Journal of Artificial Intelligence and Capsule Networks (ISSN: 2582-2012)
www.irojournals.com/aicn/

Comparative Analysis of Machine Learning Algorithms for Early Prediction of Parkinson's Disorder based on Voice Features

C. D. Anisha[1], N. Arulanand[2]

[1]Research Scholar, CSE, PSG College of Technology, Coimbatore, Tamil Nadu, India
[2]Professor, CSE, PSG College of Technology, Coimbatore, Tamil Nadu, India

E-mail: [1]ani.c.dass@gmail.com, [2]naa.cse@psgtech.ac.in

Abstract

Parkinson Disorder (PD) is a neurological disorder which is progressive in nature and has no cure. Early diagnosis of PD plays a key role in delaying the progression of the disorder. Dysphonia is the most prominent early symptom which is exhibited by approximately 90% of PD patients. Voice features based early diagnosis with the integration of Artificial Intelligence plays a prominent role in providing accurate, non-invasive, and robust predictions to PD patients. This paper focuses on providing comparative and experimental analysis of Machine Learning (ML) algorithms for the prediction of PD based on the voice features dataset retrieved from the UCI repository. This paper presents the results from the four sampling experiments conducted with different traditional ML algorithms for the retrieved voice dataset. The results of this study make it evident that Naïve Bayes provides a highest accuracy of 89% when compared to other ML algorithms. This study helps in identifying the best ML algorithm among the traditional ML algorithms for PD prediction based on voice features dataset.

Keywords: Machine learning, Parkinson disorder, feature extraction, voice analysis, artificial intelligence

1. Introduction

Parkinson Disorder (PD) is the neurological disorder affecting the region of Substantia Nigra which is mainly associated with dopamine secretion and responsible for motor movements. The prominent symptoms which is vivid across of the stages of PD are

DOI: https://doi.org/10.36548/jaicn.2022.4.007
Received: 19.11.2022, received in revised form: 21.12.2022, accepted: 07.01.2023, published: 18.01.2023

slurred speech, Bradykinesia (slowness of movements), tremor (includes rest tremor), rigid muscles, sleep problems, reduced facial expression, etc.

The challenges regarding PD and its diagnosis are:

- PD does not have any cure.
- PD diagnosis is not specific.
- PD diagnosis based on imaging techniques are not affordable.
- PD diagnosis mostly does not cover the early symptoms such as slurred speech.

The solutions to the above challenges regarding PD are:

- Providing an early diagnosis for PD in order to delay the progression.
- Providing PD diagnosis based on specific symptoms.
- Providing non-invasive and affordable diagnosis by integrating Artificial Intelligence (AI) which aids in accurate and robust diagnosis in addition.
- Providing Voice Modalities based diagnosis system to cover the early symptom of PD.

The paper is organized as follows. Section 2 provides the related works associated with PD prediction based on ML algorithms for voice features modality and other modalities. Section 3 presents the research methodologies incorporated in this study, which mainly includes the main workflow and the techniques implemented. Section 4 illustrates the result analysis which focuses on providing the graphical and tabular representation of the results along with the insights and inferences procured from the results. Section 5 summarizes the conclusion which addresses the challenges in the present study, briefs the results of the highest performed ML algorithm, and elaborates the future prospects of the domain for further exploration.

2. Related works

2.1 ML algorithms for PD prediction based on Voice Modality

Sakar et al. [1], presented the early prediction based on voice modality by providing the real time acquisition data procured from healthy subjects and PD subjects integrated with novel feature extraction parameter namely tunable – Q factor wavelet transform and signal processing algorithms. Two hybrid Machine Learning models based on the voice features namely Principal Component Analysis integrated with Support Vector Machine (SVM) and Auto Sparse Encoder combined with SVM were proposed in paper [2]. Acoustic based

analysis to predict the PD severity was presented in paper [3], which is based on ML regression algorithms namely Support Vector Regression, Random Forest Regression and Multiple Linear Regression. Iqra Nissar et al. [4], explained a PD detection system which integrates voice features data, various ML techniques and feature selection techniques. The various ML techniques used in the system were K Nearest Neighbour, Decision Tree, Logistic Regression, Multi-Layer Perceptron, Naïve Bayes, SVM and Ensemble Classifier. An Ensemble ML model was proposed in paper [5] for PD prediction based on the voice measurements.

