Explainable Machine Learning Framework for Parkinson's Disease Detection Using Voice Features and Recursive Feature Elimination

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), a chronic neurodegenerative condition, is still a serious clinical problem because of its multifactorial etiology and insidious early-stage symptoms. This work introduces an explainable machine learning (ML) approach that combines Recursive Feature Elimination (RFE) and Logistic Regression to achieve reliable and interpretable PD classification based on vocal biomarkers. Using the UCI Parkinson's Disease dataset of speech-derived features from 188 PD patients and 64 healthy controls, the model involves preprocessing, handling class imbalance through KMeansSMOTE, and dimensionality reduction using RFE. The suggested model exhibits better testing accuracy (96.46%) and balanced performance on precision, recall, and F1-score metrics compared to traditional models like Random Forest, XGBoost, and SVM. Model explainability is improved by SHAP (SHapley Additive exPlanations), which offers clear and transparent explanations of feature importance. The findings confirm that integrating feature selection and explainable AI yields highly performing and reliable PD detection models with good possibilities for early diagnosis and clinical decision support.

Keywords: *Parkinson's Disease; Heritability; Machine Learning; Mental Health Disorders; Genetic-Environmental Interaction.*

Introduction: Parkinson's disease is a neurological movement disease that is progressive. It causes weakening, damage, and death of nerve cells (neurons) in certain parts of the brain, leading to symptoms such as stiffness, tremor, trouble moving, and impaired balance. Individuals who have Parkinson's disease (PD) can experience it being more difficult to walk, talk, or do other everyday activities as their symptoms get worse [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 disease is a 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 monitoring of Parkinson's Disease (PD) have seen significant advancements through the integration of ML, DL, and multi-modal data analysis. Table 1 below provides a summary of recent studies, highlighting their objectives, methodologies, advantages, and limitations.

Srinivasan et al. (2024) sought to identify Parkinson's Disease based on voice 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 evaluated model performance in terms of the area under the precision-recall curve (AUC-PR). 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 voting on voice datasets	Used 19 ML models + ANN; cross-validation; ensemble voting classifier; optimal hyperparameter tuning	High accuracy (up to 97.35%); use of two acoustic datasets;	Limited to voice biomarkers; potential overfitting with multiple models

			improved reliability through cross-validation	
[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 detect PD and	Multi-modal data from 504 participants using smartwatch + smartphone;	Balanced accuracy: PD vs HC (91.16%), PD vs DD (72.42%);	Lower accuracy in distinguishing similar disorders (PD vs DD);

differentiate from similar disorders	classical + deep learning; cross-validated	large real-world dataset; home-based assessment	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 (188) and healthy controls (64), 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.

Proposed Architecture: The framework being proposed is a solid pipeline for Parkinson's Disease (PD) detection from speech data incorporating a mix of data preprocessing, feature selection, model training, and explainability. The method starts with data preprocessing, where the removal of irrelevant columns and standardization of data to maintain homogeneity are carried out. At the same time, exploratory data analysis (EDA) is done to visualize distributions of classes and identify missing and outlier values.

For handling class imbalance, KMeansSMOTE is utilized, which balances the dataset synthetically through clustering-based oversampling. The dataset is then partitioned into a training and a test set. Recursive Feature Elimination (RFE) using Logistic Regression is utilized to perform dimensionality reduction and retain the most important features, enhancing model performance as well as generalizability.

Model training is done with hyperparameter tuning by GridSearchCV to optimize the Logistic Regression classifier. After training, the model is tested against major performance metrics like accuracy, precision, recall, F1-score, and ROC-AUC.

To facilitate the interpretability and transparency of the predictions, SHAP (SHapley Additive exPlanations) is embedded as part of the explainable AI module. SHAP offers both local and global explanations of model choices, allowing clinicians and researchers to see the impact of specific features on predictions.

The ultimate output is a binary judgment—either classifying the subject as having PD or healthy—with explainable information presented in each judgment that facilitates trust and clinical usefulness.

