PARKINSON'S DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

By

Purna Chakraborty
Student ID: 193000512

Department of Computer Science & Engineering
School of Science, Engineering & Technology

Supervised by
Joydwip Mohajon
Lecturer
East Delta University

In partial fulfillment of the requirement for the degree of

Bachelor of Science in *Computer Science and Engineering*

December, 2023

East Delta University
Noman Society, East Nasirabad, Chittagong-4209

PARKINSON'S DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

This thesis is submitted in partial fulfillment of the requirement for the degree of

Bachelor of Science in Computer Science & Engineering.

By

Purna Chakraborty
Student ID: 193000512

Supervised by
Joydwip Mohajon
Lecturer
East Delta University

Department of Computer Science & Engineering

School of Science, Engineering & Technology

East Delta University, Noman Society, East Nasirabad, Chittagong-4209

The undergraduate thesis titled "**Parkinson's Disease Prediction Using Machine Learning Algorithms''** submitted by Purna Chakraborty, Student ID: 193000512 has been accepted as satisfactory in fulfilment of the requirement for the degree of Bachelor of Science in Computer Science and Engineering to be awarded by East Delta University.

Board of Examiners

Md. Ishtiaque Aziz Zahed, PhD — **Chairman**
Associate Professor
School of Science, Engineering and Technology
East Delta University

Arifa Sultana — **Member**
Assistant Professor
School of Science, Engineering and Technology
East Delta University

Tanvir Azhar — **Member**
Lecturer
School of Science, Engineering and Technology
East Delta University

Joydwip Mohajon — **Supervisor**
Lecturer
School of Science, Engineering and Technology
East Delta University

Declaration

It is hereby declared that the work contained in this thesis is original. The information derived from the literature or work has been duly acknowledged and presented in the reference section. No part of this thesis has been submitted elsewhere for the degree, diploma or other similar title of recognition.

Date:

Purna Chakraborty
Student ID: 193000512

ACKNOWLEDGMENTS

I wish to acknowledge the effort of my supervisor for his guidance and support in conducting the research and preparation of the report. I am so thankful to my parents, and friends for their support. Finally, I am very thankful to East Delta University for allowing me to complete my degree of Bachelor of Science in Computer Science and Engineering.

TABLE OF CONTENTS

‘

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

WHO	: World Health Organization
GNB	: Gaussian Naive Bayes
SVM	: Support Vector Machine
K-NN	: K Nearest Neighbour
LR	: Logistic Regression
RF	: Random Forest
DT	: Decision Tree
PD	: Parkinson’s Disease
ML	: Machine Learning
SMOTE	: Synthetic Minority Oversampling Technique
TP	: True Positive
FP	: False Positive
FN	: False Negative
TN	: True Negative

ABSTRACT

Parkinson’s disease or PD is an age-related degenerative brain condition, meaning it causes parts of your brain to deteriorate. It’s best known for causing slowed movements, tremors, balance problems and more. Most cases happen for unknown reasons, but some are inherited. In simpler words, Parkinson's disease is a condition where a part of your brain deteriorates, causing more severe symptoms over time. While this condition is best known for how it affects muscle control, balance and movement, it can also cause a wide range of other effects on your senses, thinking ability, mental health and more. While Parkinson's disease is usually age-related, it can happen in adults as young as 20. A study titled "Global Burden of Disease" in 2015, roughly estimated the reoccurrence of PD to be around 6.2 million in the world. The modern era allowed us to combine medical research and the progressive era of computer science to solve or somewhat attain a goal towards the ultimate solution to diagnose silent or invisible diseases and disorders similar to PD. A database published by the UN's World Health Organization (WHO) in 2020 shows that there have been around 4000 deaths due to PD; the death rate being 3.49 per 100,000 of the population. This journal shows the detection of PD using machine learning algorithms - Support Vector Machine (SVM), Logistic Regression (LR), K- Nearest Neighbour (K-NN), Gaussian Naive Bayes (GNB), Random Forest (RF) and finding out the best amongst the similar algorithms. The prime objective of this journal is to create, test and train a machine that will produce a result of the prediction of PD in Bangladesh and will be judged with other similar journals in the same area in a comparative way.

Keywords— Parkinson’s disease, Alzheimer’s Dementia, Machine Learning.

CHAPTER 1

INTRODUCTION

According to Dr. Kristine Domingo, Neurologist at M Health Fairview, Parkinson's disease or PD in short is "A long-term and progressive disorder that affects the dopamine-producing cells in the brain, leading to the motor and non-motor symptoms experienced." In simpler words, PD is a neurological problem that usually in most cases is seen in older people; this commonness is seen in patients who are fifty years old or older. However, there are certain special cases of young patients who are also diagnosed with PD. In cases of monogenic PD, three to five percent in most populations have shown a link to having PD genes. On the other hand, sixteen to thirty six percent of the heritable risk of non-monogenic PD is explained through ninety risk variants [1]. This section is composed of the motivation of the research as well as the research challenges and objectives of the research. It also includes a detailed discussion about the Machine learning algorithms used in this research.

1.1 Motivation of the research

As people's life expectancy in Bangladesh rises, so does the number of persons suffering from Parkinson's disease. According to a 2006 assessment by de Lau and Breteler, an estimated 10 million individuals worldwide and 1% of those over 60 are afflicted by Parkinson's disease, making it the second most common neurological illness after Alzheimer's. From September 2014 to March 2016, 76 early Parkinson's disease patients

were included in this cross-sectional observational study at the Department of Neurology at Bangabandhu Sheikh Mujib Medical University in Dhaka, Bangladesh; where the percentage of males and females was 69.7% and 30.3% respectively and an age range from twenty four to seventy seven years [2].

1.2 Research Challenges

People with Parkinson's disease may also differ greatly in terms of their symptoms, the severity of those symptoms, disease progression, responsiveness to treatment, and risk of consequences. Researchers have significant difficulty in better understanding what causes the intricacy and diversity of Parkinson's disease.

Doctors and clinicians discover Parkinson's disease by identifying symptoms. They may administer clinometric tests to measure movement-related and non-movement-related symptoms. These tests, however, are solely dependent on the doctor's experience and analysis. These tests can only imply that the applicant has Parkinson's disease but cannot prove it. Manual assessment of Parkinson's disease involves more time, money, and skill, which might be difficult for the patient.

Only upper limb limitations are detected by FTT and handwriting sketching. As a result, upper limb disability alone cannot be utilized to detect Parkinson's disease. Also, the procedure for creating a PET or SPECT picture is intrusive, as radioactive traces are implanted in the patient's body before the test. It is harmful to the patient's body. It is also not appropriate for expectant moms.

