

Prediction of Parkinson's disease using feature selection and ensemble learning techniques

Sharan T. D., Sujata Joshi

Department of Computer Science and Engineering, Nitte Meenakshi Institute of Technology, Bengaluru, India

Article Info

Article history:

Received Dec 13, 2024

Revised Apr 3, 2025

Accepted Jul 2, 2025

Keywords:

Feature selection
Parkinson's disease
SMOTE
Speech biomarkers
XGBoost

ABSTRACT

Parkinson's disease (PD) is a progressive neurodegenerative disorder that significantly impacts quality of life and healthcare systems. Early detection is crucial for timely interventions that can mitigate disease progression and improve patient outcomes. This study leverages advanced machine learning (ML) techniques to detect PD using speech features as non-invasive biomarkers. A dataset containing 754 features derived from sustained vowel phonations of 252 individuals (188 PD patients, 64 healthy controls) was analyzed. The dataset, originally collected by Istanbul University and publicly hosted via the UCI ML repository, was accessed through Kaggle for preprocessing and analysis. To identify the most predictive features, we employed recursive feature elimination (RFE), random forest importance, lasso regression, and the Boruta algorithm—ensuring robust feature selection while reducing dimensionality. The XGBoost model, optimised using the synthetic minority oversampling technique (SMOTE) for class balancing, achieved an accuracy of 96.69%, a recall of 96%, and an F1-score of 98%. Model robustness was validated through 5-fold cross-validation, yielding an average accuracy of 89.54%. These findings establish a scalable, cost-effective, and non-invasive framework for early PD detection, demonstrating the potential of speech analysis and ML in neurodegenerative disease management.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Sharan T. D.

Department of Computer Science and Engineering, Nitte Meenakshi Institute of Technology

Yelahanka, Bengaluru, India

Email: shrn282@gmail.com

1. INTRODUCTION

Parkinson's disease (PD) is a progressive neurodegenerative disorder affecting approximately 10 million individuals worldwide [1]. It primarily impairs the central nervous system, leading to motor dysfunctions such as tremors, stiffness, bradykinesia, and postural instability [2]-[4]. Additionally, non-motor symptoms, including cognitive decline, speech impairments, sleep disturbances, and depression, significantly impact patients' quality of life and pose challenges for caregivers and healthcare systems [5]-[7].

Early diagnosis of PD is crucial to initiate timely therapeutic interventions, mitigate symptom progression, and prolong functional independence [8]-[9]. However, traditional diagnostic methods rely heavily on subjective clinical assessments, leading to inter-observer variability and late-stage detection [10]-[12]. This underscores the need for objective, scalable, and early diagnostic methodologies. Speech analysis has emerged as a promising non-invasive biomarker for PD diagnosis [13]-[16]. The disease affects phonatory and motor control systems, leading to subtle alterations in jitter, shimmer, and harmonicity, which often precede overt motor symptoms [17], [18]. Recent advancements in machine learning (ML) and deep

learning (DL) enable automated analysis and interpretability of these speech features, offering a pathway for robust and consistent PD prediction [19]-[21].

This paper proposes an ML-based PD detection framework utilizing the full “Parkinson’s disease speech features” dataset, originally collected by the Department of Neurology at Istanbul University and hosted via the UCI ML repository [22]. The dataset includes 754 extracted features from sustained vowel phonation recordings of 252 individuals (188 PD patients and 64 healthy controls), covering diverse demographic backgrounds. By addressing challenges like high-dimensionality, class imbalance, and model interpretability, this work integrates multi-method feature selection, ensemble learning techniques, and interpretable AI to deliver a robust, scalable, and clinically relevant solution for early PD detection.

2. RELATED WORK

Govindu and Palwe [23] applied SVM, KNN, logistic regression (LR), and random forest (RF) to the MDVP audio dataset for PD classification. Their RF model achieved 91.83% accuracy, but the small dataset (30 patients) increased susceptibility to overfitting, limiting generalisability.

Makarious *et al.* [24] developed a multi-modality ML model integrating PPMI and PDBP datasets, utilising RF, SVM, and DL techniques. Their model initially achieved an AUC of 89.72%, later optimised to 85.03%. However, while their approach incorporated genetic, transcriptomic, and clinical markers, its reliance on omics-based testing and imaging data reduces accessibility for real-world clinical use.

Zhang *et al.* [25] compared decision tree (DT), KNN, Naïve Bayes, RF, SVM, and XGBoost on the PPMI dataset. Their penalised LR model reached an AUC of 0.94, while XGBoost achieved 0.92. Although their study categorised PD risk factors based on cost and accessibility, it lacked explicit class balancing and external validation beyond the dataset constraints.