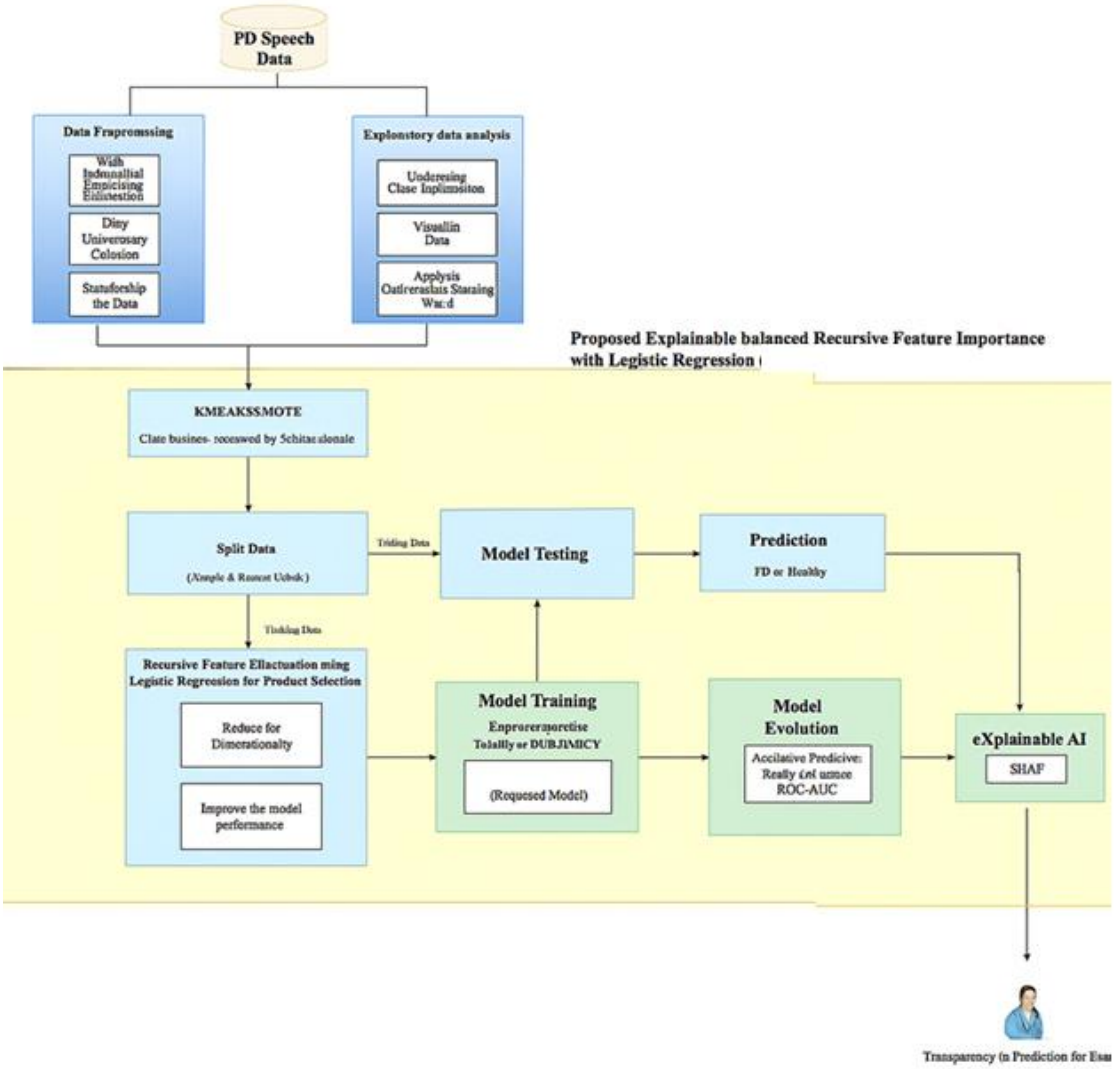

Figure 2: Proposed model framework

Algorithm 1 outlines the step-by-step workflow of the proposed model for feature selection using Recursive Feature Elimination (RFE) with Logistic Regression. The process begins by loading the input dataset D, which includes feature set F and target labels YYY. The data is first preprocessed by eliminating irrelevant or non-informative columns and standardizing the remaining features to ensure uniformity. A Logistic Regression model is then initialized with appropriate hyperparameters (e.g., setting max_iter to 1000). The desired number of features to retain is specified as $n\_features\_to\_select$, after which the RFE process is initiated. In this iterative loop, the model is trained on the current feature set, feature importance scores (typically model coefficients) are calculated, and the least important feature is removed. This cycle continues until the specified number of features remains. Finally, the algorithm outputs the optimal subset of features F_{opt}, which can be used for further model training or analysis.

Algorithm 1: Proposed model step by step flow
Input: Dataset D with features F and target label Y
Output: Selected optimal subset of features F_opt

1. Start
2. Load Dataset D
3. Preprocess Data:
a. Remove irrelevant or non-informative columns
b. Standardize the remaining features in F
4. Initialize Logistic Regression Model:
a. Set base estimator to Logistic Regression
b. Set model parameters (e.g., max_iter = 1000)
5. Specify the number of features to select → n_features_to_select
6. Begin Recursive Feature Elimination:
while (number of current features > n_features_to_select):
a. Fit the Logistic Regression model on the current feature set
b. Evaluate feature importances (e.g., coefficients)
c. Identify and remove the least important feature
d. Repeat model training with updated feature set
7. End loop when the desired number of features is selected
8. Output the selected feature subset F_opt
9. End

Result Analysis:

The Figure 3 shows a comparative comparison of several machine learning models against major performance indicators: training accuracy, testing accuracy, precision, recall, and F1-score. Out of the models compared, the suggested classifier—Recursive Feature Elimination (RFE) using Logistic Regression—has the most generalization power, recording an overall performance with a testing accuracy, precision, recall, and F1-score of 96.46% despite a low training accuracy of 97.23%. Although other models such as Random Forest, MLP, and XGBoost achieve perfect training accuracy, their generalization for testing data is relatively lower, suggesting overfitting. The KNN classifier is the weakest in all measures. In general, the proposed model ranks better than others in terms of high accuracy and consistency in all measures of evaluation, affirming its strength and applicability to trusted prediction tasks.

