World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Predicting the Parkinson's disease using machine learning algorithms

Rebina Ferdous [1,*], Arif Hossain [2] and Mahinor Afroza [2]

[1] *Genome Research Centre, Bangladesh Jute Research Institute, Dhaka-1207, Bangladesh.*
[2] *Department of Computer Science and Engineering, Jagannath University, Dhaka-1100, Bangladesh.*

World Journal of Advanced Research and Reviews, 2025, 26(02), 3342-3346

Publication history: Received on 12 April 2025; revised on 23 May 2025; accepted on 26 May 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.26.2.1977

Abstract

Parkinson's disease (PD) is a neurodegenerative movement disease where the symptoms gradually develop start with a slight tremor in one hand and a feeling of stiffness in the body and it became worse over time. Parkinson's is considered one of the deadliest and progressive nervous system diseases that affect movement. It is the second most common neurological disorder that causes disability, reduces the life span, and still has no cure. Nearly 90% of affected people with this disease have speech disorders. In real-world applications, the information is been generated by using various Machine Learning techniques. Machine learning algorithms help to generate useful content from it. To increase the lifespan of elderly people the machine learning algorithms are used to detect diseases in the early stages. Speech features are the main concept while taking into consideration the term 'Parkinson's'. In this paper, we are using various Machine Learning techniques like Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree Classifier, Support Vector Machine (SVM), Bernoulli Naive Bayes(BNB), Gaussian Naive Bayes(GNB), Random Forest Classifier and how these algorithms are used to predict Parkinson's based on the input taken from the user and the input for algorithms is the dataset. The data set contains 24 attributes and 195 instances. From the results, It shows the predicting accuracy of the algorithms. To recover the patients from early stages, prediction is important. This process can be done with the help of Machine Learning.

Keywords: Parkinson's Disease; Logistic Regression; K-Nearest Neighbors; Decision Tree; Support Vector Machine; Random Forest; Bernoulli naive bayes; Data Mining; Machine Learning; Parkinson's Disease

1. Introduction

Parkinson's disease is one of the most lethal diseases of this world. It is a universal public health problem of massive measurement. It is a dynamic neurodegenerative disorder influencing over 6 million people worldwide. It takes the lives of million people every year. There is no cure or prevention for PD. Parkinson's the disease can be controlled in early stage. The data mining techniques is used as a effective way for early detection and diagnosis of the health disease. Machine learning can be used to generate hidden patterns from those data and to observe and predict for the future. These investigated concealed examples in clinical datasets can be utilized for clinical findings. Not with standing, clinical datasets are broadly scattered, heterogeneous, and colossal in nature. These datasets should be coordinated and incorporated with the medical clinic management systems. In this work different machine learning algorithms and data mining techniques are used to develop models by which I can predict cardiovascular disease efficiently so that more patients can get medicines inside a more limited time frame, bringing about saving great much life. In this work, Machine learning models is designed using different machine learning algorithms to the UCI data set (Saleh et al. 2024). Algorithms that are have used in this research are: Logistic Regression, Decision Tree, Random Forest(information gain), Random Forest (Entropy), Support Vector Machine, K-Nearest Neighbors, Gaussian Naïve Bayes and, Bernoulli Naïve Bayes Decision tree. The reason of using different machine learning algorithms is find out the best algorithm for this problem and it will allow me to compare among the models.

* Corresponding author: Rebina Ferdous

2. Literature Review

Machine learning is a domain of artificial intelligence (AI). It gives the systems the capability to spontaneously improve by learning from the past experiences. Basically machine learning's aim to develop computer programs to access the data and use the data for learning themselves. In modern age machine learning is the most used buzzword. Firstly with observations from huge number of data the learning process begins. It looks for patterns in data from direct experiences, instructions, or examples. It makes better decisions from those pattern for the future. Allowing the computes or programs to learn spontaneously and also without any kind of human or manual interference is the main goal, after learning the programs also should take accurate decisions and actions according to the learning. The aim of machine learning is to program computers to use data or past experience to solve a problem and predict accordingly. In present era there are lots of applications of machine learning around us, including systems that analyze customer previous buying data and predict customer behavior for future, optimize workers behavior so that they can complete a task using minimum resources, extract knowledge from data and many more. The most useful use of machine learning is in medical fields, it can be used for solving a lot of medical treatment related problems which can decrease the death rate and provide healthy and secure life to the human being. It also used in many more fields like statistics, pattern recognition, neural networks, signal processing, control, artificial intelligence, In order to present a unified treatment of machine learning problems and solutions, it discusses many methods from different fields, including statistics, pattern recognition, neural networks, artificial intelligence, signal processing, control, and data mining (Keserwani, Das, and Sarkar 2024).

