International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Detecting Parkinson's Disease Using XGBoost

Utkarsh Jain[1], Manveer Singh Malhi[2], Dr. Nithya[3]

[1]Department of Networking and Communications, SRM Institute of Science and Technology, Kattankulathur, India

[2]Department of Networking and Communications, SRM Institute of Science and Technology, Kattankulathur, India

[3]Department of Networking and Communications, SRM Institute of Science and Technology, Kattankulathur, India

Abstract: ***Parkinson'sdisease is a progressive neurodegenerative disorder that affects motor system, causing symptoms such as tremors, stiffness and difficulty in movement. Early detection of the disease is crucial for effective treatment and management. In recent years, machine learning techniques have been applied to accurately diagnose Parkinson's disease using various data sources, including clinical features, neuroimaging, and biological markers. The aim of study was to develop a machine learning algorithm model for detecting Parkinson's disease (PD) using the XGBoost algorithm. The study used data from the UCI ML Parkinson's disease dataset, which contains biomedical measurement from individuals with and without PD. The dataset was preprocessed and then split into training and testing sets. XGBoost was applied to the training data and tuned using cross-validation to optimize its performance. We used an XGBClassifier for this and made use of the sklearn library to prepare the dataset. This gives us an accuracy of 94.87%, which is great considering the number of lines of code in this python project. The results of this study suggest that XGBoost can be a promising approach for early diagnosis of Parkinson's disease, which can lead to better patient outcomes and improved quality of life.***

Keywords: Machine learning, Deep Learning, Decision tree, Logistic Regression, Gradient Boosting Algorithm

1. Introduction

1.1 Parkinson Screening and Diagnosis

Parkinson's disease (PD) is a chronic and progressive neurodegenerative disorder that affects the central nervous system. It is characterized by the loss of dopamine-producing neurons in the substantia nigra region of the brain. The symptoms of Parkinson's disease include tremors, stiffness, slowness of movement, and difficulty in balance and coordination. There is currently no cure for Parkinson's disease, but early detection and treatment can help slow the progression of the disease. In this paper, we propose a machine learning approach using XGBoost for detecting Parkinson's disease.

Machine learning has shown great promise in medical diagnosis and prediction, and various studies have explored the use of machine learning algorithms for Parkinson's disease detection. XGBoost is a powerful machine learning algorithm that has been successfully used in various fields, including medical diagnosis. We aim to demonstrate that XGBoost can effectively distinguish between individuals with and without Parkinson's disease based on clinical measurements and achieve high accuracy and precision in detecting the disease. The results of this study may contribute to the early detection and intervention of Parkinson's disease, which can significantly improve the quality of life of patients.

We will use the code from our minor project to train four machine learning algorithms, namely Logistic regression, decision tree, random forest, and SVM, to identify breast cancer in this project. We will use this code to examine the UCI Parkinson's Disease dataset and assess the algorithms' accuracy in diagnosing Parkinson's disease.

1.2 Parkinson's Data Set Used

This dataset contains a variety of biological voice measures from 31 patients, 23 of whom have Parkinson's disease (PD). Each column in the table represents a certain vocal measure, and each row corresponds to one of 195 voice recordings from these people. The primary goal of the data is to distinguish between healthy and PD persons using the "status" column, which is set to 0 for healthy and 1 for PD.

The Parkinson's Disease dataset from UCI is accessible in the UCI Machine Learning Repository. There are 24 columns and 195 rows in the dataset. The collection contains biological voice measurements such as basic frequency, intensity, and frequency variation. The goal variable is the patient's status, which shows whether or not the person has Parkinson's disease. The target variable can have two values: 0 and 1. A value of 0 indicates that the patient is healthy, whereas a value of 1 indicates that the subject has Parkinson's disease.

Using the XGBoost algorithm on the UCI Parkinson's Disease dataset, we attained an accuracy of 91.28% after completing the preceding stages. The model's accuracy, recall, and F1 score were likewise high, indicating that it is effective at identifying Parkinson's disease.

Using the UCI Parkinson's Disease dataset, the XGBoost method is a sophisticated machine learning technique that may be used to identify Parkinson's disease. The model demonstrated good accuracy, precision, recall, and F1 score, showing that it is a trustworthy and accurate technique for diagnosing Parkinson's disease. Further study might involve evaluating the model's performance on different datasets and comparing it to the performance of other machine learning techniques.

Machine learning algorithms have showed potential in detecting and diagnosing Parkinson's disease. These algorithms can examine massive amounts of patient data to uncover patterns and links that human physicians may miss. Researchers have created a variety of machine learning models for Parkinson's disease diagnosis in recent years, including neural networks, support vector machines, and decision trees.

