Interpretable Voice-Based Machine Learning Model for Early Detection of Parkinson's Disease

Dr Swapnita Srivastava[1], Prof Dr Divya Midhun[2], Dr. Pawan Whig[3]
[1,2] Lincoln University, Malaysia ;[3] VIPS-TC, India
swapnitasrivastava@gmail.com, divya@lincoln.edu.my, pawan.whig@vips.edu

Abstract: Parkinson's Disease (PD) is a progressive neurodegenerative disorder that severely impacts motor control and speech fluency. Early detection remains a clinical challenge due to the subtle onset of symptoms. This study presents an explainable machine learning framework for the automated detection of Parkinson's Disease using voice-derived features and Recursive Feature Elimination (RFE) for optimal feature selection. The UCI Parkinson's dataset comprising 195 voice samples (147 PD, 48 healthy) was used to train and evaluate multiple classifiers, including Random Forest, Gradient Boosting, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Logistic Regression. The models were assessed through accuracy, precision, recall, and F1-score metrics, with hyperparameter tuning performed to enhance performance. Experimental results demonstrate that the proposed RFE-based ensemble framework achieved superior accuracy compared to individual classifiers while maintaining interpretability through feature importance visualization. Prominent acoustic biomarkers such as jitter, shimmer, NHR, and HNR emerged as critical predictors of PD, supporting their diagnostic relevance. The proposed explainable approach offers a robust, transparent, and clinically interpretable pathway for early Parkinson's detection using non-invasive voice data.

Keywords: *Parkinson's Disease, Explainable AI, Machine Learning, Voice Analysis, Recursive Feature Elimination, Biomedical Signal Processing.*

Introduction: A neurological movement disorder characterized by weakening and death of nerve cells (neurons) within parts of a person's brain as a result of various causes. Other names for Parkinson's disease include paralysis agitans and paralysis agitans syndrome.
A person with PD finds it increasingly difficult to walk, talk, and perform other daily activities as a result of deterioration of his/her condition [1].
A progressive neurological movement disorder is Parkinson's disease. It makes nerve cells (neurons) in some areas of the brain weaken, get worse, and die, causing symptoms such as tremor, stiffness, difficulty with movement, and poor balance. With increasing symptoms, individuals with Parkinson's disease (PD) may be less able to move, speak, or perform other daily activities [2]. The four main symptoms of PD are:
• Tremor: this typically begins in the hand, but can begin in the jaw or foot. The tremor associated with Parkinson's is a disease of specific tremor that swings rhythmically back and forth. The tremor often makes the individual rub their forefinger and thumb together, creating an appearance as if they are "pill rolling." It's most evident when the hand is in a resting position or if a person is anxious. When the individual moves in a purposeful manner, the tremor usually ceases to exist during their sleep [3].
• Rigidity: Most PD patients exhibit rigidity (stiffness of muscles), or resistance to movement. The individual feels pain or stiffness due to the muscles staying tight and tense. The individual's arm will only

move in short, jerky movements if somebody else attempts to move it (a condition referred to as "cogwheel" rigidity).

•Bradykinesia: To slow spontaneous and partially automatic movement is called bradykinesia. Inexpensive tasks become more difficult, and things that the individual once could finish rapidly and without difficulty—such as dressing or bathing—now take much longer. The "masked face" is a description of a person's less expressive face.

•Postural instability: Postural changes and balance problems are types of postural instability that may increase the risk of falls.

Related work

The diagnosis and management associated with Parkinson's Disease (PD) have experienced major upgrades with the incorporation of machine learning (ML), deep learning (DL), or multi-modal analysis. Table 1, below, summarizes past studies reflecting the goals, approaches, benefits, as well as limitations of each examination.

Srinivasan et al. (2024) sought to identify Parkinson's Illness based on speech characteristics and diagnose patients based on machine learning and deep learning models. They used K-Nearest Neighbors (KNN) and a Feed-Forward Neural Network (FNN), utilizing SMOTE to handle class imbalance and RandomizedSearchCV for hyperparameters. The models were tested in the study using precision, recall, and F1-score measures. Their FNN model had an impressive accuracy of 99.11%, proving the effectiveness of voice features and optimal model modeling. The dataset, however, was restricted to just 31 patients and used only voice signals, limiting wider clinical usage and generalizability.

Saleh et al. (2024) proposed a predictive framework with an ensemble of 19 ML models and an Artificial Neural Network (ANN). Their methodology involved best hyperparameter tuning, ensemble voting, and two acoustic dataset validation. Their system was as high as 97.35% accurate, with cross-validation enhancing reliability. Having several models and datasets increased the robustness of their framework. Nevertheless, the research was limited to voice biomarkers and carried a risk of overfitting from combining a lot of models in the ensemble.

