Parkinson's Disease Detection

SNEHA VERMA

B.Sc. (Hons.) Physics, 3rd Year

Bangabasi College, under Calcutta University

Period of Internship: **25th August 2025 – 19th September 2025**

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

1. Abstract

This project applies machine learning to detect Parkinson's disease using voice features such as frequency, jitter, and shimmer. Logistic Regression and Random Forest models were used to classify patients as healthy or diseased. The aim is to compare their performance and demonstrate how machine learning can support early diagnosis. Results indicate that Random Forest achieves higher accuracy as it handles complex patterns more effectively.

2. Introduction

Parkinson's disease is a neurological disorder that affects movement and speech. Early detection is crucial for timely treatment, but traditional diagnosis can be slow and subjective. Machine learning offers a faster and data-driven approach to medical diagnosis. In this project, we utilize a voice-based Parkinson's dataset and apply Logistic Regression and Random Forest algorithms. Logistic Regression provides a simple linear model, while Random Forest captures non-linear relationships effectively. By comparing these approaches, we explore which model is more suitable for predicting Parkinson's disease and emphasize the transformative role of machine learning in healthcare applications.

3. Project Objectives

The objectives of this project are:

- To analyze the Parkinson's dataset and identify patterns that distinguish patients from healthy individuals.
- To preprocess the dataset by handling missing values, scaling features, and performing feature selection for improved accuracy.
- To apply different machine learning models (Logistic Regression, Random Forest, SVM, KNN) for classification tasks.

- To compare the performance of these models and identify the most effective predictive approach.
- To test the hypothesis that machine learning can effectively classify Parkinson's disease with high accuracy.

4. Methodology

Data Collection

- The dataset was obtained from the UCI Machine Learning Repository.
- It contains biomedical voice measurements from individuals, including both healthy subjects and Parkinson's patients.

Data Preprocessing

- Checked for missing values and removed inconsistencies from the dataset.
- Normalized the data to maintain features within a standard range for optimal model performance.
- Encoded labels as binary values (1 = Parkinson's, 0 = Healthy) for classification purposes.

Exploratory Data Analysis (EDA)

- Created histograms, box plots, and heatmaps to observe data distribution patterns.
- Analyzed correlations among features such as jitter, shimmer, and fundamental frequency.
- Identified key variables that are most significant in predicting Parkinson's disease.

Model Development

- Split the dataset into training (80%) and testing (20%) sets to ensure proper evaluation.

- Applied multiple models including Logistic Regression, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN).
- Implemented k-fold cross-validation to prevent overfitting and ensure model reliability.

Model Evaluation

- Compared models using accuracy, precision, recall, F1-score, and ROC-AUC metrics.
- Selected the best model based on the highest prediction accuracy and balanced performance metrics.

Tools Used

- **Python** programming language (executed in Google Colab environment)
- **Libraries:** NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn

5. Data Analysis and Results

Descriptive Analysis

- The dataset contained approximately 195 records with 23 distinct features.
- About 75% of records belonged to patients diagnosed with Parkinson's disease.
- Features such as **MDVP:Fo(Hz), Jitter, Shimmer, and Harmonics-to-Noise Ratio (HNR)** showed significant differences between healthy and diseased subjects.

Inferential Analysis (Hypothesis Testing)

- **Null Hypothesis (H_0):** There is no significant difference in voice features between Parkinson's patients and healthy individuals.
- **Alternative Hypothesis (H_1):** There is a significant difference in voice features between the two groups.

- Using t-tests and correlation analysis, H_0 was rejected, confirming that vocal measurements are strong indicators of Parkinson's disease.

Model Performance Summary

MODEL	ACCURACY	PRECISION	RECALL	F1-SCORE
LOGISTIC REGRESSION	88%	0.87	0.89	0.88
RANDOM FOREST	92%	0.91	0.93	0.92
SUPPORT VECTOR MACHINE	90%	0.89	0.91	0.90
K-NEAREST NEIGHBORS	85%	0.83	0.87	0.85

Best Model: Random Forest achieved the highest accuracy of 92%.

Visualizations

- Histograms revealed distinct feature distributions between patient classes.
- Correlation heatmaps highlighted strongly related features that contribute to classification.
- ROC curves demonstrated that Random Forest achieved the largest Area Under the Curve (AUC).

Comparative Analysis of Models

- **Logistic Regression:** Performed adequately but showed limitations with non-linear relationships.
- **Support Vector Machine:** Provided good generalization capabilities but was computationally expensive.
- **K-Nearest Neighbors:** Underperformed due to the high-dimensional nature of the data.
- **Random Forest:** Emerged as the most reliable model with superior accuracy and interpretability.

6. Conclusion

This project successfully demonstrated that **machine learning can accurately detect Parkinson's disease using voice features**. The best-performing model was **Random Forest**, achieving **92% accuracy** in classification tasks. This confirms our hypothesis that biomedical voice measurements serve as effective indicators of Parkinson's disease progression.

The results suggest significant potential for implementing machine learning-based diagnostic tools in clinical settings, providing healthcare professionals with objective, data-driven support for early Parkinson's detection.

Future Work

- Implement deep learning models such as Neural Networks to potentially improve accuracy further.
- Collect larger and more diverse datasets to enhance model generalization across different populations.
- Develop and deploy the model into web or mobile applications for real-time diagnostic support in clinical environments.

7. Appendix

References

- UCI Machine Learning Repository (Parkinson's Dataset)
- Scikit-learn Documentation – https://scikit-learn.org/stable/
- Research articles on biomedical voice analysis and Parkinson's disease detection
- 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.

GitHub Repository

(Project codes and documentation will be uploaded to GitHub repository)

https://github.com/vermasneha828/IRIS-Flower-classification-Parkison-

Additional Resources

- Dataset and presentation materials can be accessed via Google Drive or GitHub repository.
- Complete Python code implementation available upon request.