Wang *et al.* [20] developed a deep learning (DNN) and ML-based PD classification system using the PPMI dataset (584 individuals: 401 early PD, 183 healthy controls). Their DNN model achieved 96.68% accuracy. However, the lack of explicit feature selection made the model prone to overfitting and less interpretable for clinical applications.

Alshammri *et al.* [26] applied KNN, SVM, DT, RF, and multi-layer perceptron (MLP) to the UCI voice dataset (195 speech samples). Their MLP model outperformed SVM (95%) with an accuracy of 98.31%. However, as their study relied solely on speech features, it failed to consider non-motor PD symptoms, limiting its real-world applicability.

Saeed *et al.* [27] applied KNN, SVM, Naïve Bayes, RF, and MLP on a UCI voice dataset (240 recordings, 46 features), using filter-based (PCA, IG) and wrapper-based (PSO, Greedy Stepwise) feature selection. Their best model (KNN with wrapper selection) achieved 88.33% accuracy. However, their smaller dataset and feature set limited generalizability.

Nahar *et al.* [28] investigated feature selection-based classification for early PD detection, applying Boruta, RFE, RF, XGBoost, Bagging, and Extra Trees Classifier on the UCI dataset. Their best model (Bagging) achieved 82.35% accuracy, but the small dataset (80 participants) and focus on speech-only features reduced its broader clinical utility.

Ali *et al.* [29] proposed an ensemble learning model (EOFSC) integrating deep neural networks (DNN) with feature selection. Their approach, focused on multiple vowel phonations, achieved 95% accuracy. However, their reliance on phonation-specific features and majority voting limited interpretability and feature generalization.

Varghese *et al.* [30] applied SVM, DTs, linear regression, and support vector regression (SVR) on the UCI Parkinson’s Telemonitoring dataset to predict motor and total UPDRS scores. Their model achieved an RMSE of 7.49 for Total UPDRS and 6.06 for Motor UPDRS, demonstrating effective severity prediction rather than direct PD classification.

Srinivasan *et al.* [31] proposed a multiclass classification approach for PD detection using voice signals, comparing KNN, KSVM, DT, RF, and feed-forward neural network (FNN) on the UCI dataset (31 patients, 195 voice samples). Their FNN model achieved 99.11% accuracy but relied entirely on DL, making it computationally expensive and harder to interpret for clinical deployment.

3. PROPOSED METHOD

Our proposed model integrates advanced feature selection techniques, data balancing methods, and optimised ML algorithms to achieve robust PD detection. The UCI Parkinson’s Speech Dataset, which includes 252 participants (188 PD, 64 healthy controls), is used to train the model. Unlike previous studies (Table 1) that rely solely on speech features, our approach incorporates a combination of speech and symptom-based features to improve accuracy and real-world applicability. XGBoost, optimised via

Randomised Search CV, is employed as the primary classification model, achieving 96.69% accuracy. The flowchart of the proposed model is as in Figure 1.

Table 1. Summary of related work

Authors	Best method	Type of modallity	Performance	Volume of dataset
Govindu and Palwe [23]	Random Forest (RF)	MDVP audio Data	91.83%	31 people
Makarious <i>et al.</i> [24] Zhang <i>et al.</i> [25]	Deep Learning (DL) Penalized LR	PPMI, PDBP PPMI	83.95% AUC of 0.94	756 people 747 people (missing values greater than 10 excluded)
Wang <i>et al.</i> [20]	Deep Neural Network	PPMI	96.68%	584 individuals
Alshammri <i>et al.</i> [26]	Multi-Layer Perceptron (MLP)	UCI dataset	98.31%	195 records of voice signal features
Saeed <i>et al.</i> [27]	KNN + Feature Selection	Kaggle	88.33%	240 recordings from 80 PD patients
Nahar <i>et al.</i> [28] Ali <i>et al.</i> [29]	Bagging Classifier Deep Learning + FScore	UCI dataset Multi-type vowel phonations dataset	82.35% 95%	80 participants 160 subjects
Varghese <i>et al.</i> [30]	Support Vector Regression	UCI ML repository	RMSE 7.49 (UPDRS)	42 candidates
Srinivasan <i>et al.</i> [31]	Feed-Forward Neural Network	UCI ML Repository	99.11%	31 people

