PARKINSON’s DISEASE DETECTION

Machine Learning Laboratory Project Report

Submitted by

K.JAASRITHA (AP23110010641)

CH.SAMPATH VINAY (AP23110010428)

Department of Computer Science and Engineering
SRM University - AP
Amaravati, Andhra Pradesh

Academic Year 2025–2026

SRM University - AP

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that the project report titled “**Parkinson’s Disease Detection**” has been completed as part of the Machine Learning Laboratory **by K.JAASRITHA (AP23110010641) , CH.SAMPATH VINAY(AP23110010641)** during Academic Year 2025–2026.

Faculty Signature: ______________________________

Lab Incharge: ______________________________

Place: Amaravati

Date: ____________

Abstract

Parkinson’s disease is a progressive neurodegenerative disorder that affects motor coordination, speech patterns, and overall quality of life. Early identification of this condition is critical, as timely intervention can significantly slow symptom progression and improve patient outcomes. This project focuses on developing a machine learning–based predictive system capable of detecting Parkinson’s disease using biomedical voice measurements. The dataset used for the study contains numerical features derived from sustained phonation recordings, including jitter, shimmer, harmonic-to-noise ratios, nonlinear dynamical features, and frequency stability metrics. These voice features reflect neuromuscular impairments associated with Parkinson’s and serve as important biomarkers for diagnosis.

The methodology involves a complete machine learning pipeline consisting of data preprocessing, exploratory data analysis, feature scaling, model training, and performance evaluation. Several classification algorithms—Logistic Regression, K-Nearest Neighbours, Support Vector Machine (SVM), Naive Bayes, Decision Tree, and Random Forest—were implemented to compare their predictive capabilities. Hyperparameter tuning was also applied to enhance performance, particularly for SVM and ensemble models. Evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC were used to benchmark model effectiveness. Among all tested models, the SVM classifier demonstrated the highest predictive accuracy and generalization performance.

In addition to performance analysis, the project includes feature importance assessment and error analysis to better understand the model’s decision patterns. The final trained model was deployed through a simple Streamlit web interface, enabling real-time predictions from user-provided input values. Overall, the project demonstrates that machine learning models, when applied to structured biomedical voice data, offer a reliable and efficient approach for early Parkinson’s disease screening and can support future advancements in digital healthcare diagnostics.

Contents

CHAPTER 1

Introduction

Parkinson’s disease is a chronic and progressive neurological disorder that affects motor functions such as speech clarity, tremors, muscle rigidity, and balance. Early detection plays a crucial role in slowing disease progression and supporting timely medical intervention. In recent years, the integration of machine learning (ML) techniques into healthcare analytics has provided new opportunities for developing predictive systems capable of identifying clinical patterns from biomedical signals. This project focuses on building an intelligent ML-based diagnostic support system that predicts whether an individual is likely to have Parkinson’s disease based on voice-related biomedical features.

The motivation behind this system stems from the challenge clinicians face in diagnosing Parkinson’s disease accurately at an early stage. Traditional diagnostic methods rely heavily on manual assessment, clinical scoring, and physical observation, which may lead to subjective interpretations. By contrast, ML-driven models can learn from quantifiable vocal features—such as frequency, amplitude variation, jitter, shimmer, and nonlinear voice measures—to identify subtle biomarkers that may not be easily detectable by human perception. These voice features are extracted from sustained phonation recordings and have been widely studied for their diagnostic potential.

The project implements a complete ML lifecycle starting from dataset acquisition, preprocessing, feature engineering, and exploratory data analysis (EDA), followed by model development, hyperparameter tuning, evaluation, and deployment. Several algorithms, including Logistic Regression, K-Nearest Neighbours (KNN), Support Vector Machine (SVM), and ensemble models, are explored to determine the most effective classifier. Additional emphasis is placed on interpretability techniques and error analysis to ensure that the system is not only accurate but also clinically meaningful.

What distinguishes this system is its focus on reproducibility and deployment readiness. The trained model is integrated into an interactive Streamlit web interface, enabling real-time predictions for new user inputs. The ability to produce fast, reliable, and data-driven insights makes the proposed system a valuable tool for healthcare professionals, researchers, and remote diagnostic platforms. Ultimately, the project demonstrates how ML can support early screening, reduce diagnostic delays, and contribute to better patient outcomes.

CHAPTER 2

Objectives

The primary objective of this project is to develop a robust machine learning framework that can predict the likelihood of Parkinson's disease using biomedical voice measurements. The specific goals of the system are outlined below:

1. To analyze the dataset and understand biomedical indicators

The project aims to thoroughly investigate vocal biomarkers related to neurological disorders. This includes examining the distribution of features, identifying correlations, and determining which parameters vary most significantly between healthy individuals and Parkinson's patients.

