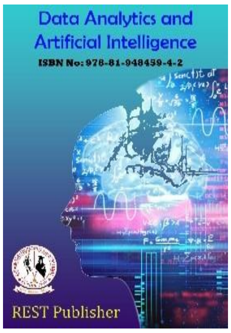
Data Analytics and Artificial Intelligence
Vol: 3(6), 2023
REST Publisher; ISBN: 978-81-948459-4-2
Website: http://restpublisher.com/book-series/daai/


Data Analytics and Artificial Intelligence
ISBN No: 978-81-948459-4-2
REST Publisher

Early Identification of Parkinson’s Disease Using Machine Learning Techniques

***T. Jemima Jebaseeli, R. Geowin Christosingh, M. Navin Kumar**
Karunya Institute of Technology and Sciences, Coimbatore, Tamil Nadu, India.
*Corresponding Author Email: jemima_jeba@karunya.edu

Abstract: *There are enormous medical datasets accessible to diagnose disorders. Parkinson's Disease (PD) is described as among the most lethal and degenerative central nervous illnesses impacting mobility. It is the second-most commonly diagnosed neurodegenerative illness that causes impairment, shortens lifetime, and does not known treatment. Almost 90% of those afflicted by this condition had communication difficulties. In practical uses, data is created utilizing a variety of Machine Learning algorithms. In the proposed system there are two data sets used to diagnose Parkinson disease. Machine learning techniques assist in the generation of meaningful information from that as well. The proposed research employs a few Machine Learning approaches such as K-Nearest Neighbors Algorithm (KNN), Naive Bayes, and Logistic Regression to forecast Parkinson's disease in response to user instruction, with the dataset serving as the insight for the procedures. The accuracy is predicted by using hyper parameter optimization method. Depending on such characteristics, the author estimates the algorithms that will provide more reliability. The accuracy attained for three techniques is 94.5% for KNN, 97.3% for Logistic Regression, and 98.9% for Naive Bayes, which is applied to identify if the individual has Parkinson's disease or not. We also employed the hyper parameter tuning approach to improve accuracy. Prediction is critical in the initial stages of clinical outcomes. This may be accomplished with the use of machine learning.*

Index Terms: *Parkinson's disease, speech disorder, KNN, Naive Bayes, logistic regression, machine learning.*

1. INTRODUCTION

According to the World Health Organization's latest study, the prevalence and illness consequence of Parkinson's disease sufferers is quickly increasing. This disease is developing so quickly in China because it is anticipated that it will affect a majority of the population over the next 10 years [1]. Classification techniques are mostly employed in the healthcare industry to categorize facts based on a variety of criteria. The conversely dangerous neurological ailment is Parkinson's disease. It causes trembling, tightness as well as trouble walk and balancing [2]. It is mostly triggered by the breakdown of cells inside the neurological system. Parkinson's disease potentially manifests both movement and non-motor signs. Slowness of movements, stiffness, mobility problems, and convulsions are examples of movement disorders [3]. This disease is developing so quickly in China because it is anticipated that it will affect a majority of the population over the next 10 years. Classification techniques are mostly employed in the healthcare industry to categorize facts based on a variety of criteria [4]. It is mostly triggered by the breakdown of cells inside the neural synapses [5]. Parkinson's disease potentially manifests both movement and non-motor signs. Slowness of movements, stiffness, mobility problems, and convulsions are examples of movement disorders [6]. If the condition progresses then the people may have difficulties walking and speaking.

