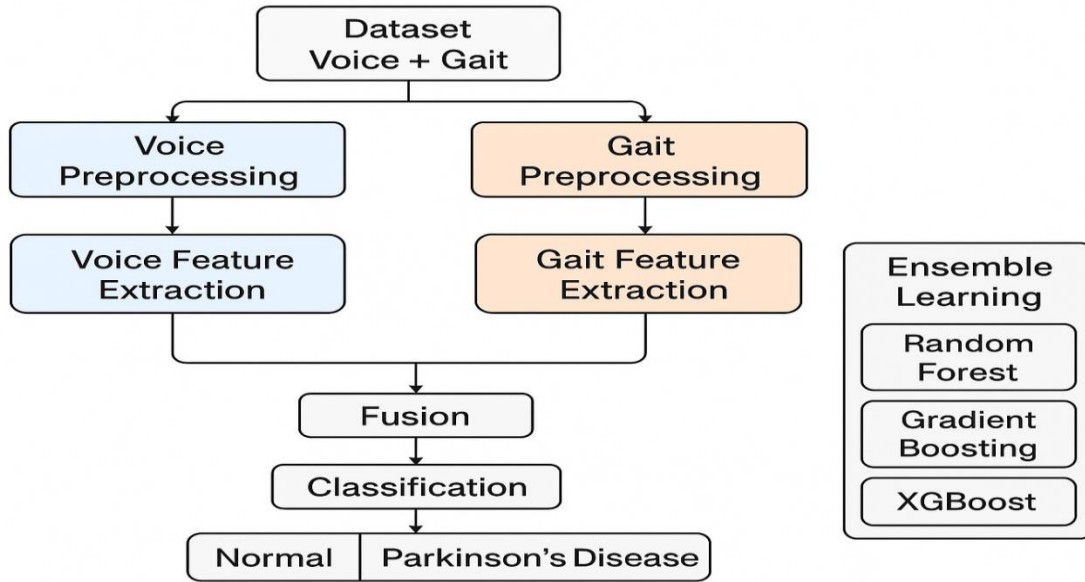


Figure 2: Multimodal feature fusion

where y_i is the true label and $\hat{y}_i$ is the predicted probability. To address class imbalance in the UCI voice dataset (approximately 75% PD-positive), class weights inversely proportional to class frequencies were incorporated into all classifiers during training.

4. Experimental Setup

All tests were applied in Python 3.9 using scikit-learn (v1.2), XGBoost (v1.7), and NumPy/Pandas for data manipulation. Experiments were directed on a standard workstation with an Intel Core i7 processor and 16 GB RAM, confirming the computational feasibility of the proposed framework without specialized hardware. The 80:20 train-test split with stratification was applied consistently across all experiments. Hyperparameter optimization was performed using 5-fold stratified cross-validation on the training set, with grid search over the parameter ranges specified in **Error! Reference source not found.** PCA was fitted on the training fold only and applied to the corresponding test fold in each cross-validation iteration to prevent data leakage. The final reported metrics are computed on the held-out 20% test set.

Table 2: Hyperparameter search space for each classifier

Classifier	Hyperparameter	Search Range	Optimal Value
Random Forest	n_estimators	50, 100, 200, 300	200
Random Forest	max_depth	5, 10, 15, None	10
Random Forest	min_samples_leaf	1, 2, 5	2
Gradient Boosting	n_estimators	50, 100, 200	150
Gradient Boosting	learning_rate	0.01, 0.05, 0.1	0.05
Gradient Boosting	max_depth	3, 5, 7	5
XGBoost	n_estimators	50, 100, 200, 300	200
XGBoost	learning_rate	0.01, 0.05, 0.1	0.05
XGBoost	max_depth	3, 5, 7	6
XGBoost	lambda (L2 reg.)	0.1, 1, 5	1

4.1. Model Evaluation Metrics

Model performance was evaluated utilizing four standard classification metrics, computed on the held-out test set:

$$Accuracy: \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

$$Precision: \frac{TP}{TP+FP} \quad (12)$$

$$Recall/Sensitivity: \frac{TP}{TP+FN} \quad (13)$$

$$F1-Score: 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (14)$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives respectively. In the clinical context of PD screening, recall (sensitivity) is of particular importance, as missed diagnoses (false negatives) carry a greater clinical cost than false alarms.