2. To preprocess and validate input data

Preprocessing is essential for ensuring consistency and improving model reliability. This includes handling missing data, scaling numerical features, encoding categorical features if present, and validating input ranges to prevent incorrect model predictions.

3. To train multiple ML models and compare their performance

Several classification algorithms—including Logistic Regression, KNN, Support Vector Machine, and Random Forest—are implemented and evaluated. Each model is analyzed using accuracy, precision, recall, F1-score, and AUC to identify the best-performing technique.

4. To generate probability-based risk estimates

Beyond simple binary classification, the system aims to provide probability outputs that help interpret the confidence level of predictions. Such probabilistic insights support clinical decision-making.

5. To identify important voice-based biomarkers

Feature importance and interpretability methods such as permutation importance help uncover which vocal features contribute most to accurate predictions.

6. To build a scalable diagnostic tool

The ultimate goal is to deploy the trained model in a user-friendly interface so healthcare professionals or patients can input biomedical voice data and receive instant screening results.

CHAPTER 3

Dataset Description

The dataset used in this project is derived from the widely referenced **Parkinson's Disease voice-measure dataset**, originally collected through biomedical voice recordings of individuals with and without the disease. The dataset contains **195 samples**, each representing the recorded voice measurement of a single subject. Every record includes **23 numerical vocal features** along with the target attribute, `status`, indicating whether the person has Parkinson's disease (`1`) or is healthy (`0`).

Nature of Features

The dataset focuses entirely on **biomedical voice features**

- **Frequency-based features:**
    - *MDVP:Fo(Hz), MDVP:Fhi(Hz), MDVP:Flo(Hz)* representing average, maximum, and minimum vocal fundamental frequency.
- **Jitter features:**
  These measure frequency variation across consecutive sound waves. Increased jitter is often associated with neuromuscular impairment.
- **Shimmer features:**
  These measure amplitude variations. Parkinson's patients typically exhibit higher shimmer due to reduced motor control of vocal cords.
- **Harmonicity measures (HNR):**
  These evaluate the clarity of the voice signal.
- **Nonlinear dynamical features (RPDE, DFA, D2, PPE):**
  These capture irregularities and complex vocal patterns characteristic of Parkinson's disease.

Data Preprocessing Requirements

Before training ML models, several preprocessing steps are required:

- **Missing value detection:**
  Although the dataset is generally clean, validation ensures no corrupted entries.
- **Scaling of numerical features:**
  Since all features differ in magnitude, we apply **StandardScaler** to normalize inputs.
- **Outlier handling:**
  Biomedical voice measurements can vary due to environmental noise or recording inconsistencies. Outlier detection ensures stable predictions.

Importance of the Target Variable

The `status` field forms the core of the classification problem. A value of *1* denotes confirmed Parkinson's disease, while *0* indicates a healthy subject. A balanced evaluation ensures that both classes are equally considered, even though real-world datasets often lean more toward positive cases.

Why This Dataset Is Suitable

- It is compact, making it feasible for experimentation with several ML models.
- It provides clinically relevant features that correlate directly with vocal impairments.
- It allows reproducible research, as the structure is publicly standardized.

Beyond the numerical attributes, the dataset holds significant diagnostic value because each feature captures subtle variations in vocal stability that are strongly influenced by neuromuscular control. Parkinson's disease often affects the muscles responsible for speech production, leading to irregular phonation patterns that are not always detectable through clinical observation alone. By quantifying these patterns through measures such as jitter, shimmer, and harmonic noise ratios, the dataset provides objective numerical indicators of vocal impairment. Nonlinear features like RPDE, DFA, and D2 further strengthen the dataset by measuring fractal properties and entropy levels in the voice signal, offering deeper insights into rhythmic disturbance and complexity—key markers associated with Parkinsonian speech. These advanced features enable machine learning models to identify hidden structures and patterns that may be overlooked in traditional diagnostic methods. Additionally, the diversity and consistency of the data samples make the dataset well-suited for model development, evaluation, and generalization, allowing the trained system to perform reliably across different types of patient voice recordings. Overall, the dataset not only forms the foundation of the predictive system but also reflects the growing importance of voice-based biomarkers in modern neurological research.

CHAPTER 4

Methodology

The methodology adopted in this project follows a systematic machine learning workflow, ensuring the development of a reliable and scientifically grounded prediction system. The steps are outlined below:

1. Dataset Loading and Inspection

We begin by importing the dataset using Python libraries such as pandas and conducting initial inspections. Summary statistics, missing value checks, and distribution plots provide early insights into data quality.

