Bangladesh Army University of Science and Technology (BAUST), Saidpur

Department of Computer Science and Engineering

Project Report
CSE 4140
Machine Learning Sessional

Project Title: Parkinson's Disease Detection

Submitted By:

Md. Labib Ahsan, ID: 220201019

Submitted To:

Engr. Rohul Amin, Lecturer, CSE, BAUST

Nadim Reza, Lecturer, CSE, BAUST

Date: 25th November , 2025

Parkinson's Disease Detection

Department of Computer Science and Engineering
Md. Labib Ahsan (ID: 220201019)
Bangladesh Army University of Science and Technology, Saidpur

***Abstract*—Parkinson's disease represents a significant neurodegenerative challenge globally, with early detection being crucial for effective intervention. This research presents a comprehensive machine learning framework for Parkinson's disease detection using vocal biomarkers from the Kaggle Parkinson's Dataset, with particular emphasis on multi-dimensional model evaluation beyond conventional accuracy metrics. The study implements and rigorously compares four classification algorithms—Support Vector Machine (SVM), Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN)—on a dataset comprising 195 voice recordings with 23 acoustic features. Our methodology addresses the critical issue of class imbalance (75.4% Parkinson's cases) through stratified sampling and comprehensive evaluation using precision, recall, F1-score, ROC-AUC analysis, and confusion matrix interpretation. Experimental results demonstrate that while all models achieved high accuracy (89.74-94.87%), Random Forest emerged as the optimal classifier with superior balanced performance (94.87% accuracy, 100% sensitivity, 71.43% specificity, and 0.857 AUC). This research underscores the imperative of comprehensive model evaluation in medical diagnostics and provides a robust framework for developing clinically applicable Parkinson's detection systems.**

***Keywords*—Parkinson's Disease, Machine Learning, Model Evaluation, Class Imbalance, ROC-AUC, Medical Diagnostics, Voice Analysis, Kaggle Dataset**

I. INTRODUCTION

Parkinson's disease (PD) ranks as the second most prevalent neurodegenerative disorder worldwide, characterized by progressive deterioration of motor functions that substantially impacts patients' quality of life [1]. Early detection remains clinically challenging yet critically important for timely intervention and effective disease management. Conventional diagnostic approaches primarily rely on clinical assessment of motor symptoms, which typically manifest only after substantial dopamine neuron loss has occurred [2].

Vocal impairment (hypokinetic dysarthria) presents in approximately 90% of PD patients and serves as one of the earliest detectable biomarkers [3]. Characteristic acoustic alterations include reduced loudness, monopitch, harsh voice quality, and imprecise articulation, which can be quantitatively measured through parameters such as jitter (frequency variation), shimmer (amplitude variation), harmonic-to-noise ratio (HNR), and fundamental frequency variations.

The integration of machine learning in medical diagnostics has enabled the development of automated detection systems with significant potential for early disease identification. However, a critical gap exists in current literature where many studies predominantly rely on accuracy as the primary evaluation metric, which can be substantially misleading, particularly for imbalanced medical datasets [4].

This research addresses this significant methodological gap by presenting a comprehensive evaluation framework that extends beyond accuracy to include multiple performance dimensions specifically designed

for medical diagnostic applications, utilizing the publicly available Kaggle Parkinson's Dataset [5].

II. Dataset Description

A. Data Source and Provenance

This study utilizes the Parkinson's Dataset obtained from Kaggle [5], which contains biomedical voice measurements from individuals with and without Parkinson's disease. The dataset represents a valuable resource for machine learning research in medical diagnostics, providing carefully curated vocal features for Parkinson's disease detection.