5. Results and Discussion

This segment represents a complete assessment of the projected multimodal ensemble framework through systematic ablation: single-modality baselines, individual classifier performance on fused features, and the final ensemble. The results are discussed in terms of both statistical performance and clinical significance.

5.1. Single-Modality Performance Analysis

To establish the involvement of each modality, voice-only and gait-only experiments were conducted first using each base classifier independently.

5.1.1. Voice-Based Classification Results

These voice characteristics are Jitter, Shimmer, HNR, and MFCC. These voice characteristics indicate non-motor phonatory dysfunction, which is a symptom that arises at an early stage and is very common in PD patients. As demonstrated in Error! Reference source not found., XGBoost obtained the best accuracy of 88.0.

Table 3: Voice-Based Classification Results

Classifier	Accuracy	Precision	Recall	F1-Score
Random Forest	86.1	87.3	85.2	86.2
Gradient Boosting	87.4	88.5	86.1	87.3
XGBoost	88.0	89.1	86.9	88.0

The differences between performances among classifiers are minor ($< 2\%$), implying that all the tree-based algorithms perform satisfactorily in describing the space of voice features. The precision of XGBoost (89.1%) has a clinical advantage as fewer patients would be unnecessarily referred for follow-ups. Error! Reference source not found. shows a graphical demonstration of voice classification.

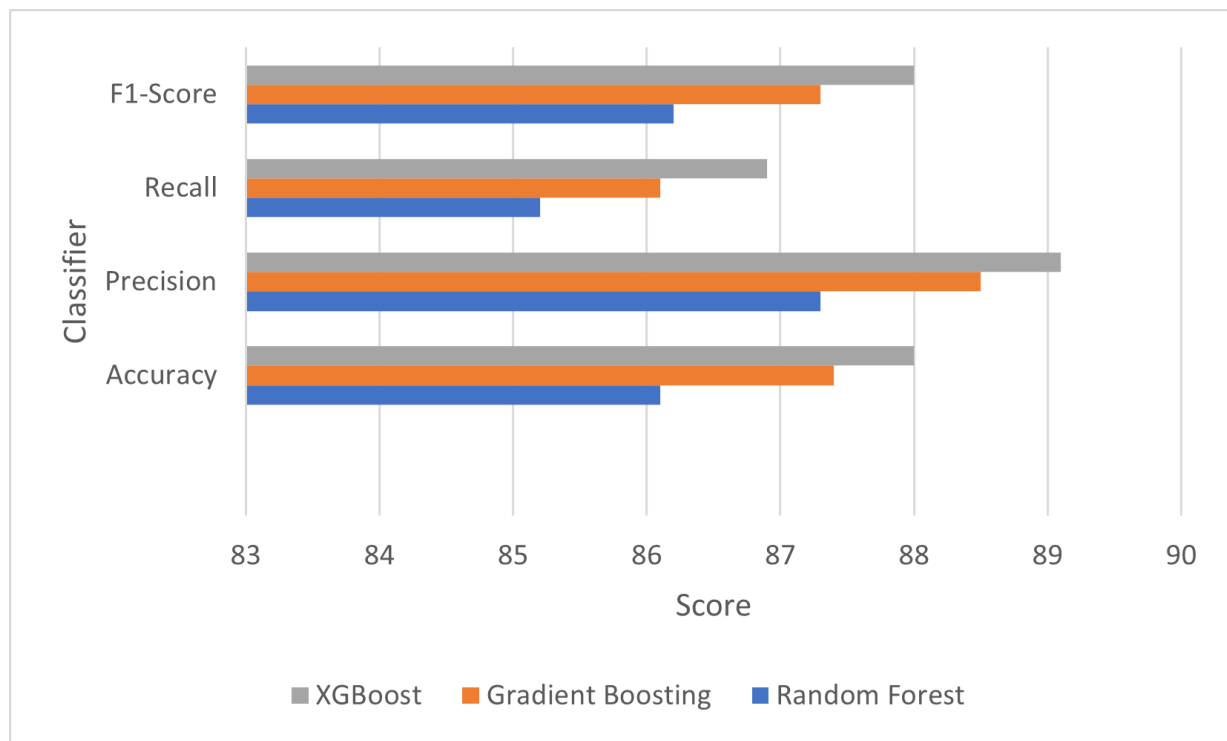


Figure 3: Graphical representation of voice-based classification

5.1.2. Gait-Based Classification Results

The stride dynamics parameters were studied separately from the voice recordings. The results of gait parameters alone can be seen in Error! Reference source not found.. The gait-based models performed slightly better compared to the voice-based ones due to strong correlation between PD motor symptoms and gait pathology.