The non-motor side effects include nausea, respiratory issues, sadness, loss of smell, and changes in speaking. If these symptoms are found in the person, the information is recorded [7]. The author evaluates the patients' speech qualities in this research, and this information is utilized to determine whether person has Parkinson's disease or not. Neurodegenerative diseases are caused by gradual shredding and synapse loss in many regions of the nervous system. Synapses are the brain’s primary components [8]. They are not uninterrupted, but instead contiguous. A nutritious neuron has dendrites or axons, an inner medulla, and nuclei that hold the DNA. DNA is the genetic chromosome and the complete genome is bundled into a 100 trillion synapses. Whenever a neuron goes ill, it lost its expansion and therefore its capability to communicate, which in itself is not good for that too, as well as its weight dropped, leading it to gather detritus, one which attempts to keep in small packets. When circumstances worsen, and if the axon is a cell cultures, it lost its growth and becomes spherical and filled with pores [9-11]. This proposed research is about the forecast of Parkinson's disease that is a rapidly expanding incurable condition. The major molecules that impact the brain activity are dopamine and acetylcholine. A number of environmental component that has been linked in PD are the stated factor which produced Parkinson's disease in a person. The article is structured as follows. Section 2 includes relevant work. Section 3 comprises a proposed methodology and discusses the dataset, the explanations of distinct characteristics collected, statistical analysis of these characteristics, and the building of prediction and classification models. contains the results and discussions. Finally, Section 5 provides the conclusion.

RELATED RESEARCH WORK

Researchers applied several traits and statistics to forecast Parkinson's disease. Kamal Nayan et al., [11] attempted to develop this classifier using multilayer perceptron, Random Forest. It was discovered that boosted logistic regression performed better and produces 97.15% accuracy and 98.9% ROC. Mhyre et al., [12] utilized a speech dataset obtained from persons to validate the hypothesis. Deep neural networks are used to identify PD. 85% accuracy is produced that is higher than the typical clinical diagnostic accuracy of non-experts. Dutkiewicz et al., [13] proposes an automated approach based on emotion change detection during pronouncing prescribed verbal tasks. Using the XGBoost method 69% accuracy was achieved. Here, the sentence that is challenging to pronounce. Because the properties can be described, it may be employed in clinical practice. Fear was indicated as the most effective emotion for PD identification in this situation. Anastasia et al., [14] employed the binary classification for each motor task independently and several feature based strategies.

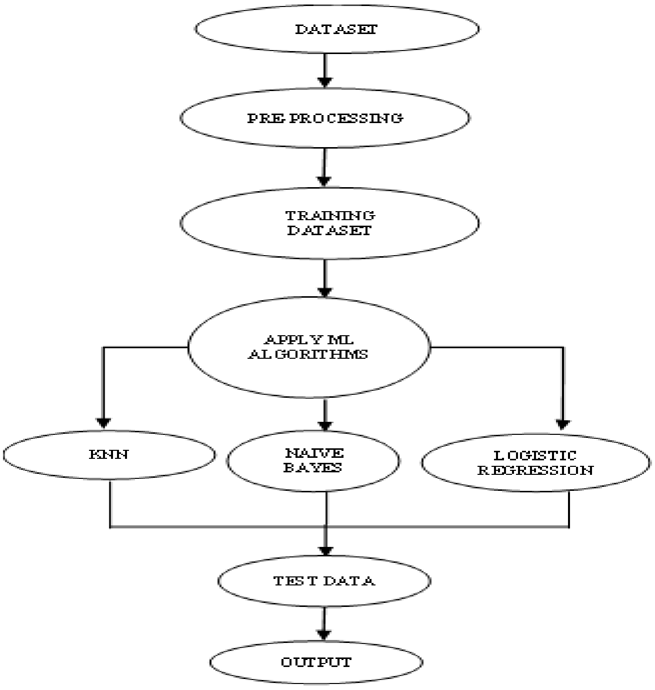


FIGURE 1. The architecture of the proposed system to identify Parkinson's disease.

2. PROPOSED METHODOLOGY

Figure 1 depicts a flowchart of the proposed research. The data was gathered, and after the necessary preprocessing, the data characteristics were retrieved. The classification job is then performed using several machine learning methods. Finally, a comparison of the accuracy offered by several machine learning models is performed.

A. Dataset: The data collection included 31 patients. In that 23 of them had PD. Each column of the dataset represents a different voice metric, and each row corresponds to one of 195 voice recordings from these people.

