Development of a Machine Learning Model for Early-Stage Parkinson's Disease Prediction Using Voice Recording Data

Ananna Biswas & Keerthi Kesavan

SAT5114: AI in Healthcare
Department of Applied Computing

Michigan Technological University (MTU)

Outlines

- Motivation and Background
- Why use Voice Data for Parkinson's Disease (PD) Detection
- Challenges in Machine Learning Model Development
- Our Approach to Address these Challenges
- Performance Comparison among proposed ML models
- Comparison with Literature Benchmarks
- Conclusion and Future Work

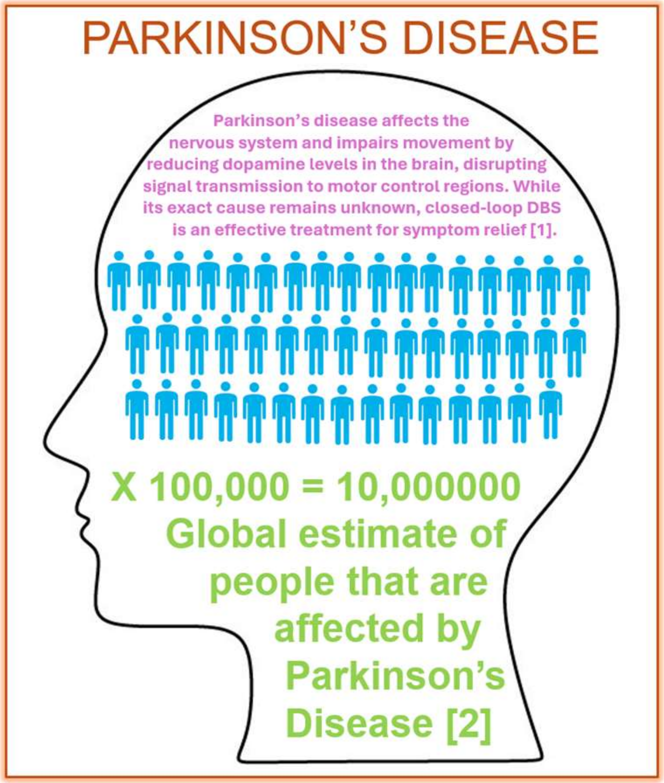
Motivation and Background

Parkinson's Disease (PD):

It is a neurodegenerative disorder that affects movement, balance, and coordination.

It leads to motor symptoms like tremors, bradykinesia, and rigidity, as well as non-motor symptoms such as cognitive impairment and mood disturbances.

While the exact cause of PD remains uncertain, **early detection** is key to improving patient outcomes and slowing disease progression.



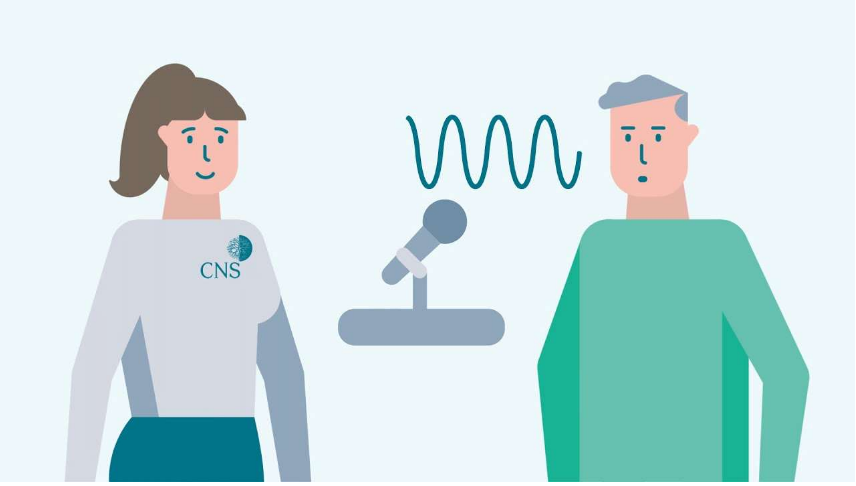
Why Use Voice Data for Parkinson's Disease (PD) Detection

Speech Abnormalities: reduced vocal intensity, imprecise articulation, and speech monotony are the earliest signs of PD.