Because Parkinson's disease affects 1% of adults over the age of 60, the available datasets are relatively tiny. Because of the tiny dataset, judging the accuracy of a technique is challenging. It frequently leads to overfitting[3].

1.3 Research Objectives

The objective of this research is to create, test and train a model using machine learning. Another objective of this research is to develop a model with a proper dataset to achieve better accuracy.

1.4 Thesis Organization

This research paper contains five chapters. The first chapter being "Introduction" that gives an overview of the topic. The second chapter is offered as a review of the literature regarding this research and is titled "Literature Review". The following third chapter titled as "Methodology", provides a detailed explanation of the techniques used. Then the fourth chapter is "Result and Discussion", compiled with this research findings. Lastly, the fifth chapter "Conclusion", provides a brief overview of this research work.

CHAPTER 2

LITERATURE REVIEW

In early studies of Parkinson's disease detection, in paper [12] authors applied ensemble learning techniques to display the greatest result compared to applied ml models that are SVM, KNN, RF, DT, MLP, SC(Stacking Classifiers), LR and the methodology of this paper involved data acquisition and preprocessing by utilizing VCM(Vote Count Model) for classification with a detailed architecture. They found 94.87% accuracy, 81.99% MCC and 94.87% f1-score after the experiment compared to other methods of comparative tools [12]. Another journal by Aditi Govinda and Sushila Palwe [11] represents an ensemble model where the model is trained to classify given audio data based on frequency variations. The model works with four algorithms– Logistic regression (LR), Support Vector Machine (SVM), Random Forest (RF), and K- Nearest Neighbor (K-NN) with the same 25% of the dataset. The accuracy table of the mentioned algorithms are LR with 83.67% accuracy, SVM with 85.71% accuracy, RF with 91.83%, K-NN with 85.71%. This research showed RF gave the best accuracy compare to other models. A further review journal by Arti Roma *et al.* [13] shows the existing ML-based research to diagnose PD in terms of handwritten patterns, voice attributes, and gait datasets and to determine the most appropriate techniques to diagnose PD with an accuracy rate. The best accuracy was for the voice feature which was done by L-1 Norm Sum with K-fold cross-validation with 99%, handwritten pattern with bagging ensemble with an accuracy of 97.96% and gait analysis with SVM algorithm with an accuracy of 100%. An additional study by Jie Mei *et*

al. [14] reviews the results from all studies that applied machine learning methods with the associated outcomes, types of clinical, behavioral and biometric data used for rendering more accurate diagnoses, potential biomarkers for assisting clinical decisions and other highly relevant information with proper database divided 448 machine learning models from the 209 studies into 8 categories: (1) support vector machine (SVM) and variants (n = 132 from 130 studies), (2) neural networks (n = 76 from 62 studies), (3) ensemble learning (n = 82 from 57 studies), (4) nearest neighbor and variants (n = 33 from 33 studies), (5) regression (n = 31 from 31 studies), (6) decision tree (n = 28 from 27 studies), (7) naïve Bayes (n = 26, from 26 studies), and (8) discriminant analysis (n = 12 from 12 studies). A small percentage of models used did not fall into any of the categories (n = 28, used in 24 studies). Out of 209 studies, 122 (58.4%) applied machine learning methods to movement-related data, i.e., voice recordings (n=55, 26.3%), movement data (n = 51, 24.4%), or handwritten patterns (n = 16, 7.7%). Imaging modalities analyzed include MRI (n = 36, 17.2%), SPECT (n = 14, 6.7%), and positron emission tomography (PET; n = 4, 1.9%). Five studies analyzed CSF samples (2.4%). In 18 studies (8.6%), a combination of different types of data was used. Ten studies (4.8%) used data that did not belong to any categories mentioned above. The result of other study by Ilkka Suuronen *et al.* [15] shows that the search algorithm indicated that with EO data, the best classification performance was obtained with only five channels. With the five-channel budget, Channels that were placed far away from each other on the scalp were selected into the classifier, possibly because this enabled sampling from somewhat independent EEG activity sources. Also, the results suggest that a small subset of electrodes of an EEG recording can suffice for detecting PD with a classification performance on par with a full set of electrodes.

Furthermore, our results demonstrate that separately collected EEG datasets can be used for pooled machine learning-based PD detection with reasonable classification performance.

CHAPTER 3

RESEARCH METHODOLOGY

In this chapter, the following sections describe various models that have been used in the PD dataset. Section 3.1 contains a figure of proposed methodology. Section 3.2 describes the data collection of this study. Section 3.3 carried out the visualization of the dataset by doing analysis and preprocessing of the dataset. The applied ML models in this study are described in section 3.4 Machine Learning Models Implementation. Section 3.5 shows how models are trained and evaluated with proper figures. Section 3.6 describes additional analysis of this study.

3.1 Proposed methodology

Figure 3.1 defines a flowchart of the proposed methodology.

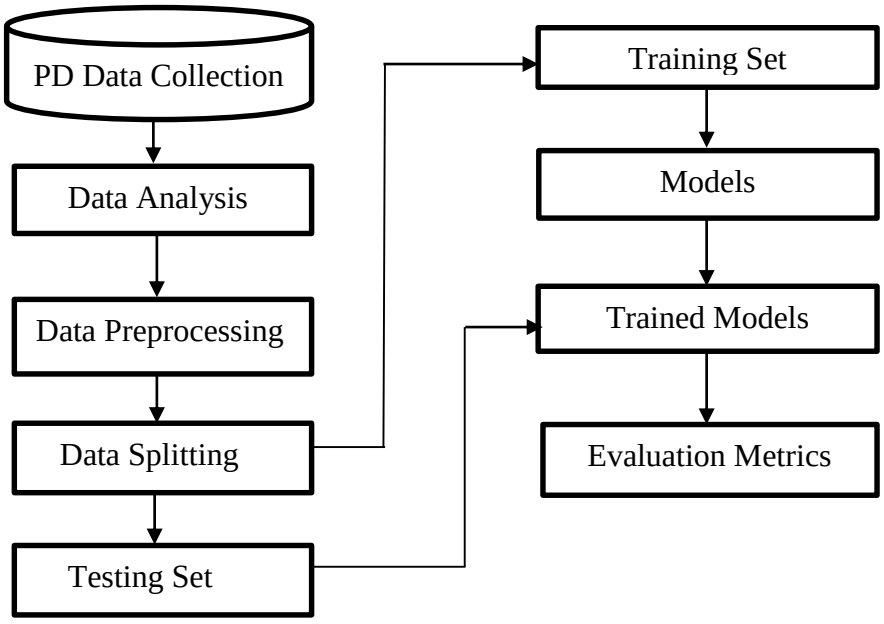

Figure 3.1: Proposed Methodology Flowchart

3.2 Dataset collection

The dataset used in the study is sourced from the reliable UCI Machine Learning Repository Little, Max. (2008)[10]. Parkinson's, which contains relevant features for binary classification. The dataset includes information about jitter, shimmer MDVP etc. of vocal background is given in the Table 3.1[11]. The data is pre-processed and visualized for a deep understanding of the functions.