B. Proposed models for identifying PD: For the classification problem in this work, four distinct machine learning models are used. The major goal is to improve the accuracy through the hyper parameter tuning. First, the dataset is preprocessed using normalize filter. 70% of data is used as training dataset and the remaining is utilized for testing.

C. Testing and Training Data: After training the PD prediction model on the pre-processed dataset, the model is evaluated using different sets of data. The model is tested for completeness and accuracy in this stage by presenting it with a test dataset. To choose the optimal model to utilize, all training techniques must be validated. The proposed model predicts the values for the test dataset after fitting it using training data. These anticipated values in testing data are utilized to compare models and calculate accurately.

D. Logistic Regression: Logistic regression is used to identify a certain variable using a group of independent variables. It turns into a categorization difficulty in and of itself. The choice on the significance of the edge value is heavily influenced by accuracy and recall values. There is a need of both precision and recall tradeoffs, thus utilizing the following considerations to decide on a threshold: Low precision and high precision. Logistic regression is given as follows.

$$\mathrm{f}(X) = \frac{1}{1+e^{X}} \qquad (1)$$

Where x is the input function and f(x) yields the output ranges from 0 to 1.

E. Naive Bayes: Naive Bayes is the straightforward supervised learning methods. The Nave Bayes classifier is a quick, efficient, and dependable method. The Nave Bayes classifier is based on the assumption that the influence of a given characteristic throughout a class is irrespective of other characteristics. Whether a patient has Parkinson's disease or not is determined by the patient's speech characteristics.

$$P\left(h/D\right) = \frac{P\left(D/h\right) * P(h)}{P(D)} \qquad (2)$$

Where $P(h)$ is the likelihood hypothesis and h will be true regardless of data d . $P(D)$ is the probability regardless of the hypothesis. $P\left(h/D\right)$ is the likelihood of given data. $P\left(D/h\right)$ is the probability of data if hypothesis is true.

F. K-Nearest Neighbor: The KNN approach is used to solve problems with classification and regression. It is straightforward to implement and understand. It is related to the field of supervised learning. Let m indicate the number of instances of training data. P is an unidentified place to maintain the sample points array with the training instances. Each array item is a tuple used to compute the Euclidean distance.

3. RESULTS AND DISCUSSION

To show the results of the proposed techniques the remaining test data has been run it through different algorithms. Following that, the proposed trained model is now able to estimate whether or not the illness is existing and the results are given in Figure 2. The accuracy test is performed in Google Colab, which itself is a Python notebook. Many criteria are used to assess the prediction findings in the proposed approach. Following are the metrics to evaluate the proposed method.

A. Confusion Matrix: The Error function is another name for the confusion matrix. It is typically used to illustrate a algorithm's efficiency. In the correlation analysis, TP denotes true positive, FP denotes false positive, FN denotes false negative, and TN denotes true negative. Figure 3 depicts the correlation matrix.

B. Accuracy: It is calculated by dividing the total number of samples by the sum of true positive and true negative cases.

$$Accuracy = \frac{(TP + TN)}{TP + FP + TN + FN} \qquad (3)$$

C. Precision: Precision is the ratio of genuine positive to projected yes cases. It is sometimes referred to as the ratio of accurate positive results to total positive results anticipated by the algorithm.

$$\mathrm{Pr}\,ecision = \frac{TP}{TP+FP} \qquad (4)$$

D. Recall (or) Sensitivity: Recall is the ratio of correct positive outcomes to the total amount of relevant samples or as the percentage of True Positive cases to True occurrences.

$$\mathrm{Re}\,call = \frac{TP}{TP+FP} \qquad (5)$$

E. F1-Score: F1-Score is the proportion of the recall and precision product to the total of the recall and accuracy parameters of classification. It is the summing of accuracy and recall. It assesses test accuracy. This measure has a scale of 0 to 1.