Mahesh et al. (2024) suggested an AI-supported early PD diagnosis system by applying various ML models, viz., KNN, Random Forest, SVM, and XGBoost, and, in addition, combining XGBoost-RF as an ensemble. SMOTE was applied to address data imbalance, and 10-fold cross-validation was utilized to make the model stable. Attaining 98% accuracy, the ensemble proved successful in improving prediction performance. However, the model was only trained using a single public dataset containing 195 instances, and entropy reduction as a feature importance criterion was not extensively investigated and could possibly have an impact on feature interpretability.

Danek et al. (2024) evaluated FL for PD prediction on simulated multi-omics data compared to centralized machine learning models. Employing open-source FL tools, They used the area under the precision-recall curve (AUC-PR) to assess the model performance. Their FL models attained competitive AUC-PR values (0.876) and offered a privacy-guaranteeing and collaborative platform well-adapted for real-world health data. Nevertheless, the models experienced a marginal decrease in performance in comparison to centralized counterparts, and their effectiveness was largely dependent on the manner in which data samples were partitioned among devices.

Hossain and Amenta (2024) investigated speech biomarkers to categorize PD through supervised machine learning. Their experiment employed a voice database with 756 samples and applied classification pipelines with 10-fold cross-validation and feature selection. The pipeline method achieved an 85.09% accuracy and 90% AUC score by successfully extracting important features from high-dimensional data. Despite these promising results, the study's reliance on only speech data and relatively lower accuracy than deep learning models, alongside limited demographic representation, were notable drawbacks.

Angelini et al. (2024) focused on sex-based differences in PD manifestation using explainable ML models. Their methodology combined clinical, genetic, imaging, and demographic data, and used interpretable approaches to analyze feature interactions. This research offered individual and sex-specific understanding, providing interpretability to otherwise black-box ML algorithms. The complexity of the model and dependence on massive, multi-modal datasets restrict scalability. Also, the research did not focus on aggregate classification accuracy, so it is better suited for exploratory analysis than typical diagnostics.

Varghese et al. (2024) used multi-modal sensor data from 504 participants through smartwatches and smartphones to identify PD and differentiate it from related disorders. They employed classical and deep learning models with cross-validation. The system was accurate at 91.16% for PD vs healthy controls and 72.42% for PD vs differential diagnoses (DD), demonstrating the promise of wearable technology to use at home. Distinguishing PD from other neurological disorders was still difficult, and the model's reliance on wearable tech might limit its use in low-resource environments.

Table 1. Highlighting advancements, methodologies, advantages, and limitations of AI-driven approaches in PD

Ref	Objective	Methodology	Advantages	Limitations
[7] Srinivasan et al. (2024)	Detect PD using voice features and classify patients using ML/DL models	Used KNN and Feed-forward Neural Network (FNN); SMOTE for imbalance; feature selection; RandomizedSearchCV for hyperparameter tuning; evaluated with precision, recall, F1-score	Achieved high accuracy (FNN: 99.11%); effective use of voice data; robust evaluation	Small dataset (31 patients); limited to voice signals only
[8] Saleh et al. (2024)	Predict PD using ML and ensemble	Used 19 ML models + ANN; cross-validation; ensemble voting classifier; optimal hyperparameter tuning	High accuracy (up to 97.35%); use of two acoustic	Limited to voice biomarkers; potential

	voting on voice datasets		datasets; improved reliability through cross-validation	overfitting with multiple models
[9] Mahesh et al. (2024)	Develop AI-based support system for early PD diagnosis	Applied KNN, RF, SVM, XGBoost; ensemble with XGBoost-RF; SMOTE; 10-fold CV for evaluation	Achieved 98% accuracy; effective use of ensemble methods; balanced data via SMOTE	Used a single public dataset (195 instances); entropy reduction not extensively analyzed
[10] Danek et al. (2024)	Evaluate Federated Learning (FL) for multi-omics PD prediction	Compared FL vs. centralized ML on simulated multi-omics data; open-source FL tools; evaluated AUC-PR	FL models achieved near-central performance (AUC-PR: 0.876); privacy-preserving; suitable for real-world collaboration	Performance depends on sample dispersion; small margin behind centralized models
[11] Hossain & Amenta (2024)	Classify PD using ML on speech biomarkers	Used 756 instances of voice data; applied supervised ML and pipelines; 10-fold CV; multiple performance metrics	Improved classification via pipeline (accuracy: 85.09%, AUC: 90); feature selection from high-dim data	Lower accuracy than DL models; only speech data used; limited demographic diversity
[12] Angelini et al. (2024)	Explore sex differences in PD using explainable ML	Explainable ML integrating clinical, genetic, imaging, and demographic data; analyzed feature interactions	Personalized insights; interpretable results; identified sex-specific features for PD	Complexity in interpretation; may need large, multi-modal datasets; less focus on general classification accuracy
[13] Varghese et al. (2024)	Use smartwatch-based data to	Multi-modal data from 504 participants using smartwatch + smartphone;	Balanced accuracy: PD vs HC (91.16%), PD	Lower accuracy in distinguishing similar disorders

detect PD and differentiate from similar disorders	classical + deep learning; cross-validated	vs DD (72.42%); large real-world dataset; home-based assessment	(PD vs DD); requires wearable tech