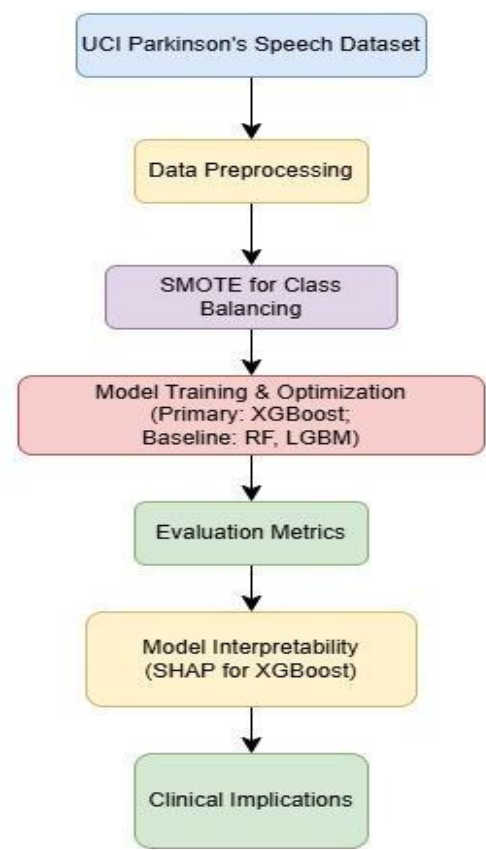


Figure 1. Flowchart of the proposed model

3.1. Dataset collection

The dataset used in this study comprises 756 samples and 754 extracted features, sourced from sustained phonation recordings of the vowel sound “/a/.” It contains data from 252 individuals: 188 PD patients (107 males, 81 females) and 64 healthy controls (23 males, 41 females). Participants ranged in age from 33 to 87 years (mean age: 65.1 ± 10.9 for PD, 61.1 ± 8.9 for controls). The dataset was originally collected by the Department of Neurology, Cerrahpaşa Faculty of Medicine, Istanbul University, under physician supervision, and made publicly available via the UCI ML repository [22]. For analysis purposes, it was accessed through Kaggle, which mirrors the original dataset. During data collection, participants were instructed to sustain the phonation of the vowel “a” in three repetitions, recorded using a standard microphone at a sampling rate of 44.1 kHz. All data participants in Table 2 is de-identified and ethically cleared for research use.

Table 2. Dataset summary

Dataset attribute	Description
Total Participants	252
PD Patients	188 (107 males, 81 females)
Healthy Controls	64 (23 males, 41 females)
Total Features	754
Sampling Rate	44.1 kHz
Data Collection	Sustained phonation of vowel /a/

3.2. Pre-processing

The Parkinson’s Speech Dataset contains 754 features extracted from sustained phonation recordings, presenting a high-dimensional and potentially redundant feature space. To address this, we applied five feature selection techniques: RF Importance, recursive feature elimination (RFE), LASSO Regression, Boruta, and principal component analysis (PCA). Our final feature set was derived using a consensus-based strategy, where features consistently identified by at least two out of the four primary techniques (RF, RFE, LASSO, Boruta) were retained. This approach balances dimensionality reduction and interpretability while reducing method-specific bias. PCA was used solely for comparison and not for final feature selection due to its lack of interpretability.

3.2.1. Feature selection techniques

To improve model efficiency and accuracy, several feature selection techniques were applied to identify the most relevant features for PD classification.

RF Feature Importance was used to rank features based on their contribution to classification, determined through Gini impurity reduction. A RF model with 30 estimators was trained, identifying energy and frequency-based speech markers as key indicators of Parkinsonian speech impairments. These high-ranking features played a crucial role in refining the dataset for classification as shown in Table 3.

Table 3. Top features selected by RF

Feature	Importance score
std_delta_delta_log_energy	0.0285
tqwt_entropy_log_dec_12	0.0160
tqwt_energy_dec_27	0.0136
tqwt_entropy_shannon_dec_12	0.0130
std_6th_delta_delta	0.0127

RFE, with LR as the base estimator, iteratively removed less significant features to retain the most discriminative ones. The top five selected features were: tqwt_entropy_log_dec_18, tqwt_entropy_log_dec_20, tqwt_entropy_log_dec_24, tqwt_entropy_log_dec_25, tqwt_entropy_log_dec_28.

Lasso Regression (L1 Regularization) further reduced dimensionality by shrinking irrelevant feature weights to zero while preserving crucial predictors. A LassoCV model with five-fold cross-validation was used to identify the most significant features. Notably, the top 10 selected features showed strong overlap with those identified by RF and RFE, reinforcing their predictive strength. These features included std_delta_delta_log_energy, tqwt_kurtosisValue_dec_31, tqwt_entropy_log_dec_28, std_7th_delta_delta, std_6th_delta_delta, tqwt_entropy_log_dec_26, tqwt_kurtosisValue_dec_27, tqwt_kurtosisValue_dec_33, tqwt_maxValue_dec_25, tqwt_entropy_log_dec_33.

