

AI-powered early detection of neurological disease: Parkinson's disease

Kavitha Soppari, Bharath Vupperpally *, Harshini Adloori, Kumar Agolu and Sujith kasula

Department of CSE (Artificial Intelligence and Machine Learning), ACE Engineering College, India.

International Journal of Science and Research Archive, 2025, 14(01), 278-282

Publication history: Received on 29 November 2024; revised on 06 January 2025; accepted on 08 January 2025

Article DOI: <https://doi.org/10.30574/ijrsra.2025.14.1.0041>

Abstract

Parkinson's disease (PD), a neurological illness that gradually compromises motor abilities. Tremors, muscle rigidity, and bradykinesia (slowness of movement) are symptoms of PD. Effective therapy depends on a prompt and accurate diagnosis, yet traditional diagnostic methods can be laborious and subjective. The goal of this study is to create a machine learning-based model that uses clinical information, like vocal characteristics, to detect Parkinson's disease. Through the use of advanced machine learning algorithms and the extraction of important data patterns, the project hopes to develop a trustworthy diagnostic tool that will help physicians identify Parkinson's disease (PD) early on, facilitating quicker interventions and improved patient care.

Keywords: Jitter; Shimmer; MDVP; CatBoost; Vocal Features; PCA

1. Introduction

Neurological diseases like Parkinson's Disease (PD) pose significant challenges due to their progressive nature and complex etiology. Early detection can significantly improve treatment outcomes, yet traditional diagnostic methods are often time-consuming and subjective. A possible technique for creating impartial, effective diagnostic models is artificial intelligence (AI). This study focuses on leveraging machine learning to detect PD using vocal biomarkers, which are non-invasive and easily collectible.

1.1. Historical Background of Parkinson's disease

Dr. James Parkinson originally provided a formal description of Parkinson's disease (PD) in his 1817 book "*An Essay on the Shaking Palsy*." In his seminal work, Parkinson characterized the disease as a condition causing involuntary tremors, impaired movement, and postural instability. Over time, the disease became more widely studied, and in the late 19th century, the phrase "Parkinson's Disease" was created by French neurologist Jean- Martin Charcot in recognition of James Parkinson's work.

The loss of dopaminergic neurons in the substantia nigra, a crucial area of the brain that produces dopamine, a neurotransmitter necessary for motor control, was later found to be the main cause of Parkinson's disease (PD). Growing neurodegeneration is the cause of Parkinson's diseases (PD) hallmark symptoms, which include:

- Motor symptoms include muscle rigidity, bradykinesia (slowness of movement), tremors, and irregular gait.
- Non-motor symptoms include autonomic dysfunction, sleep issues, depression, cognitive deterioration, and speech abnormalitie

* Corresponding author: Bharath Vupperpally

1.2. Importance of Early Detection of Parkinson's Disease

Early detection of PD is critical because the disease is **progressive** and worsens over time. By the time motor symptoms appear, approximately **60-80% of dopaminergic neurons** are already lost. Early diagnosis allows for:

- **Timely intervention** with pharmacological treatments (e.g., Levodopa) and therapies that slow disease progression.
- Improved **quality of life** through symptom management and lifestyle modifications.
- Enhanced ability to participate in **clinical trials** for emerging treatments.

However, traditional diagnostic methods rely heavily on clinical assessments and subjective observations of motor symptoms, which are prone to variability and late-stage diagnosis. Techniques such as neuroimaging and biomarker analysis are available but are often costly, invasive, or inaccessible for widespread clinical use.

2. Literature Review

Due to the possibility of an early, precise, and non-invasive diagnosis, research into machine learning (ML) techniques for Parkinson's disease (PD) detection and progression monitoring has accelerated. Numerous studies have investigated different methods and algorithms, each with unique limits and areas of interest.