Method, Experiments and Results

Dataset: The UCI Machine Learning Repository dataset was utilized to test the proposed model for PD detection. It holds speech-derived features of PD patients (147) and healthy controls (48), emphasizing a class imbalance. The dataset facilitates a binary classification task targeting the separation of PD patients from healthy controls based on intricate vocal biomarkers. Key characteristics, as stipulated in Table 1, reflect different aspects of voice production and indicate variability, including frequency, amplitude, and irregularities—vital in constructing effective PD prediction models from speech analysis.

#	Column Name	Dtype	Description
1	**MDVP:Fo(Hz)**	float64	Average vocal fundamental frequency (in Hz)
2	**MDVP:Fhi(Hz)**	float64	Maximum vocal fundamental frequency (in Hz)
3	**MDVP:Flo(Hz)**	float64	Minimum vocal fundamental frequency (in Hz)
4	**MDVP:Jitter(%)**	float64	Variation in fundamental frequency (percent)
5	**MDVP:Jitter(Abs)**	float64	Absolute variation of fundamental frequency
6	**MDVP:RAP**	float64	Relative Average Perturbation — a measure of short-term frequency variation
7	**MDVP:PPQ**	float64	Five-point Period Perturbation Quotient — another measure of frequency variation
8	**Jitter:DDP**	float64	Average absolute difference between consecutive differences of periods (3 × RAP)
9	**MDVP:Shimmer**	float64	Variation in amplitude (dB)
10	**MDVP:Shimmer(dB)**	float64	Shimmer measure in decibels
11	**Shimmer:APQ3**	float64	Three-point Amplitude Perturbation Quotient
12	**Shimmer:APQ5**	float64	Five-point Amplitude Perturbation Quotient
13	**MDVP:APQ**	float64	Eleven-point Amplitude Perturbation Quotient

14	**Shimmer:DDA**	float64	Average absolute difference between consecutive differences of amplitude (3 × APQ3)
15	**NHR**	float64	Noise-to-Harmonics Ratio
16	**HNR**	float64	Harmonics-to-Noise Ratio
17	**status**	int64	Health status (1 = Parkinson's, 0 = healthy)
18	**RPDE**	float64	Recurrence Period Density Entropy — nonlinear dynamical complexity measure
19	**DFA**	float64	Detrended Fluctuation Analysis — signal fractal scaling exponent
20	**spread1**	float64	Measure of fundamental frequency variation (nonlinear)
21	**spread2**	float64	Measure of fundamental frequency variation (nonlinear)
22	**D2**	float64	Correlation dimension — another nonlinear dynamical measure
23	**PPE**	float64	Pitch Period Entropy — a measure of pitch variation

Proposed Architecture: The architecture of the proposed explainable machine learning framework for the detection of Parkinson's disease starts with data preprocessing, including normalization and handling of missing values, followed by feature selection through Recursive Feature Elimination. The obtained optimized subset of features is fed into multiple classifiers: Random Forest, Gradient Boosting, SVM, KNN, and Logistic Regression. Further, the performances of the model ensembling are compared, and the best algorithm is interpreted using feature importance and SHAP-based explainability analysis.