Boruta Feature Selection, a wrapper-based technique using RF, validated the importance of jitter, shimmer, and period-based features, which are widely recognized in Parkinson's speech pathology. These features, commonly linked to phonatory and acoustic changes, included meanPeriodPulses, locPctJitter, locAbsJitter, rapJitter, ppq5Jitter, ddpJitter, apq11Shimmer.

Additionally, PCA was explored to transform the dataset into 50 principal components, enabling faster computation. However, PCA was not used in the final selection, as direct feature selection methods (RF, RFE, Lasso, and Boruta) provided better classification performance and interpretability. By combining these feature selection techniques, we made sure that only the most relevant features were kept in the dataset, increasing model accuracy while lowering computational cost and redundancy. Table 4 highlights the comparative results of feature selection techniques.

The final feature subset was derived through a consensus-based approach, selecting features identified by at least two of the four techniques: RF, RFE, LASSO, and Boruta. This method ensured that only the most consistently ranked and biologically relevant features were retained, improving model robustness and interpretability. Table 4 explicitly highlights these selected features, which include key biomarkers like std_delta_delta_log_energy, std_6th_delta_delta, and tqwt_entropy_log_dec_28, commonly associated with Parkinsonian vocal impairments.

Table 4. Comparative feature selection results

Feature	Random forest rank	RFE selected	Lasso coefficient	Boruta selected	Selected in final subset
std_delta_delta_log_energy	1	Yes	0.0285	Yes	Yes
tqwt_entropy_log_dec_12	2	No	0.0160	No	No
tqwt_entropy_dec_27	3	No	0.0136	No	No
tqwt_entropy_shannon_dec_12	4	No	0.0130	No	No
std_6th_delta_delta	5	Yes	0.0127	Yes	Yes

3.2.2. Model training and optimization

After selecting the most relevant features, the next step involves training and optimizing the ML model for PD classification. We employed XGBoost as the primary classification algorithm due to its high accuracy, robustness against imbalanced datasets, and efficiency in handling high-dimensional data. To optimize the XGBoost model, Randomized Search Cross-Validation (CV) was used to fine-tune key hyperparameters, improving generalization and reducing overfitting. After 50 iterations, the best hyperparameters were determined and were then used in the final XGBoost model training. Table 5 contains the best hyperparameters after 50 iterations.

Once the best hyperparameters were selected, the final XGBoost model was trained on the balanced dataset (after SMOTE was applied). To ensure the robustness and reliability of the proposed model, 5-fold stratified cross-validation was performed, allowing for an unbiased evaluation across different data splits. The model achieved a mean accuracy of 89.54%, demonstrating consistent performance and generalizability in distinguishing PD from healthy controls. By integrating Randomized Search CV for hyperparameter tuning, SMOTE for class balancing, and a refined feature selection process, the final XGBoost model achieved an overall accuracy of 96.69%, reinforcing its effectiveness as a highly reliable and interpretable solution for PD classification.

Table 5. Final optimized model parameters

Hyperparameter	Optimised values
n_estimators	300
max_depth	7
learning_rate	0.1
subsample	0.8
colsample_bytree	0.8
min_child_weight	1
gamma	0.1

4. RESULTS AND DISCUSSION

To evaluate the effectiveness of the proposed XGBoost model, its performance was compared with RF, LightGBM (LGBM), and a Voting Classifier. RF was used as a baseline for feature selection but struggled with overfitting and inefficiency in high-dimensional data. LGBM, a gradient-boosting model, was

optimized for speed and memory efficiency, making it suitable for large datasets. To enhance robustness, a Voting classifier combining XGBoost and LGBM was implemented to improve predictive accuracy. Table 6 shows the models performance comparison.