$$F1Score \quad = 2\times\frac{1}{\left(\frac{1}{\mathrm{Pr}\,ecision}\right)+\left(\frac{1}{\mathrm{Re}\,call}\right)} = \frac{2PR}{(P+R)} \qquad (6)$$

F. Specificity: The specificity of a test is a measure of its ability to detect genuine negatives.

$$Specificity = \frac{TN}{TN+FP} \qquad (7)$$

G. Odd Score: Odd Score is described as determining the likelihood of an event's occurrence when there are three alternative possibilities and a casual influence.

$$Odd\ Score \quad = \frac{Sensitivity \big/ 1-Specificity}{1-Sensitivity \big/ Specificity}$$

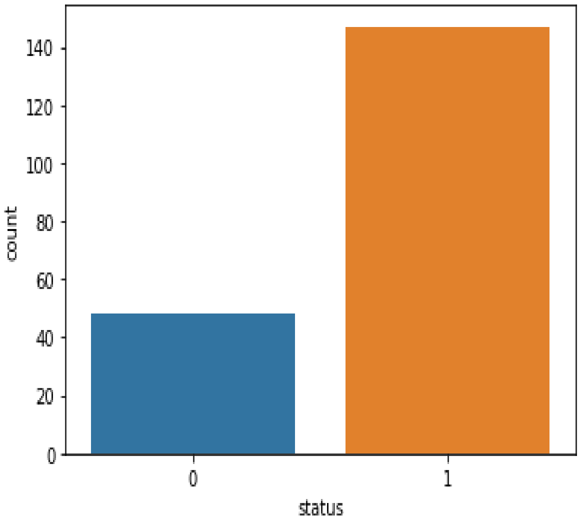


FIGURE 2. Plotting the class instance (Status Count).

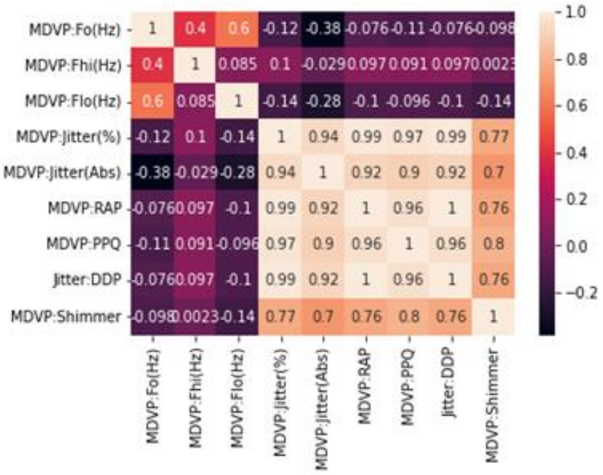


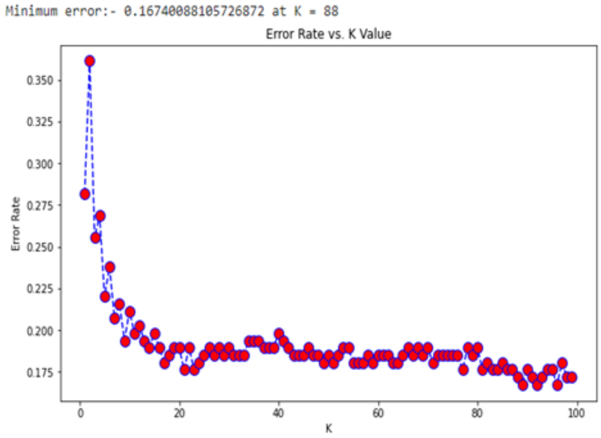
FIGURE 3. Correlation matrix

H. Youden Score: The Youden index is defined as the distinction between both the proportions of class 0 of class 1.

$$Youden = Sensitivity + Specificity - 1 \qquad (9)$$

I. Error rate: The error rate is determined as the total number of two inaccurate predictions on the dataset. Figure 4 shows the projected error rate of 0.16.