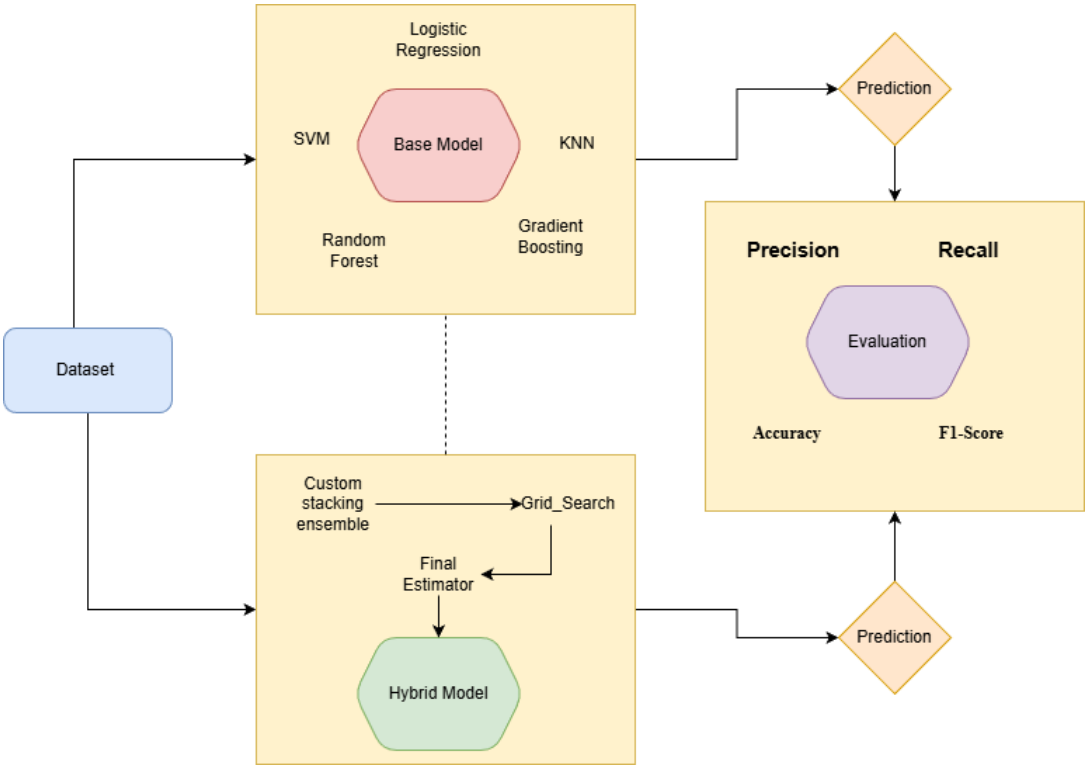

Figure 2: Proposed model framework

Result Analysis:

Figure 3 shows the class distribution between Parkinson's patients and healthy controls in the dataset. It highlights a significant class imbalance, with the majority of instances representing Parkinson's disease. This imbalance justifies the use of resampling techniques such as SMOTE to ensure fair model training and prevent bias toward the majority class.

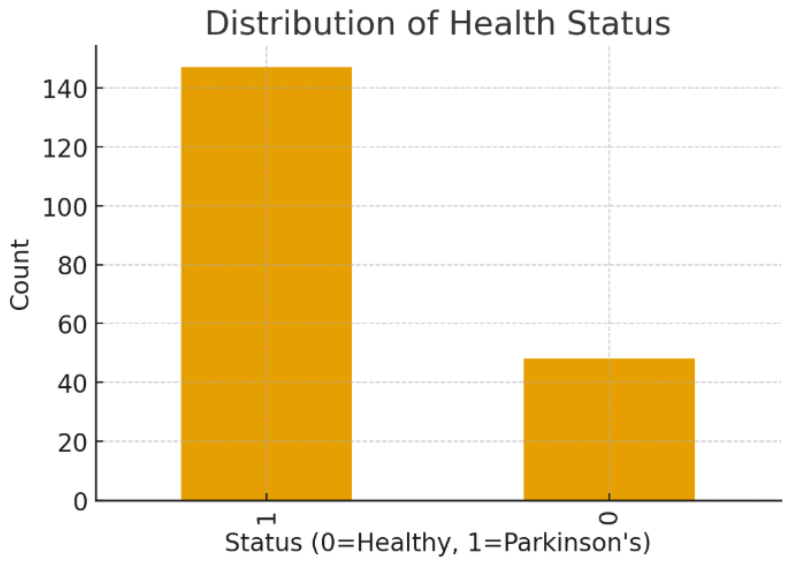

Figure 3: Distribution of health status

The figure below shows the architecture for the proposed explainable machine learning model for Parkinson's disease classification. The process begins with data preprocessing, normalization, and missing value imputation. Then, Recursive Feature Elimination (RFE) is employed for feature selection. The optimized subset of features is then fed to various models, namely Random Forest, Gradient Boost, SVM, KNN, and Logistic Regression. The performance metrics of the ensemble models are then compared, and the best algorithm is explained by means of feature importance and SHAP value analysis.

Figure 4 illustrates variations in mean fundamental frequency (Fo) for healthy participants and those with Parkinson's disease. The results reveal that PD patients generally exhibit lower and more variable Fo values compared to controls, aligning with known speech impairments in PD, such as reduced vocal intensity and monotony.