Table 6. Model performance comparison

Model	Accuracy	Precision	Recall (PD - 1)	Recall (Healthy - 0)	F1 Score
Random Forest	84.1%	82%	96%	52%	83%
LightGBM	88.9%	85%	97%	67%	88%
Voting Classifier	90.7%	95%	99%	69%	90%
XG Boost (Optimised)	96.69%	99%	96%	97%	98%

The results clearly demonstrate that the XGBoost model achieves superior performance across all evaluation metrics, with an accuracy of 96.69%, precision of 99%, and an F1-score of 98%. Unlike the baseline models, XGBoost maintains high recall across both classes (PD: 96%, Healthy: 97%), indicating excellent sensitivity and specificity. In contrast, RF exhibited significant class imbalance bias, as reflected in its recall disparity (96% vs. 52%). LightGBM and the Voting Classifier showed improved balance but failed to match the overall discriminative power of XGBoost. These metrics (Figure 2) underscore XGBoost's robustness, particularly in handling high-dimensional, imbalanced data scenarios common in clinical applications.

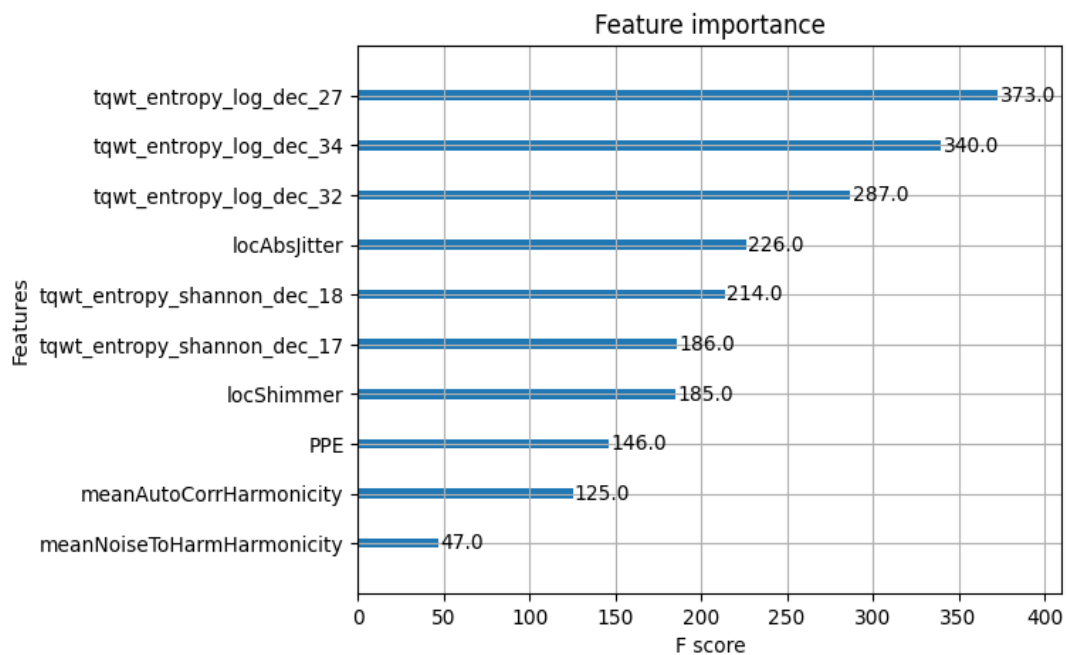


Figure 2. Top-ranked features by XGBoost's built-in importance metric, highlighting the key acoustic markers used in Parkinson's classification

SHapley Additive exPlanations (SHAP) analysis was conducted to enhance model transparency and interpretability. Figure 3 illustrates how individual features contribute to classification outcomes, quantifying each feature's impact on the final prediction. Notably, std_delta_log_energy and tqwt_entropy_log_dec_12 exhibited the highest SHAP values, reinforcing their status as dominant predictors. These features correspond to variability and entropy in frequency-modulated speech patterns, which are known to deteriorate early in PD patients due to phonatory muscle control loss. By elucidating the model's decision process, SHAP enables clinicians to trace predictions back to explainable acoustic biomarkers, thereby bridging the gap between AI output and clinical intuition. This interpretability is essential for clinical trust and regulatory validation of AI-assisted diagnostic systems.