The following table represents the used features or attributes and their descriptions.

Table 3.1: Used Features Description of UCI dataset [11]

Features	Descriptions
Name	The patient's name and recording numbers are sorted in ASCII CSV[11]
MDVP:Fo(HZ)	Fundamental frequency of pitch record[11]
MDVP:Fhi(Hz)	Maximum threshold of voice modulation[11]
MDVP:Flo(Hz)	Minimum threshold of voice modulation[11]
MDVP: Jitter, Abs, RAP, PPQ, DDP	These are the measurements of various multi-dimensional voice programs which is a traditional measure of the frequency of vibrations in vocal folds at the pitch period to vibrations at the start of the next cycle called pitch mark. [10][11]
Jitter & Shimmer	Measurements of absolute difference within frequencies of each cycle, after normalizing the average[11]
NHR & HNR	Measures signal-to-noise and tonal ratio that signifies robustness of surroundings to noise[11]
Status	1 means patients with PD while 0 means patients without PD[11]
D2	To find dysphonia in speech using fractal objects, correlation dimension is used which is a nonlinear dynamic attribute. [11]
RPDE	Recurrence Period Density Entropy quantifies the extent to which the signal is periodic. [11]
DFA	DFA measures the extent of stochastic self-similarity of noise in speech signals. [11]
PPE	Pitch Period Entropy is used to estimate the abnormal variations in speech on a scale of logarithms. [11]
Spread1, spread2	Analysis of extent in speech concerning MDVP: Fo(Hz) [11]

3.3 Data Distribution

The dataset consists of multivariate characteristics and real feature types which have a total of 197 instances. This dataset consists of multiple biomedical sound measurements of 31 individuals, 23 of whom suffer from PD. In this dataset table, each column is a specific sound measure, and each row corresponds to one of the 195 sound recordings of those individuals (the "name" column). The purpose of the data is to distinguish healthy people from patients with PD following the "status" column where 0 is for Healthy and 1 for PD [10]. To gain insights into the distribution and characteristics of this dataset, the EDA (exploratory data analysis) was performed. To acknowledge the patterns and the interdependency within the features, visualization and statistical analysis were applied. The basic statistical measures are computed to acknowledge the central tendency and dispersion, visualizing target variable distribution and individual sound measurements to reveal patterns and potential variance between patients without PD and patients with PD individuals. Figure 3.2 represents a graphical view of class distribution before balancing the imbalance class.

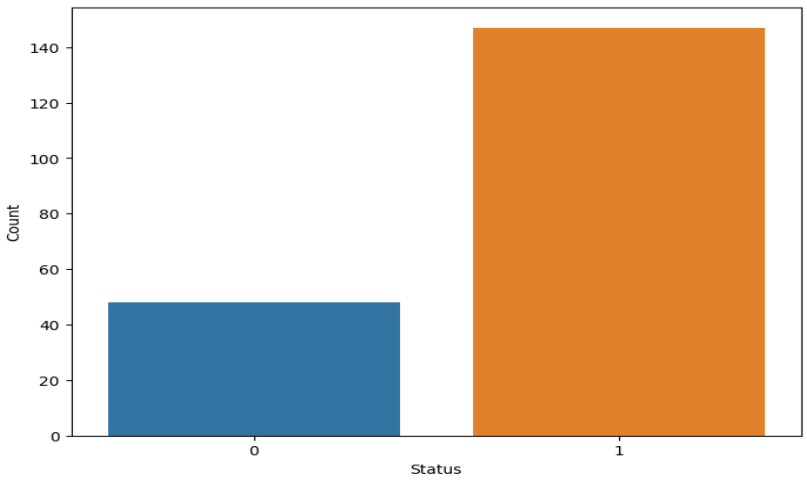

Figure 3.2: Class Distribution

3.4 Data Analysis

Figure 3.3 represents a boxplot graph of some features around a total of 22 features of the dataset. These boxplots are used to analyze the outliers of the features due to the noise of the recorded speech.