Studies show that voice is the most commonly and **severely affected feature** in the **early stages** of Parkinson's disease [3].

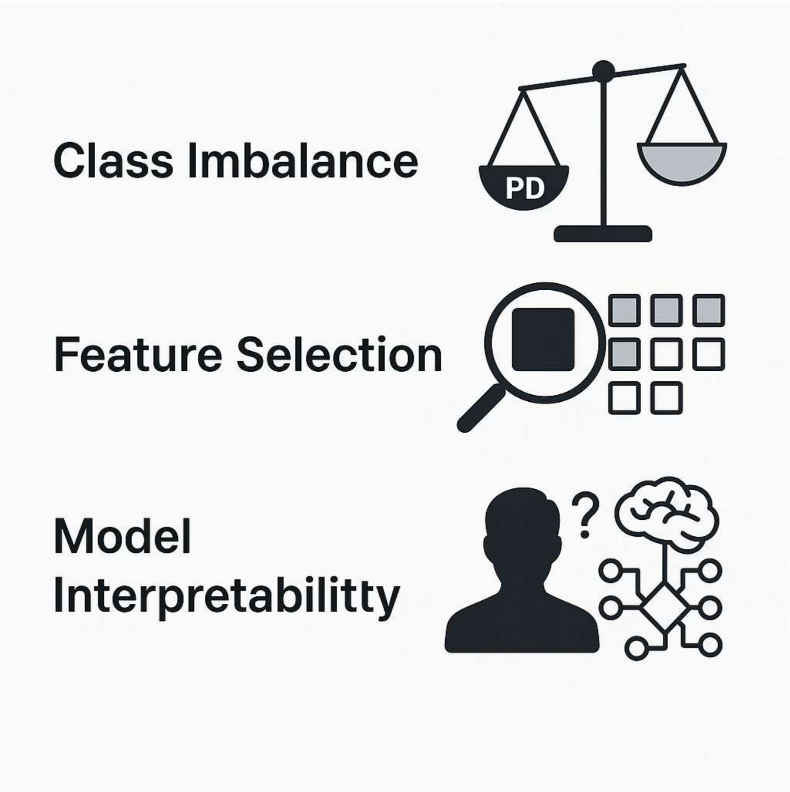
Voice analysis for PD is a promising tool: non-invasive, cost-effective, and can be monitored remotely.

Voice Data For PD Prediction



Challenges in ML based PD Detection

- Class Imbalance: PD samples greatly outnumber healthy controls, risking biased model predictions [4].
- Feature Selection: Irrelevant or redundant features can reduce model accuracy.
- Model Interpretability: Clinically useful models must provide understandable decision logic for healthcare professionals.



PD Detection using ML Models

Our Approach to Address the Challenges:

- **Handling Class Imbalance:** Applied **SMOTE** (Synthetic Minority Over-sampling Technique) to balance the dataset and reduce prediction bias.

- **Improving Feature Selection:** We used **Recursive Feature Elimination (RFE)** to retain only the most relevant features.

- **Ensuring Robust Model Performance:** Trained and compared three ML models: **Random Forest, SVM, and KNN**, with and without RFE.

PD Detection using ML Models

Handling Data Imbalance for Voice Dataset:

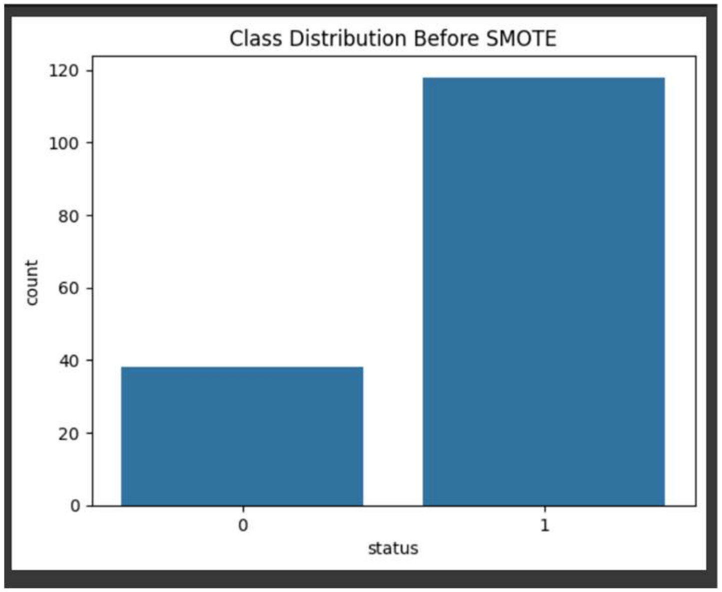
The publicly available voice data of PD and Healthy Subjects, accessed via the UCI Machine Learning Repository [5].