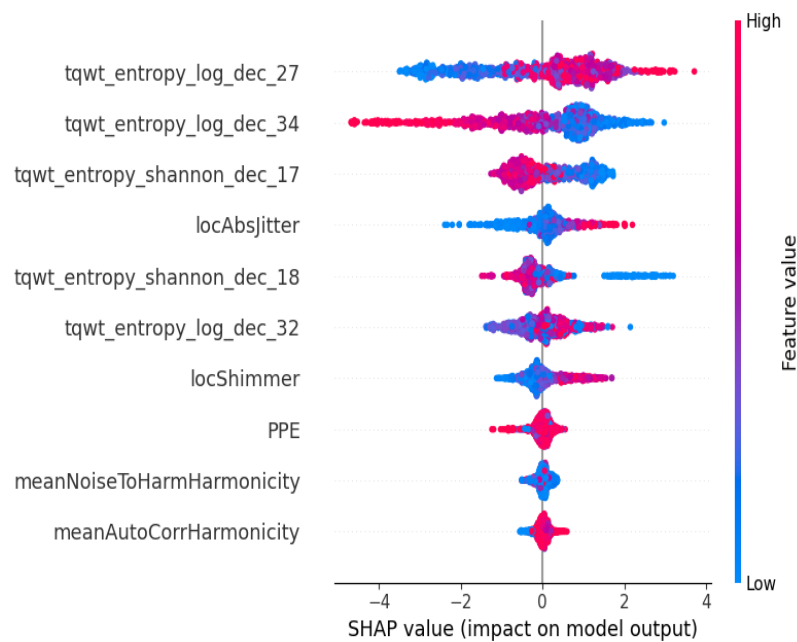


Figure 3. SHAP summary plot showing the marginal contribution of each feature toward model predictions. Features with higher SHAP values contribute more significantly to classification decisions

5. CONCLUSION

This study presents a ML-based approach for PD detection, utilizing speech biomarkers extracted from the UCI Parkinson's Speech Dataset. By integrating feature selection techniques, data balancing using SMOTE, and hyperparameter optimization through Randomized Search CV, an XGBoost-based classification model was developed, achieving 96.69% accuracy with high recall and precision across both PD and healthy classes. A comparative analysis against RF, LightGBM, and an ensemble Voting Classifier demonstrated that XGBoost outperforms traditional classifiers, making it the most effective model for PD classification. The feature importance analysis emphasized the significance of energy-based and time-frequency speech features, reinforcing the role of acoustic biomarkers in PD screening. The application of SHAP explainability techniques further enhanced the model's interpretability, increasing its potential for clinical deployment and integration into decision-support systems. Despite the promising results, there are several areas for future improvement. One key limitation is the dataset size, which, while sufficient for initial validation, requires further expansion and external validation on larger, more diverse, and multi-center datasets. Future studies should focus on integrating multimodal biomarkers, incorporating motor-based features (e.g., handwriting patterns, gait analysis), clinical symptoms, and wearable sensor data to develop a more comprehensive diagnostic model. Additionally, deep learning architectures, such as convolutional neural networks (CNNs) or Transformer-based models, could be explored to capture complex patterns in voice data more effectively. To facilitate real-world clinical deployment, future research should focus on developing real-time Parkinson's detection systems, integrating the model into mobile applications or telemedicine platforms. Such advancements could enable remote patient monitoring, early intervention, and continuous disease progression tracking, enhancing the management of PD. Furthermore, improving model generalization and robustness through federated learning could allow secure collaboration across different healthcare institutions while preserving patient privacy.

Overall, this study demonstrates the potential of AI-driven, non-invasive PD screening tools, paving the way for future advancements in ML-based neurodegenerative disease diagnostics. By expanding dataset diversity, integrating multimodal features, and deploying real-time detection systems, AI-based Parkinson's detection can become a more reliable, accessible, and clinically useful tool for early diagnosis and improved patient.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Sharan T. D.	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Sujata Joshi		✓				✓		✓	✓	✓	✓	✓		

C : Conceptualization	I : Investigation	Vi : Visualization
M : Methodology	R : Resources	Su : Supervision
So : Software	D : Data Curation	P : Project administration
Va : Validation	O : Writing - Original Draft	Fu : Funding acquisition
Fo : Formal analysis	E : Writing - Review & Editing	

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] A. Haahr, H. Groos, and D. Sørensen, "'Striving for normality' when coping with Parkinson's disease in everyday life: A metasynthesis," *International Journal of Nursing Studies*, vol. 118, p. 103923, Jun. 2021, doi: 10.1016/j.ijnurstu.2021.103923.

[2] J. Moini, A. Logalbo, and J. G. Schnellmann, "Pharmacology of Parkinson's disease," in *Neuropsychopharmacology*, Elsevier, 2023, pp. 257–274.

[3] G. Söderbom, "Status and future directions of clinical trials in Parkinson's disease," in *International Review of Neurobiology*, vol. 154, Elsevier, 2020, pp. 153–188.

[4] S. Y. Kim, T. M. Jeitner, and P. M. Steinert, "Transglutaminases in disease," *Neurochemistry International*, vol. 40, no. 1, pp. 85–103, Jan. 2002, doi: 10.1016/S0197-0186(01)00064-X.