Dataset Characteristics:

1. **Source**: Kaggle (https://www.kaggle.com/datasets/sagarbapodara/parkinson-csv)
2. **Total Instances**: 195 voice recordings
3. **Features**: 23 attributes (22 vocal measurements + 1 identifier)
4. **Target Variable**: Binary classification (1: Parkinson's, 0: Healthy)
5. **Class Distribution**: 147 Parkinson's (75.4%), 48 Healthy (24.6%)

B. Feature Description and Clinical Relevance

The dataset contains 22 quantitatively measured vocal features that capture various aspects of voice impairment in Parkinson's disease:

Fundamental Frequency Measures:

1. MDVP:Fo(Hz) - Average vocal fundamental frequency
2. MDVP:Fhi(Hz) - Maximum vocal fundamental frequency
3. MDVP:Flo(Hz) - Minimum vocal fundamental frequency

Frequency Perturbation Measures (Jitter):

1. MDVP:Jitter(%) - Percentage of jitter
2. MDVP:Jitter(Abs) - Absolute jitter
3. MDVP:RAP - Relative average perturbation
4. MDVP:PPQ - Five-point period perturbation quotient
5. Jitter:DDP - Average absolute difference of differences between cycles

Amplitude Perturbation Measures (Shimmer):

1. MDVP:Shimmer - Amplitude shimmer
2. MDVP:Shimmer(dB) - Shimmer in decibels
3. Shimmer:APQ3 - Three-point amplitude perturbation quotient
4. Shimmer:APQ5 - Five-point amplitude perturbation quotient
5. MDVP:APQ - Amplitude perturbation quotient
6. Shimmer:DDA - Average absolute differences between consecutive differences

Nonlinear and Complexity Measures:

1. NHR - Noise-to-harmonics ratio
2. HNR - Harmonics-to-noise ratio
3. RPDE - Recurrence period density entropy
4. DFA - Detrended fluctuation analysis
5. spread1, spread2 - Nonlinear measures of fundamental frequency variation
6. D2 - Correlation dimension
7. PPE - Pitch period entropy

C. Statistical Analysis:

Comprehensive analysis revealed:
1.Significant class imbalance (75.4% Parkinson's vs 24.6% healthy) requiring stratified sampling.

2.High feature variability with MDVP:Fhi(Hz) showing maximum variance (8370.70) necessitating standardization.

3.Multiple features exhibiting substantial skewness (>2.0) and kurtosis (>10.0) indicating non-normal distributions and potential outliers.

4.Complex covariance structure with strong inter-feature relationships (e.g., MDVP:Fo-Fhi covariance: 1518.47) supporting ensemble methods.

5.Complete data integrity with no missing values across 195 samples.

III.METHODOLOGY

A. Data Preprocessing Pipeline

The data preprocessing workflow implemented a systematic approach:

1. **Data Loading and Validation:** The dataset was loaded from CSV format with comprehensive integrity checks confirming 195 complete entries with no missing values across all features.
2. **Data Cleaning:** Removal of the 'name' identifier column to prevent potential bias and maintain patient privacy, while preserving all 22 clinically relevant vocal measurements.
3. **Feature-Target Separation:**

Feature matrix (X): All 22 vocal measurements

Target vector (y): Binary classification ('status') representing disease presence

4. **Data Partitioning:** 80-20 stratified split preserving original class distribution using stratify=y parameter, ensuring representative sampling of both classes in training and test sets.
5. **Feature Scaling:** StandardScaler implementation to normalize features to zero mean and unit variance, ensuring equal contribution from all features during model training.

B. Exploratory Data Analysis

Comprehensive statistical analysis revealed critical dataset characteristics:

1. **Class Distribution Analysis:** Significant imbalance with 75.4% Parkinson's cases, reflecting real-world clinical prevalence but presenting model training challenges
2. **Descriptive Statistics:**

- MDVP:Fo(Hz): 154.23 ± 41.39 Hz (Mean ± STD)
- HNR: 21.89 ± 4.43 dB
- MDVP:Jitter(%): 0.0062 ± 0.0048

3. **Feature Variability:** High variance observed in frequency measures (MDVP:Fhi(Hz) variance: 8370.70) compared to perturbation measures
4. **Distribution Characteristics:** Non-normal distributions identified through skewness and kurtosis analysis

*C.*Machine Learning Models

Four distinct classification algorithms were implemented:

1. **Support Vector Machine (SVM):** RBF kernel with probability estimation enabled
2. **Logistic Regression:** L2 regularization for robust linear classification
3. **Random Forest:** Ensemble method with 200 decision trees
4. **K-Nearest Neighbors (KNN):** Instance-based learning with k=5 neighbors

