

Parkinson's Disease Prediction Using Machine Learning

Manoj M¹, Prof. K. Sharath²

¹Student, Department of MCA, Bangalore Institute of Technology, Bangalore, India

²Assistant Professor, Department of MCA, Bangalore Institute of Technology, Bangalore, India

Abstract

Parkinson's Disease (PD) is a progressive neurodegenerative disorder that primarily impairs motor functions, speech, and cognitive ability. Detecting PD at an early stage is challenging because symptoms overlap with other neurological conditions and often manifest late, delaying medical intervention. This paper proposes a machine learning-based Parkinson's Disease Prediction System that leverages clinical datasets and supervised algorithms for early detection. The system is developed using Support Vector Machine (SVM) and Random Forest classifiers, with Random Forest demonstrating superior accuracy and robustness. Clinical data is preprocessed, normalized, and feature-selected to optimize classification performance. The model is deployed as a Streamlit-based application that provides real-time predictions through an interactive interface. The system is further enhanced with a self-assessment module and an AI-powered chatbot using the Cohere API to improve user engagement and awareness. Experimental results reveal that Random Forest outperforms SVM in terms of accuracy, precision, recall, and F1-score, thereby establishing its reliability in predicting Parkinson's Disease. This study demonstrates the practical role of machine learning in assisting healthcare professionals with early detection and supporting patients with accessible digital tools.

Keywords: Parkinson's Disease, Machine Learning, Random Forest, Support Vector Machine, Clinical Data, Early Diagnosis, Streamlit, Artificial Intelligence

1. Introduction

Parkinson's Disease (PD) is one of the most common neurodegenerative disorders globally, affecting nearly ten million individuals according to recent health reports. It is caused by the gradual degeneration of dopamine-producing neurons in the substantia nigra region of the brain, leading to tremors, rigidity, slowed movement, and impaired balance. While the disease progresses slowly, its symptoms significantly impact the quality of life and lead to substantial social and economic burdens. Early diagnosis is crucial, as

medical treatment and lifestyle interventions introduced at earlier stages can slow progression and enhance patient outcomes. However, existing diagnostic methods primarily rely on clinical examinations, medical imaging, and specialist expertise, which are resource-intensive and often limited by late symptom visibility.

Machine learning (ML) offers a promising alternative for addressing this challenge by analyzing patient data and identifying patterns that may not be visible to human experts. By leveraging clinical features such as biomedical voice recordings, motor assessments, and symptom-based data, ML models can classify patients with high accuracy. Several studies have demonstrated the potential of machine learning in predicting PD, but there remains a need for systems that are accurate, interpretable, and accessible to both medical professionals and patients.

This paper presents a Parkinson's Disease Prediction System that employs SVM and Random Forest classifiers, with preprocessing techniques applied to ensure high-quality feature representation. Random Forest is identified as the most effective model, offering higher predictive accuracy compared to SVM. The system is deployed through a Streamlit-based application that allows users to input test parameters and obtain real-time predictions. Additionally, the application integrates a self-assessment questionnaire for preliminary evaluation and a chatbot assistant to provide guidance, thereby making the system both interactive and patient-friendly.

2. Literature Survey

Research on Parkinson's Disease prediction has evolved significantly over the past decade. Early work primarily focused on statistical models and clinical evaluations, but recent studies have demonstrated the effectiveness of machine learning approaches. Studies have shown that voice recordings can be analyzed using ML models to differentiate between PD patients and healthy individuals with high accuracy. Other works have highlighted the effectiveness of ensemble models such as Random Forest and Gradient Boosting, which

outperform traditional classifiers in handling high-dimensional medical datasets.

Several authors have emphasized the importance of preprocessing and feature scaling in improving prediction accuracy. Support Vector Machines have been applied to voice and motor function datasets, delivering competitive accuracy but facing limitations with large, imbalanced datasets. Ensemble models such as Random Forest have been reported to handle such challenges better due to their ability to capture non-linear relationships. Recent studies have also explored deep learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), though these often require large datasets and computational resources, which may not be feasible in all healthcare settings.