In (Senturk 2020) design and develop a prediction system for Parkinson's disease prediction problem. Classification and Regression Trees, Artificial Neural Networks, and Support Vector Machines were used for the classification of Parkinson's patients in the experiments. Support Vector Machines with Recursive Feature Elimination was shown to perform better than the other methods. Their models performances have been given. 93.84% accuracy was achieved with the least number of voice features for Parkinson's diagnosis.

(Rahman, Khan, and Raza 2020) examines the multiple types of vowel samples collected from PD patients and healthy subjects and utilize state-of-the-art signal processing algorithms like Perceptual Linear Prediction (PLP) and Realitive Spectral PLP (RASTA-PLP) for feature extraction purposes. In this study, they applied two state-of-the-art signal processing algorithms i.e. Perceptual Linear Prediction (PLP) and ReAlitive SpecTrAl PLP (RASTA-PLP) for feature extraction purposes and SVM model with four different types of kernels for classification task. It was observed that highest classification accuracy of 74% was achieved using PLP based features extraction and SVM model developed with RBF kernel, followed by RASTA-PLP with 68% of classification accuracy obtained under SVM model with MLP kernel. In (Govindu and Palwe 2023) applied build three different machine learning models using three different machine learning algorithms named logistic regression, Random forests and Support Vector Machine to predict Parkinson's disease prediction. Their models performance shows that Random forests performed better with accuracy 80% in compared to other algorithms whereas logistic regression achieved 74% and SVM achieved 70% of accuracy. In (Wang et al. 2020) learning algorithms proposed a model for Parkinson's disease. They selected 7 feature for Support Vector machine Algorithm. In 2020, (Anila and Pradeepini 2020) applied build three different machine learning models using three different machine learning algorithms named Random forests, K-Nearest Neighbors and Support Vector Machine to predict Parkinson's disease prediction. Random forests performed better with accuracy 90.26% in compared to other algorithms [5].

3. Methodologies

Figure 1 shows the block diagram of summarizing the methodology adopted in this research, after that the descriptions will also be given.

