Parkinson’s Disease Detection Using Machine Learning

Report by: MD Alimuzzaman Haris
ID: 0952320005101019
Dept: CSE
Trust University, Barishal

1. Introduction

Parkinson’s disease is a progressive neurological disorder that primarily affects movement and speech. It occurs due to the degeneration of dopamine-producing neurons in the brain, leading to symptoms such as tremors, stiffness, slowness of movement, and difficulty in speech. Early detection of Parkinson’s disease is crucial, as it helps in managing symptoms and improving the quality of life of patients.

With the advancement of Artificial Intelligence and Machine Learning, automated systems can assist medical professionals in detecting diseases at an early stage. This project focuses on developing a machine learning-based system to detect Parkinson’s disease using biomedical voice measurements.

---

2. Objective of the Project

The main objectives of this project are:

- To analyze a Parkinson’s disease dataset provided in CSV format
- To preprocess and prepare the data for machine learning
- To train a supervised machine learning model for disease detection
- To evaluate the performance of the model using standard metrics

---

3. Dataset Description

The dataset used in this project contains biomedical voice measurements collected from both healthy individuals and patients diagnosed with Parkinson’s disease.

Dataset characteristics:

- Format: CSV (Comma Separated Values)
- Total samples: 195
- Number of features: 22
- Target variable: `status`
    - `0` → Healthy person
    - `1` → Parkinson’s patient

Each row represents a patient’s voice data, and each column represents a specific biomedical feature such as frequency, jitter, shimmer, and noise-to-harmonics ratio.

---

4. Methodology

The project follows a supervised machine learning approach. The steps involved are described below:

4.1 Data Preprocessing

- The dataset was loaded using the Pandas library.
- Non-numerical columns such as patient names were removed.
- The dataset was divided into features (independent variables) and target labels (dependent variable).

4.2 Train-Test Split

- The dataset was split into training and testing sets.
- 80% of the data was used for training and 20% for testing.
- This helps evaluate the model's performance on unseen data.

4.3 Feature Scaling

- Feature scaling was applied using StandardScaler.
- Scaling ensures that all features contribute equally to the model.
- This step is especially important for distance-based algorithms like SVM.

4.4 Model Selection

The **Support Vector Machine (SVM)** algorithm was chosen for this project because:

- It performs well on small and medium-sized datasets
- It handles high-dimensional data efficiently
- It provides good accuracy for medical classification problems

The Radial Basis Function (RBF) kernel was used to handle non-linear relationships.

---

5. Implementation

The system was implemented using Python in Google Colab.
The following libraries were used:

- Pandas
- NumPy
- Scikit-learn

The model was trained using the training dataset and evaluated using the test dataset.

---

6. Results and Evaluation

The performance of the trained model was evaluated using the following metrics:

- **Accuracy**
- **Precision**
- **Recall**
- **F1-Score**
- **Confusion Matrix**

6.1 Accuracy

The model achieved an accuracy of approximately **89.7%**, which indicates that the system can correctly classify most patients.

6.2 Confusion Matrix Analysis

- True Positives: Parkinson’s patients correctly identified
- True Negatives: Healthy individuals correctly identified
- False Positives and False Negatives were minimal

The high recall value for Parkinson’s patients is especially important, as missing a disease case can be critical in medical diagnosis.

---

7. Discussion

The results demonstrate that machine learning can effectively assist in detecting Parkinson's disease using voice features. The SVM model showed strong performance with high accuracy and reliable classification. Feature scaling played a significant role in improving the model's performance.

However, the system depends heavily on the quality and size of the dataset. Real-world performance may vary, and the model should be used as a decision-support tool rather than a replacement for medical professionals.

---

8. Conclusion

This project successfully implemented a machine learning-based Parkinson's disease detection system using Support Vector Machine. The system achieved nearly 90% accuracy, demonstrating the potential of artificial intelligence in healthcare applications.

Machine learning techniques can assist doctors in early diagnosis and improve patient outcomes when used responsibly and ethically.

---

9. Limitations

- The dataset size is relatively small
- The model relies only on voice data
- Real-time clinical validation was not performed

---

10. Future Work

Future improvements may include:

- Using deep learning models for higher accuracy
- Combining multiple medical data sources

- Developing a real-time diagnostic web or mobile application
- Expanding the dataset with more patient records

---

11. Tools and Technologies Used

- Programming Language: Python
- Platform: Google Colab
- Libraries: Pandas, NumPy, Scikit-learn
- Algorithm: Support Vector Machine (SVM)

---

12. Acknowledgment

I would like to express my sincere gratitude to my course instructor for providing guidance and the dataset for this project. This project helped enhance my understanding of machine learning concepts and their real-world applications.

---

13. References

1. Scikit-learn Documentation
2. Parkinson’s Disease Dataset (UCI Machine Learning Repository)
3. Introduction to Machine Learning – Academic Resources