**ABSTRACT**

Parkinson’s disease(PD) is a chronic and progressive neurodegenerative disorder.The “Parkinson’s disease detection using Machine Learning” project focus on detecting Parkinson’s disease using the voice recordings of the patient. The relevance of this project lies in its potential to provide a non-invasive, cost-effective, and accessible method for Parkinson's disease detection.

From the three papers,we get to know that different models were used for the detection of parkinson’s disease.These papers proposes a machine learning model for the early detection of PD using voice and speech data.The reaserchers applied various data processing techniques and dimensionality reduction methods and then compared different machine learning classifiers.

The proposed system is the comparative study of two algorithms:Random Forest and Support Vector Machine algorithm.The model will classify the disease under two classes.By comparing these algorithms,the project aims to determine which model offers the highest accuracy in diagnosing the disease.

The dataset is taken from Kaggle repository.The dataset contains sample 6916 observations and has 17 columns and 1 identifier,1 class variable and 14 features.The dataset contains numeric and boolean values.

Dataset: [parkinsons-detector/data at master · adachille/parkinsons-detector · GitHub](https://github.com/adachille/parkinsons-detector/tree/master/data)

**References**

* Govindu, Aditi, and Sushila Palwe. "Early detection of Parkinson's disease using machine learning." *Procedia Computer Science* 218 (2023): 249-261.
* Pramanik, Anik, and Amlan Sarker. "Parkinson’s disease detection from voice and speech data using machine learning." *Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020*. Springer Singapore, 2021.
* Lamba, Rohit, et al. "A hybrid system for Parkinson’s disease diagnosis using machine learning techniques." *International Journal of Speech Technology* (2022): 1-11.
* Erdogdu Sakar, B., Isenkul, M., Sakar, C.O., Sertbas, A., Gurgen, F., Delil, S., Apaydin, H., Kursun, O., 'Collection and Analysis of a Parkinson Speech Dataset with Multiple Types of Sound Recordings', IEEE Journal of Biomedical and Health Informatics, vol. 17(4), pp. 828-834, 2013.

**Faculty Guide: Submitted by:**

Prof Biju Skaria Sandriya Soman

Asso.Professor & HOD MAC23MCA-2049

MCA Dept,MACE

**Signature of project coordinator**

**INTRODUCTION**

Parkinson’s disease(PD) is a chronic and progressive neurodegenerative disorder.It affects mobility,speech and posture.It occurs due to the death of neurons,resulting in a decrease in dopamine levels in the brain.The “Parkinson’s disease detection using Machine Learning” project focus more on detecting Parkinson’s disease and diagnose it using the voice recordings of the patient.

Changes in speech are among the early symptoms of Parkinson's. Speech characteristics such as voice tremor, reduced volume can serve as biomarkers for early detection.90% of patients display signs of vocal cord injuries as a symptom in stage 0.It is not only easy to measure,but also falls under the category of telemedicine.Patient need not travel pysically to a doctor instead;they can record audio using phones and perform a simple test at home.Common voice modulation symptoms are called dysphonia.Patients can be asked to hold a single vowel’s pitch for as long as possible,also known as sustained phonation .These test is used to diagnose PD in stage 0.

The relevance of this project lies in its potential to provide a non-invasive, cost-effective, and accessible method for Parkinson's disease detection. Traditional methods often involve expensive imaging techniques, which may not be readily available in all healthcare settings, especially in low-resource environments. A machine learning-based approach using voice data can facilitate early diagnosis,and potentially slowing disease progression.In this project two algorithms are compared.They are Support Vector Machine and Random Forest Algorithm.