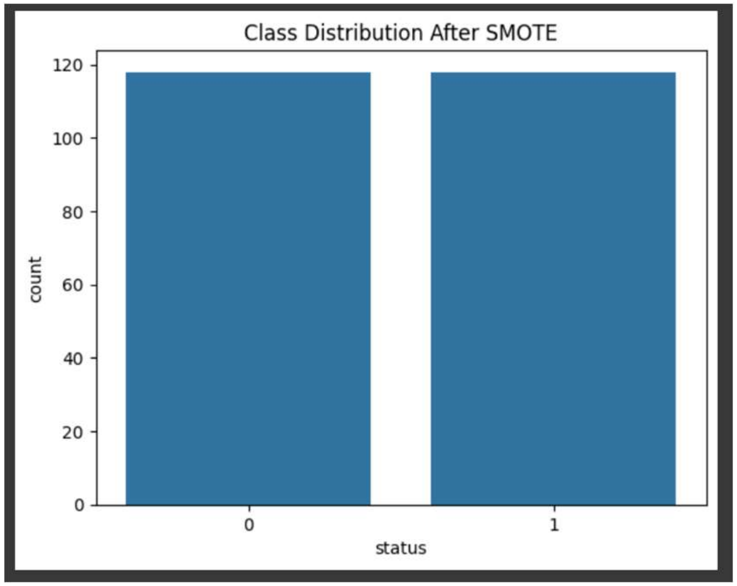
This dataset has gained significant attention in Parkinson's disease research due to its extensive range of speech-related features.

It comprises 195 voice recordings from 31 participants, with an imbalanced class distribution (23 PD patients and 8 healthy subjects).

PD Detection using ML Models

Handling Data Imbalance for Voice Dataset: SMOTE generates synthetic examples of the minority class to create a balanced training set. Healthy Status = 0, PD status = 1