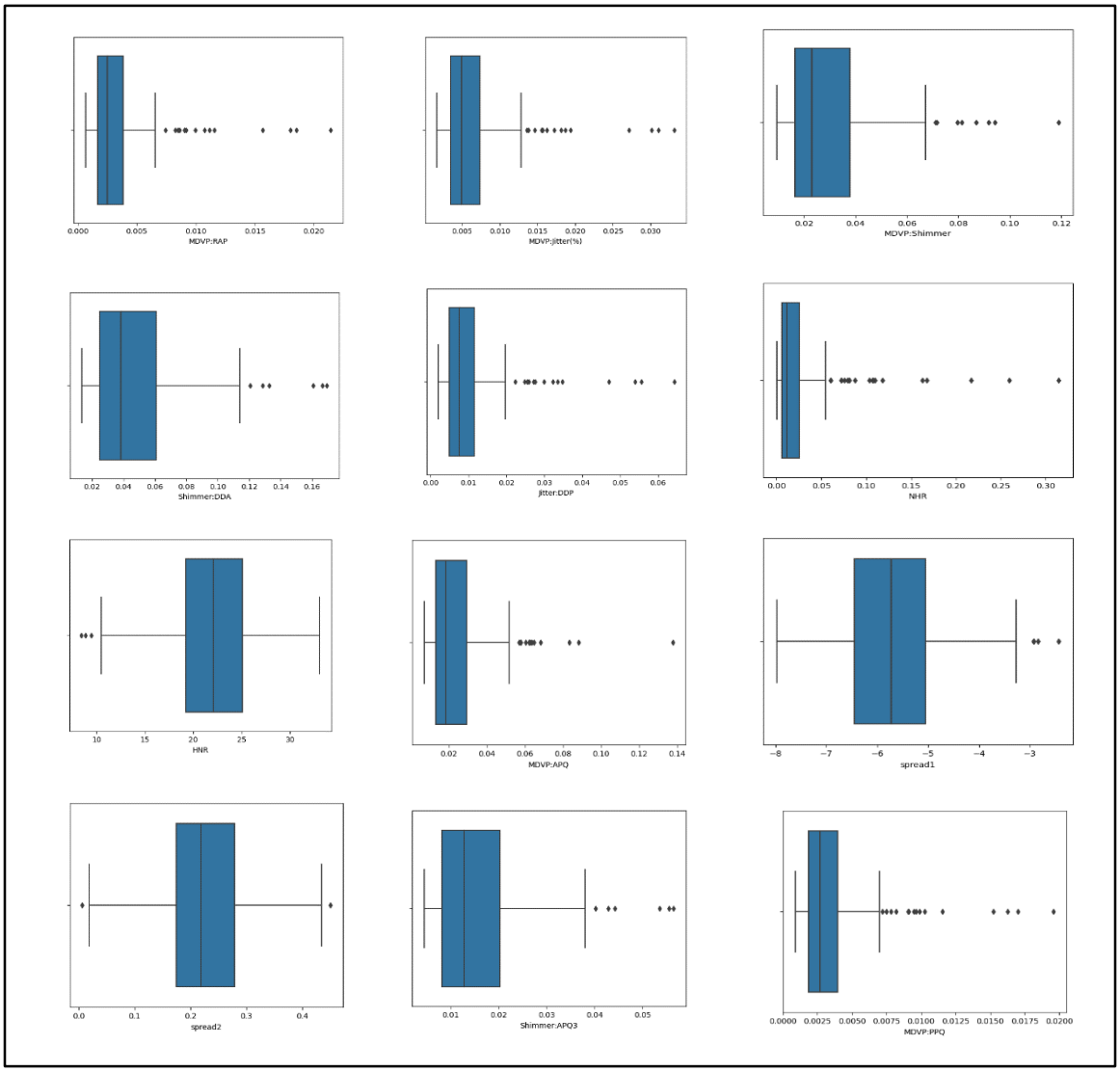

Figure 3.3: Boxplot of 22 features.

Figure 3.4 illustrates a pair plot by using a scatterplot graphical figure to check the correlation between all of the features gradually which is an essential step of feature

extraction. It shows most of the features are lowly correlated. It applies to prevent the model from being biased and avoid redundancy.

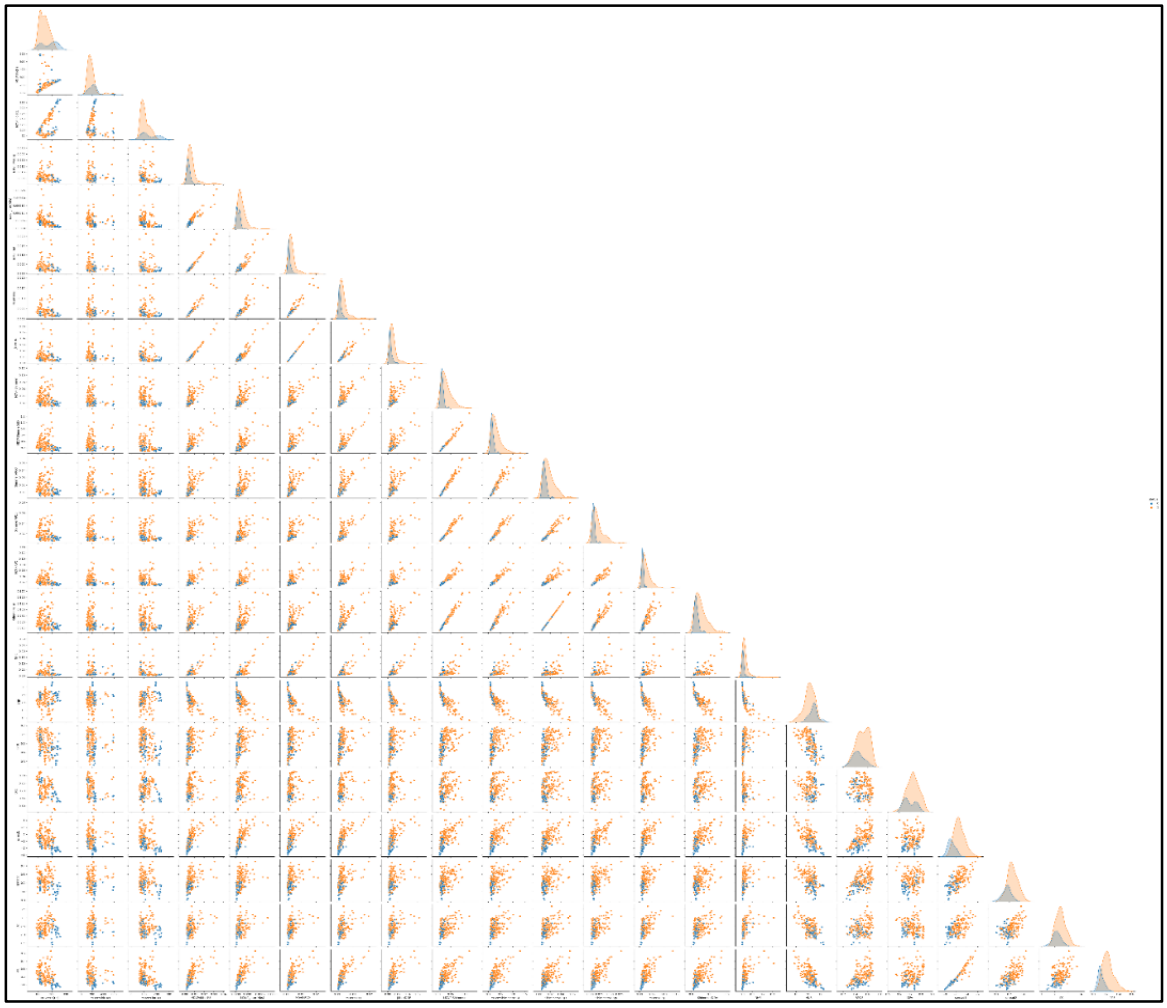

Figure 3.4: Scatterplot Graphical View.

3.5 Data Preprocessing

In this section, features and target variable were separated to use the feature for training and use the target variable for prediction. This step make the model more organized. Then SMOTE-TOMEK was applied for oversampling for handling the imbalanced dataset of the 'status' column where zero (0) means healthy/patients without PD and one (1) means

patients with PD that are shown in Figure 3.5 in below. Also scaling the features through standardization by using standard scaler on train data.

Figure 3.5 represents a graphical view of class distribution after applying SMOTE-Tomek which balancing the imbalance class of the dataset.