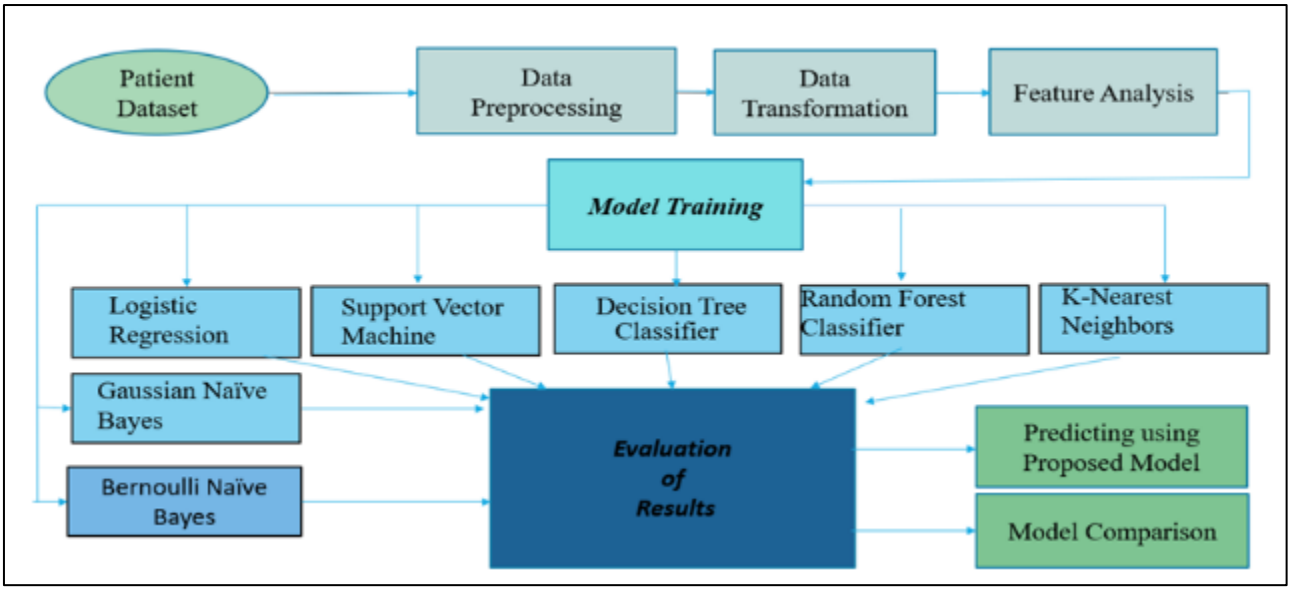

Figure 1 Block diagram of summarizing the methodology

The Parkinson's disease dataset is collected UCI dataset, which has been taken from UCI Machine Learning Repository. This dataset is the so far best dataset on this problem and it is the most used dataset by the experts so far. This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). This dataset contains total 24 attributes and 195 instances. Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from these individuals ("name" column). The main aim of the data is to discriminate healthy people from those with PD, according to "status" column which is set to 0 for healthy and 1 for PD. After collecting the dataset the first task is to preprocess the data before using it. Before starting let me give a brief description of data preprocessing. Data Preprocessing is a data processing technique that involves several steps to complete the process, including changing the data format and some other steps. In general real-world data is often inconsistent, incomplete and missing in certain behaviors, and contain many errors and outliers. By data preprocessing we can solve all of these problems and make dataset ready for using in any machine learning model. When we make a machine learning model train that model with dataset which is not preprocessed that model will not be able to give better result and the performance of that model eventually decrease a lot. The following preprocessing are applied, Visualize the data frame, Checked Null Values, Outliers, Incomplete Data in dataset, Identifying and removing duplicate rows is the first step of the data preprocessing, Checked missing values in dataset, Differentiate Categorical and Continuous Data. In data preprocessing it has checked whether there is any missing data, outliers, or inconsistent data in the data set. The categorical and continuous values are separated and plotted the correlation matrix and have seen how the features or attributes are interrelated with each other. In figure 2 the correlation matrix is given, here each cell of the matrix shows the relationship between the features of x-axis and y-axis, 1 means highest relation and -0.4 and lower than that means the lowest relation.

Feature Engineering means working with the features, modifying them and choosing the best features for the model training. Using the domain knowledge to extract features from the raw data via data mining techniques and machine learning knowledge is feature engineering. To improve the accuracy and performance of the machine learning algorithms these feature can be used. In this work we applied PCA for feature engineering. The Principal Component Analysis is a popular unsupervised learning technique for reducing the dimensionality of data. It increases interpretability yet, at the same time, it minimizes information loss. The total dataset have splitted in two parts, one is training data set and another is testing data set. The ratio was 80:20, 80% of the full data set is used for training purpose and 20% is used for testing purpose.