- Yadav and Jain (2022) utilized Support Vector Machines (SVM), Random Forest (RF), Decision Trees (DT), K-Nearest Neighbours (KNN), and Logistic Regression (LR) to classify PD using the UCI Parkinson's Voice Dataset. Their results highlighted RF as the most effective model, achieving an accuracy of 92%, emphasizing non-invasiveness and computational efficiency. However, their reliance on a small dataset and single-modality input (voice data) limited generalizability. Expanding the dataset size and integrating multimodal features, such as gait or neuroimaging data, was proposed to address these limitations.
- Mei Jie et al. (2021) conducted a comprehensive review of ML techniques, including SVM, Artificial Neural Networks (ANN), RF, and ensemble methods, applied to multimodal data. While the review provided valuable insights into different approaches, it lacked performance comparisons and omitted practical considerations for real-world implementation. Future work was suggested to include comparative studies and evaluate methods in real-world settings with diverse datasets.
- Yang et al. (2022) focused on gait-based classification using a Convolutional Neural Network (CNN) variant, ResNet, trained on wearable sensor data. Their approach achieved high accuracy, demonstrating the potential of deep learning (DL) in gait analysis. However, the study faced challenges with high computational costs and variability in data due to different wearable devices. Optimizing models for resource-limited devices and ensuring data standardization across devices were recommended for future research.
- Rao et al. (2022) employed dimensionality reduction techniques like Principal Component Analysis (PCA) in combination with SVM, KNN, and DT for voice data classification from the UCI Parkinson's Voice Dataset. PCA enhanced accuracy by reducing dimensionality, but the small dataset and exclusive focus on voice features were significant constraints.

Combining PCA with DL models for multimodal data was suggested to improve robustness.

Other notable studies explored diverse modalities and techniques. For instance, DL methods like CNNs, 3D-CNNs, and VGGNet were applied to neuroimaging data (MRI and PET scans) to analyze structural brain changes. These methods provided accurate diagnostics but required expensive, labeled neuroimaging data. Wearable sensors paired with RF, LSTM, and SVM were utilized for real-time, non-invasive symptom monitoring, though device variability and user compliance posed challenges. Hybrid models combining SVM and neural networks were explored to integrate voice and gait data, yielding high accuracy but at the cost of increased computational complexity. Ensemble methods such as RF, AdaBoost, and Gradient Boosting enhanced performance but remained resource-intensive and required testing on large, multimodal datasets.

Recent trends also include the application of deep reinforcement learning (Deep RL) with CNNs for predicting disease progression using longitudinal data, and multimodal data integration strategies employing CNNs, SVMs, and ensemble methods to combine voice, gait, and neuroimaging data. While these approaches show promise in increasing diagnostic accuracy, they also face challenges related to data heterogeneity and computational demands.

Overall, the literature highlights a diverse array of algorithms and methodologies, each contributing to specific aspects of PD detection and monitoring. Future directions emphasize the need for large-scale multimodal datasets, lightweight and resource-efficient models, and robust methods for combining heterogeneous data sources.

2.1. Architecture

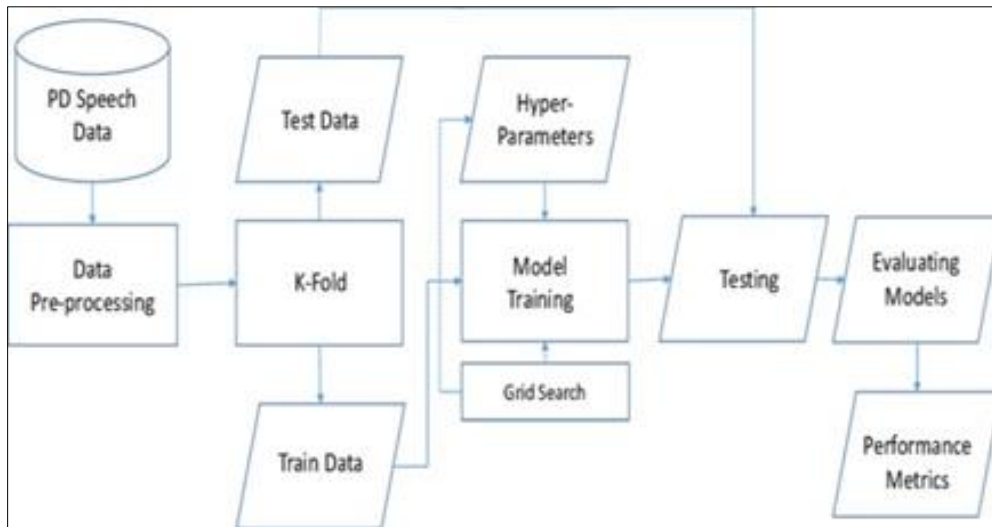


Figure 1 Architecture for Prediction of Parkinson's Disease

The first part of Architecture is collecting a Data from the Parkinson's Official Website and Performing the required Data Preprocessing, Checking for the required data cleaning values like `isnull().sum()` etc..