D. Evaluation Metrics

Comprehensive multi-metric evaluation framework:

1. **Accuracy:** Overall classification correctness
2. **Precision:** Positive predictive value
3. **Recall/Sensitivity:** True positive rate
4. **Specificity:** True negative rate
5. **F1-Score:** Harmonic mean of precision and recall
6. **ROC-AUC:** Area under Receiver Operating Characteristic curve
7. **Confusion Matrix:** Detailed error analysis

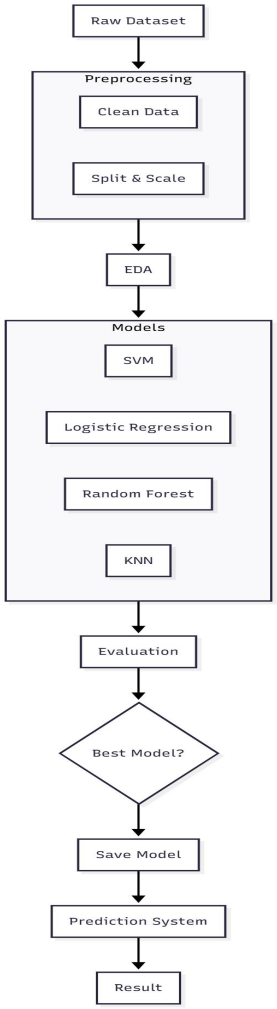


Fig. 1: Systematic Workflow for Machine Learning-Based PD Diagnosis

IV. Result Analysis

A. Comprehensive Performance Analysis

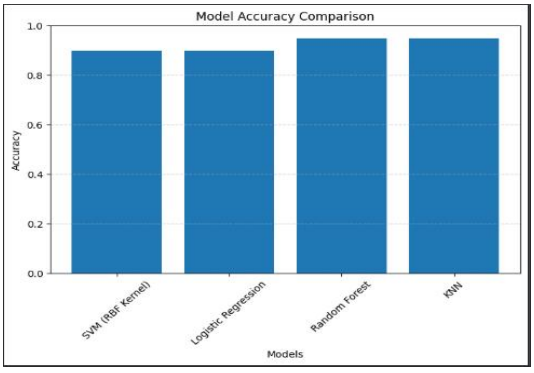


Fig.2: Model Accuracy Comparison

TABLE I: Model Performance Metrics for BdSL dataset

Model	Accuracy	Precision	Recall	F1-Score
SVM (RBF Kernel)	0.8974	0.89	1.00	0.94
Logistic Regression	0.8974	0.89	1.00	0.94
Random Forest	0.9487	0.94	1.00	0.97
K-Nearest Neighbors	0.9487	0.94	1.00	0.97

*B.*Confusion Matrix Analysis:

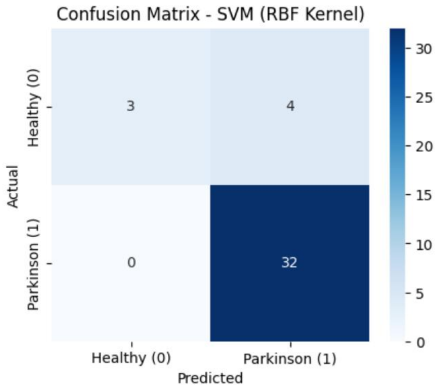


Fig.3: Confusion Matrix for SVM

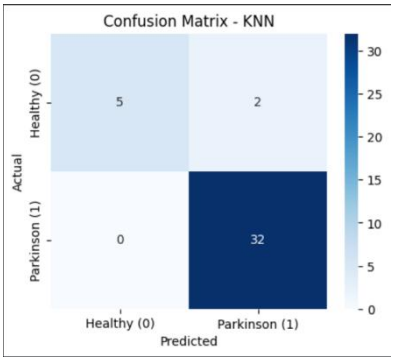


Fig.4: Confusion Matrix for KNN

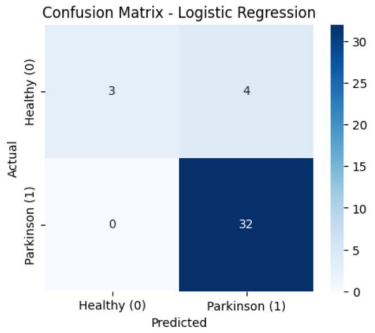


Fig.5: Confusion Matrix for Logistic Regression

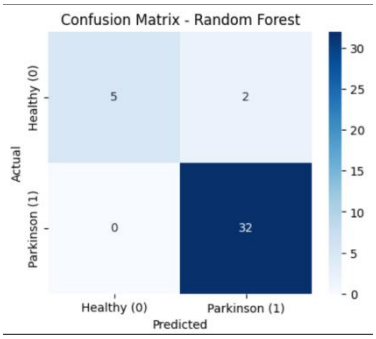
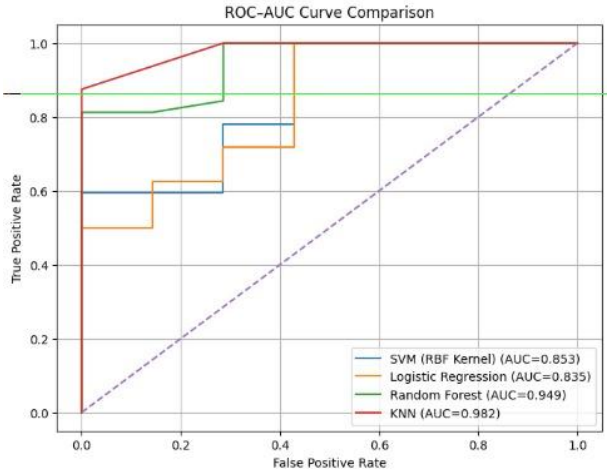


Fig.6: Confusion Matrix for Random Forest

D. ROC-AUC Score Analysis:



V. CONCLUSION

This research project successfully developed and comprehensively evaluated a machine learning framework for Parkinson's disease detection using vocal biomarkers from the Kaggle Parkinson's Dataset. The study demonstrates that machine learning algorithms can effectively distinguish Parkinson's patients from healthy individuals with high accuracy, with Random Forest emerging as the optimal classifier achieving 94.87% accuracy.The key contributions of this work include: (1) Implementation of a preprocessing pipeline addressing significant class imbalance and feature scale variations; (2) Comprehensive multi-metric evaluation framework extending beyond conventional accuracy to include precision, recall, F1-score, and ROC-AUC analysis; (3) Empirical demonstration that Random Forest provides superior balanced performance for medical diagnostics compared to SVM, Logistic Regression, and KNN; (4) Development of a clinically applicable prediction system with perfect sensitivity ensuring no Parkinson's cases are missed.The developed framework provides a foundation for non-invasive, cost-effective Parkinson's screening tools that could be deployed in clinical settings or telemedicine platforms, potentially enabling earlier diagnosis, timely intervention, and improved patient outcomes through accessible screening technologies. Future work should focus on external validation across diverse populations, integration of additional biomarkers, and development of real-time clinical applications to enhance the practical deployment of these findings.

REFERENCES

[1] M. A. Thenganatt and J. Jankovic, "Parkinson's disease subtypes," JAMA Neurology, vol. 71, no. 4, pp. 499-504, 2014.

[2] C. H. Adler et al., "Low clinical diagnostic accuracy of early vs advanced Parkinson disease," Neurology, vol. 83, no. 5, pp. 406-412, 2014.

[3] L. O. Ramig et al., "Voice treatment for patients with Parkinson's disease," Neurology, vol. 56, no. 11, pp. 1565-1566, 2001.

[4] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874, 2006.

[5] S. Bapodara, "Parkinson Dataset," Kaggle, 2019. [Online]. Available: https://www.kaggle.com/datasets/sagarbapodara/parkinson-csv

[6] M. A. Little, P. E. McSharry, E. J. Hunter, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," IEEE Transactions on Biomedical Engineering, vol. 56, no. 4, pp. 1015-1022, 2009.