2.2 ML algorithms for PD prediction based on other Modalities

A Telemonitoring system was provided for PD in the paper [6], wherein ML was incorporated and integrated with voice and tremor data. In paper [7], a comparative analysis of ML algorithms namely Logistic Regression, SVM, and K-Nearest Neighbor was presented with the three prominent recording data namely static winding test, dependability test score and dynamic winding test.

3. Research Methodologies

Figure 1 presents the architecture of the proposed study of comparative analysis of ML algorithms with different sampling experiments.

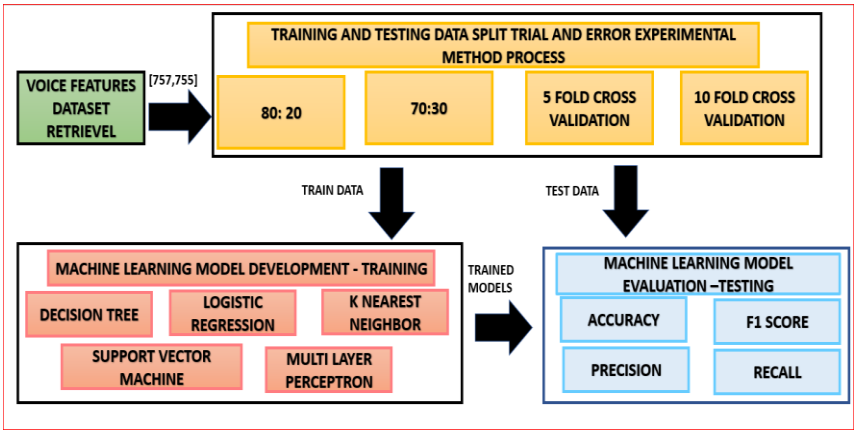

Figure 1. Architecture of the Proposed System – Comparative Analysis of ML Algorithms for PD Prediction

3.1 Dataset Description

The voice features dataset considered in this study is retrieved from the University of California Irvine (UCI) Repository. The dimension of the dataset is 757 instances and 755 features. The dataset consists of data from 64 healthy subjects and 188 PD subjects [1].

3.1.1 Training and Testing Set Split Process

The training and testing set split process is carried out as experiments in this study. There are four experiments of sampling performed which are as follows:

- Experiment 1- Standard Sampling [80: 20]: The training and testing set is split in the ratio 80 :20 in the standard sampling procedure. The total number of training samples with respect to the particular dataset considered after splitting are 605 samples and total number of testing samples with respect to the particular dataset considered after splitting are 151 samples.

- Experiment 2 – Standard Sampling [70:30]: The training and testing set is split in the ratio 70:30 in the standard sampling procedure. The total number of training samples with respect to the particular dataset considered after splitting are 529 samples and total number of testing samples with respect to the particular dataset considered after splitting are 227 samples.

- Experiment 3 – Equivalent Sampling [5-Fold Cross Validation]: The training and testing set is split as 5-fold cross validation, wherein all samples of the dataset are equally split into 5 folds and all samples in the dataset get an equal opportunity to act as training set sample as well as testing set sample in each fold. The total number of training samples with respect to the particular dataset considered after splitting are 605 samples in each fold and total number of testing samples with respect to the particular dataset considered after splitting are 151 samples in each fold.

- Experiment 4 – Equivalent Sampling [10 Fold Cross Validation]: The training and testing set is split into 10-fold cross validation, wherein all samples of the dataset are equally split into 10 folds and all samples in the dataset get an equal opportunity to act as training set sample as well as testing set sample in each fold. The total number of training samples with respect to the particular dataset considered after splitting are 529 samples in each fold and total number of testing samples with respect to the particular dataset considered after splitting are 227 samples in each fold.