The goal is to employ multiple assessment measures such as Accuracy, precision, recall, specificity, F1-Score, LR+, LR-, and Youden score to efficiently forecast illness. The proposed system utilized a speech dataset that incorporates patient voice attributes and is available on the Kaggle website. The dataset includes about 700 attributes collected from 750 patients. The models are designed using the five best characteristics discovered through feature selection. The proposed model works well are described through various statistical comparisons. The comparison of the three models based on the assessment metrics are shown in Figure 5. The dataset was then tested using the remaining test data. KNN model gives 80% accuracy. The training dataset is used to train the Naive Bayes algorithm, which is subsequently tested using the remaining test data. This method gives 81% accuracy. Lastly, the logistic regression algorithm is trained using the training dataset and then evaluated using the remaining test data.The proposed system employed the hyper parameter tuning approach to fine-tune the algorithm's accuracy. The utilized trained model variable makes a prediction result. The proposed techniques made the classification predictions as the patient has PD or the patient doesn't have PD. Table I shows the performance improvement of KNN, Naive Bayes, and Linear Regression Techniques through the proposed parameter fine tuning approaches.



```
Minimum error:- 0.16740088105726872 at K = 88
```

FIGURE 4. Error Rate

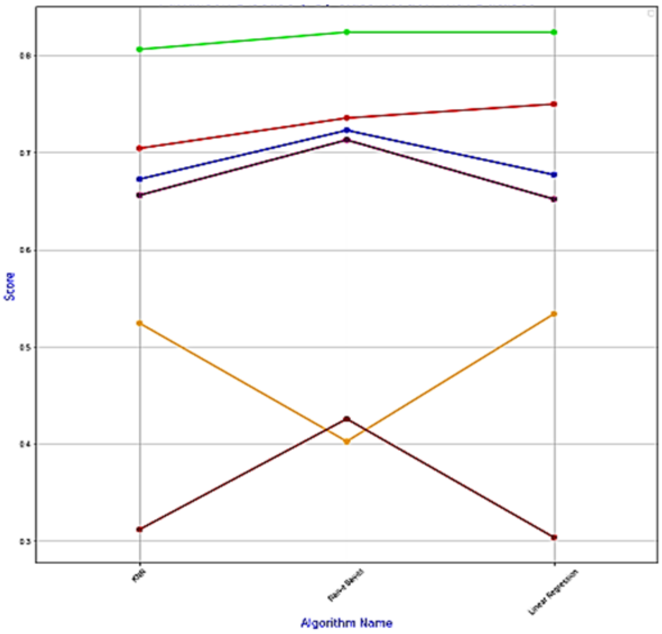


FIGURE 5. Parkinson’s disease classifications with Kaggle dataset.

TABLE 1. Performance measures for various classifiers used in the proposed research.

ALGORITHMS	**PRECISION**	**RECALL**	**F1_SCORE**	**AUC**	**LR +**	**LR-**	**ODD SCORE**	**YOUDEN**	**SPECIFICITY**
KNN	0.95	0.93	0.66	0.94	2.01	0.49	4.06	0.33	0.88
Naive Bayes	0.98	0.96	0.70	0.97	2.38	0.41	5.68	0.40	0.95
Linear Regression	0.97	0.94	0.55	0.96	1.25	0.79	1.57	0.11	0.94

4. CONCLUSION

Parkinson's disease is the neurological illness and its prognosis is critical. Hence, the proposed system employed three different prediction models to predict Parkinson's disease which includes Machine Learning methods such as KNN, Naive Bayes, and Logistic Regression. This model employed a speech dataset that included patient voice characteristics. The dataset is trained using these models, and compared with all other models produced using different approaches to identify the best model that fits for this operation. The hyper parameter tuning method improves the accuracy. Based on these findings, Nave Bayes outperforms the other two machine learning algorithms with an accuracy of 98.9% to predict the PD.