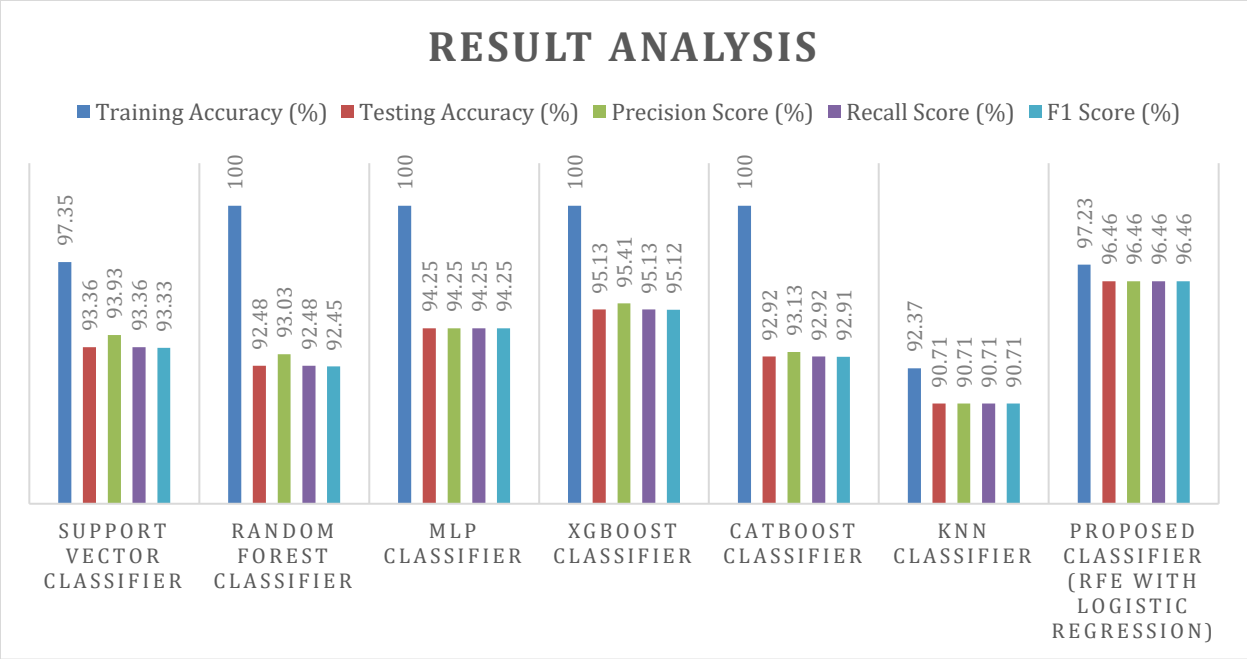

Figure 3: Comparative result analysis between different ML model with proposed model

Figure 4 shows a Receiver Operating Characteristic (ROC) curve plot between different machine learning classifiers. The suggested Logistic Regression model, optimized with feature selection, has a high AUC of 0.99, which indicates its high capacity to differentiate classes with low false positives. Similar high performance is observed in the MLP Classifier (AUC = 0.99), followed closely by Random Forest, SVM, CatBoost, and XGBoost with AUC values of 0.98, an observation that reflects high predictive ability in these models. The KNN classifier, while slightly lower with an AUC of 0.97, also reflects good classification ability. Overall, the ROC curves affirm the Logistic Regression model's robustness and reliability, showing that it offers competitive, if not superior, discriminative performance compared to more complex models.

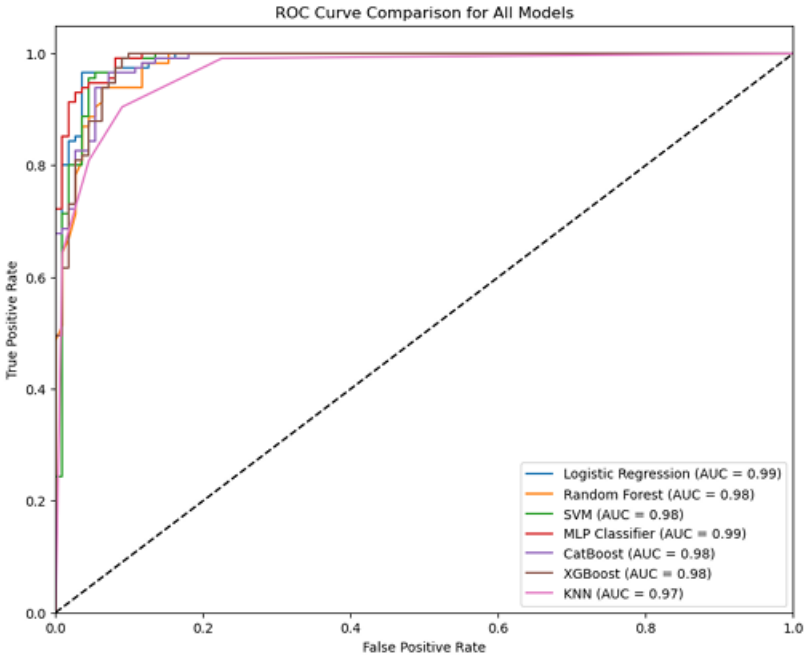

Figure 4: ROC curve comparison for all models

Figure 5 presents the SHAP (SHapley Additive exPlanations) summary plot, indicating the contribution of each feature to the output of the model both in magnitude and direction. Dots refer to a SHAP value for each feature in a prediction, where colors represent feature values (red for high, blue for low). Features are ordered in priority, with Feature 41, Feature 25, and Feature 26 having the highest influence on the model prediction. The horizontal range of each feature indicates variability in influence across samples, while color distribution indicates how high or low values nudge the model towards a specific class. The story also shows that there are positive as well as negative SHAP values, which means that features may drive predictions towards Parkinson's or healthy classes. This interpretability helps us comprehend the effect of certain speech characteristics toward the final prediction, helping in transparency and ensuring clinical confidence in the AI system.