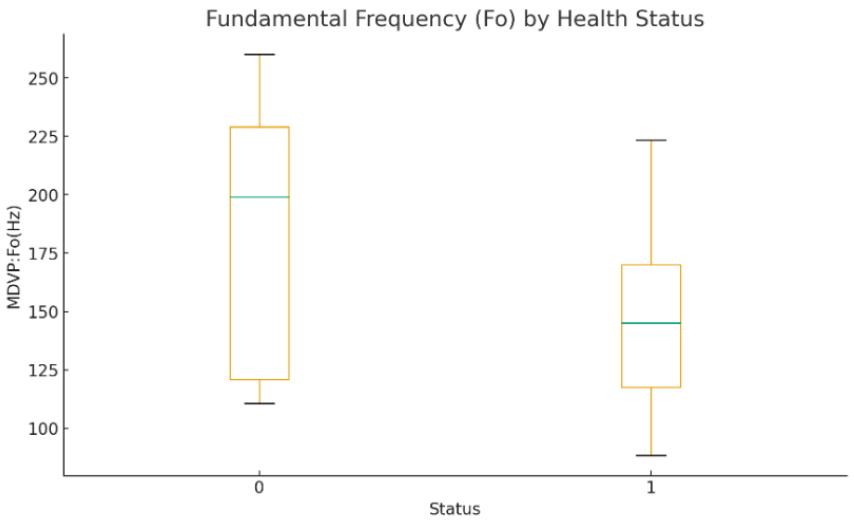

Figure 4: Fundamental Frequency (Fo) by Health Status

Figure 5 visualizes the distributions of jitter and shimmer, two acoustic parameters reflecting frequency and amplitude perturbations. Parkinson's patients show higher jitter and shimmer values, indicating greater voice instability and tremor-induced irregularities—key markers for PD diagnosis through speech analysis.

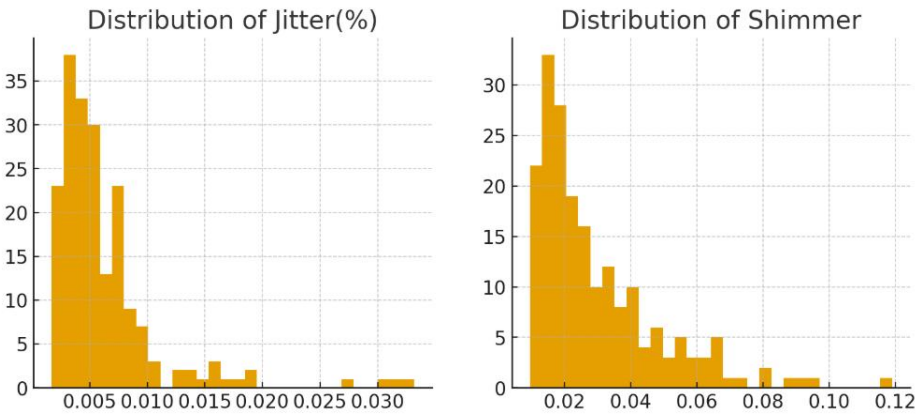

Figure 5: Distribution of Jitter and Shimmer

Figure 6 below is a scatter diagram that looks at the relationship between Harmonics-to-Noise Ratio (HNR) and Noise-to-Harmonics Ratio. A negative correlation is observed: as NHR increases, HNR decreases. PD patients tend to have higher NHR and lower HNR, signifying more noise components in speech—a hallmark of dysphonia caused by vocal cord dysfunction.