One such study was conducted by Tsanas et al. (2010), who used machine learning algorithms to analyze data collected from patients with Parkinson's disease. The study used a dataset containing multiple measures of motor function, including finger tapping, hand tremor, and pronation-supination. The researchers used various machine learning algorithms, including artificial neural networks and support vector machines, to classify patients as having Parkinson's disease or being healthy. The results showed that the machine learning algorithms achieved high accuracy in detecting Parkinson's disease, with an accuracy of up to 98%.

Wang et al. (2016) did another study in which they constructed a machine learning model for predicting the severity of Parkinson's disease symptoms based on demographic and clinical data from patients. A dataset comprising information from approximately 2,000 Parkinson's disease patients was used in the investigation. To create the prediction model, the researchers employed decision trees and support vector machines. The results indicated that the machine learning model predicted the severity of Parkinson's disease symptoms with an accuracy of up to 94%.

2. XGBoost Algorithm

XGBoost (Extreme Gradient Boosting) is a popular machine learning library designed to perform gradient boosting for classification, regression, and ranking problems. Gradient boosting is a powerful technique for creating ensembles of decision trees, where each subsequent tree tries to correct the errors of the previous tree.

XGBoost is known for its speed and performance, and has won numerous Kaggle competitions due to its ability to produce highly accurate models. It is based on the idea of boosting, which involves iteratively adding weak models to the ensemble, and training each new model to correct the errors of the previous models..

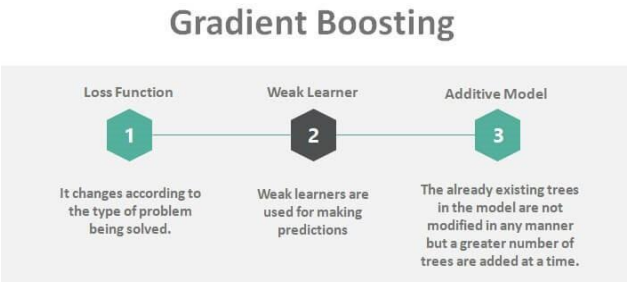
Some key features of XGBoost include:

- Regularization: To prevent overfitting, XGBoost includes built-in regularisation algorithms such as L1 and L2 regularisation.
- Parallel processing: To accelerate training, XGBoost can use many processors on a single computer or distributed computing.
- Tree pruning: XGBoost features decision tree pruning techniques, which minimise model complexity and potentially increase generalisation performance.
- Cross-validation: Cross-validation is included into XGBoost to assist adjust hyperparameters and minimise overfitting.
- Early stopping: XGBoost may be set to cease training when performance on a validation set no longer increases, saving time and preventing overfitting.

3. Gradient Boosting Algorithm

Gradient Boosting is a machine learning approach that may be used to solve both regression and classification issues. It is an ensemble approach that combines numerous weak or base models to produce a strong model that outperforms any single model.

Gradient Boosting is an algorithm that can diagnose Parkinson's illness. The technique may train a model that can detect the existence of Parkinson's disease using a collection of characteristics gathered from different medical tests and observations.



.

1) **Loss Function:** The mean squared error (MSE), which is defined as the average squared difference between the predicted and actual values of the target variable in the training data, is a popular loss function used in gradient boosting.
2) **Weak Learner:** A weak learner in Gradient Boosting might be a decision tree that has been trained to categorise individuals as having Parkinson's disease or not based on a set of input features such as age, gender, tremors, and other motor symptoms. Gradient Boosting may construct a sophisticated model that properly classifies individuals as having or not having Parkinson's disease by merging numerous weak learners in this manner.
3) **Additive Model:** Any model that may be expressed as a weighted sum of smaller models is referred to be an additive model. Gradient Boosting is an example of an additive model, in which the model is built up iteratively by adding weak learners (e.g., decision trees) to it. The additive model may be described as follow in the context of diagnosing Parkinson's disease using Gradient Boosting:

