Parkinson's Disease Rate (PD)

Roopa, Rajendra G E

[1]*Assistant Professor, Computer Science Department, JSS College of Arts Commerce and Science Ooty-Road, Mysuru.*

[2]*U.G. Student, Computer Science Department, JSS College of Arts Commerce and Science Ooty-Road, Mysuru*

***Abstract*—Parkinson's Disease (PD) is a progressive neurolog- ical disorder that affects millions of people around the world. Detecting it early can make a huge difference in how well patients respond to treatment and manage their symptoms. In this study, we explored how machine learning (ML) can help identify PD earlier and more accurately by analyzing clinical and genetic data.We gathered a rich dataset that included detailed medical and genetic information from both individuals diagnosed with PD and healthy participants. Using this data, we applied several ma- chine learning techniques—specifically Random Forest, Decision Tree, and AdaBoost—to build predictive models for Parkinson's Disease. We also used feature selection and engineering methods to improve the accuracy and efficiency of these models. Among the algorithms we tested, the Random Forest model stood out, achieving an impressive AUC-ROC score of 0.95. This means it was highly effective at distinguishing between PD patients and healthy individuals. Through feature importance analysis, we also identified specific clinical signs and genetic markers that played a major role in predicting the disease. Our findings show that machine learning can be a powerful tool in the early detection of Parkinson's Disease. These models have the potential to support doctors in making more informed decisions, allowing for earlier intervention and more personalized care for those at risk.**

***Key Words* — Dopaminergic neurons, Tremor, Substantia ni- gra, Alpha-synuclein, Carbidopa, Neurodegenerative disorder, Lewy bodies.**

I. INTRODUCTION

Parkinson's disease (PD) is a widely recognized neurode- generative disorder, though its exact cause remains unclear. One of the main challenges in identifying PD early lies in its symptoms, which are often subtle and can resemble the natural effects of aging. As a result, early indicators may go unnoticed or be mistaken for unrelated health concerns, increasing the risk of misdiagnosis.

Traditionally, PD diagnosis has been based on a clinician's expertise through observation, patient history, and neurological examinations. While these methods are helpful, they often fall short—particularly in the early stages of the disease when symptoms are mild or atypical. Advanced imaging and biomarker tests can provide additional insights but are not always accessible or definitive.

In recent years, the rapid advancement of artificial in- telligence (AI) has introduced new opportunities for early PD detection. With the support of sophisticated algorithms, especially deep learning models, researchers have started to explore more accurate, scalable diagnostic tools. Among these, ensemble methods have shown encouraging results in improv- ing classification performance.

Despite growing innovations, there is still no cure for Parkinson's. This reality places even greater importance on early and accurate detection. Interestingly, research has found that nearly 90% of individuals with PD experience some form of vocal impairment—often as early as stage 0 of the disease. Since voice changes can be measured remotely with minimal effort, this opens doors for telemedicine-based screening methods. Patients can simply use a mobile phone to record their voice and undergo basic phonation tests from home.

Common vocal symptoms such as dysphonia and dysarthria provide useful indicators of PD-related motor issues. Tests like sustained vowel phonation or running speech analysis are straightforward to administer and can serve as early screening tools for detecting Parkinson's in its earliest stages.

II. LITERATURE SURVEY

Research into Parkinson's Disease (PD) has historically con- centrated on analyzing MRI scans, gait patterns, and genetic data. However, studies that explore speech impairments as early indicators of PD remain limited. For example, Bilal et al. [**?**] utilized genetic data in conjunction with an SVM model to predict the onset of PD in older adults, achieving an accuracy of 88.9%. In contrast, our research demonstrates a refined SVM model that attains a higher accuracy of 91.83%, reinforcing the potential of vocal features as a reliable predic- tor of PD.

Other researchers have also explored alternative approaches. Raundale, Thosar, and Rane [**?**] examined keystroke dynamics from the UCI telemonitoring dataset and trained a Random Forest classifier to estimate the severity of PD in elderly individuals. Similarly, Cordella et al. [**?**] investigated the use of audio features for classification but relied heavily on MATLAB, which can limit accessibility and scalability. Our study addresses this gap by employing open-source Python- based models that offer improved speed and efficiency.