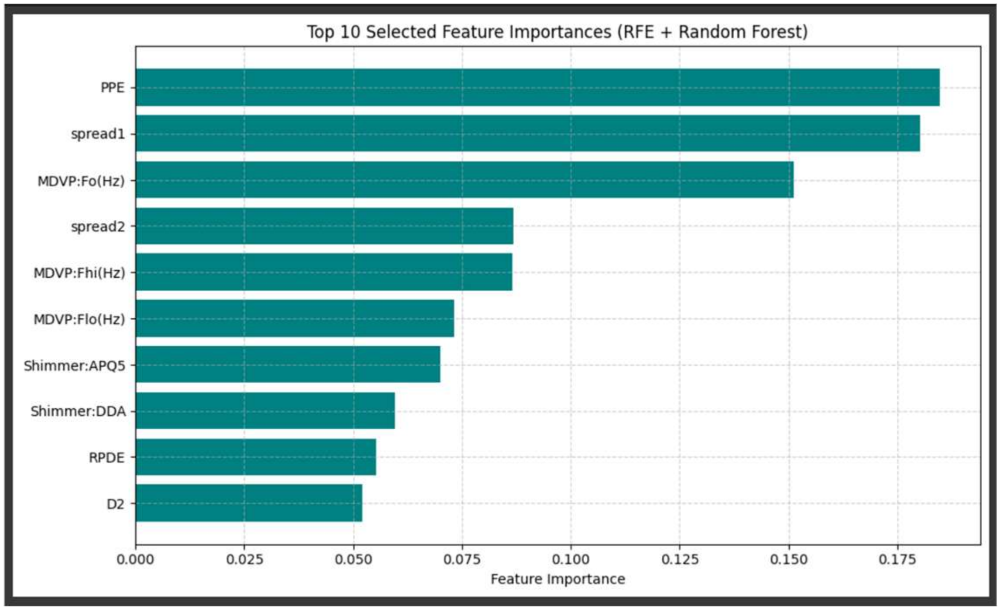




PD Detection using ML Models

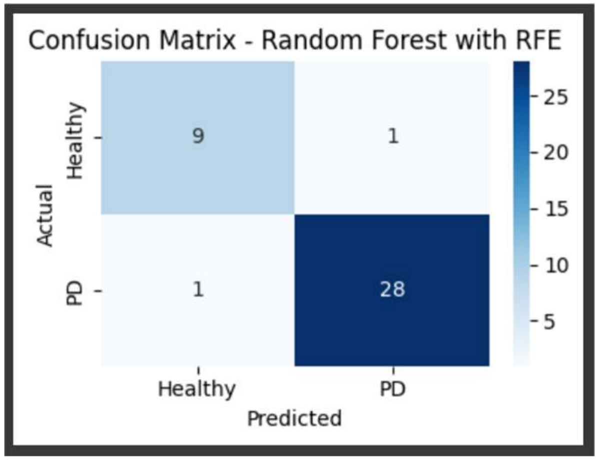
Feature Selection and Training with Hyperparameter Tuning:

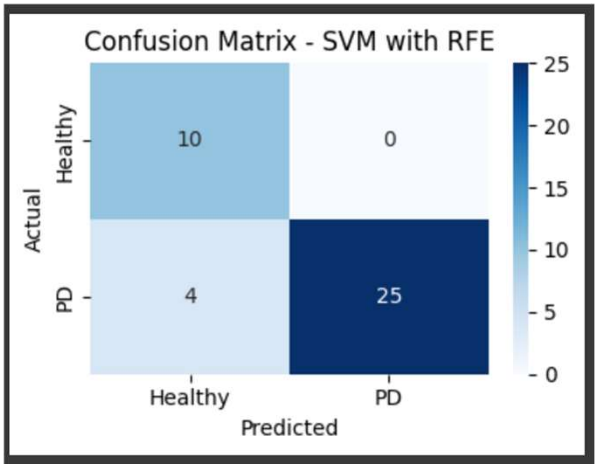
- We have used RFE (Recursive Feature Elimination) for improving Feature Selection.
- RFE helps in selecting a subset of the most significant features by recursively eliminating the least important ones.
- To improve our models' performance, we have fine-tuned hyperparameters using Grid Search CV.

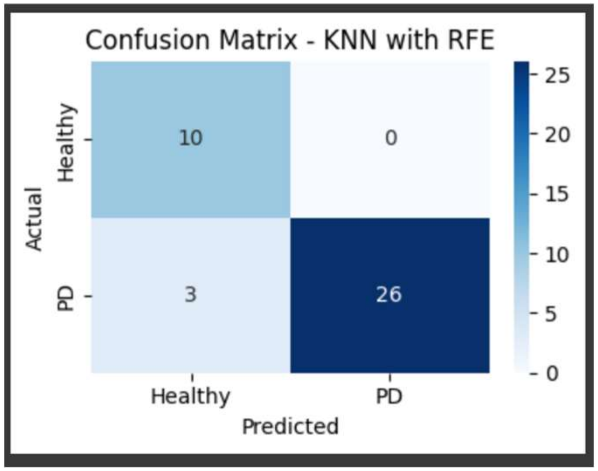


Performance Comparison of ML Models

Confusion Matrix Comparison among Random Forest, SVM, and KNN of **SMOTE with RFE Model:**