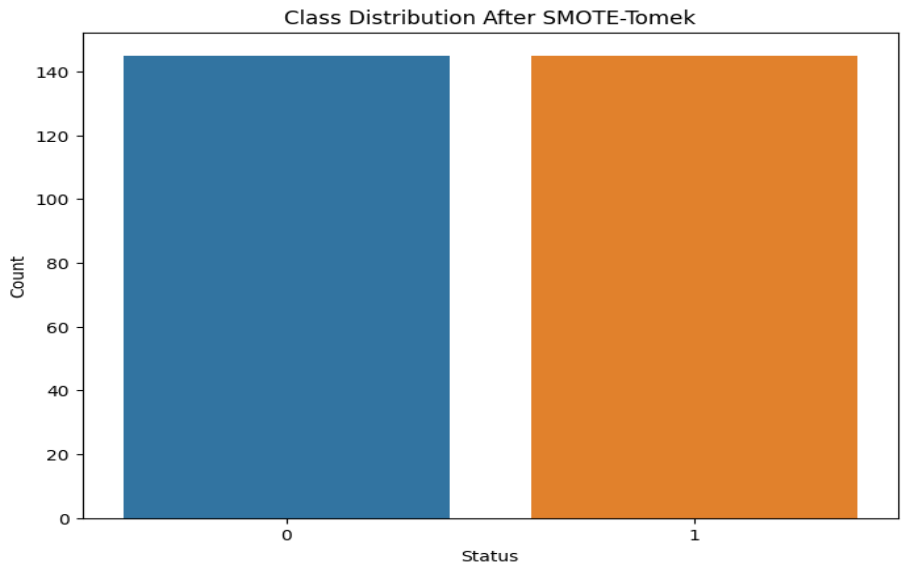

Figure 3.5: Class Distribution After SMOTE-Tomek

3.6 Machine Learning Models Implementation

For this study, the implementation of multiple models – Support Vector Machine, Logistic Regression, Random Forest Classifier, K-Nearest Neighbors, and Gaussian Naïve Bayes to advantageously/strongly distinguish between patients without PD and patients with PD. Before the application of the models, the dataset was in 2 sets which are training data and testing data. A total of 80% of the dataset was trained by the following models. The workflow of the models:

3.6.1 SVM: A Support Vector Machine (SVM) is a machine learning algorithm based on forced minimization problems. As one of the ML models of classifications, SVM is used in many medical studies including the analysis of PD. The goal is to find the maximum separation distance between support vectors and objects by calculating dot products and to find the largest separation distance between classes [4][5]. The idea is to transform a non-linear differentiable dataset into a better dimensional space where we can find a hyperplane that isolates objects. Like that in this work SVM is used to find an optimal hyperplane which maximally separates different classes in a high-dimensional feature space. By using the linear kernel this model is trained on the oversampled data.

3.6.2 LR: The goal of Logistic Regression (LR) is to predict the likelihood of an event occurring in a given set of data, with the result being either a yes or a no. Logistic regression can easily be extended to include more than one independent (predatory) variable. This allows researchers to investigate the relationship of each independent variable to the binary (dual) outcome while keeping the values of other independent variables constant [6]. This model is a simple but effective algorithm for binary classification tasks that is also trained on the oversampled training data.

3.6.3 RF: Random Forest (RF) is a flexible and easy-to-use machine learning algorithm that produces great results without hyperparameter tuning. It is also one of the most used algorithms due to its simplicity and versatility. In recent years, machine learning techniques have been considered an effective method for the accurate diagnosis of various diseases, among which Random Forest (RF) has been widely used [9]. In various ensemble learning methods one of the best methods like RF classifier which builds 10 multiple decision trees and consolidates the trees together on the oversampled training data in this work.

3.6.4 KNN: The K Nearest Neighbor(KNN) algorithm, which is used for classification and regression, belongs to the Supervised Learning category. It's a versatile calculation likewise utilized for crediting missing qualities and checking datasets. K Nearest Neighbor, as the name suggests, uses K nearest neighbors—also known as data points—to predict the new data point's class or continuous value [7]. For a simple yet effective algorithm, KNN is based on the distance between data points of the dataset that classifies with 3 nearest neighbors on following sampled training data used in this work.

3.6.5 GNB: Naive Bayes is an algorithm of convenient learning that makes use of the rule of Bayes together with an excessive premise or assumption whose traits rely on the independence supplied through the class. Though Naïve Bayes brings the accuracy of competitive classification time and again, whilst assuming this independently is often invaded in practice [8]. For all predictions, Gaussian Naïve Bayes uses the entire attributes. A probabilistic algorithm based on Bayes' theorem is applied to the following oversampled training data to emphasize the specific considerations like assumptions of feature independence in this binary classification task of PD.

3.7 Model Training

This section is crucial in the process of ML which provides several benefits that support the creation and application of successful models. Model training ensure accurateness, robustness and generalizable models by learning through the meaningful patterns, optimization of patterns, determining the performance and conducting the decision-making process. On the preprocessed dataset, each model was trained after the oversampling using SMOTE-Tomek to confront class imbalance. Splitting the dataset into training and testing sets. 80% of the dataset is used for training and 20% of the dataset is used for testing set.

CHAPTER 4

RESULT AND DISCUSSION

4.1 Evaluation

The accuracy of each model was estimated on both the training and testing part of the datasets. Additionally, the classification reports such as precision, recall, and F1 score are calculating a comprehensive evaluation to find out the best model with the best accuracy. This study strategy compares the results of 5 trained models. To differentiate, the use of 5-fold cross-validation of the best 3 models is also an essential task.

4.1.1 Classification Report

This segment covers the mathematical formulas of confusion matrix of classification reports. These are formulated in equation (4.1-4.4):

$$Precision = \frac{TP}{TP + FP} \tag{4.1}$$

$$F1score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4.2}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.3}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{4.4}$$

In this formula section,

- **TP** means true positive rate.
- **TN** convey true negative rate.
- **FP** signify false positive rate.
- **FN** signify false negative rate.
- **Precision** measures the ratio of correctly predicted positive data instances among all instances predicted as positive, denoting the model's ability to minimize the incorrect positive predictions.
- **F1-score** provides a balance between the harmonic mean of both precision and recall to assess the model's overall outcomes on positive class prediction.
- **Recall** calculates the ratio of correctly predicted positive instances among all actual positive instances, indicating to determine all positive instances.
- **Accuracy** gives an overall evaluation of the model's reliability rate in prediction by measuring the ratio of accurately classified data instances.

The classification report tables of each model are given below:

Table 4.1.1 shows, the classification report of the Support Vector Machine (SVM) model with the following metrix- precision, recall, f1-score and accuracy.

Table 4.1.1: Accuracy of SVM

Status	Precision	Recall	F1- Score	Accuracy
0	0.87	0.97	0.92	-
1	0.97	0.85	0.91	-
-	-	-	-	91%

The table below, Table 4.1.2 displays, the classification report of the Logistic Regression (LR) model with the following metrics- precision, recall, f1-score and accuracy.