Deep learning continues to play a major role in PD de- tection. For instance, Ali et al. [**?**] employed ensemble deep learning models on phonation data to track disease progres- sion. However, their methodology lacked feature selection, which can be crucial for enhancing model performance. In our approach, Principal Component Analysis (PCA) was used to reduce the original 22 features to the 7 most impactful voice-related attributes, leading to more efficient classification.

In efforts to reduce reliance on wearable devices, Huang et al. [**?**] trained a traditional decision tree model using 12

complex speech attributes from the MDVR-KCL dataset [**?**]. Wodzinski et al. [**?**] approached the problem differently, us- ing a ResNet architecture trained on spectrogram images of audio data. While innovative, this approach did not take full advantage of the rich frequency information contained in raw audio. Wroge et al. [**?**] attempted to remove clinician bias in PD diagnosis through machine learning but only achieved a maximum accuracy of 85%.

Motivated by these findings, our research focuses on de- veloping a robust voice-based PD detection model. We ex- perimented with four widely used classification algorithms: K-Nearest Neighbors (KNN), Logistic Regression, Random Forest, and SVM. Among these, the KNN model outperformed the rest, yielding an accuracy of 91.83% and a sensitivity score of 0.95. These results suggest that voice data, when analyzed correctly, can serve as a powerful tool for early and accessible diagnosis of Parkinson's disease.

III. METHODOLOGY

In this study, we utilized vocal data obtained from the PPMI and UCI datasets, which include audio recordings of individuals both with and without Parkinson's Disease (PD). These datasets feature a variety of acoustic measurements such as jitter, shimmer, and MDVP values, all captured from sustained vowel phonations. Our objective was to investigate whether these vocal features could serve as reliable indicators for early PD detection.

We began by cleaning and preprocessing the data. This involved visualizing feature distributions and examining class imbalances. Following this, we trained four machine learning models: Logistic Regression, Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (KNN). Each model was trained on 75% of the dataset and tested on the remain- ing 25%. We evaluated model performance using sensitivity, precision, accuracy, ROC-AUC score, and confusion matrices. The overall architecture follows a typical machine learning pipeline, encompassing data collection, feature scaling, model training, and performance evaluation. This approach not only aims to identify the best-performing model but also examines

the influence of feature selection and class balancing.

To thoroughly analyze the classification potential of these models, we implemented three distinct experimental ap- proaches:

A. Approach 1: Baseline Training on All Features

This initial approach uses all 22 available voice features for model training, serving as a baseline for performance comparison.

- Collected MDVP audio data from the PPMI and UCI datasets.
- Conducted exploratory data analysis to assess feature distribution and class imbalance.
- Applied Standard Scaler to normalize the feature values.
- Split the dataset into 75% training and 25% testing sets.

- Trained Logistic Regression, SVM, Random Forest, and KNN models.

B. Approach 2: Feature Selection Using PCA

This approach implements Principal Component Analysis (PCA) to reduce dimensionality and enhance model focus on key voice features.

- Used the same preprocessing steps as in Approach 1.
- Applied PCA to extract the top 5 features that explain the highest variance.
- Retrained all four models on the reduced feature set.
- Compared results using accuracy, confusion matrix, and ROC-AUC score.

C. Approach 3: Addressing Class Imbalance

Here, we handled dataset imbalance using upsampling tech- niques to equalize the number of PD and non-PD samples.

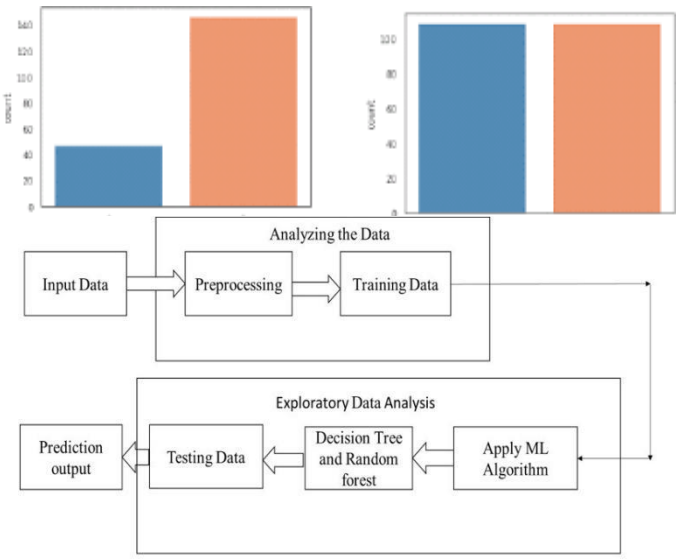
- Identified imbalance: 109 PD samples vs. 40 healthy samples.
- Applied upsampling to increase the number of healthy samples to 109.
- Normalized the dataset using Standard Scaler.
- Split the dataset into training and testing sets.
- Retrained the four models and evaluated their perfor- mance.