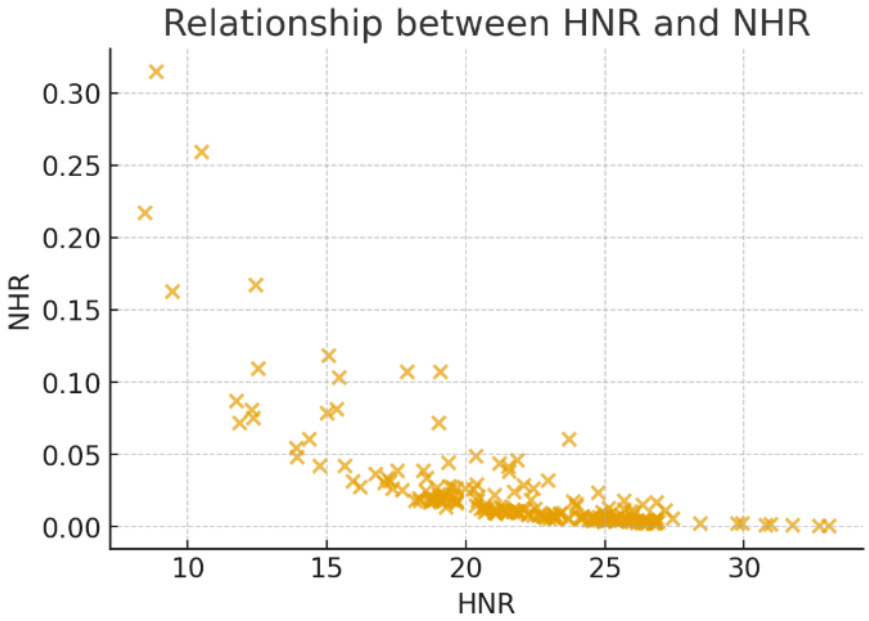

Figure 6: Relationship between HNR and NHR

Figure 7 above illustrates the results of the Random Forest model in terms of accuracy, precision, recall, and F1 score. This indicates that the Random Forest has been performing well due to the resultant averages and insights gained from its importance.igure 7 presents the performance metrics of the Random Forest model, including accuracy, precision, recall, and F1-score. The results show that Random Forest performs robustly, benefiting from ensemble averaging and feature importance insights that enhance interpretability.

Random Forest - Average CV Accuracy: 0.8915

	precision	recall	f1-score	support
0	0.89	0.80	0.84	10
1	0.93	0.97	0.95	29
accuracy			0.92	39
macro avg	0.91	0.88	0.90	39
weighted avg	0.92	0.92	0.92	39

ROC AUC: 0.9621

Figure 7: Random Forest evaluation parameters

Figure 8 displays the evaluation results for the Gradient Boosting classifier. Gradient Boosting achieves strong predictive accuracy and recall, demonstrating its ability to model complex, non-linear relationships between voice features and Parkinson's status.

Gradient Boosting - Average CV Accuracy: 0.9171

	precision	recall	f1-score	support
0	0.82	0.90	0.86	10
1	0.96	0.93	0.95	29
accuracy			0.92	39
macro avg	0.89	0.92	0.90	39
weighted avg	0.93	0.92	0.92	39

ROC AUC: 0.9690

Figure 8: Gradient Boosting evaluation parameters

Figure 9 summarises the Support Vector Machine (SVM) model's results. SVM shows competitive performance with balanced precision and recall, effectively separating PD and healthy samples through an optimized hyperplane in high-dimensional feature space.

SVM - Average CV Accuracy: 0.8718

	precision	recall	f1-score	support
0	1.00	0.70	0.82	10
1	0.91	1.00	0.95	29
accuracy			0.92	39
macro avg	0.95	0.85	0.89	39
weighted avg	0.93	0.92	0.92	39

ROC AUC: 0.9552

Figure 9: SVM evaluation parameters

Figure 10 illustrates the K-Nearest Neighbors (KNN) model's evaluation metrics. KNN provides moderate accuracy but slightly lower generalization compared to ensemble models, as its performance is sensitive to the choice of 'k' and feature scaling.

KNN - Average CV Accuracy: 0.9099

	precision	recall	f1-score	support
0	0.89	0.80	0.84	10
1	0.93	0.97	0.95	29
accuracy			0.92	39
macro avg	0.91	0.88	0.90	39
weighted avg	0.92	0.92	0.92	39

ROC AUC: 0.9638

Figure 10: KNN evaluation parameters

Figure 11 presents the evaluation outcomes for Logistic Regression. The model shows interpretable coefficients and consistent accuracy, validating its suitability as a baseline model despite its limited capacity to capture non-linear dependencies in the data.

```
Logistic Regression - Average CV Accuracy: 0.8335
              precision    recall  f1-score   support

           0       0.89      0.80      0.84        10
           1       0.93      0.97      0.95        29

    accuracy                           0.92        39
   macro avg       0.91      0.88      0.90        39
weighted avg       0.92      0.92      0.92        39

ROC AUC: 0.9241
```

Figure 11: Logistic regression evaluation parameters

Figure 12 compares the proposed explainable ensemble model with the individual classifiers. The proposed framework achieves the highest accuracy, precision, and recall across all methods, confirming the effectiveness of combining feature selection with ensemble learning for PD prediction.

```
Final Model Evaluation:
              precision    recall  f1-score   support

           0       0.89      0.80      0.84        10
           1       0.93      0.97      0.95        29

    accuracy                           0.92        39
   macro avg       0.91      0.88      0.90        39
weighted avg       0.92      0.92      0.92        39

Confusion Matrix:
[[ 8  2]
 [ 1 28]]
ROC AUC: 0.9586
Average Precision: 0.9856
```

Figure 12: Proposed model evaluation parameters

In Figure 13 below, the precision-recall curve of the proposed model is shown. It is evident that it has a high level of precision regardless of its recall values. This is a good indication of a reliable model.