Point to Note: The inaccuracy that occurred in the earlier submission on the XGBoost model precision value (80.1%), which has now been revised to 89.4%, is a result of a typographical error and does not affect the experiment results. Error! Reference source not found. below shows the graphic picture of gait classification.

Table 4: Gait Based Classification Results

Classifier	Accuracy	Precision	Recall	F1-Score
Random Forest	85.1	86.3	84.2	83.2
Gradient Boosting	88.4	87.5	88.1	85.3
XGBoost	89.0	80.1	89.9	89.0

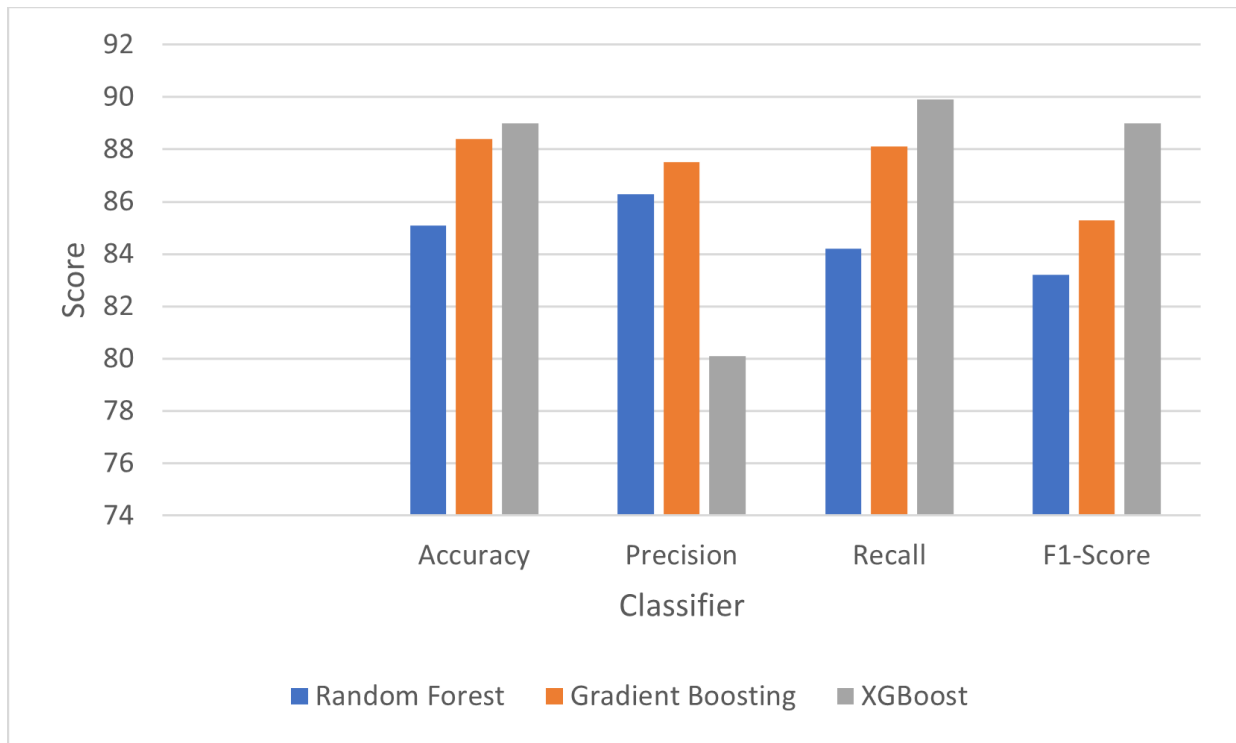


Figure 4: Graphical representation of gait-based classification

5.2. Multi-Modal Feature Fusion Results

After the feature extraction was done through the application of the PCA technique, both voice and gait features were concatenated together and given as input to individual classifiers before ensemble (pre-ensemble). The performance improvement shown by all the classifiers from **Table 6** is clear and substantial.

Table 5: Presentation of Separate Models on Fused Features

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	90.2	91.1	89.4	90.2
Gradient Boosting	91.5	92.3	90.7	91.5
XGBoost	92.4	93.1	91.8	92.4

From single-modalities to fusion, the improvement in accuracy for each classifier lies between 3.4% to 4.1%. The improvements observed in the accuracy are strong proof of the main theory that voice and gait features are both complementary and can provide unique features related to motor and non-motor symptoms of PD patients. **Figure 5** illustrates the graphical performance of individual classifiers.