Each approach provided unique insights into the effective- ness of audio features for PD detection, allowing us to compare how dimensionality reduction and data balancing affect model outcomes.

Fig. 1: (a) Imbalanced data with 40 normal records; (b) Balanced data after up sampling

I. SYSTEM DESIGN

II. Fig. 2: System Architecture Diagram



IV. EXPERIMENTS AND RESULT

To validate the effectiveness of our proposed approach, we carried out experiments using two well-established datasets. We focused on comparing Random Forest (RF) and Support Vector Machine (SVM) models, as both have consistently performed well in related research. Although many metrics exist to evaluate classification models, we chose to prioritize classification accuracy in our analysis, since it is widely used and provides a clear measure of how well the models are performing.

A. Data

To train and evaluate the performance of our proposed method, we relied on two carefully selected datasets, each offering a unique set of speech-based features relevant to Parkinson's Disease (PD) detection.

Dataset 1 – Parkinson's Disease Classification Dataset (D1):

This dataset consists of 756 samples, out of which 192 belong to individuals diagnosed with Parkinson's Disease. It includes 754 features that capture a wide range of voice characteristics, such as Time-Frequency features, Mel-Frequency Cepstral Coefficients (MFCC), wavelet-based descriptors, and vocal fold dynamics.

Dataset 2 – Constructed Parkinson's Disease Dataset (D2):

The second dataset was collected from Dongguan Songshanhu Central Hospital and includes real-world doctor-patient con- versations recorded in Cantonese and Mandarin. After prepro- cessing, the dataset

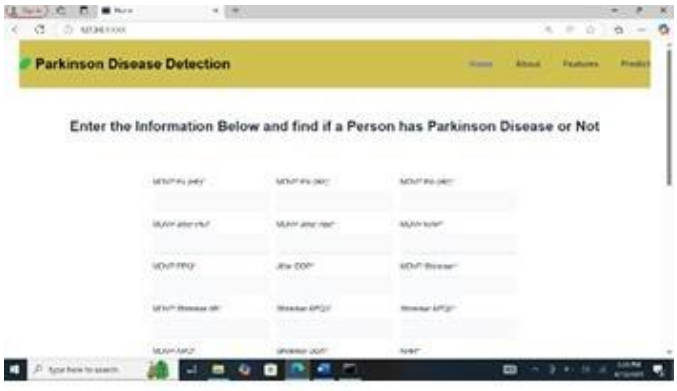
contains 3614 samples, with 2267 samples representing PD patients and 1347 from healthy individuals. The 50 extracted features include variations in pitch (e.g., minimum and maximum fundamental frequencies), jitter and shimmer measures, and 39 MFCC-based features spanning the original, first-order, and second-order derivatives.

B. Results

The *Health Assistant* is an intelligent application that uti- lizes machine learning techniques to support the prediction of various health conditions based on user-provided inputs. Users can enter specific health parameters, and the application provides instant predictions by running pre-trained models through the Streamlit platform.

To evaluate the performance and reliability of each disease prediction model, standard evaluation metrics are used. These include accuracy, precision, recall, and the F1-score. Each of these metrics gives insight into how well the model is identifying and classifying different conditions. The results help validate the system's effectiveness in real-time health assessment scenarios and demonstrate its potential for practical use in a telemedicine or health-monitoring environment

PARKINSON DISEASE PREDICTION



I. MODEL PERFORMANCE EVALUATION

The performance of the model was evaluated using real-world datasets. The table below shows the accuracy of the model used in this application:

Parkinson's Prediction Model	Accuracy
Support Vector Machine (SVM)	87.1%

TABLE I: Model Accuracy for Parkinson's Prediction

VII DISCUSSION AND CONCLUSION

This study introduces a new machine learning approach called the Local Class-Centered Boundary (LCCB) method for identifying Parkinson's Disease. The core idea behind LCCB is to first identify the boundary between two classes and then calculate how far a test sample is from the nearest neighbors of each class's boundary. The class with the shortest distance is then predicted.