Table 4.1.2: Accuracy of LR

Status	Precision	Recall	F1- Score	Accuracy
0	0.84	0.84	0.84	-
1	0.84	0.84	0.84	-
-	-	-	-	84%

Table 4.1.3 displays also, a classification report of the Random Forest (RF) model with the following metrix- precision, recall, f1-score and accuracy.

Table 4.1.3: Accuracy of RF

Status	Precision	Recall	F1- Score	Accuracy
0	1.00	0.99	1.00	-
1	0.99	1.00	1.00	-
-	-	-	-	100%

Moreover, Table 4.1.4 shows a classification report of the K Nearest Neighbors (KNN) model with the following metrix- precision, recall, f1-score and accuracy.

Table 4.1.4: Accuracy of K-NN

Status	Precision	Recall	F1- Score	Accuracy
0	0.96	0.99	0.97	-
1	0.99	0.96	0.97	-
-	-	-	-	97%

Below, Table 4.1.5 also displays a classification report of the Gaussian Naïve Bayes (GNB) model with the following metrix- precision, recall, f1-score and accuracy.

Table 4.1.5: Accuracy of GNB

Status	Precision	Recall	F1- Score	Accuracy
0	0.73	0.95	0.83	-
1	0.93	0.66	0.77	-
-	-	-	-	80%

4.1.2 Hyperparameter Tuning

Hyperparameters are used to specify the learning capacity and complexity of the model during training process. These are also used for performance optimization in this model. Table 4.1.6 shows the hyperparameters used in this model.

Table 4.1.6: Hyperparameters of the model

Sl. No.	Algorithm	Hyperparameter	Value
01	SVM	kernel	linear
02	LR	default by scikit-learn	-
03	RF	criterion n_estimators	entropy 10
04	KNN	n_neighbors	3
05	GNB	default by scikit-learn	-

4.1.3 Confusion Matrix

The given Figure 4.1.1 below presents the confusion metrics of KNN where the true negative rate is 115 which indicates the model predicts correctly the patients without PD/healthy, false positive rate is 1 which means the model predicts incorrectly to detect the patients with PD, false negative rate is 5 that indicates the model predicts incorrectly the patients without PD and true positive rate is 111 that indicates the model predicts correctly the patients with PD.

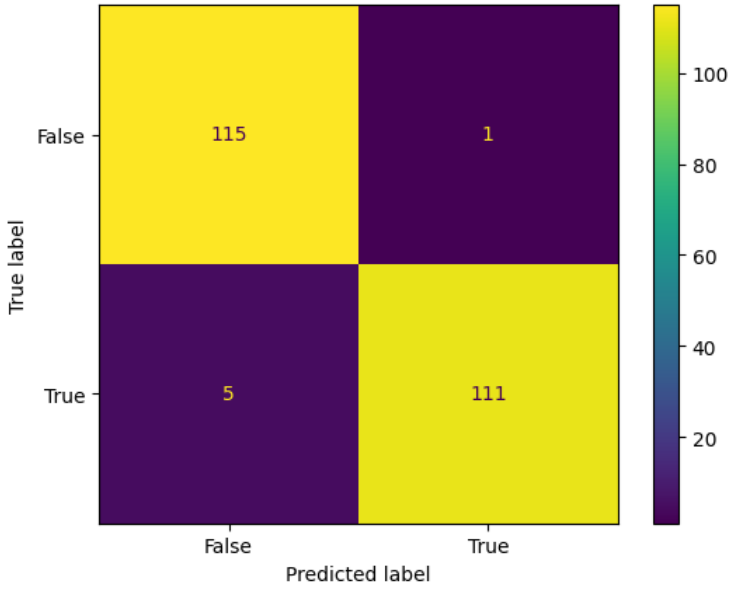

Figure 4.1.1: Confusion matrix of KNN

4.1.4 5-fold Cross Validation

For comparison of the implemented models apply 5-fold cross-validation on 3 models that are having highest accuracy within the 5 models. In Figure 4.1.2, the bar chart shows a graphical view of the 5-fold cross-validation of the Random Forest classifier. In Figure 4.1.3, the bar chart illustrates the view of the 5-fold cross-validation of the K-nearest neighbour classifier. Figure 4.1.4 also displays the view of cross-validation of the Logistic Regression classifier in a graphical state.