The second part of the Architecture is Resampling the data so that the model can avoid the Probabilistic bias.

The Third part of the Architecture is Augmentation of the dataset and increasing the size of it for a better accuracy.

The final part of Architecture is Evaluating the PD dataset using Different Machine Learning Algorithms and considering a High Accurate prediction.

3. Proposed Methodology

3.1. CatBoost

High-performance gradient boosting algorithms like CatBoost (Categorical Boosting) are especially well-suited for structured datasets, such as those employed in Parkinson's disease prediction, because they are adept at handling categorical information natively. Unlike traditional models such as XGBoost and LightGBM, which require manual preprocessing like one-hot encoding, CatBoost simplifies this process by directly processing categorical data, reducing preprocessing time and improving efficiency. Its "ordered boosting" technique minimizes overfitting, which is especially valuable for small medical datasets, where other models may struggle. While XGBoost and LightGBM perform well on larger datasets, they often require meticulous parameter tuning and preprocessing, and Random Forest, though simpler, generally lacks the accuracy and interpretability of boosting models. CatBoost also provides robust interpretability tools like SHAP values and feature importance, making it ideal for clinical applications where understanding model decisions is crucial. Its user-friendly nature and strong baseline performance with minimal tuning make it a standout choice for predicting Parkinson's disease compared to other models.

3.2. Architecture

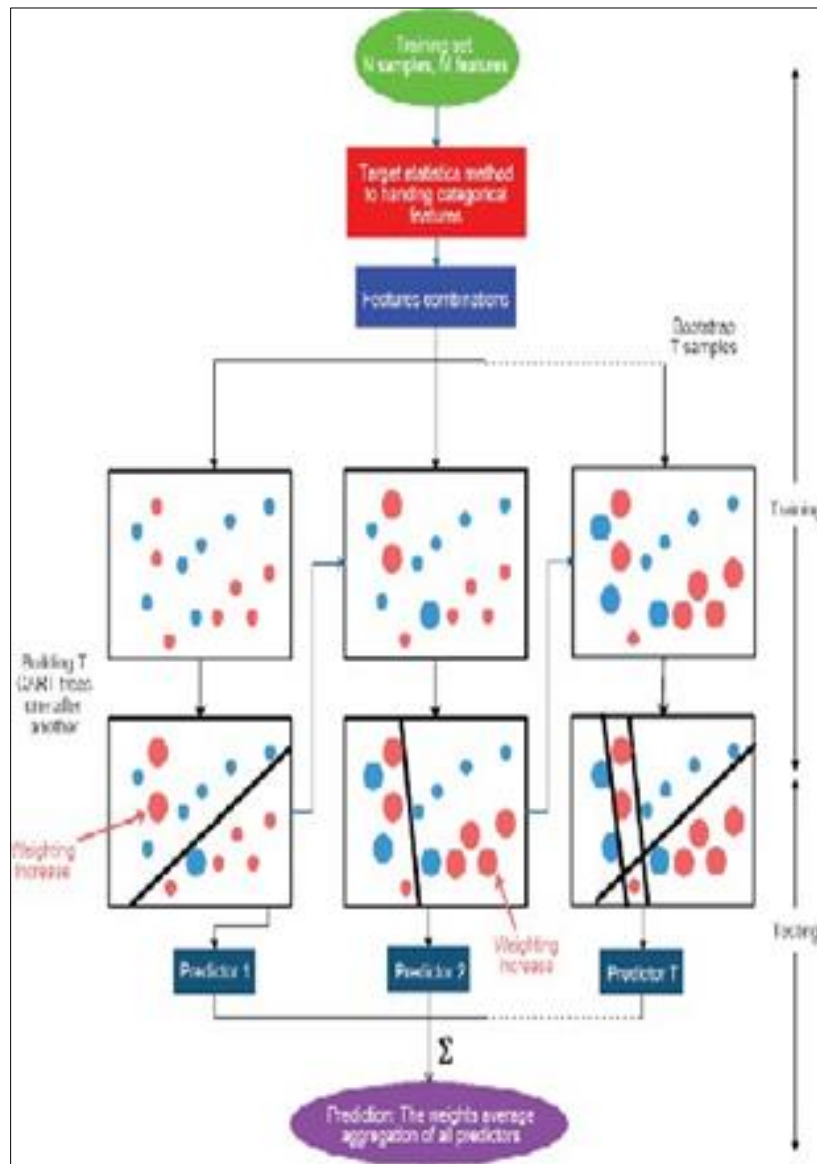


Figure 2 Architecture for CatBoost Algorithm for prediction of Parkinson's Disease