3.1.2 Model Training using Machine Learning Algorithms

The implementation of ML model algorithm is performed with the help of the python framework executed in the Jupyter IDE. The scikit package of python framework is utilized for quick implementation of the following ML algorithms.

a) Decision Tree

Decision Tree is a supervised learning method and ML classification algorithm which classifies the instance by constructing tree from root node to the leaf node with the branches and parses though the tree based on the attributes or features and makes decision at each level of tree and final decision (classification or prediction) is made at the terminal node (leaf node). Leaf nodes holds the final decision of a particular tree. The scikit package DecisionTreeClassifier provides 12 parameters which aids in the implementation of decision tree.

b) Support Vector Machine

SVM is a supervised ML model which creates decision boundaries to classify the dimensions into classes. The hyperplane is known as the best decision boundary. Support Vectors are extreme points near the best decision boundary which helps in effective decision making. The scikit package SVC provides 15 parameters which aids in the implementation of SVM

c) Logistic Regression

Logistic Regression is a supervised ML model which has a predictive modelling similar to regression models but used for classification. The mathematical function namely sigmoid function is used to map the probabilities to the values 0 and 1, as the range for the classification of logistic regression is between 0 and 1 and it expands to 2 and 3 when it is a multiclass classification. The scikit package LogisticRegression provides 15 parameters which aids in the implementation of logistic regression.

d) Naïve Bayes

Naïve Bayes classifier is a supervised ML model which is a probabilistic based model. It is based on Bayesian Theorem. The basic assumption made by Naïve Bayes classifier is the "Independence of features". The scikit package GaussianNB provides 3 parameters which aids in the implementation of Naïve bayes.

e) K Nearest Neighbor Algorithm

K Nearest Neighbor (KNN) algorithm is a supervised ML model which classifies a new instance to a label based on the label of the nearest instances. The distance between new instance and the other instances is computed using Euclidean Distance. KNN comes under

the category of non-parametric and also known as lazy learner algorithm. The scikit package KNeighborsClassifier provides 8 parameters which aids in the implementation of KNN.

f) Artificial Neural Network

Artificial Neural Network (ANN) is a supervised ML model. It consists of input layer, hidden layer and output layer. Each layer consists of n number of neurons, and the input layer consists of input neurons which are fed to the neurons of the hidden layer along with the weights, to perform computation based on the activation function, and the values from the hidden layer are fed to the neurons of the output layer. The scikit package MLPClassifier provides 8 parameters which aids in the implementation of Multi-Layer Perceptron, a type of ANN.

4. Result Analysis and Discussion

The testing and evaluation of the model is performed using the evaluation metrics namely Accuracy, Precision, Recall and F1 Score. Figure 3-6 depict the confusion matrix of all classifiers in the four experiments. Figure 2 depicts the comparison accuracy graph.

Table 1. Comparison of all experiments and classifiers

Exp. No.	Decision Tree	KNN	Logistic Regression	Multi-layer Perceprton	Naïve Bayes	Support Vector Machine
1	81	66	71	72	70	70
2	81	68	71	33	69	72
3	77	82	82	84	87	82
4	82	82	84	84	89	84

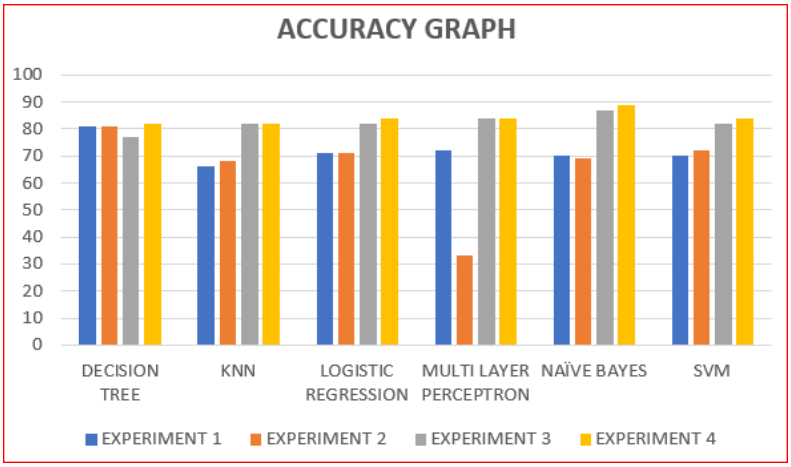

Figure 2. Accuracy Graph - Comparison of same classifier in same experiment

Figure 3. Confusion Matrices of Experiment 1

(Top Left Clockwise: Decision Tree, K Nearest Neighbor, Logistic Regression, Multi-Layer Perceptron, Naïve Bayes, Support Vector Machine)