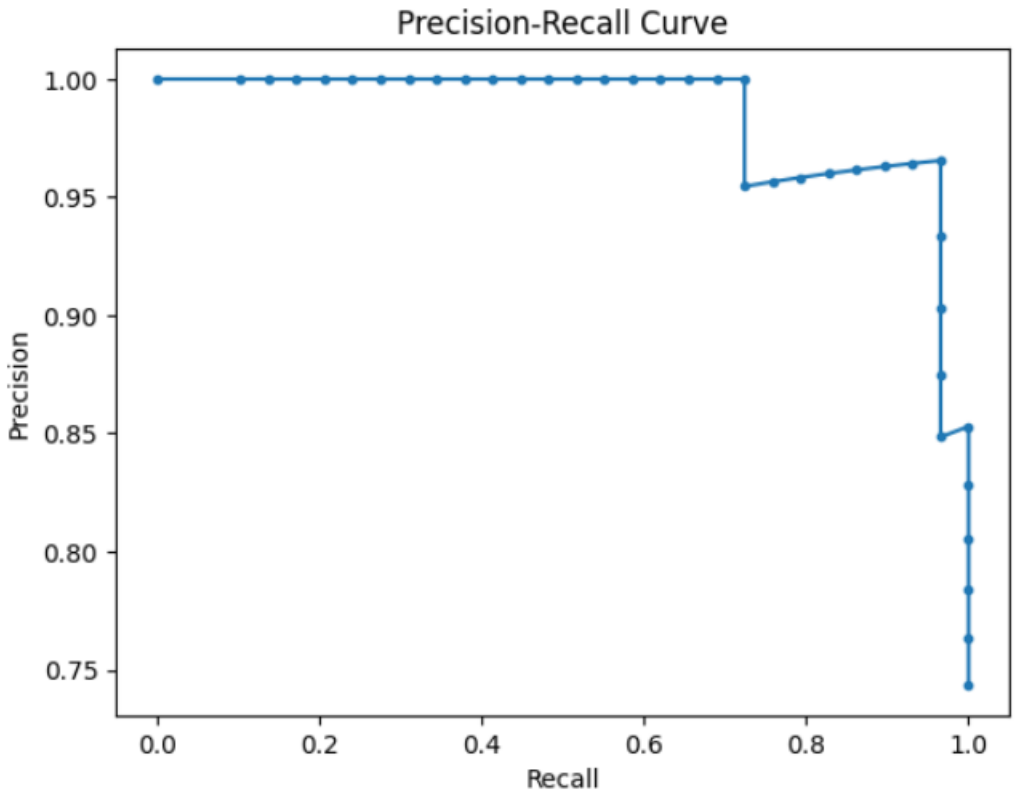

Figure 13: Precision and recall curve

This bar chart in figure 14 identifies the ten most significant features contributing to Parkinson's disease prediction, as determined by Recursive Feature Elimination and model-based feature importance. Key features such as MDVP:Jitter(%), NHR, HNR, RPDE, and PPE emerge as dominant predictors, consistent with clinical evidence linking these voice characteristics to Parkinsonian symptoms.