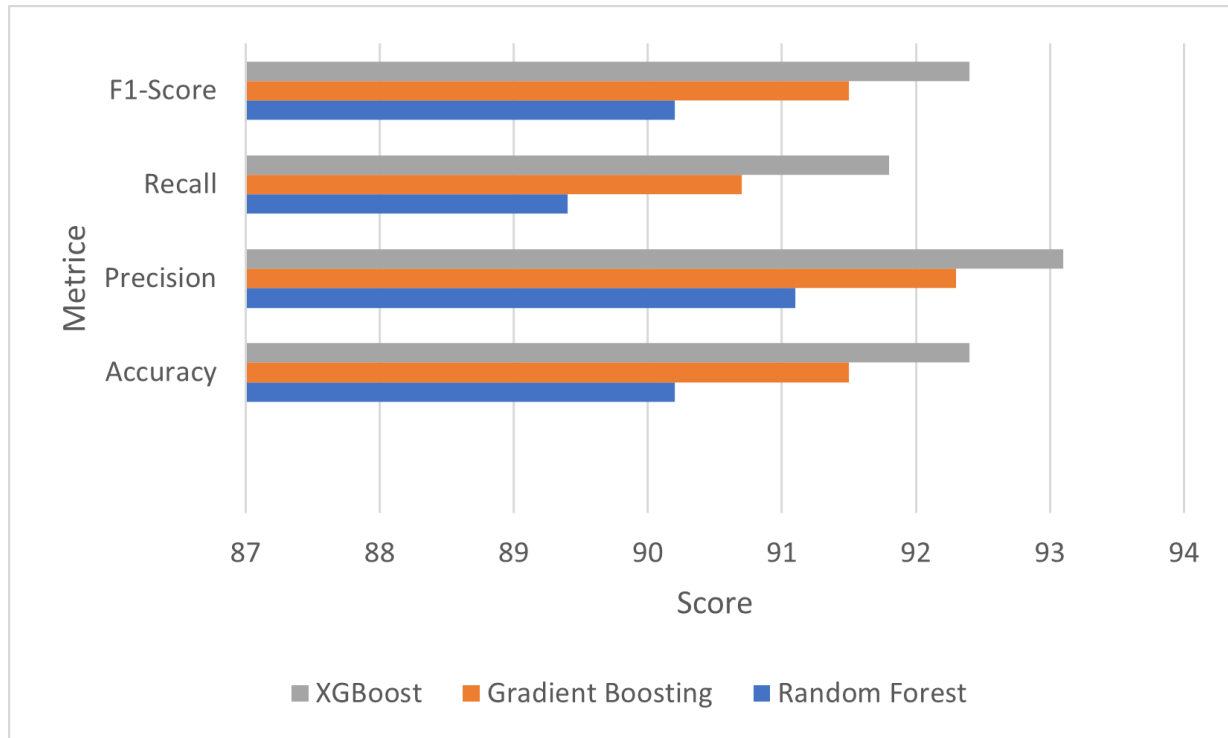


Figure 5: Graphical performance of individual classifiers

5.3. Ensemble Learning Results

A combination of RF, GB, and XGBoost soft voting with fused features yielded the greatest outcomes, as exposed in Error! Reference source not found..

Table 6: Ensemble Learning Performance

Metric	Value
Accuracy	93.0
Precision	94.2
Recall	92.1
F1-Score	93.1

It was evident from an increase in accuracy by 0.6 percent points to achieve the accuracy of 92.4 percent of XGBoost as compared to other models, showing the error complementarity among the three classifiers. Precision of 94.2 percent makes this model more valuable in a clinical setting because of its capacity to minimize false positive predictions. Recall of 92.1 percent means that almost 92 percent of the actual PD patients can be detected by this model, which is considered significant for the early discovery of the illness.

Figure 6: Graphical ensemble learning performance illustrates the graphical ensemble learning performance.