Figure 4. Confusion Matrix of Experiment 2

(Top Left Clockwise: Decision Tree, K Nearest Neighbor, Logistic Regression, Multi-Layer Perceptron, Naïve Bayes, Support Vector Machine)

Figure 5. Confusion Matrix of Experiment 3

(Top Left Clockwise: Decision Tree, K Nearest Neighbor, Logistic Regression, Multi-Layer Perceptron, Naïve Bayes, Support Vector Machine)

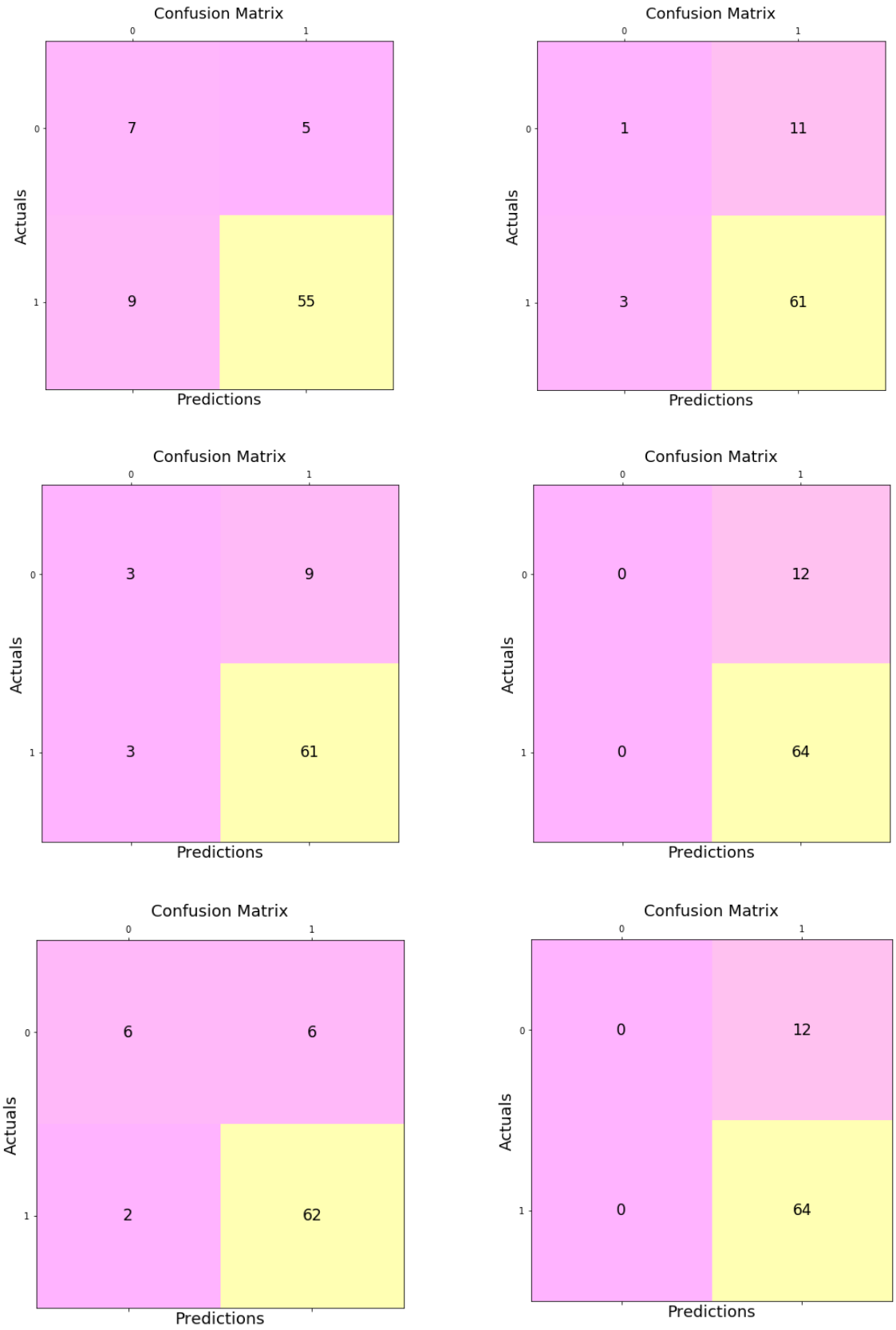

Figure 6. Confusion Matrix of Experiment 3

(Top Left Clockwise: Decision Tree, K Nearest Neighbor, Logistic Regression, Multi-Layer Perceptron, Naïve Bayes, Support Vector Machine)