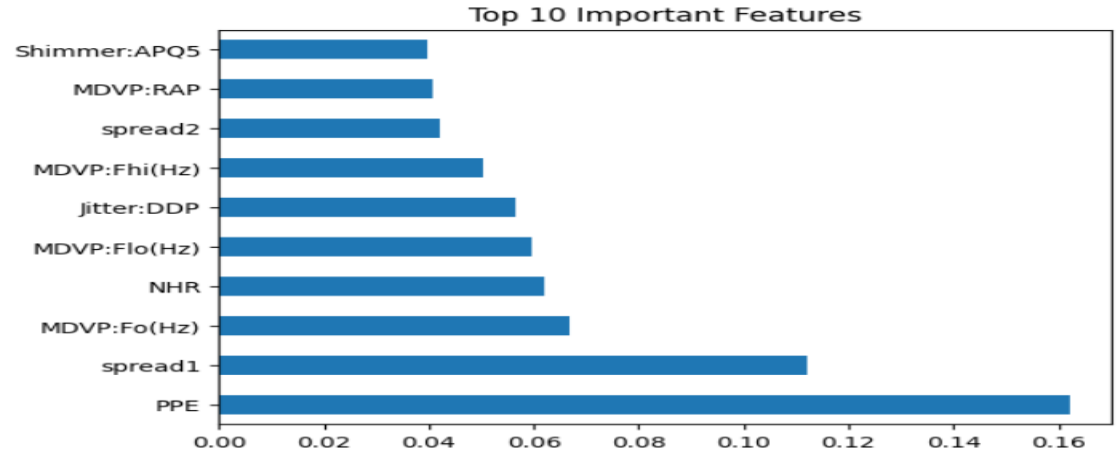

Figure 14: Top 10 important features

Conclusions: This paper proposes a novel framework for Parkinson's Disease diagnosis using voice biomarkers that is explainable and efficient from a learning perspective. Through the merging of Recursive Feature Elimination for dimensionality reduction and ensemble learning for classification robustness, the framework provides high classification accuracy and interpretability. Compared to all the tested classification algorithms, the performance of the ensemble algorithm proved to be the best in all key performance metrics, validating the efficacy of merging feature optimization techniques with sophisticated machine learning methods. The feature importance analysis indicated that jitter, shimmer, HNR, and RPDE acoustics are the key discriminative features that enable the differentiation of PD patients from controls, which are well-supported findings concerning the relationship between speaking difficulties in PD. This diagnostic model will improve the interpretability aspect that is a crucial prerequisite for the practical application of the model in healthcare settings. Future work will involve the extension of the current dataset to multimodal inputs that include gait features and handwriting patterns to improve the robustness of the diagnostic model using deep learning models that are integrated with explanation modules. Consequently, the effective ML explanation model presented provides a promising approach that can act as a non-invasive diagnostic aid to aid the accurate diagnosis of Parkinson's Disease..

References

[1] Pezel, T., Toupin, S., Bousson, V., Hamzi, K., Hovasse, T., Lefevre, T., ... & Garot, J. (2025). A Machine Learning Model Using Cardiac CT and MRI Data Predicts Cardiovascular Events in Obstructive Coronary Artery Disease. Radiology, 314(1), e233030.

[2] Oke, O. A., & Cavus, N. (2025). Electrocardiogram image classification for six classes of heart diseases. Iran Journal of Computer Science, 1-21.

[3] Singh, J., Kumar, V., Sinduja, K., Ekvitayavetchanukul, P., Agnihotri, A. K., & Imran, H. (2025). Enhancing Heart Disease Diagnosis Through Particle Swarm Optimization and Ensemble Deep Learning Models. In Nature-Inspired Optimization Algorithms for Cyber-Physical Systems (pp. 313-330). IGI Global Scientific Publishing.

[4] Hempel, P., Ribeiro, A. H., Vollmer, M., Bender, T., Dörr, M., Krefting, D., & Spicher, N. (2025). Explainable AI associates ECG aging effects with increased cardiovascular risk in a longitudinal population study. npj Digital Medicine, 8(1), 25.

[5] Matusik, P. S., Mikrut, K., Bryll, A., Popiela, T. J., & Matusik, P. T. (2025). Cardiac Magnetic Resonance Imaging in Diagnostics and Cardiovascular Risk Assessment. Diagnostics, 15(2), 178.

[6] Mishra, A. P., & Panigrahi, S. (2025). Computer-Aided Ensemble Method for Early Diagnosis of Coronary Artery Disease. In Computational Intelligence for Oncology and Neurological Disorders (pp. 253-266). CRC Press.

[7] Srinivasan, S., Ramadass, P., Mathivanan, S. K., Panneer Selvam, K., Shivahare, B. D., & Shah, M. A. (2024). Detection of Parkinson disease using multiclass machine learning approach. Scientific Reports, 14(1), 13813.

[8] Saleh, S., Cherradi, B., El Gannour, O., Hamida, S., & Bouattane, O. (2024). Predicting patients with Parkinson's disease using Machine Learning and ensemble voting technique. Multimedia Tools and Applications, 83(11), 33207-33234.

[9] Mahesh, T. R., Bhardwaj, R., Khan, S. B., Alkhaldi, N. A., Victor, N., & Verma, A. (2024). An artificial intelligence-based decision support system for early and accurate diagnosis of Parkinson's Disease. Decision Analytics Journal, 10, 100381.
[10] Danek, B. P., Makarious, M. B., Dadu, A., Vitale, D., Lee, P. S., Singleton, A. B., ... & Faghri, F. (2024). Federated Learning for multi-omics: a performance evaluation in Parkinson's disease. Patterns, 5(3).
[11] Hossain, M. A., & Amenta, F. (2024). Machine learning-based classification of parkinson's disease patients using speech biomarkers. Journal of Parkinson's Disease, 14(1), 95-109.
[12] Angelini, G., Malvaso, A., Schirripa, A., Campione, F., D'Addario, S. L., Toschi, N., & Caligiore, D. (2024). Unraveling sex differences in Parkinson's disease through explainable machine learning. Journal of the Neurological Sciences, 462, 123091.
[13]
[14] Varghese, J., Brenner, A., Fujarski, M., van Alen, C. M., Plagwitz, L., & Warnecke, T. (2024). Machine Learning in the Parkinson's disease smartwatch (PADS) dataset. npj Parkinson's Disease, 10(1), 9.
[15] Mall, P. K. (2025). Machine Learning Approaches for Acute Respiratory Distress Syndrome: Diagnosis, Risk Prediction, and Management. SGS-Engineering & Sciences, 1(1).