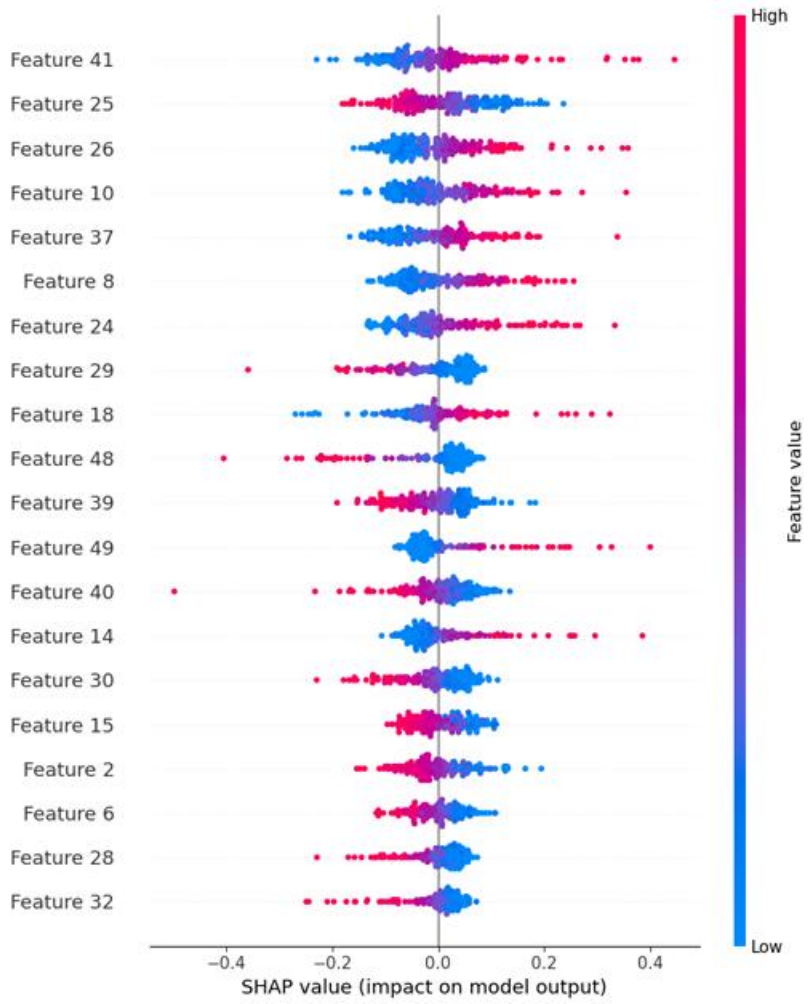

Figure 5: XAI (SHAP) impact on model output

Conclusions: The paper offers a stable and explainable machine learning pipeline for identifying Parkinson's Disease from speech features. Utilizing Recursive Feature Elimination with Logistic Regression, the model is able to produce a test accuracy of 96.46% with high precision, recall, and F1-scores with better generalization ability than more sophisticated models such as Random Forest, XGBoost, and MLP. The combination of KMeansSMOTE resolves data imbalance well, and SHAP-based explainability promotes transparent, clinician-acceptable predictions. The suggested model not only retains competitive accuracy but also focuses on interpretability—important for healthcare. The strategy could be the basis for scalable, non-invasive, and interpretable PD screening tools. Future research will continue to validate the model in varied datasets and incorporate multi-modal biomarkers for extended applicability.

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] 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.
[14] Mall, P. K. (2025). Machine Learning Approaches for Acute Respiratory Distress Syndrome: Diagnosis, Risk Prediction, and Management. SGS-Engineering & Sciences, 1(1).