REFERENCES

[1]. Vu, T.C., Nutt, J.G, and Holford, N.H, “Progression of motor and nonmotor features of Parkinson's disease and their response to treatment”, British journal of clinical pharmacology, vol. 74(2), pp.267-283, 2012.

[2]. Prashanth, R., Roy, S.D., Mandal, P.K, and Ghosh, S., “High-accuracy detection of early Parkinson's disease through multimodal features and machine learning”, International Journal of Medical Informatics, vol. 90, pp.13-21, 2016.

[3]. Wroge, T.J., Ozkanca, Y., Demiroglu, C., Si, D., Atkins, D.C, and Ghomi, R.H., “Parkinson’s disease diagnosis using machine learning and voice”, In 2018 IEEE signal processing in medicine and biology symposium, pp. 1-7, 2018.

[4]. Skibinska, J and Burget, R., “Parkinson’s disease detection based on changes of emotions during speech. In 2020 12th International Congress on Ultra-Modern Telecommunications and Control Systems and Workshops” pp. 124-130, 2020.

[5]. Moshkova, A., Samorodov, A., Voinova, N., Volkov, A., Ivanova, E. and Fedotova, E., “Parkinson’s disease detection by using machine learning algorithms and hand movement signal from Leap Motion sensor”, In 2020 26th Conference of Open Innovations Association, pp. 321-327, 2020.

[6]. Wang, W., Lee, J., Harrou, F, and Sun, Y., “Early detection of Parkinson’s disease using deep learning and machine learning. IEEE Access, 8, pp.147635-147646.

[7]. Shukla, A., Mani, A., Bhattacharya, A. and Revilla, F., 2014. “Understanding Postural Response of Parkinson’s Subjects Using Nonlinear Dynamics and Support Vector Machines”, Austin J Biomed Eng, vol. 1(1), 2020.

[8]. Silveira-Moriyama, L., Carvalho, M.D.J., Katzenschlager, R., Petrie, A., Ranvaud, R., Barbosa, E.R. and Lees, A.J., “The use of smell identification tests in the diagnosis of Parkinson's disease in Brazil”, Movement disorders: official journal of the Movement Disorder Society, vol.23(16), pp.2328-2334, 2008.

[9]. Ponsen, M.M., Stoffers, D., Booij, J., van Eck-Smit, B.L., Wolters, E.C, and Berendse, H.W., “Idiopathic hyposmia as a preclinical sign of Parkinson's disease”, Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society, vol.56(2), pp.173-181, 2004.

[10]. Doty, R.L., Shaman, P, and Dann, M., “Development of the University of Pennsylvania Smell Identification Test: a standardized microencapsulated test of olfactory function”, Physiology & behavior, vol.32 (3), pp.489-502, 1984.

[11]. Challa, Kamal Nayan Reddy, Venkata Sasank Pagolu, Ganapati Panda, and Babita Majhi, “An improved approach for prediction of Parkinson's disease using machine learning techniques”, In 2016 International Conference On Signal Processing, Communication, Power and Embedded System, pp. 1446-1451, 2016.

[12]. Mhyre TR, Boyd JT, Hamill RW, and Maguire-Zeiss KA. “Parkinson's disease”, Subcell Biochem, vol. 65, pp.389-455, 2012.

[13]. Dutkiewicz J and Friedman A, “Diagnosis of autonomic disorders in Parkinson's disease”, Wiad Lek, vol.73 (4), pp.809-813, 2020.

[14]. Anastasia M Bougea, Nikolas Papagiannakis, Athina-Maria Simitsi, and Leonidas Stefanis, “Ambiental Factors in Parkinson’s disease Progression: A Systematic Review”, Medicina (Kaunas, Lithuania), vol. 59(2), 2023.

[15]. Cummings, J.L., Henchcliffe, C., Schaier, S., Simuni, T., Waxman, A, and Kemp, P, “The role of dopaminergic imaging in patients with symptoms of dopaminergic system neurodegeneration. Brain”, vol.134(11), pp.3146-3166, 2011.