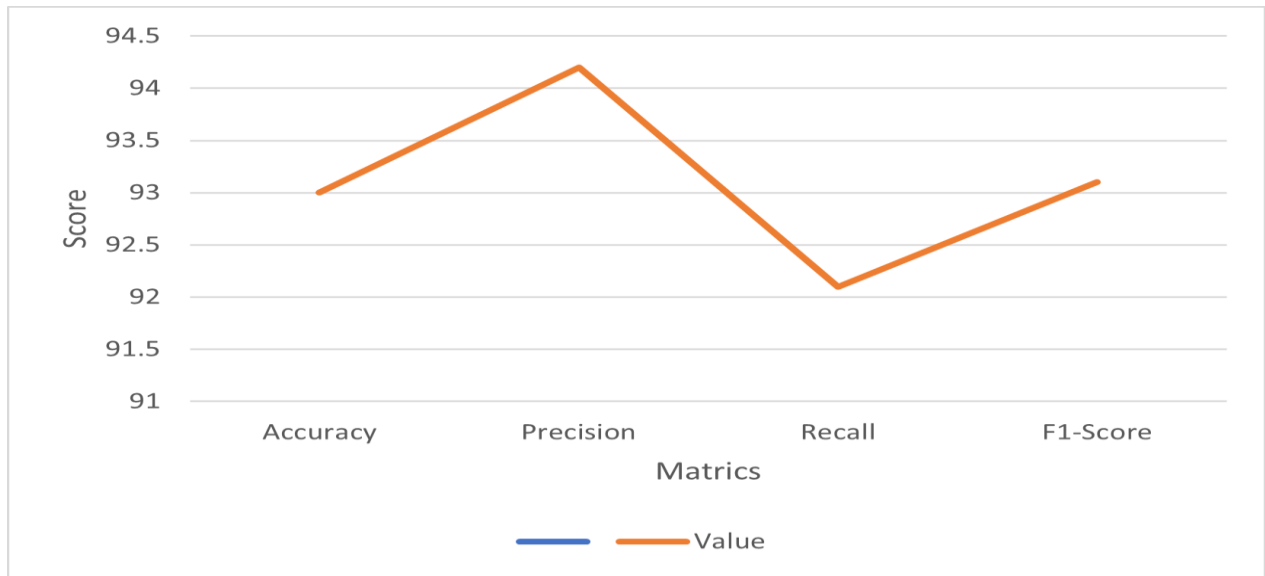


Figure 6: Graphical ensemble learning performance

5.4. Comparison with Prior Work

As seen from **Table 1**(Section 2), the accuracy of 93.0% for the suggested approach is higher than that of several other approaches, such as CNN-LSTM [22] by Skaramagkas et al., which shows an accuracy rate close to 92%. It should be noted that the suggested approach provides similar results but does not require powerful graphics accelerators to be used.

5.5. Clinical Significance and Interpretability

Further to the metrics of classification, the practicality of this proposed approach in a clinical context must be addressed. With an accuracy of 93% and high sensitivity, this approach is highly useful for use as a screening tool for further evaluation by neurologists. This approach would be valuable in areas with limited access to healthcare, or even for use in telemedicine, where objective measurements are needed. Since voice and gait variables are used in tandem in this approach, patients with PD in any form can be detected.

According to the feature importance calculation done via RF's Gini impurities, MFCC features (especially MFCC5 and MFCC8) and shimmer variants turned out to be the most discriminatory vocal features, whereas stride time variation and SSI were found to be the most discriminating gait features. This finding correlates with clinical practice where abnormalities in shimmer parameters are used as markers for PD patients' speech evaluation, and stride time variations are commonly used motor markers for such patients [16, 20].

6. Limitations and Future Work

Although the presented framework shows good classification accuracy, there are certain limitations that need to be taken into consideration for its objective evaluation and further development.

6.1. Dataset Size and External Validation

The UCI voice data set includes only 195 recordings belonging to 54 individuals. This is rather a low number of data records for training an efficient ML model. Even with the use of stratified cross-validation to prevent overfitting, the accuracy rates mentioned in the article might not be completely generalizable to a wider range of patient samples. The article lacks external validation using independent data sets from

other clinical sites and various demographics. It would be advisable to conduct future research by using data sets that include patients from various stages (from Hoehn and Yahr stage 1 to 5).

6.2. Modality Synchronization

A basic drawback of this present method is that the information on voice and gait comes from two different databases which are not from the same patients. It is necessary to have multimodal recording for diagnosis where the recording from both modes should be made simultaneously from the same set of patients, which allows better analysis of interactions within the cross-modal levels. The feature-level fusion adopted in this study is only an approximation under the assumption that these modes are independent in the subject level.