$$F(x) = \Sigma\_i \, f \, i\_(x)$$

4. AI for Parkinsons Detection: Clinical Implementation Options

4.1 Logistic regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:

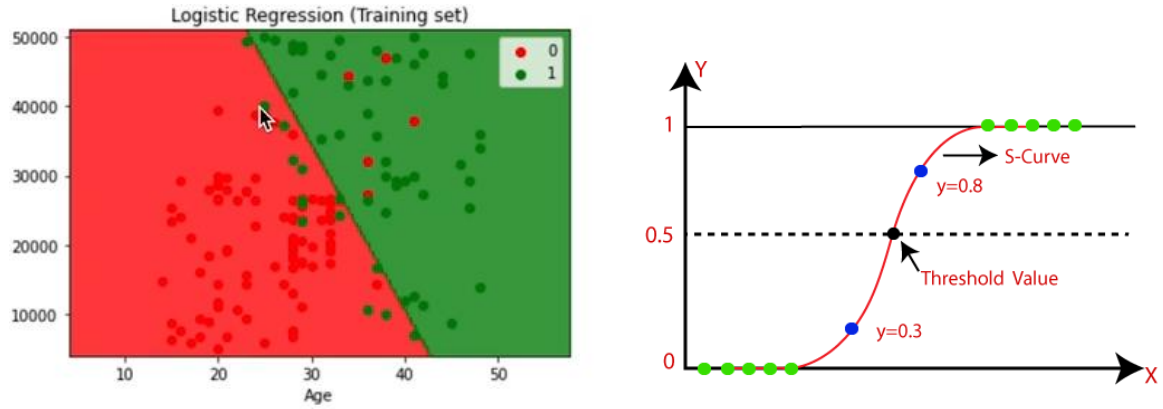


Figure 1: Logistic Regression

4.2 Decision Tree

Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. bThe decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

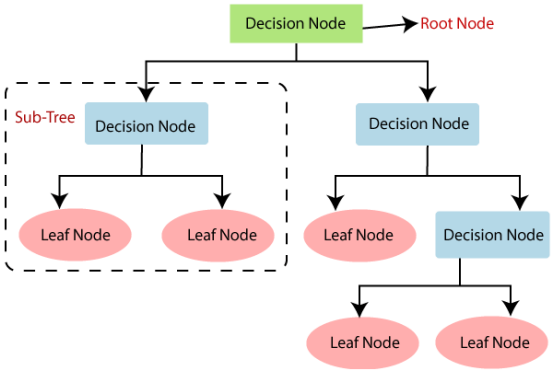


Figure 2: Decision Tree

4.3 Support Vector Machine (SVM)

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

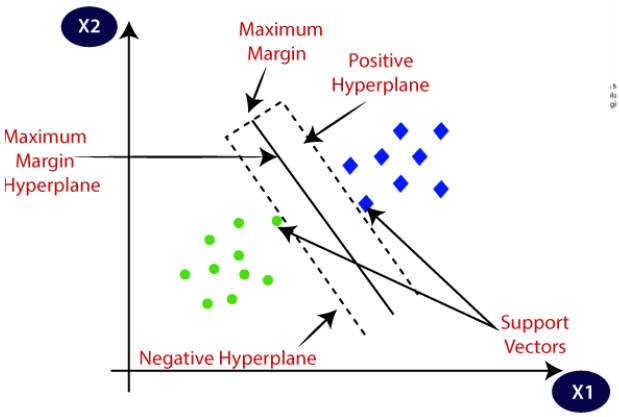


Figure 3: Support vector Machine

4.4 Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is

based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

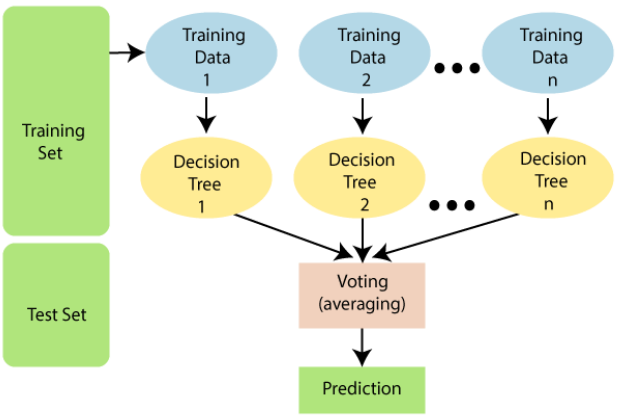


Figure 4: Random Forest Algorithm

5. Analysis

We separated the dataset into two parts: training data and testing data. The XGB Classifier was then initialised, and the model was trained in the ML Ensemble Learning group. After that, we discovered,

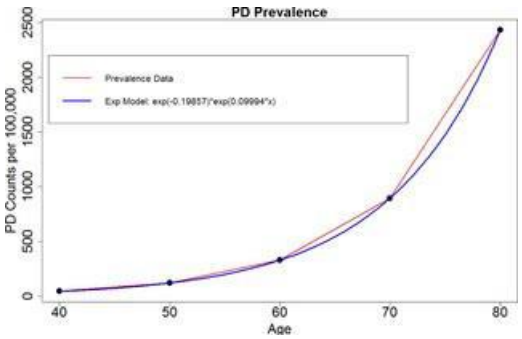


Figure: Increasing PD patient count with age

Finally, we generated y_pred (predicted values for x_test) and calculated the accuracy of our model. Comparing XGBooster with other algorithms the accuracy, precision, recall, etc. is very praiseworthy. XGBooster is not only able to keep up with all those other algorithms but exceeds them in performance which led us to pick this algorithm as the base of our project which deals with detecting Parkinson's disease.