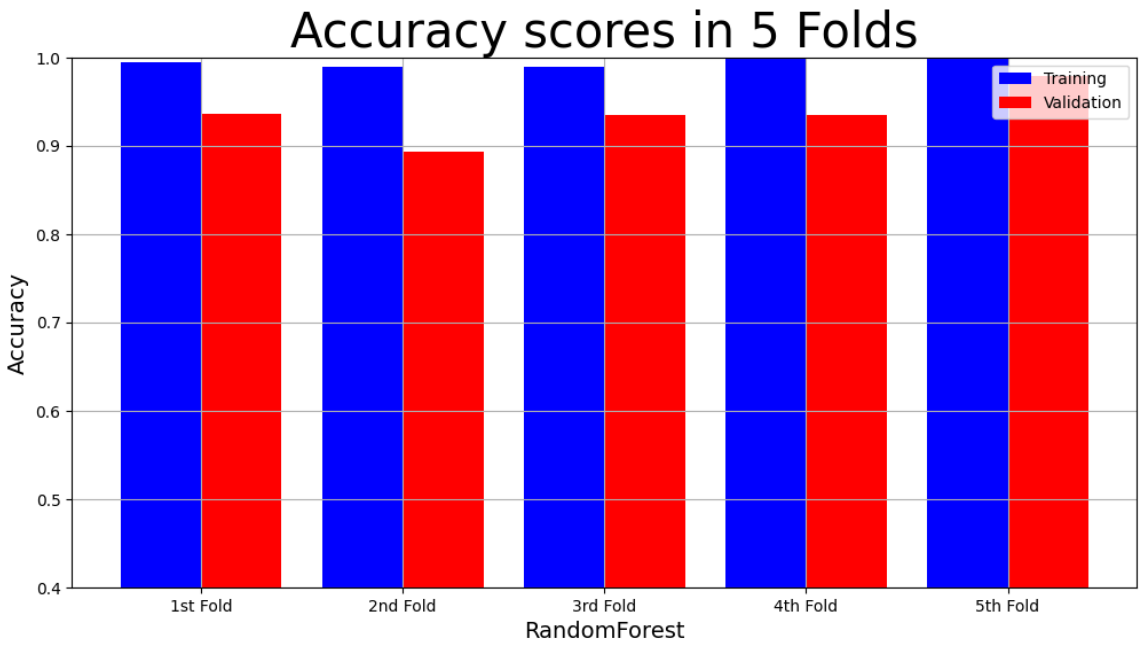

Figure 4.1.2: 5-fold Cross-validation of RF

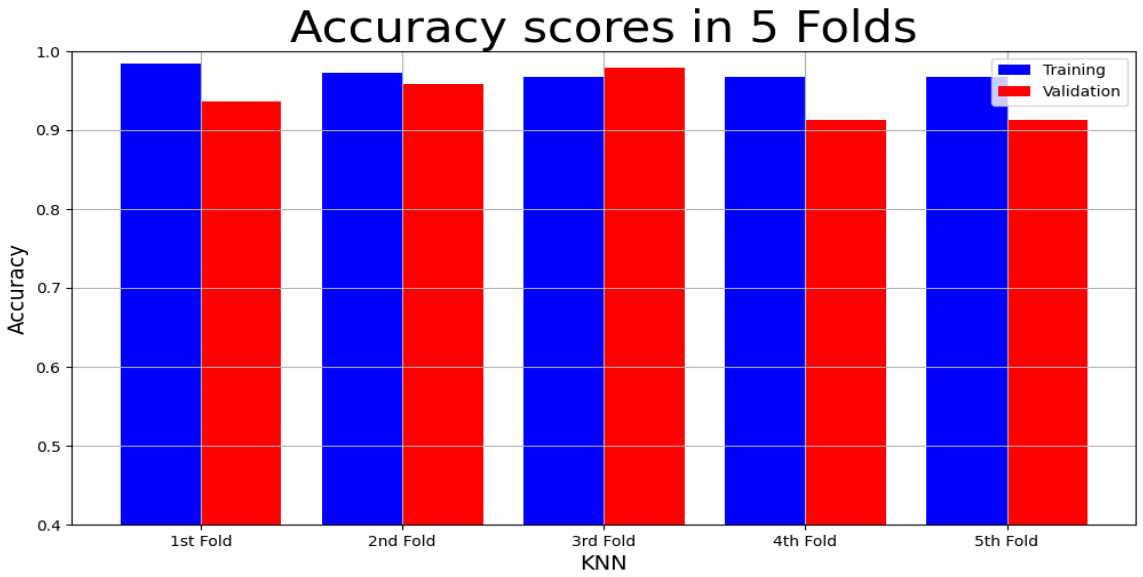

Figure 4.1.3: 5-fold Cross-validation of KNN

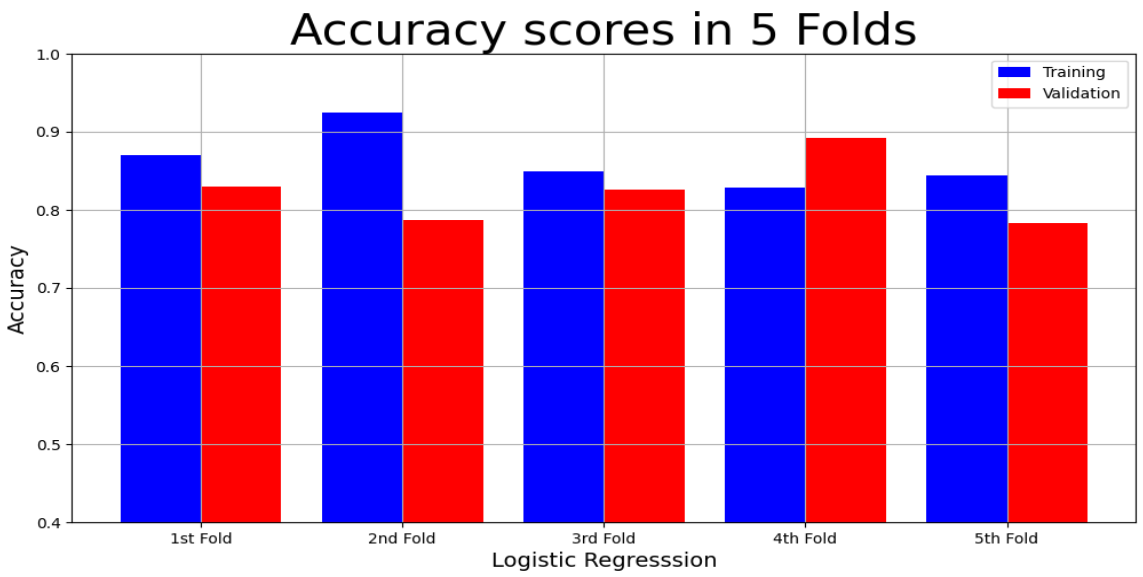

Figure 4.1.4: 5-fold Cross-validation of LR

Overall, these graphical representations of cross-validation results recommend that the KNN model exhibit stable and higher accuracy across the folds than other models by making them strong candidates for this study strategy.

4.2 Result

This section may vary as an overview of the comparison of the implemented models with the outcomes of those models. By using the 22 features of voice data of 195 samples of the PD dataset, the KNN classifier gives 97% accuracy with a 0.96 recall value on patients with PD. On the other hand, the SVM classifier also gives a better accuracy result which is 91% with 0.85 recall value. There is another model LR classifier also gave a good accuracy which is 84% whereas the GNB classifier has an 80% accuracy result. The best model of this study is KNN as easy to acknowledge and a simple algorithm than the other 3 models SVM, LR and GNB. SVM is also a simple algorithm but it's less interpretable than LR.

4.3 Result Comparison

This sub-section represents a comparative analysis view between similar ideas with different approaches and the proposed model in a tabular form. Further discussion about this table is discussed in the 4.4 discussion section.

Table 4.3: Comparison Analysis

Name	Dataset Used	Models	Highest Accuracy
Aditi Govinda and Sushila Palwe[11] (2023)	PD speech data from UCI[10]	LR, SVM, RF, KNN	91.83% with RF
Rohit Lamba *et al.*[16](2021)	PD speech data from UCI[10]	NB, KNN, RF	95.58% with the combination of genetic algorithm and RF
Proposed model	PD speech data from UCI[10]	SVM, LR, RF, KNN, GNB	97% with KNN

4.4 Discussion

To handle the imbalance of the dataset which was used in the proposed model to train and test, SMOTE-TOMEK was applied for oversampling of the dataset. The dataset was divided in 80-20 portion to train and test with the five algorithms respectively. According to the figures in section 4.1.4, after 5-fold cross-validation of the implemented algorithms, K-NN showed the highest accuracy (97%) while GNB showed the lowest (80%). Compared to the previous research on same dataset from Table 4.3 Aditi Govinda *et al*. and Rohit Lamba *et al.* both authors found highest accuracy on RF respectively 91.83% and 95.58% where this research get the highest accuracy on KNN which is 97%.