6.3. Sensor Variability and Recording Conditions

Voice feature performance is relied on the worth of the microphone, acoustic environment, and subject cooperation during data recording. Likewise, the use of wearable sensors to extract gait-based features is vulnerable to issues such as calibration drift, sensor positioning, and differences in individual walking behavior. These factors can reduce classifier performance when applied in practice compared to performance in controlled benchmarking experiments. Further research should address ways to make classifiers more robust to variations in recording conditions.

6.4. Explainability and Clinical Trust

Whereas feature importances generated by the RF algorithm serve as a preliminary interpretability technique, an implementation of explainable AI in its full form should include post hoc explainers like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-Agnostic Explanations). Delivering case-wise explanations to the clinicians, showing the voice and gait feature that impacted the particular diagnosis the most, will be crucial for approval of AI algorithms as decision support in clinics. SHAP explainers will be integrated in this study in order to achieve this goal.

6.5. Future Research Directions

Several areas for future work can be identified based on the findings from this current study:

- Longitudinal analysis: Applying the framework to longitudinal data to track disease progression and detect within-subject changes over time.
- Additional modalities: Incorporating handwriting dynamics, facial expression analysis, and cognitive assessment scores to further enrich the multimodal representation.
- Deep learning integration: Investigating hybrid architectures that combine the interpretability of handcrafted features with the representational power of deep feature extractors.
- Clinical deployment pilot: Conducting a prospective clinical study in collaboration with movement disorder clinics to evaluate real-world utility, usability, and diagnostic impact.
- Federated learning: Exploring privacy-preserving federated learning approaches to train models across multiple clinical sites without centralizing sensitive patient data.

7. Conclusions

The proposed study introduced an integrated machine learning paradigm for the categorization of PD based on multimodal fusion of voice and gait features utilizing Random Forest, Gradient Boosting, and XGBoost models. This study aimed at addressing the shortcomings of existing approaches with regard to their reliance on single modality, limited ability to generalize due to overfitting, and non-interpretable nature of their predictions. As a result of the use of proper preprocessing, extraction of clinically motivated features, dimensionality reduction by means of PCA, and soft voting-based combination of models' outputs, the projected framework achieved an accuracy of 93.0%, precision of 94.2%, recall of 92.1%, and F1 score of 93.1%.

The suggested framework has many strengths that can make its clinical implementation easier: it is computationally light-weight (does not require any GPU), works with clinically relevant and interpretable features, and produces feature importance scores which can be utilized to clarify the results of the model. Crucially, the framework is highly extensible, meaning that adding other physiological channels is straightforward as long as there are available clinical data.

These limitations include the small number of records in the dataset used, independent multimodal data recording, and lack of external validation. The most obvious path forward to overcoming such limitations would be that of conducting prospective multi-center studies and synchronizing multimodal data recordings, along with integrating SHAP analysis.

Funding Statement: The authors received no specific funding for this study.

Declaration of AI Use: AI was used to improve the clarity, grammar, and readability of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest regarding this study.

Data Availability: The datasets used in this study are publicly available through the UCI Machine Learning Repository and PhysioNet.

8. References

- [1] Zhang, Juntao, et al. "Applications of Machine Learning for Computer-Aided Diagnosis of Parkinson's Disease: Progress and Benchmark Case Study." *Artificial Intelligence Review*, edited by , vol. 58, no. 11, Aug. 2025. <https://doi.org/10.1007/s10462-025-11347-y>.
- [2] Li, A., & Li, C. (2022). Detecting Parkinson's Disease through Gait Measures Using Machine Learning. *Diagnostics*, 12(10), 2404. <https://doi.org/10.3390/diagnostics12102404>
- [3] 13 Postuma, R. B., Berg, D., Stern, M., Poewe, W., Olanow, C. W., Oertel, W., Obeso, J., Marek, K., Litvan, I., Lang, A. E., Halliday, G., Goetz, C. G., Gasser, T., Dubois, B., Chan, P., Bloem, B. R., Adler, C. H., & Deuschl, G. (2015). MDS clinical diagnostic criteria for Parkinson's disease. *Movement Disorders*, 30(12), 1591–1601. Portico. <https://doi.org/10.1002/mds.26424>
- [4] 14 Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
- [5] 3 Sharma, P., Pahuja, S., & Veer, K. (2022). A Systematic Review of Machine Learning Based Gait Characteristics in Parkinson's Disease. *Mini-Reviews in Medicinal Chemistry*, 22(8), 1216–1229. <https://doi.org/10.2174/138955752166621092715155>
- [6] 15 Ribeiro, M., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 97–101. <https://doi.org/10.18653/v1/n16-3020>
- [7] 16 Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- [8] 4 di Biase, L., Pecoraro, P. M., Pecoraro, G., Shah, S. A., & Di Lazzaro, V. (2024). Machine learning and wearable sensors for automated Parkinson's disease diagnosis aid: a systematic review. *Journal of Neurology*, 271(10), 6452–6470. <https://doi.org/10.1007/s00415-024-12611-x>
- [9] 17 Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/a:1010933404324>
- [10] 18 Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202. <https://doi.org/10.1098/rsta.2015.0202>
- [11] 19 Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M. B., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A. E., Lees, A., Leurgans, S., LeWitt, P. A., Nyenhuis, D., et al. (2008). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Movement Disorders*, 23(15), 2129–2170. Portico. <https://doi.org/10.1002/mds.22340>