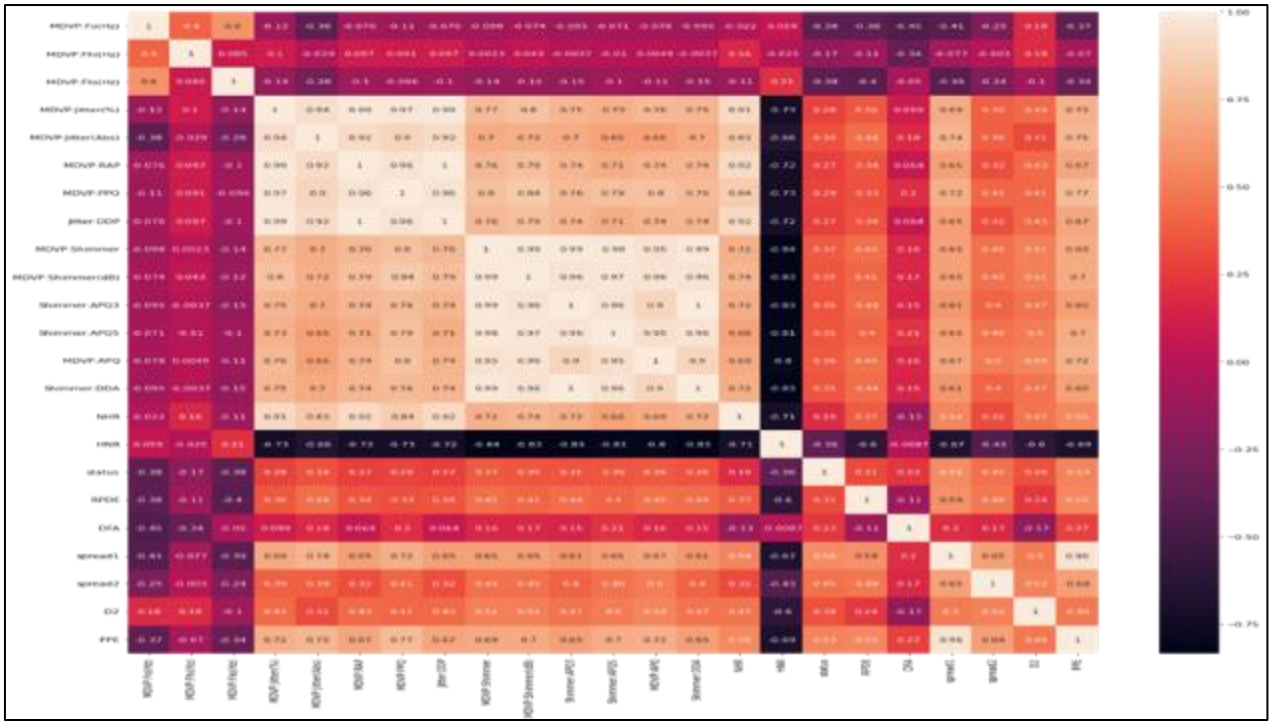

Figure 2 Correlation Matrix

4. Results and Analysis

From all the above In the Parkinson's disease prediction project, preprocess the data; we have fit the data into various models. We have also printed the matrics and evaluations parameter. We have seen from the model performance tables that which model is best suitable for this particular project. From the accuracy value we find that Logistic Regression performs poorly with accuracy 76.26% and the, Decision Tree Classifier, Support Vector Machine (SVM) and others perform good. Gaussian Naive Bayes, Bernoulli Naive Bayes) perform not well with accuracy 76.44% and 86.51%. K-Nearest Neighbors (KNN) perform best result with accuracy 96.61%, which can be chosen for the Parkinson's disease prediction. We would not just keep the accuracy as the major perameter for this project. This is a classification task this is being done a medical project so we need to make sure that keep into consideration the true positive and false positive values as well. We see the observation of the confusion matrix that the logistic regression confusion matrix is poor. Decision Tree also performs poorly.

Table 1 Showing the Models performance

Models	Accuracy (in %)	Confusion Matrix	ROC
Logistic Regression	76.27	6 false negative and 2 false positive	AUC-0.89
K-Nearest Neighbors (KNN)	96.61	0 false negative and 0 false positive	AUC-1
Support Vector Machine (SVM)	94.91	0 false negative and 3 false positive	AUC-0.98
Decision Tree Classifier	91.52	2 false negative and 2 false positive	AUC-0.92
Random Forest Classifier	94.91	0 false negative and 2 false positive	AUC-0.96
Bernoulli Naive Bayes(BNB)	76.44	7 false negative and 3 false positive	AUC-0.86
Gaussian Naive Bayes(GNB)	86.51	5 false negative and 2 false positive	AUC-0.89

It's found that in Random Forest classifier perform 0 false negative that is good. K-Nearest Neighbors (KNN) perform best for 0 false negative and 0 false positive result. In ROC curve we also find the best result for K-Nearest Neighbors (KNN). Others models also works very fine for this dataset. My models performed better than many other existing works on these fields. Table 4 shown the comparison between models.

5. Conclusion

The main purpose of this work is to comparing the accuracy of my machine learning models and also analyzing themselves and find the reason behind the variation of different algorithms. We also make sure that keep into consideration the true positive and false positive values as well. It has used Cleveland data set which is collected from UCI machine learning repository. The data set contains 24 attributes and 195 instances. It also tried to compare the ROC curve for these different machine learning algorithms. At the end of the implementation of all the algorithms and also including the confusion matrix though all the models gave very good performance but surprisingly' K-Nearest Neighbors(KNN)' gave the highest accuracy 96.61% and after that 'Support Vector Machine' gave 94.91%. The data set contains less instances, if there I could use a data set with more instances I could have get better result, but this UCI data set is so far most used and reliable data set for this type of research.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Anila, M, and GD Pradeepini. 2020. "A review on parkinson's disease diagnosis using machine learning techniques." *IJERT* 9:330-334.

[2] Govindu, Aditi, and Sushila Palwe. 2023. "Early detection of Parkinson's disease using machine learning." *Procedia Computer Science* 218:249-261.

[3] Keserwani, Pankaj Kumar, Suman Das, and Nairita Sarkar. 2024. "A comparative study: prediction of parkinson's disease using machine learning, deep learning and nature inspired algorithm." *Multimedia Tools and Applications* 83 (27):69393-69441.

[4] Rahman, Atiqur, Aurangzeb Khan, and Arsalan Ali Raza. 2020. "Parkinson's disease detection based on signal processing algorithms and machine learning." *CRPASE: Transactions of Electrical, Electronic and Computer Engineering* 6 (3):141-145.

[5] Saleh, Shawki, Bouchaib Cherradi, Oussama El Gannour, Soufiane Hamida, and Omar Bouattane. 2024. "Predicting patients with Parkinson's disease using Machine Learning and ensemble voting technique." *Multimedia Tools and Applications* 83 (11):33207-33234.

[6] Senturk, Zehra Karapinar. 2020. "Early diagnosis of Parkinson's disease using machine learning algorithms." *Medical hypotheses* 138:109603.

[7] Wang, Wu, Junho Lee, Fouzi Harrou, and Ying Sun. 2020. "Early detection of Parkinson's disease using deep learning and machine learning." *IEEE Access* 8:147635-147646.