Random Forest Model outperformed with **0.9847 accuracy.**

Performance Comparison of ML Models

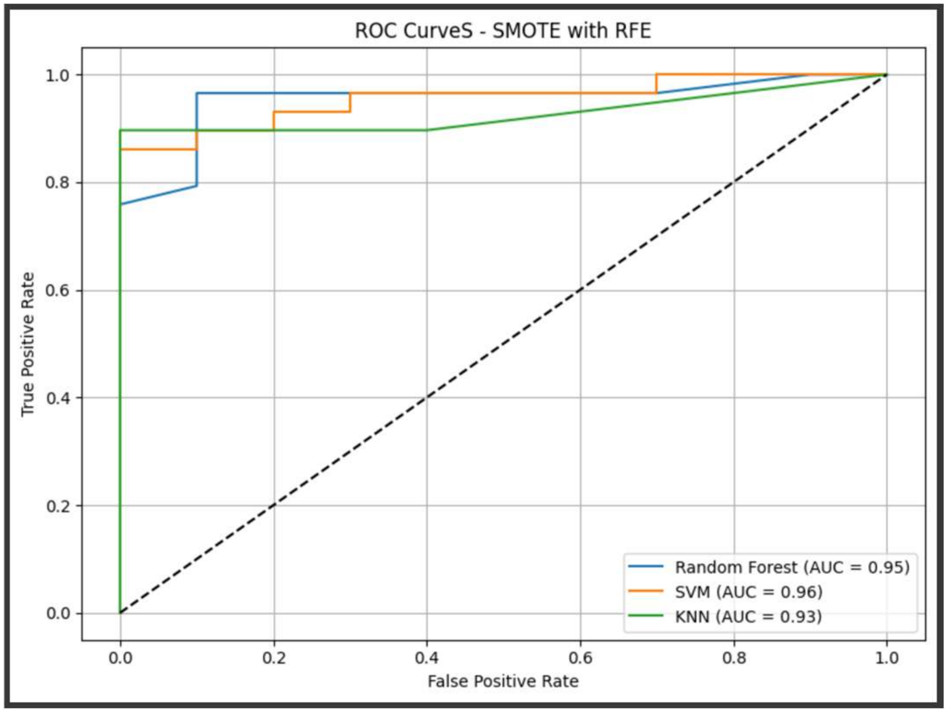
Comparison of Performance Matrices between **SMOTE with RFE** and **SMOTE without RFE Model**

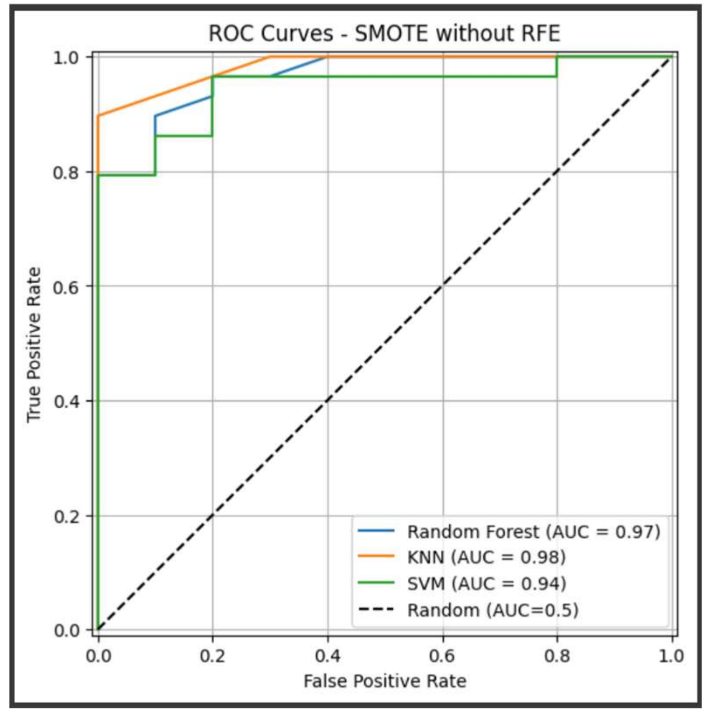
SMOTE with RFE Model	Accuracy	Precision	Sensitivity (Recall)	F1 Score
Random Forest	0.9487	0. 9487	0. 9487	**0. 9487**
KNN	0.9231	0.9408	0.9231	**0.9260**
SVM	0.8974	0.9267	0.8974	**0.9022**