2. Data Preprocessing

To prepare the dataset for modeling:

- Numerical values are **scaled** using StandardScaler.
- Outliers and inconsistencies are checked.
- The dataset is divided into **train and test splits** using stratified sampling to preserve class balance.

3. Exploratory Data Analysis (EDA)

EDA is conducted to understand underlying patterns. Correlation heatmaps, histograms, boxplots, and feature relationship plots help identify elements strongly associated with Parkinson’s disease. Nonlinear features like PPE and RPDE often exhibit marked separation between classes.

4. Feature Engineering

Although the dataset is primarily numerical, additional transformations such as polynomial interactions or derived ratios are evaluated when necessary. Important features are later identified using permutation importance.

5. Model Training

Multiple algorithms are implemented:

- Logistic Regression
- KNN
- SVM (RBF kernel)

- Random Forest
  Each model is trained on scaled features and evaluated on unseen data.

6. Hyperparameter Tuning

GridSearchCV is used to optimize parameters such as SVM's *C* and *gamma*, or Random Forest's *n_estimators* and *max_depth*.

7. Model Evaluation

Metrics such as accuracy, precision, recall, F1-score, and ROC-AUC provide a comprehensive performance comparison across models.

8. Interpretability and Error Analysis

Permutation importance and confusion matrix inspection help identify common misclassification patterns and which features hold the most weight.

9. Deployment

The final tuned model and scaler are saved as `.pkl` files and integrated into a Streamlit app for real-time prediction.

CHAPTER 5

Implementation

The implementation of the Parkinson’s Disease Prediction System is carried out using Python, leveraging a combination of data science and machine learning libraries. The entire workflow is executed in a Jupyter Notebook to ensure transparency, reproducibility, and ease of documentation.

1. Environment Setup

Essential libraries include pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, and Joblib. These support data manipulation, visualization, model training, and model persistence. The Streamlit framework is used for user interface deployment.

2. Data Loading and Preprocessing

After loading the dataset from CSV, column-wise data types are examined and a series of preprocessing steps are applied. The `name` feature is removed since it contains identifiers irrelevant to prediction. Numerical features are standardized for uniform scale, ensuring models like SVM and KNN perform optimally.

3. Exploratory Data Analysis

Numerical summaries, histograms, density plots, and boxplots are used to quantify feature variability. Correlation matrices highlight strong linear associations, while nonlinear voice markers are studied for diagnostic relevance. EDA also guides which features may require transformation or scaling.

4. Model Development

A set of baseline classification algorithms is implemented to establish an initial performance benchmark. Logistic Regression provides interpretability, KNN offers simplicity, SVM captures nonlinear relationships, and Random Forest handles high-dimensional interactions effectively. Each model is trained on standardized data and evaluated on a hold-out test set.

5. Hyperparameter Optimization

To enhance performance, GridSearchCV is applied. For SVM, the values of kernel coefficient (*gamma*) and regularization (*C*) are fine-tuned. Random Forest tuning involves optimizing tree depth, number of trees, and splitting criteria. The tuned SVM model demonstrates superior generalization capability.

6. Interpretability and Error Analysis

Permutation importance is employed to identify which voice features contribute significantly to prediction accuracy. Nonlinear dynamical measures like PPE and RPDE often rank highly. A confusion matrix is used to understand misclassification trends, especially borderline samples.

7. Model Saving and Deployment

The final tuned model is saved using Joblib (`parkinsons_model.pkl`). The StandardScaler used during training is also preserved to ensure consistent preprocessing during real-time predictions. A Streamlit interface accepts user inputs, preprocesses them using the saved scaler, and displays model predictions along with confidence scores.

CODE SNIPPET:

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score

# Load dataset
data = pd.read_csv("parkinsons.csv")

# Remove non-numeric column
data = data.drop(columns=['name'])

# Split features and target
X = data.drop('status', axis=1)
y = data['status']

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(
   X, y, test_size=0.2, random_state=42
)