[5] P. A. LeWitt and K. R. Chaudhuri, "Unmet needs in Parkinson disease: Motor and non-motor," *Parkinsonism and Related Disorders*, vol. 80, pp. S7–S12, Nov. 2020, doi: 10.1016/j.parkreldis.2020.09.024.

[6] A. Todorova, P. Jenner, and K. Ray Chaudhuri, "Non-motor parkinson's: Integral to motor parkinson's, yet often neglected," *Practical Neurology*, vol. 14, no. 5, pp. 310–322, Apr. 2014, doi: 10.1136/practneurol-2013-000741.

[7] L. A. Uebelacker, G. Epstein-Lubow, T. Lewis, M. K. Broughton, and J. H. Friedman, "A survey of Parkinson's disease patients: Most bothersome symptoms and coping preferences," *Journal of Parkinson's Disease*, vol. 4, no. 4, pp. 717–723, 2014, doi: 10.3233/JPD-140446.

[8] M. O. Oyovwi, K. H. Babawale, E. Jeroh, and B. Ben-Azu, "Exploring the role of neuromodulation in neurodegenerative disorders: Insights from Alzheimer's and Parkinson's diseases," *Brain Disorders*, vol. 17, p. 100187, Mar. 2025, doi: 10.1016/j.dscb.2025.100187.

[9] X. Li, Z. Y. Dong, M. Dong, and L. Chen, "Early dopaminergic replacement treatment initiation benefits motor symptoms in patients with Parkinson's disease," *Frontiers in Human Neuroscience*, vol. 18, May 2024, doi: 10.3389/fnhum.2024.1325324.

[10] R. Pratihari and R. Sankar, "Advancements in Parkinson's disease diagnosis: a comprehensive survey on biomarker integration and machine learning," *Computers*, vol. 13, no. 11, p. 293, Nov. 2024, doi: 10.3390/computers13110293.

[11] A. J. Espay *et al.*, "Technology in Parkinson's disease: Challenges and opportunities," *Movement Disorders*, vol. 31, no. 9, pp. 1272–1282, Sep. 2016, doi: 10.1002/mds.26642.

[12] L. Sigcha *et al.*, "Deep learning and wearable sensors for the diagnosis and monitoring of Parkinson's disease: A systematic review," *Expert Systems with Applications*, vol. 229, p. 120541, Nov. 2023, doi: 10.1016/j.eswa.2023.120541.

[13] S. Moradi, L. Tapak, and S. Afshar, "Identification of novel noninvasive diagnostics biomarkers in the parkinson's diseases and improving the disease classification using support vector machine," *BioMed Research International*, vol. 2022, no. 1, Jan. 2022, doi: 10.1155/2022/5009892.

[14] A. Ratnakar, "Evaluating speech analysis techniques for Parkinsons disease detection: a comparison of machine learning and deep learning algorithms," *International Journal of Advanced Research*, vol. 12, no. 05, pp. 1118–1137, May 2024, doi: 10.21474/ijar01/18827.

[15] K. P. Swain *et al.*, "Towards early intervention: detecting Parkinson's disease through voice analysis with machine learning," *The Open Biomedical Engineering Journal*, vol. 18, no. 1, Apr. 2024, doi: 10.2174/0118741207294056240322075602.

[16] J. Rusz, P. Krack, and E. Tripoliti, "From prodromal stages to clinical trials: The promise of digital speech biomarkers in Parkinson's disease," *Neuroscience and Biobehavioral Reviews*, vol. 167, p. 105922, Dec. 2024, doi: 10.1016/j.neubiorev.2024.105922.





[17] A. Favaro *et al.*, "Multilingual evaluation of interpretable biomarkers to represent language and speech patterns in Parkinson's disease," *Frontiers in Neurology*, vol. 14, Mar. 2023, doi: 10.3389/fneur.2023.1142642.

[18] Q. C. Ngo, M. A. Motin, N. D. Pah, P. Drotár, P. Kempster, and D. Kumar, "Computerized analysis of speech and voice for Parkinson's disease: A systematic review," *Computer Methods and Programs in Biomedicine*, vol. 226, p. 107133, Nov. 2022, doi: 10.1016/j.cmpb.2022.107133.





