Content Checked For Plagiarism
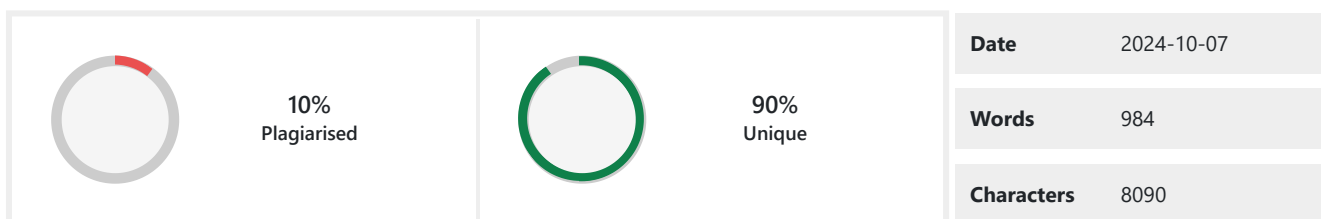
XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

Comparison of different Machine Learning algorithms for Parkinson's disease detection

Lokesh Khedekar
Computer Science And
Engineering (AI)
BRACT'S Vishwakarma
Institute of Technology
Pune, India
lokesh.khedekar@vit.edu

Tanishq Thuse
Computer Science And
Engineering (AI)
BRACT'S Vishwakarma
Institute of Technology
Pune, India
tanishq.thuse23@vit.edu

Atharva Saltri
Computer Science And
Engineering (AI)
BRACT'S Vishwakarma
Institute of Technology
Pune, India
atharva.saltri23@vit.edu

Tripti Mirani
Computer Science And
Engineering (AI)
BRACT'S Vishwakarma
Institute of Technology
Pune, India
tripti.mirani23@vit.edu

Shrey Rupnavar
Computer Science And

Engineering (AI)
BRACT'S Vishwakarma
Institute of Technology
Pune, India
shrey.rupnavar23@vit.edu

Abstract— Parkinson's disease (PD) is a progressive neurodegenerative condition that affects millions worldwide, and early diagnosis remains critical for effective treatment. Over the past few years, multiple innovative types of machine learning approaches have been developed for PD diagnosis based on data in the form of speech patterns, handwriting samples, sensor data, and many more. However, the performance of these algorithms varies significantly with the choice of data and features used.
In this paper, authors have compared the performance of a number of machine learning algorithms to diagnose Parkinson's disease. The dataset for this study was acquired from Oxford UCI Machine repository. We obtained the dataset for the study from the Oxford UCI Machine repository. After preprocessing the data and extracting relevant features, we applied various algorithms, including logistic regression, support vector machines, random forests to classify individuals as either PD patients or healthy controls.
The performance evaluation parameters, including accuracy, precision, recall, F1 score, and Precision-Recall curve (PR curve), were used to compare the algorithms. Our findings revealed that random forests achieved the highest accuracy and other performance metric scores, while logistic regression and support vector machines offered greater interpretability. Overall, our research underscores the potential of machine learning in early PD detection and emphasizes the importance of comprehensive datasets and effective feature selection. The long-term goal of this work is to increase the sensitivity and specificity diagnostic methods and potentially improve the quality of life for individuals with Parkinson's disease along with better patient outcomes.
Comparison of the classification results of Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, KNN, Gaussian Naive Bayes, and Bernoulli Naive Bayes led to the conclusion of Random Forest as an ideal machine learning (ML) technique for the detection of Parkinson's Disease. The detection accuracy of the Random Forest model is 98.3051%

I.INTRODUCTION

Parkinson's disease has destroyed the lives of 10 million people around the world and is the second most deadly neurodegenerative disease after Alzheimer's disease. The symptoms include: "frozen" facial expression, bradykinesia or slowness of movement, akinesia or impairment of voluntary

movement, tremor, and impairment of the voice. By the time a diagnosis is made, typically 60% of nigrostriatal neurons have degenerated, as does 80% of striatal dopamine. There is no single test in which this condition can be diagnosed. Doctors must perform careful clinical analysis of the patient's medical history. However, this result was inaccurate. According to the National Institute of Neurological Disorders, early diagnosis (with symptoms for ≤ 5 years) is only 53% accurate. This is hardly much better than expected, but early diagnosis makes all the differences in the world for effective treatment.