In addition to model selection, researchers have also studied explainability and interpretability in medical applications. Feature importance analysis and interpretable decision trees have been used to enhance trust among clinicians, while other works propose hybrid approaches that combine traditional ML with modern AI techniques. Furthermore, advances in natural language processing have inspired the development of chatbot systems to provide patient education and support. These insights highlight the growing role of machine learning in improving the diagnosis and management of PD, while also motivating the approach taken in this study.

3. Existing System

The existing systems for diagnosing Parkinson's Disease are primarily based on traditional medical practices such as neurological examinations, patient history analysis, and clinical imaging techniques like MRI or PET scans. Although these methods are considered reliable, they are time-consuming, expensive, and often limited to specialized healthcare facilities. Moreover, diagnosis using conventional methods is typically made at later stages of the disease, when symptoms such as tremors, muscle stiffness, and speech impairments have already become noticeable. This late diagnosis restricts the possibility of early intervention and reduces the effectiveness of treatment strategies.

In addition to clinical evaluations, some research-oriented systems have applied statistical models or basic machine learning algorithms such as Naïve Bayes, Decision Trees, or k-Nearest Neighbors (k-NN) on

small datasets. While these approaches provide a proof of concept, they often suffer from limited scalability, lower accuracy, and challenges in handling noisy or imbalanced data. Furthermore, most existing automated systems rely on single-model classifiers without comparative analysis, which restricts their adaptability across different datasets. Many of these systems also lack interpretability and user-friendly interfaces, making them less suitable for practical deployment in real healthcare environments.

Thus, while existing methods provide valuable insights into Parkinson's diagnosis, they face limitations such as late detection, low accuracy in automated predictions, high dependency on clinical expertise, and minimal accessibility for common users. These shortcomings highlight the need for a more robust, scalable, and interactive system that combines advanced machine learning techniques with user-centric deployment platforms.

4. Proposed System

The proposed Parkinson's Disease Prediction System is designed as a comprehensive pipeline consisting of preprocessing, feature selection, model training, evaluation, and deployment. Clinical data containing patient features is first preprocessed by handling missing values, removing inconsistencies, and applying normalization. Feature selection is performed to retain the most relevant parameters, improving classification accuracy. Both SVM and Random Forest classifiers are implemented to analyze the dataset. While SVM provides a strong baseline for classification, Random Forest is found to deliver superior results due to its ensemble nature, which reduces overfitting and improves generalization.

The trained Random Forest model is integrated into a Streamlit-based web application that allows real-time predictions. Users can enter clinical parameters, and the system instantly predicts whether the patient is likely to have Parkinson's Disease. To make the system more practical, a self-assessment module is included, enabling individuals to evaluate their symptoms through a questionnaire. Additionally, an AI-powered chatbot using the Cohere API is integrated to provide guidance and awareness regarding PD. This design ensures that the system is not only accurate but also user-friendly and scalable.

5. Implementation

The implementation begins with dataset importation, followed by preprocessing steps such as normalization and feature extraction. Once the data is prepared, two supervised learning models—SVM and Random Forest—are trained using an 80-20 train-test split. Performance metrics such as accuracy, precision, recall, and F1-score are computed for comparative evaluation. Experimental results reveal that Random Forest consistently outperforms SVM, achieving higher accuracy while minimizing false positives and false negatives.

The application is developed using Streamlit, offering a graphical user interface where users can input patient test results for prediction. The integration of the self-assessment module ensures accessibility for individuals who may not have clinical data readily available, while the chatbot provides real-time responses to queries about PD. The modular structure of the implementation ensures that additional models or datasets can be incorporated without disrupting the overall workflow.

6. Results

The system was evaluated using benchmark Parkinson's datasets, and results demonstrated that Random Forest achieved superior accuracy compared to SVM. Confusion matrices and classification reports confirmed that Random Forest minimized both false positives and false negatives more effectively, making it a more reliable choice for medical applications. Accuracy scores indicated that Random Forest achieved approximately 94%, while SVM achieved around 89%. Precision and recall values were also higher for Random Forest, confirming its robustness.

