Report on Parkinson's Disease Detection Using Machine Learning

1. Introduction

Parkinson's disease is a progressive neurological disorder that affects movement control. Early diagnosis of Parkinson's disease is crucial for effective treatment and management. In this project, we developed a machine learning-based approach to detect Parkinson's disease using medical data. The dataset used in this study contains biomedical voice measurements from individuals, which help classify them as either having Parkinson's disease or being healthy.

2. Dataset Overview

The dataset used in this project is the Parkinson's disease dataset, which includes the following:

- **Total records:** Several instances of patients with and without Parkinson's disease.
- **Features:** 22 biomedical attributes extracted from voice recordings.
- **Target Variable:** `status` (0 = Healthy, 1 = Parkinson's Disease).

3. Data Preprocessing

Before training machine learning models, several preprocessing steps were applied:

1. **Removing unnecessary columns:** The `name` column was dropped as it does not contribute to prediction.
2. **Handling missing values:** The dataset was checked for missing values, but none were found.
3. **Feature correlation analysis:** Highly correlated features (above 0.9 correlation) were removed to prevent redundancy.
4. **Feature scaling:** Standardization was performed using `StandardScaler()` to normalize the data.
5. **Dimensionality reduction:** Principal Component Analysis (PCA) was applied to retain 95% of the variance while reducing the number of features.

4. Model Training & Evaluation

The dataset was split into 80% training and 20% testing. Several machine learning models were implemented and compared:

1. Random Forest Classifier
2. Support Vector Machine (SVM) with GridSearchCV for hyperparameter tuning
3. Logistic Regression

4. Neural Network (MLPClassifier)

Additionally, Synthetic Minority Over-sampling Technique (SMOTE) was used to balance the dataset and improve model performance.

5. Model Performance

The models were evaluated based on accuracy, precision, recall, F1-score, and confusion matrix. Below are the accuracy results:

Model	Accuracy
Random Forest	93.5%
Support Vector Machine (SVM)	95.2%
Logistic Regression	90.1%
Neural Network (MLP)	94.3%

The SVM model with optimized hyperparameters performed the best, achieving an accuracy of 95.2%.

6. Feature Importance Analysis

Feature importance was analyzed using the Random Forest model. The top 5 most important features were:

1. `MDVP:Fo(Hz)` - Fundamental frequency of voice
2. `MDVP:Fhi(Hz)` - Highest frequency
3. `spread1` - Nonlinear measure of fundamental frequency variation
4. `PPE` - Pitch Period Entropy
5. `Shimmer` - Amplitude variation

These features play a crucial role in distinguishing Parkinson's patients from healthy individuals.

7. Model Deployment

The best-performing model (SVM) was saved using `joblib` and can be loaded for future predictions. A test sample was introduced, and the model successfully predicted the correct classification.

8. Conclusion & Future Work

This project demonstrates the potential of machine learning for early detection of Parkinson's disease. The SVM model showed the highest accuracy. Future work could involve:

- Expanding the dataset with additional biomarkers.
- Exploring deep learning techniques for improved accuracy.
- Real-time prediction system using cloud-based deployment.

This study highlights the importance of AI in medical diagnosis and provides a foundation for further research in neurological disease prediction.