4.1 Discussion

- Decision Tree provides a highest accuracy of 81% in Experiment 1 when compared to all other classifiers.
- Decision Tree provides a highest accuracy of 81% in Experiment 2 when compared to all other classifiers.
- From (a) and (b) points, it is clear that Decision Tree classifier outperforms other classifiers in Experiment 1 and Experiment 2 because Decision Tree classifier performs well with less data preparation and more training set data as the condition of Experiment 1 and Experiment 2 are standard sampling which is 80:20 and 70:30 respectively.
- Naïve Bayes classifier provides a highest accuracy of 87% in Experiment 3.
- Naïve Bayes classifier provides a highest accuracy of 89% in Experiment 4.
- From (d), (e) points, it is clear that Naïve Bayes classifier has the highest accuracy as this classifier performs well with less samples as in experiment 3, and experiment 4 focuses on equivalent sampling namely 5-fold cross validation and 10-fold cross validation.
- The overall highest accuracy of 89% is obtained by Naïve Bayes classifier compared to other classifiers in the Experiment 4 with 10-fold cross validation data sampling technique.

5. Conclusion

Parkinson Disorder (PD) is the second highest prevailing neurodegenerative disorder after Alzheimer Disorder. The early voice symptom can be used as a base to develop early diagnosis system of PD by integrating the voice modality with Artificial Intelligence (AI) system for accurate and robust diagnosis. The proposed study performs the comparative analysis of Machine Learning (ML) models, a type of AI models on the retrieved benchmark voice features dataset. The proposed study is performed under four sampling experiments in order to overcome the imbalance problem present in the dataset. The results of the proposed system make it evident that the ML models perform at a higher rate in experiment 4 i.e., equivalent sampling category specifying to 10-fold cross validation. The highest accuracy of

89% is procured from Naïve Bayes Classifier in experiment 4. The future work focuses on the development of optimized hybrid-based ensemble classifier for PD diagnosis based on the voice features modality.

References

[1] Sakar, C.O., Serbes, G., Gunduz, A., Tunc, H.C., Nizam, H., Sakar, B.E., Tutuncu, M., Aydin, T., Isenkul, M.E. and Apaydin, H., 2018. A comparative analysis of speech signal processing algorithms for Parkinson disease classification and the use of the tunable Q-factor wavelet transform. Applied Soft Computing. Hoq M, Uddin MN, Park SB.

[2] Vocal Feature Extraction-Based Artificial Intelligent Model for Parkinson's Disease Detection. Diagnostics (Basel). 2021 Jun 11;11(6):1076. doi: 10.3390/diagnostics11061076. PMID: 34208330; PMCID: PMC8231105.

[3] Hemmerling D, Wojcik-Pedziwiatr M. Prediction and Estimation of Parkinson's Disease Severity Based on Voice Signal. J Voice. 2020 Aug 14:S0892-1997(20)30231-9. doi: 10.1016/j.jvoice.2020.06.004. Epub ahead of print. PMID: 32807590

[4] Iqra Nissar et al, Voice-Based Detection of Parkinson’s Disease through Ensemble Machine Learning Approach: A Performance Study, EAI Endorsed Transactions on Pervasive Health and Technology 05 2019 - 08 2019 | Volume 5 | Issue 19 | e2.

[5] Sheibani R, Nikookar E, Alavi SE. An ensemble method for diagnosis of Parkinson’s disease based on voice measurements. J Med Sign Sens 2019; 9:221-6.

[6] Sajal, M.S.R., Ehsan, M.T., Vaidyanathan, R. et al. Telemonitoring Parkinson’s disease using machine learning by combining tremor and voice analysis. Brain Inf. 7, 12 (2020). https://doi.org/10.1186/s40708-020-00113-1

[7] F.M. Javed Mehedi Shamrat et al, A Comparative Analysis of Parkinson Disease Prediction Using Machine Learning Approaches, International Journal of Scientific & Technology Research Volume 8, Issue 11, November 2019 ISSN 2277-8616