SMOTE without RFE Model	Accuracy	Precision	Sensitivity (Recall)	F1 Score
Random Forest	0.8974	0.9630	0.8966	**0.9286**
KNN	0.9231	1.0000	0.8966	**0.9455**
SVM	0.8205	0.9583	0.7931	**0.8679**

Performance Comparison of ML Models

Comparison of ROC Curve between **SMOTE with RFE** and **SMOTE without RFE Model**





Michigan Tech

Comparison with Literature Benchmarks

Literature	Model(s) Used	Reported Accuracy	Our Proposed Model	Our Accuracy
Tsanas et al., 2012 [6]	SVM, Regression Models	**93.00%**	Random Forest (SMOTE + RFE)	**94.87%**
Little et al., 2009 [7]	Kernel-based SVM	**91.40%**	SVM (SMOTE + RFE)	**90.22%**
K. Velu et al., 2025 [8]	Interpretable Feature Ranking IFRX (XGBoost), SVC	96.61% (IFRX) 90.29% (SVC)	Random Forest (SMOTE + RFE) KNN (SMOTE + RFE) SVM (SMOTE + RFE)	94.87% 92.30% 90.22%

Michigan Tech

Conclusion and Future Work

- Our study successfully applied multiple machine learning models (Random Forest, SVM, KNN) to Parkinson's Disease detection using voice features.
- Feature selection using RFE and class balancing with SMOTE significantly improved performance.
- The Random Forest model achieved the highest accuracy (94.87%), outperforming several established benchmarks.
- In the future, we want to apply CNNs or RNNs to automatically learn features from raw audio data. To further validate the model's performance and generalizability, we also plan to test it on additional datasets [9].
- Code repository: https://github.com/Anannabiswas/PD_Classification_Using_Voice_Data.git

References:

[1] Ito, Hidefumi. "Symptoms and signs of Parkinson's disease and other movement disorders. "*Deep Brain Stimulation for Neurological Disorders: Theoretical Background and Clinical Application*. Cham: Springer International Publishing, 2014. 21-37.

[2] Fang, Congcong, et al. "Cognition deficits in Parkinson's disease: mechanisms and treatment. "Parkin*son's disease* 2020.1 (2020): 2076942.

[3] Ho, Aileen K., et al. "Speech impairment in a large sample of patients with Parkinson's disease." Behavioural neurology 11.3 (1999): 131-137.

[4] K. Velu and N. Jaisankar, "Design of an Early Prediction Model for Parkinson's Disease Using Machine Learning," IEEE Access, vol. 13, pp. 17457-17472, 2025, doi:
10.1109/ACCESS.2025.3533703.

[5] Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., & Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Transactions on Biomedical Engineering,56(4), 1015-1022. Dataset: https://archive.ics.uci.edu/dataset/174/parkinsons

[6] Tsanas, A., Little, M. A., McSharry, P. E., Spielman, J., & Ramig, L. O. (2012). Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE Transactions on Biomedical Engineering, 59(5), 1264-1271. https://doi.org/10.1109/TBME.2012.2183367

[7] M. A. Little ∗, P. E. McSharry, E. J. Hunter, J. Spielman and L. O. Ramig, "Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease," in IEEE Transactions on Biomedical Engineering, vol. 56, no. 4, pp. 1015-1022, April 2009, doi: 10.1109/TBME.2008.2005954.

[8] K. Velu and N. Jaisankar, "Design of an Early Prediction Model for Parkinson's Disease Using Machine Learning," IEEE Access, vol. 13, pp. 17457-17472, 2025, doi:10.1109/ACCESS.2025.3533703

[9] Dataset: "Voice Samples for Patients with Parkinson's Disease and Healthy controls", https:// doi.org/ 10. 6084/ m9. figsh are. 23849 127

Team Members

Keerthi Kesavan
MS Student in Health Informatics
Department of Applied Computing,
Michigan Technological University (MTU)

Ananna Biswas
PhD Student in Computer Engineering,
Dept of Electrical and Computer Engineering,
Michigan Technological University (MTU)

Ananna Biswas, anannab@mtu.edu
Website: https://anannabiswas.github.io/

Keerthi Kesavan, kkesavan@mtu.edu