**Faculty Guide: Submitted by:**

Prof Biju Skaria Sandriya Soman

Asso.Professor & HOD MAC23MCA-2049

MCA Dept,MACE

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Title** | **Year** | **Journal Name** | **Summary** |
| **Paper 1** | Early detection of Parkinson’s disease using machine learning | 2023 | International Conference on Machine Learning and Data Engineering  (Elsevier) | Logistic  Regression 83.67%    Random  Forest 91.83%    KNN 85.71%    SVM 85.71% |
| **Paper 2** | Parkinson’s Disease Detection from voice and speech data using Machine Learning | 2021 | Parkinson’s Disease Detection from voice and Speech Data using Machine Learning  (ResearchGate) | SVM 94.1%    K-NN 86.3%    AdaBoost 90.4% |
| **Paper 3** | A hybrid system for Parkinson’s disease diagnosis using maching learning. | 2021 | Internal Journal of SpeechTechnology  (Springer) | Hybrid system accuracy:95.58% |

**PROJECT PROPOSAL**

This project aims to develop and evaluate machine learning models for early detection of Parkinson's disease using voice recordings, with the goal of enabling remote screening and diagnosis. Parkinson's disease is a progressive neurological disorder affecting movement and speech. Early detection can enable better treatment and management. Voice impairment is an early symptom that can be measured non-invasively, and machine learning on voice data offers potential for automated screening.

The proposed methodology involves the following steps. First, we will use the MDVP audio dataset of vowel phonations from PD patients and healthy controls, which contains 14 voice measurement features for 6916voice recordings. Data preprocessing will include cleaning data, handling missing values, exploring data distributions and correlations, and applying scaling/normalization. We will use Principal Component Analysis for feature selection to identify the most relevant voice features.

For model development, we will implement and compare two machine learning models: Random Forest, and Support Vector Machine. Model evaluation will use metrics like accuracy, precision and recall.

Expected outcomes include a comparative analysis of ML model performance for PD detection, identification of the most relevant voice features, insights on handling imbalanced medical data, and recommendations for the optimal model and approach. The potential impact of this project is significant, as it could enable remote, non-invasive screening for Parkinson's disease, improve early detection rates to allow earlier intervention, and advance telemedicine capabilities for neurological disorders.

**DATSET**

Dataset is taken from Kaggle repository.It contains 6916 instances.Dataset has 1 identifier,14 features and 1 class variable.Identifier is the subject and the class label is Class.The main aim is to discriminate healthy people from those with the PD,according to the Class which is set to 0 for healthy and 1 for PD

* **Study**: Identifier for the study from which the data was collected.
* **Subject**: Unique identifier for the individual subjects in the study.
* **Jitter(%)**: Variation in frequency from cycle to cycle in the voice signal.
* **Jitter(Abs)**: Absolute variation in frequency from cycle to cycle.
* **Jitter**
  + : Relative Average Perturbation; the average absolute difference between a period and the average of it and its two neighbors.
* **Jitter**
  + : Five-point period perturbation quotient; similar to RAP but averaged over five periods.
* **Jitter**
  + : Average absolute difference between consecutive differences in periods.
* **Shimmer**: Amplitude variation from cycle to cycle in the voice signal.
* **Shimmer(dB)**: Amplitude variation in decibels.
* **Shimmer**
  + : Three-point amplitude perturbation quotient; the average absolute difference between amplitude and the average of it and its two neighbors.
* **Shimmer**
  + : Five-point amplitude perturbation quotient; similar to APQ3 but averaged over five periods.
* **Shimmer**
  + : Eleven-point amplitude perturbation quotient; similar to APQ5 but averaged over eleven periods.
* **Shimmer**
  + : Average absolute difference between consecutive differences in amplitude.
* **NHR**: Noise-to-Harmonics Ratio; a measure of the noisiness of the voice signal.
* **HNR**: Harmonics-to-Noise Ratio; a measure of the harmonic strength in the voice signal.
* **RPDE**: Recurrence Period Density Entropy; a nonlinear dynamic parameter.
* **DFA**: Detrended Fluctuation Analysis; a measure of the long-term correlation in the time series.
* **spread1**: Nonlinear measure; related to the fundamental frequency variation.
* **spread2**: Nonlinear measure; related to the fundamental frequency variation.
* **D2**: Correlation dimension; a measure of the complexity of the voice signal.
* **PPE**: Pitch Period Entropy; a measure of the irregularity in the voice signal.
* **Class**: The target variable indicating whether the subject has Parkinson's disease (1) or not (0).

**Dataset:**

[parkinsons-detector/data/multiple-sound-recording/TE\_MSRtrain\_combined.csv at master · adachille/parkinsons-detector · GitHub](https://github.com/adachille/parkinsons-detector/blob/master/data/multiple-sound-recording/TE_MSRtrain_combined.csv)