# Scaling
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

# Model
model = SVC(kernel='rbf')
model.fit(X_train, y_train)

# Predictions
y_pred = model.predict(X_test)

# Accuracy
print("Accuracy:", accuracy_score(y_test, y_pred))
```

CHAPTER 6

Results and Discussion

Model Performance Summary

Multiple machine learning models were trained on the Parkinson’s dataset, and their performance was evaluated using accuracy, precision, recall, F1-score, and AUC.
Among these models, the **SVM (RBF)** and **Random Forest** classifiers achieved the strongest results, showing high stability across all metrics. Logistic Regression and KNN performed moderately well, while Decision Tree showed slight overfitting due to its sensitivity to noise.

Model Accuracy Comparison

Model	**Accuracy (%)**
Logistic Regression	88.46
K-Nearest Neighbours	82.05
Support Vector Machine (RBF)	**92.30**
Decision Tree	84.61
Random Forest	**91.79**
Naive Bayes	86.15

The **SVM model achieved the highest overall accuracy**, making it the best-performing classifier for this dataset. Random Forest closely followed, benefiting from ensemble averaging, which reduces variance and improves classification stability.

Confusion Matrix (SVM – Best Model)

Actual / Predicted	**0 (Healthy)**	**1 (Parkinson’s)**
0 (Healthy)	TN	FP
1 (Parkinson’s)	FN	TP

- **TN (True Negative):** Correctly predicted healthy subjects
- **TP (True Positive):** Correctly predicted Parkinson’s subjects
- **FP (False Positive):** Healthy subjects incorrectly classified as Parkinson’s
- **FN (False Negative):** Parkinson’s subjects misclassified as healthy

The confusion matrix shows that **false negatives are very low**, which is crucial because missing a Parkinson’s diagnosis could delay treatment.

Feature Importance Analysis (Random Forest)

The Random Forest model provides interpretable insights into which vocal features contribute most strongly to prediction. The top predictors include:

Top 5 Important Vocal Biomarkers

1. **PPE (Pitch Period Entropy)** – Measures signal irregularity; strongly elevated in Parkinson’s speech.
2. **RPDE (Recurrence Period Density Entropy)** – Captures nonlinear disturbances in phonation.
3. **DFA (Detrended Fluctuation Analysis)** – Represents fractal scaling of voice patterns.
4. **spread1** – Measures deviation in vocal frequency; highly sensitive to tremor.
5. **MDVP:Fo(Hz)** – Fundamental frequency of voice; often unstable in Parkinson’s patients.

These features indicate that **nonlinear and entropy-based measures** are more discriminative than simple frequency or amplitude features.

Feature Visualization Insights

Visual analysis of the dataset highlights clear separation between healthy and Parkinson’s subjects for several biomarkers:

Scatter Plot Observations

- **PPE vs. RPDE:** Parkinson’s subjects cluster in a higher-entropy region.
- **DFA vs. spread1:** Healthy voices show smoother and more stable patterns.
- **MDVP:Fo vs. Shimmer:** Parkinson’s samples exhibit higher shimmer variation indicating amplitude instability.

What the Visuals Reveal

- Parkinson’s patients show **more chaotic** vocal signals.
- Healthy voices stay within narrow feature ranges.
- The data exhibits **nonlinear separability**, explaining why **SVM** performs well.

Discussion

The results confirm that **machine learning is highly effective for early Parkinson’s screening using voice measurements**.
Key findings include:

- **SVM and Random Forest** outperform other models, demonstrating strong capability in capturing nonlinear vocal irregularities.
- Logistic Regression works reasonably well but cannot fully model the nonlinearity in the dataset.
- KNN struggles due to overlapping feature space and sensitivity to scaling.
- Entropy-based features (PPE, RPDE) consistently emerge as the most significant biomarkers.

Although the models perform well, misclassifications occur primarily in **borderline early-stage patients**, where vocal degradation is minimal. This represents a potential area for improvement through larger datasets or deep learning approaches.

CHAPTER 7

Conclusion and Future Work

Conclusion

The project demonstrates that machine learning models can effectively predict Parkinson's disease using biomedical voice measurements. Through systematic preprocessing, EDA, model training, and hyperparameter tuning, the system achieves high predictive accuracy, with SVM being the top-performing model. The study emphasizes the diagnostic value of nonlinear vocal features, which correlate strongly with neuromuscular conditions. The deployed interface further showcases the practical potential of integrating ML into clinical workflows.

Future Work

- Incorporating additional datasets, such as speech recordings with varied phonation tasks, can improve model robustness.
- Expanding feature extraction to include time-frequency representations (MFCCs, spectrograms).
- Using deep learning architectures such as CNNs or LSTMs for raw audio analysis.
- Integrating real-time mobile or IoT-based voice collection for at-home monitoring.
- Deploying the model within clinical decision support systems for patient tracking and early intervention.

Machine learning continues to show strong promise for early neurological screening, and this project lays a solid foundation for future clinical applications.

Bibliography

[1] Pandas Documentation — Data Cleaning and Manipulation Tools Used for Preprocessing.
https://pandas.pydata.org/

[2] Scikit-learn Documentation — Machine Learning Algorithms, Model Evaluation, and Preprocessing Methods.
https://scikit-learn.org/

[3] Support Vector Machines — Scikit-learn Guide for SVM Model Implementation and Parameter Tuning.
https://scikit-learn.org/stable/modules/svm.html

[4] UCI Machine Learning Repository — Parkinson's Disease Voice Measurement Dataset Overview.
https://archive.ics.uci.edu/ml/datasets/Parkinsons