3.3. Algorithm

- Input: features_pca, target, iterations=1000, learning_rate=0.1, depth=6, loss_function='Logloss', verbose=False, n_splits=5, shuffle=True, random_state=42
- Initializemodel=CatBoostClassifier (iterations, learning_rate, depth, loss_function, verbose)
- Initializekf= KFold (n_splits, shuffle, random_state)
- Scores = cross_val_score (model, features_pca, target, cv=kf, scoring='roc_auc')
- average_auc = mean(scores)
- std_auc = standard_deviation(scores) Set accuracy_catboost = average_auc

4. Results and Analysis

- Cross-validated scores: [1. 0.98823529
- 0.98617512 0.98809524 0.98335315]
- Average AUC: 0.989171759686192
- Standard deviation of AUC: 0.005694255699235215

4.1. Comparative Visual Analysis

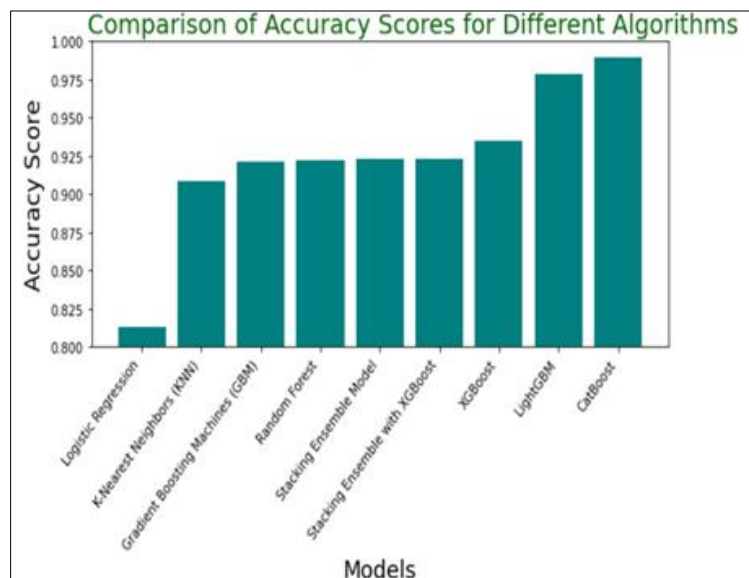


Figure 3 Comparative Analysis of Machine Learning Algorithms to predict Parkinson's Disease

5. Conclusion

The Final Conclusion from above result analysis is, by comparing the different Machine Learning Algorithms The CatBoost and LightGBM have outstanding Accuracy compared to other Algorithms, Hence CatBoost Algorithm is Best Suited to predict that parkinson disease effected or not.

5.1. Future Scope

To increase early detection and accuracy, complex patterns in both structured and unstructured data can be found using sophisticated feature engineering techniques, such as automated tools or hybrid models that combine CatBoost and deep learning. Real-time monitoring systems using IoT and wearable devices can enable continuous tracking of symptoms for timely interventions. Additionally, improving interpretability through tools like SHAP or LIME and ensuring adaptability with online learning methods will build clinician trust and maintain system relevance as new data becomes available. Scaling the system through cloud integration and achieving regulatory compliance for clinical trials will pave the way for broader adoption and personalized patient care.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

- [1] "Comparative Analysis of Machine Learning Algorithms for Parkinson's Disease Prediction" (Yadav & Jain, 2022)
- [2] "Machine Learning for the Diagnosis of Parkinson's Disease: A Review of Literature" (Mei Jie et al., 2021)
- [3] "PD-ResNet for Classification of Parkinson's Disease from Gait" (Yang et al., 2022)
- [4] "Diagnosis of Parkinson's Disease using Principal Component Analysis and Machine Learning algorithms with Vocal Features" (Rao et al., 2022).