What makes LCCB unique is its ability to combine the advantages of different models—specifically, MFCS, SVM, and HKNN. This combination allows it to handle complicated decision boundaries more effectively while also making the model easier to interpret. Our experimental results show that LCCB performs better than several more complex approaches in both speed and accuracy.

As part of this research, we also introduced MFCS to further improve model performance. When we applied MFCS

to the second dataset (D2), all the models showed noticeable improvement. This outcome confirms that MFCS plays an important role in improving the overall quality of data and helping the models learn more effectively.

Our proposed LCCB model reached an impressive 96% accuracy on the first dataset (D1), demonstrating its strong potential in handling voice-based Parkinson's detection tasks. Overall, this research not only highlights the effectiveness of LCCB but also its potential to perform well even in more optimized and challenging data environments.

FUTURE WORK

Looking ahead, these models could be trained using larger and more detailed datasets that include a wider variety of fea- tures. This could potentially improve their prediction accuracy even further.

As the models become more reliable, they may be integrated into clinical environments—helping doctors and labs identify Parkinson's disease earlier and more efficiently.

There's also potential to apply the same machine learning approach to other medical datasets, broadening its usefulness beyond Parkinson's disease.

Future studies might focus on building a hybrid model that can identify multiple diseases using a dataset with overlapping or shared features.

Another exciting direction would be to develop models that can automatically highlight the most important features within a dataset, allowing for smarter and more accurate predictions.

ACKNOWLEDGEMENTS

We are truly grateful to our professors and coordinators for their constant guidance, thoughtful feedback, and encour- agement throughout the course of this project. Their support played a key role in shaping the direction and quality of our work.

We'd also like to thank our families, friends, and peers for their unwavering support, motivation, and belief in us. Their words of encouragement helped us stay focused and driven from start to finish.

REFERENCES

1. Singh, S., and Xu, W. (2020). Robust detection of Parkinson's disease using harvested smartphone voice data: A telemedicine approach. *Telemedicine and e-Health*, 26, 327–334. https://doi.org/10.1089/tmj.2018.0271

2. Tsanas, A., Little, M., McSharry, P., and Ramig, L. (2009). Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests. *Nature Pre- cedings*, 57, 884–893. https://doi.org/10.1038/npre.2009.3920.1

3. Moumita, P., Pradhan, R., Nandy, P., Borah, S., Dey, N., and Gupta, P. (2021). Biomarkers for detection of Parkinson's disease using machine learning—a short review. In S. Borah, R. Pradhan, N. Dey, P. Gupta (Eds.), *Soft Computing Techniques and Applications* (pp. 461–475). Springer Singapore. https://doi.org/10.1007/978-981-15-7394-1_43

4. Nilashi, M., et al. (2020). Remote tracking of Parkinson's disease progression using ensembles of deep belief network and self-organizing map. *Expert Systems with Applications*, 159, 113562. https://www.sciencedirect.com/science/article/pii/S0957417420303869

5. Yang, L., et al. (2022). Changes in facial expressions in patients with Parkinson's disease during the phonation test and their correlation with disease severity. *Computer Speech Language*. https://doi.org/10.1016/j.csl.2021.101286

6. Hireš, M., et al. (2022). Convolutional neural network ensemble for Parkinson's disease detection from voice recordings. *Computers in Biology and Medicine*, 141, 105021. https://www.sciencedirect.com/science/article/pii/S0010482521008155

7. Trivedi, D., Jaeger, H., and Stadtschnitzer, M. (2019). Mobile Device Voice Recordings at King's College London (MDVR-KCL) from both early and advanced Parkinson's disease patients and healthy controls. https://doi.org/10.5281/zenodo.2867216

8. Wang, W., Lee, J., Harrou, F., and Sun, Y. (2020). Early detection of Parkinson's disease using deep learning and machine learning. *IEEE Access*, 8, 147635–147646. https://doi.org/10.1109/ACCESS.2020.3016062