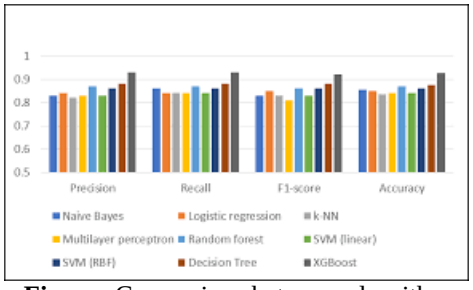


Figure: Comparison between algorithm

6. Conclusions

In conclusion, the results of this project show that machine learning algorithms can be used to accurately diagnose Parkinson's disease. The random forest algorithm performed the best in terms of accuracy, followed by the decision tree, SVM, and logistic regression algorithms. Future work can include testing the algorithms on other datasets to evaluate their performance and comparing their performance with other machine learning algorithms.

To summarise, XGBoost is a sophisticated machine learning algorithm that has demonstrated encouraging results in the detection of Parkinson's disease. XGBoost may be used to create reliable prediction models that can uncover trends and essential illness traits. Researchers may use XGBoost to accurately categorise people as healthy or afflicted by Parkinson's disease, which can aid in early identification and treatment of the condition. XGBoost's performance, like that of any other machine learning technique, is dependent on the quality and quantity of data used to train the model, as well as the selection of relevant features. As a result, when training an XGBoost model to identify Parkinson's disease, the data preparation and feature selection procedures should be carefully considered.

Acknowledgment
We want to convey our thanks to our Panel Head, Dr Naresh, Professor, Department of Networking and Communications, SRM Institute of Science and Technology, for their inputs during the project reviews and support. We register our immeasurable thanks to our Faculty Advisor, Dr. KNimala, Asst Professor, Network and Communication, SRM Institute of Science and Technology, for leading and helping us to complete our course.

Our inexpressible respect and thanks to my guide, Dr. Naresh, Asst Professor, Network and Communication, SRM Institute of Science and Technology, for providing me with an opportunity to pursue my project under his/her/their mentorship. He/She/They provided me with the freedom and support to explore the research topics of my interest. Her/His/Their passion for solving problems and making a difference in the world has always been inspiring.

We sincerely thank the Networking and Communications Department staff and students, SRM Institute of Science and Technology, for their help during our project. Finally, we would like to thank parents, family members, and friends for their unconditional love, constant support, and encouragement.

References

[1] J. S. Hawley, "What is Parkinson ' s disease ?," *Park. Dis. Improv. Patient Care*, vol. 501, no. c, pp. 1–6, 2014.
[2] J. Browniee, "A Gentle Introduction to XGBoost for Applied Machine Learning," *Mach. Learn. Mastery*, pp. 1–20, 2016, [Online]. Available: https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/.
[3] S. Bind, A. K. Tiwari, and A. K. Sahani, "A Survey of Machine Learning Based Approaches for Parkinson Disease Prediction," *Int. J. Comput. Sci. Inf. Technol.*, 2015.
[4] M. Ene, "Neural network-based approach to discriminate healthy people from those with Parkinson's disease," *Ann. Univ. Craiova, Math. Comp. Sci. Ser*, 2008.
[5] S. Shetty and Y. S. Rao, "SVM based machine learning approach to identify Parkinson's disease using gait analysis," 2016, doi: 10.1109/INVENTIVE.2016.7824836.
[6] D. Chang, M. Alban-hidalgo, and K. Hsu, "Diagnosing Parkinson ' s Disease From Gait," *Stanford*, 2015.
[7] S. V. Perumal and R. Sankar, "Gait monitoring system for patients with Parkinson's disease using wearable sensors," 2016, doi: 10.1109/HIC.2016.7797687.
[8] F.-T. S., G. N., P. C., H. T., G. L., and H. J.M., "Effect of gait speed on gait rhythmicity in Parkinson's disease: Variability of stride time and swing time respond differently," *J. Neuroeng. Rehabil.*, 2005.
[9] Nikisha Jadhav *1, Sangita Phad*2, SnehalKamble*3, Detecting Parkinson's Disease using xgboost classifier machine.
[10] G. Litjens, T. Kooi, B.E. Bejnordi, et al. A survey on deep learning in medical image analysis
[11] Raksha Sharma, Detecting Parkinson's Disease using machine learning.