The deployed Streamlit application successfully provided real-time predictions with minimal latency. Usability testing demonstrated that the interface was intuitive, while performance testing confirmed that the system remained stable under repeated queries. The integration of the chatbot and self-assessment module further enhanced the user experience, making the application both interactive and informative.

7. Conclusion

This paper presents a machine learning-based Parkinson's Disease Prediction System that applies supervised learning algorithms to clinical datasets. The study demonstrates that Random Forest outperforms SVM in terms of accuracy, precision, and recall, making it the preferred model for early detection of Parkinson's Disease. The system's deployment through a Streamlit application ensures accessibility, while the integration of a self-assessment module and chatbot makes the tool more practical for real-world use.

The results highlight the potential of machine learning to support healthcare professionals in diagnosing PD and assisting patients in understanding their condition. By combining predictive accuracy with usability, the system bridges the gap between academic research and clinical practice.

8. Future Enhancement

Future improvements may include the integration of deep learning models such as BERT, CNNs, and LSTMs to capture more complex relationships in patient data. Voice data analysis in real time could be incorporated to detect speech-related symptoms more effectively. Expanding the system to support multilingual and regional languages would increase accessibility across diverse populations. Integration with wearable devices such as smartwatches could enable continuous monitoring of motor symptoms, providing richer data for predictions. Cloud deployment and API integration could further scale the system for use in hospitals, research institutions, and mobile health applications. Additionally, the incorporation of explainable AI frameworks such as SHAP or LIME would enhance transparency and build trust among clinicians and patients.

9. References

- [1] Smith, J., et al. "Parkinson's Disease Prediction Using Voice Features and Machine Learning." *Journal of Medical Systems*, 2017.
- [2] Zhang, Y., et al. "Machine Learning Models for Early Diagnosis of Parkinson's Disease." *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2018.

[3] Prashanth, R., et al. "High-Accuracy Detection of Parkinson's Disease Using Random Forest Classifier." *Pattern Recognition Letters*, 2016.

[4] Shah, S., et al. "Application of Support Vector Machines in Parkinson's Disease Detection." *Biomedical Signal Processing and Control*, 2019.

[5] Arora, S., et al. "Voice Analysis for Parkinson's Disease Detection Using Machine Learning." *Procedia Computer Science*, 2020.

[6] Das, R. "A Comparison of SVM and Random Forest for Parkinson's Prediction." *Expert Systems with Applications*, 2010.

[7] Pereira, C., et al. "Parkinson's Disease Diagnosis Using Ensemble Machine Learning." *Artificial Intelligence in Medicine*, 2020.

[8] Little, M.A., et al. "Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease." *IEEE Transactions on Biomedical Engineering*, 2009.

[9] Oung, Q., et al. "Deep Learning for Parkinson's Disease Classification." *Frontiers in Neurology*, 2021.

[10] Kumar, A., et al. "Hybrid Machine Learning Approaches for Parkinson's Disease Prediction." *Computer Methods and Programs in Biomedicine*, 2022.

[11] Chatterjee, P., et al. "Random Forest-Based Classification of Parkinson's Patients Using Clinical Data." *Springer Nature Computer Science*, 2021.

[12] Khare, V., et al. "A Comparative Analysis of Machine Learning Algorithms for Parkinson's Disease Detection." *Journal of Healthcare Informatics Research*, 2021.

[13] Tsanas, A., et al. "Accurate Telemonitoring of Parkinson's Disease Progression Using Nonlinear Speech Features." *IEEE Transactions on Biomedical Engineering*, 2012.

[14] Ali, L., et al. "Predicting Parkinson's Disease Using Supervised Learning and Neural Networks." *Applied Sciences*, 2019.

[15] Singh, R., et al. "Cloud-Based Parkinson's Disease Detection System Using Random Forest and SVM." *IEEE Access*, 2022.