These problems motivated us to investigate a machine learning approach to accurately diagnose Parkinson's, using a dataset of various speech features (a non-invasive yet characteristic tool) from the University of Oxford.

The purpose of this project is to discover a machine learning technique that can be effectively used for the prediction of Parkinson's disease using relevant data. This study will conduct a comparative analysis of feature selection and representation techniques to identify the most relevant and enlightening features from the available data useful in the treatment of patients.

II. METHODOLOGY

➢ Dataset

We have collected the required Parkinson's disease dataset, created by the University of Oxford, called the Oxford Parkinson's Disease Detection dataset, containing 197 recordings and 23 features, created by Max Little from the University of Oxford, along with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals.

The data contains biomedical voice recordings from 31 individuals, 23 of whom had Parkinson's disease (PD). Each column in the table is a voice measurement, and each row corresponds to one of the 195 voices recorded by that individual ("name row"). For patients with Parkinson's disease, the "Status" column is set to 0 for healthy patients and 1 for Parkinson's disease. The file is in ASCII CSV format.

The rows of the CSV file contain examples for a recording. There are approximately six entries for each patient, and the patient's name is identified in the first row.

TABLE I. OVERVIEW OF VOCAL ANALYSIS ATTRIBUTES AND THEIR PURPOSES

Attribute Purpose

Name Data is stored in ASCII CSV format where patient name and recording number is stored

MDVP:

Fo (Hz)

Fundamental frequency of pitch period

MDVP:

Fhi (Hz)

Upper limit of fundamental frequency or maximum
threshold of voice modulation

MDVP:

Flo (Hz)

Lower limit or minimal vocal fundamental frequency

MDVP:

Jitter,

Abs,

RAP,

PPQ,

DDP

Various measures from Kay Pentax's multi-dimensional
voice program (MDVP), assessing frequency of vibrations
in vocal folds at pitch periods

Jitter

and

Shimmer

Measures of absolute difference between frequencies of
each cycle, after normalizing the average

NHR and

HNR

Signal to noise and tonal ratio measures that indicate
robustness of the environment to noise

Status 0 indicates healthy person while 1 indicates person with
Parkinson's disease (PWP)

D2 Correlation dimension used to identify dysphonia in speech
using fractal objects; a nonlinear, dynamic attribute

RPDE Recurrence Period Density Entropy quantifies the extent to
which the signal is periodic

DFA Detrended Fluctuation Analysis measures the extent of
stochastic self-similarity of noise in speech signals

PPE Pitch Period Entropy is used to assess abnormal variations
in speech on a logarithmic scale

Spread1,

Spread2

Analysis of extent or range of variations in speech with
respect to MDVP: Fo (Hz)

Matched Source

Similarity 25%

Title:Sum-Rate Analysis of Intelligent Reflecting Surface Aided ...

by SN Sur · 2021 · Cited by 3 — Here in this paper authors have compared the performance of a IRS assisted. MU network considering optimized and non-optimized active/passive beamforming ...

Similarity 5%

Title:anushka23g/Parkinson-Disease-Classification

... dataset of various speech features (a non-invasive yet characteristic tool) from the University of Oxford. Why speech features? Speech is very predictive ...

Similarity 4%

Title:Parkinsons - dataset by uci - data.world

The dataset was created by Max 🔎Little ✔ of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals. The original study published the feature extraction methods for general voice disorders.

https://data.world/uci/parkinsons

Similarity 3%

Title:Early detection of Parkinson's disease using machine learning

WEB · RPDE Recurrence Period Density Entropy quantifies the extent to which signal is periodic DFA Detrended Fluctuation Analysis or DFA measures the extent of ...

https://www.sciencedirect.com/science/article/pii/S1877050923000078

Similarity 1%

Title:Incubation period, clinical and lung CT features for early ...

by H Peng · 2022 · Cited by 4 — ... accuracy, precision, recall, F1 score, and precision-recall curve (PRC). We wanted our risk model to avoid a missed diagnosis of ...

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9095818/

Similarity 1%

Title:A Machine Learning Approach for an Early Prediction of ...

... Analysis of extent or range of variations in speech with respect to MDVP: Fo(Hz). D2. Correlation dimension is used to identify dysphonia in speech using ...

https://indjst.org/download-article.php?Article_Unique_Id=INDJST13720