CHAPTER 5

CONCLUSION

Known as a neurodegenerative disorder, Parkinson’s disease or PD mainly happens from the loss of dopaminergic neurons characterized by associated motor symptoms. Here, the term "Neurodegenerative" refers to the progressive and irreversible loss of neurons. Although the most common motor symptoms are rigidity, tremors, gate instability etc.; signs and symptoms of this disorder appear when 50% of the neurons have been lost. Most cases of PD are known to have no specific cause but some rare cases have shown genetics and environmental factors as the main reason for PD. PD is most common amongst the people who are in the age range of 60-65 years but 10-15% of the cases exist where the ages are less than 40. According to the World Health Organization (WHO), in 2019, PD resulted in 5.8 million disability-adjusted life years (DALYs), an increase of 81% since 2000, and caused 329,000 deaths, an increase of over 100% since 2000. It can be said that the prevalence of PD patients has doubled over the past 25 years. Coping with PD in daily life could be very difficult for an individual. Consequently, a good screening technique can be useful, especially in circumstances in which a normal remedy isn't essential. For that reason, for the prognosis of PD, ML algorithms were evaluated. The main aim of this evaluation became identifying existing ML-based research frequently, diagnosing PD in terms of handwritten patterns, voice attributes, and gait dataset and daily deciding the maximum suitable technique to diagnose PD with proper accuracy. In this journal, from the dataset which contains 22(excluding the "status" column) voice attributes of 195

samples, the author found the most appropriate process of ml which is KNN as a simple recognized classifier with a high accuracy-97% to detect normal and patients whom may be affected by this disease. This assessment addressed diverse challenges and additionally provided a few future suggestions and possibilities, as it is determined that there may nonetheless be several works that must be achieved in the future. This assessment is also significant for the developments in neural networks and associated learning systems, which offer valuable insights and guidelines for future development.

REFERENCES

[1] B. R. Bloem, M. S. Okun, and C. Klein, "Parkinson's Disease," in *The Lancet*, vol. 397, no. 10291, pp. 2284–2303,2021. doi: 10.1016/S0140-6736(21)00218-X

[2] M. M. Emran, H. Z. Rahman, M. A. Habib, M. A. Hoque, K. M. Sobhan, G. K. Paul, N. Fatema, "Clinical Profile of Parkinson's Disease Patients in a Tertiary Hospital.," in *Mymensingh Medical Journal: MMJ*, vol. 31, no. 4, pp. 1073–1076, 2022. Accessed: Dec. 12, 2023. Available: https://pubmed.ncbi.nlm.nih.gov/36189554.

[3] K. Khanna, S. Gambhir, and M. Gambhir, "Current Challenges in Detection of Parkinson's Disease," in *Journal of Critical Reviews*, vol. 7, no. 18, pp. 1461–1467,2020. Accessed: Dec. 12, 2023. Available:https://www.researchgate.net/publication/357870234_current_challenges_in_detection_of_parkinson's_disease.

[4] P. Gupta and S. Garg, "Breast Cancer Prediction Using Varying Parameters of Machine Learning Models*" in Proc. Third International Conference on Computing and Network Communications (CoCoNet'19)*, vol. 171. Elsevier B.V, 2020, pp. 593–601. doi: 10.1016/j.procs.2020.04.064.

[5] A. Mert, N. Kilic, and A. Akan, "Breast Cancer Classification by Using Support Vector Machines with Reduced Dimension" in *Proc. 53rd International Symposium ELMAR-2011,* Zadar, Croatia. IEEE, 14-16 September 2011, pp. 37-40.

[6] P. Schober and T. R. Vetter, "Logistic Regression in Medical Research," in *Anesthesia & Analgesia*, vol. 132, no. 2, pp. 365–366, 2021, doi: 10.1213/ane.0000000000005247.

[7] S. Patwardhan, “Simple Understanding and Implementation of KNN Algorithm!,” *Analytics Vidhya*, [Online]. Accessed: Dec. 12, 2023.Available: https://www.analyticsvidhya.com/BLOG/2021/04/simple-understanding-and-implementation-of-knn-algorithm/

[8] T. R. Gadekallu N. Khare, S. Bhattacharya, S. Singh, P. K. R. Maddikunta, I. H. Ra, M. Alazab, “Early Detection of Diabetic Retinopathy Using PCA-Firefly Based Deep Learning Model,” in *Electronics*, vol. 9, no. 2, p. 274, 2020, doi: 10.3390/electronics9020274.

[9] Z. Huang and D. Chen, Eds., “A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering Random Forest Algorithm,” in *IEEE Access*, vol. 10, pp. 3284–3293, 2022, doi: 10.1109/access.2021.3139595.

[10] “Parkinsons,” *UCI Machine Learning Repository*. Little, Max, 2008. Accessed: Dec. 12, 2023. Available: https://archive.ics.uci.edu/dataset/174/parkinsons

[11] A. Govindu and P. Sushila, “Early Detection of Parkinson’s Disease Using Machine Learning”, in Proc. *International Conference on Machine Learning and Data Engineering*, vol. 218. Elsevier B.V, Jan. 2023. doi: 10.1016/j.procs.2023.01.007

[12] P. Kumar Mall, R. Kumar Yadav, A. Kumar Rai, V. Narayan, and S. Srivastava, “Early Warning Signs of Parkinson’s Disease Prediction Using Machine Learning Technique,” in *Journal of Pharmaceutical Negative Results*, vol. 13, no. 10, pp. 4784–4792, 2022, doi: 10.47750/pnr.2022.13.S10.579.

[13] J. Mei, C. Desrosiers, and J. Frasnelli, “Machine Learning for The Diagnosis of Parkinson’s Disease: A Review of Literature,” in *Frontiers in Aging Neuroscience*, vol. 13, p. 633752, 2021, doi: 10.3389/fnagi.2021.633752.

[14] I. Suuronen, A. Airola, T. Pahikkala, M. Murtojärvi, V. Kaasinen, and H. Railo, “Budget-Based Classification of Parkinson’s Disease from Resting State EEG,” in *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 8, pp. 3740–3747, 2023, doi: 10.1109/JBHI.2023.3235040.

[15] R. Lamba, T. Gulati, H. F. Alharbi, and A. Jain, “A Hybrid System for Parkinson’s Disease Diagnosis Using Machine Learning Techniques,” in *International Journal of Speech Technology*, vol. 25, no. 3, pp. 583–593, 2021, doi: 10.1007/s10772-021-09837-9.

[16] I. Ahmed, S. Aljahdali, M. S. Khan, and S. Kaddoura, “Classification of Parkinson Disease Based on Patient’s Voice Signal Using Machine Learning,” in *Intelligent Automation & Soft Computing*, vol. 32, no. 2, pp. 705–722, 2022, doi: 10.32604/iasc.2022.022037.