- [19] M. A. Islam, M. Z. Hasan Majumder, M. A. Hussein, K. M. Hossain, and M. S. Miah, "A review of machine learning and deep learning algorithms for Parkinson's disease detection using handwriting and voice datasets," *Heliyon*, vol. 10, no. 3, p. e25469, Feb. 2024, doi: 10.1016/j.heliyon.2024.e25469.
- [20] W. Wang, J. Lee, F. Harrou, and Y. Sun, "Early detection of Parkinson's disease using deep learning and machine learning," *IEEE Access*, vol. 8, pp. 147635–147646, 2020, doi: 10.1109/ACCESS.2020.3016062.
- [21] C. D. Prithvi Achar, M. P. Anvesh, A. Kodipalli, and T. Rao, "Exploring computational models for Parkinson's disease diagnosis: unveiling insights with LIME and SHAP explainability techniques," in *2024 International Conference on Knowledge Engineering and Communication Systems, ICKECS 2024*, Apr. 2024, pp. 1–6, doi: 10.1109/ICKECS61492.2024.10616831.
- [22] "Parkinson's disease classification - UCI machine learning repository." <https://archive.ics.uci.edu/dataset/470/parkinson+s+disease+classification> (accessed Apr. 02, 2025).
- [23] A. Govindu and S. Palwe, "Early detection of Parkinson's disease using machine learning," *Procedia Computer Science*, vol. 218, pp. 249–261, 2022, doi: 10.1016/j.procs.2023.01.007.
- [24] M. B. Makarious *et al.*, "Multi-modality machine learning predicting Parkinson's disease," *npj Parkinson's Disease*, vol. 8, no. 1, p. 35, Apr. 2022, doi: 10.1038/s41531-022-00288-w.
- [25] J. Zhang *et al.*, "Prediction of Parkinson's disease using machine learning methods," *Biomolecules*, vol. 13, no. 12, p. 1761, Dec. 2023, doi: 10.3390/biom13121761.
- [26] R. Alshammri, G. Alharbi, E. Alharbi, and I. Almubark, "Machine learning approaches to identify Parkinson's disease using voice signal features," *Frontiers in Artificial Intelligence*, vol. 6, Mar. 2023, doi: 10.3389/frai.2023.1084001.
- [27] F. Saeed *et al.*, "Enhancing Parkinson's disease prediction using machine learning and feature selection methods," *Computers, Materials and Continua*, vol. 71, no. 2, pp. 5639–5657, 2022, doi: 10.32604/cmc.2022.023124.
- [28] N. Nahar, F. Ara, M. A. I. Neloy, A. Biswas, M. S. Hossain, and K. Andersson, "Feature selection based machine learning to improve prediction of Parkinson disease," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12960 LNAI, Springer International Publishing, 2021, pp. 496–508.
- [29] L. Ali, C. Chakraborty, Z. He, W. Cao, Y. Imrana, and J. J. P. C. Rodrigues, "A novel sample and feature dependent ensemble approach for Parkinson's disease detection," *Neural Computing and Applications*, vol. 35, no. 22, pp. 15997–16010, Mar. 2023, doi: 10.1007/s00521-022-07046-2.
- [30] B. K. Varghese, D. G. B. Amali, and K. S. U. Devi, "Prediction of Parkinson's disease using machine learning techniques on speech dataset," *Research Journal of Pharmacy and Technology*, vol. 12, no. 2, pp. 1–5, 2019, doi: 10.5958/0974-360X.2019.00114.8.
- [31] S. Srinivasan, P. Ramadass, S. K. Mathivanan, K. P. Selvam, B. D. Shivahare, and M. A. Shah, "Detection of Parkinson disease using multiclass machine learning approach," *Scientific Reports*, vol. 14, no. 1, Jun. 2024, doi: 10.1038/s41598-024-64004-9.

BIOGRAPHIES OF AUTHORS



Sharan T. D.     is an undergraduate student at Nitte Meenakshi Institute of Technology, currently pursuing a Bachelor of Engineering (B.E.) in Computer Science. His academic interests lie in the domains of machine learning, artificial intelligence, and computational healthcare. Passionate about leveraging data-driven approaches for medical diagnostics, he is particularly focused on predictive modeling and feature selection techniques to enhance disease detection accuracy. He can be contacted at email: shrn282@gmail.com.



Dr. Sujata Joshi     is currently working as a Professor in the Department of Computer Science and Engineering, Nitte Meenakshi Institute of Technology, Bangalore, India. She has completed a Ph.D. from Visvesvaraya Technological University, Bangalore. She has authored and co-authored several journal and conference papers and book chapters. She has guided many students for their project work and student research publications. Her research interests include predictive modelling, machine learning, and artificial intelligence, particularly in the medical domain. She has been a member of the organizing committee for many conferences and also on the review committee of conferences and journals. She can be contacted at email: sujata.joshi@nmit.ac.in.