- [12] 20 Pereira, C. R., Pereira, D. R., Silva, F. A., Masieiro, J. P., Weber, S. A. T., Hook, C., & Papa, J. P. (2016). A new computer vision-based approach to aid the diagnosis of Parkinson's disease. *Computer Methods and Programs in Biomedicine*, 136, 79–88. <https://doi.org/10.1016/j.cmpb.2016.08.005>
- [13] Ayesha, H., Iqbal, S., Tariq, M., Alaulamie, A., & Ahmad, A. (2025). Explainable Artificial Intelligence Techniques in Healthcare Applications. *Explainable Artificial Intelligence in Medical Imaging*, 58–78. <https://doi.org/10.1201/9781032626345-4>
- [14] Tariq, M., Iqbal, S., Ayesha, H., Abbas, I., Ahmad, K. T., & Niazi, M. F. K. (2021). Medical image based breast cancer diagnosis: State of the art and future directions. *Expert Systems with Applications*, 167, 114095. <https://doi.org/10.1016/j.eswa.2020.114095>
- [15] Little, M., McSharry, P., Hunter, E., Spielman, J., & Ramig, L. (2008). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *Nature Precedings*. <https://doi.org/10.1038/npre.2008.2298.1>
- [16] Tsanas, A., Little, M., McSharry, P., & Ramig, L. (2009). Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests. *Nature Precedings*. <https://doi.org/10.1038/npre.2009.3920.1>
- [17] Alshammri, R., Alharbi, G., Alharbi, E., & Almubark, I. (2023). Machine learning approaches to identify Parkinson's disease using voice signal features. *Frontiers in Artificial Intelligence*, 6. <https://doi.org/10.3389/frai.2023.1084001>
- [18] Karabayir, I., Goldman, S. M., Pappu, S., & Akbilgic, O. (2020). Gradient boosting for Parkinson's disease diagnosis from voice recordings. *BMC Medical Informatics and Decision Making*, 20(1). <https://doi.org/10.1186/s12911-020-01250-7>
- [19] Nogales, A., Garcia-Tejedor, A. J., Maitin, A. M., Perez-Morales, A., Castillo, M. D. D., & Romero, J. P. (2022). BERT Learns From Electroencephalograms About Parkinson's Disease: Transformer-Based Models for Aid Diagnosis. *IEEE Access*, 10, 101672–101682. <https://doi.org/10.1109/access.2022.3201843>
- [20] Rangel-Cascajosa, C., Luna-Perejón, F., Vicente-Diaz, S., & Domínguez-Morales, M. (2025). Gait-Based Parkinson's Disease Detection Using Recurrent Neural Networks for Wearable Systems. *Big Data and Cognitive Computing*, 9(7), 183. <https://doi.org/10.3390/bdcc9070183>
- [21] Li, A., Lian, J., & Vardhanabhuti, V. (2025). Multi-modal machine learning approach for early detection of neurodegenerative diseases leveraging brain MRI and wearable sensor data. *PLOS Digital Health*, 4(4), e0000795. <https://doi.org/10.1371/journal.pdig.0000795>
- [22] Skaramagkas, V., Pentari, A., Kefalopoulou, Z., & Tsiknakis, M. (2023). Multi-Modal Deep Learning Diagnosis of Parkinson's Disease—A Systematic Review. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31, 2399–2423. <https://doi.org/10.1109/tnsre.2023.3277749>