WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Digital Report Prediction

[1]Bharath SRS, [2]Chowdri SS, [3]Aravind M, [4]Aravinth P,[5] Vimala P

[1,2,3,4,] Student, CSE Department, Dhirajlal Gandhi College Of Technology Salem, India

[5] Assistant Professor of CSE Department , Dhirajlal Gandhi College Of Technology, Salem, India

[1] srsbharath557@gmail.com, [2] sschowdri@gmail.com, [3] luciferaravind1811@gmail.com, [4] pparavind813@gmail.com, [5] vimala.cse@dgct.ac.in

ABSTRACT:

Around the globe, thousands of people worldwide are suffering by Parkinson's Disease (PD), a central nervous system degenerative condition. Early detection and diagnosis of PD is crucial for successful treatment and management of the disease. In past few years, Machine learning (ML) algorithms has shown great potential in predicting PD based on various physiological and neurological markers. In this disease prediction system, a system is proposed using ML-based approach to predict the presence of PD in patients. The system employs various machine learning models, including Gradient Boosted Tree, random forest, and logistic regression, to identify key markers and patterns associated with the disease. Overall, this disease prediction system provides a valuable tool for early detection and diagnosis of PD, which can lead to better management and treatment of the disease. The proposed approach can also be extended to other neurological disorders, providing a general framework for disease prediction and diagnosis.

Keywords – Machine Learning Prediction, Gradient Boosted Tree, Random Forest, Logistic Regression Model, Parkinson Disease

1. Introduction :

Thousands of people globally suffer from Parkinson's disease (PD), a devastating neurological disorder. Dopaminergic neurons gradually disappear in PD, which causes a some of physical and non- physical symptoms include tremors, rigidity, and poor balance. Early detection and diagnosis of PD are critical for successful treatment and disease management.

The use of machine learning technology emerged as a promising approach to predicting PD in patients. With the increasing availability of large-scale datasets and advancements in machine learning algorithms, it is now possible to create reliable prediction algorithms that can identify individuals at high risk of acquiring PD before symptoms even manifest.

In [1] Study, these models are typically based on a range of clinical, genetic, and imaging data that can be used to identify key biomarkers associated with PD. Machine learning algorithms can then be trained to analyze these biomarkers and identify patterns and associations between them and the development of PD. In last few years, machine learning (ML) algorithms has shown great potential in predicting and diagnosing PD based on various physiological and neurological markers.

These ML-based approaches can analyze large and complex datasets, identify patterns and relationships between different variables, and make accurate predictions about the presence of PD. One popular approach to predicting PD using machine learning is through the use of support vector machines (SVMs). An example of a supervised learning algorithm is the SVM, which can be trained to categorize data based on particular features. SVMs can be trained on enormous datasets of patient data in the case of PD prediction in order to pinpoint specific traits that are indicative of PD.

2. Related Work

This paper [3] surveys various machine learning algorithms for predicting Parkinson's disease. Among them are Decision tree models, random forests, machines with support vectors, artificial neural networks and other algorithms. The algorithms' accuracy spans from 70% to 99%, with certain algorithms performing better than others.

The study in paper [4] found that SVM showed good accuracy (88.9%) compared to other algorithms, and Random Forest had the highest accuracy of 90.26% while Naïve Bayes had the lowest level accuracy of 69.23%. Hierarchical clustering and SOM were also used, predicting higher numbers of clusters in healthy datasets.

In the paper [5], with 34 support vectors, the Nu-SVM model depending on the Gaussian method was shown to have the maximum sensitivity and overall accuracy. The research presents an ensemble learning approach for utilising machine learning to predict early warning signals of Parkinson's disease. The proposed model surpasses existing approaches such as SVM, KNN, RF, DT, MLP, SC, and LR, with an accuracy of 94.87%.

In this study [6], functional MRI (fMRI) data were used to discover brain activity patterns linked to optimal and non optimal deep brain stimulation (DBS) settings in Parkinson's disease (PD) patients. achieving 88% accuracy in forecasting optimal vs. non-optimal circumstances.

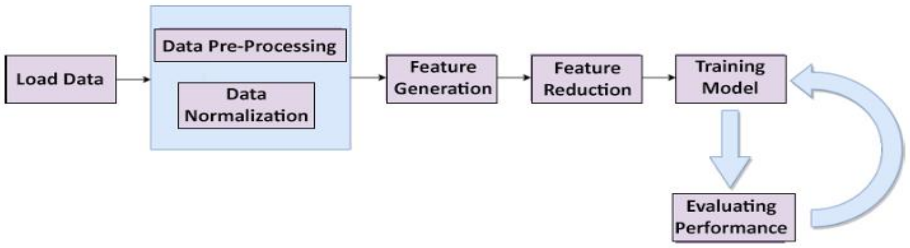
In paper [7] The analysis reveals that patients can be classified into three subtypes of PD: slow progressors, moderate progressors, and fast progressors. The approach can aid in the interpretability of clinical features and disease progression. The algorithms used were unsupervised learning and mathematical projection.

This paper [8] presents a study of Parkinson's Disease (PD) diagnosis using voice and tremor data. For tremor data, kNN achieved the highest accuracy of 98.5% for 2-level classification and 90% for 5-level classification. By combining both voice and tremor data, an accuracy of 99.8% was achieved using ensemble averaging of kNN, SVM, and naive Bayes for PD detection. The study uses kNN, SVM and Naive Bayes algorithms for classification. The highest accuracy for male voice samples was found to be 90.3% in kNN, and for female voice samples, it was 95.8% in kNN. In tremor data, the maximum accuracy for PD vs non-PD classification was 98.5% in kNN.

This paper [9] presents a multimodal machine learning model for predicting the risk of Parkinson's disease. The model was developed using an open-source auto-ML package called GenoML and was validated in an external cohort. The model outperformed previous efforts with an accuracy of 89.72% and was based on a combination of clinico- demographic, genetic, and transcriptomic data.

3. System Design

Ensemble Classifier, Support Vector Machine, and Decision Tree were employed in the study [15]. The classifiers' performance is measured using accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F-score. The suggested method is developed in MATLAB 2018 with multiple classifier parameter values. The Ensemble Classifier with 30 learners produced the maximum accuracy of 94.7%. The flow of building the Parkinson system is depicted in Fig. (2).



3.1 Load Data

Loading data in Spark involves various steps such as creating a SparkSession, specifying the data source format, defining the schema or letting Spark infer it, specifying options like delimiter, header, encoding, etc., and reading the data into a DataFrame. Once the data is loaded, it can be processed and transformed using Spark's distributed computing capabilities. It's important to ensure the data is clean, consistent, and in the right format before loading it into Spark for optimal performance and accurate results.

3.2 Data Preprocessing

A format is machine learning systems can easily analyse. Noise, inconsistencies, missing values, and irrelevant characteristics in raw data can all have an influence on the performance of machine learning algorithms. Data preparation is necessary for operations such as purifying the data and preparing it for a machine learning model, which improves the model's accuracy and efficacy. Data preprocessing entails a procedure of cleaning, transforming, and reducing data. Mathematical concepts such as statistics, linear algebra, and probability theory are used to perform these operations. For example, mean, median, mode, standard deviation, correlation matrices, and matrix operations are used to handle missing values, identify outliers, and perform scaling and normalization.

3.3 Data Normalization

Data normalization is a common data preprocessing technique used in machine learning to transform the data to a common scale or range. This is important because many machine learning algorithms assume that all input variables are on the same scale. If variables are not on the same scale, some variables may have a greater impact on the algorithm, leading to inaccurate or biased results. One common technique for data normalization is min-max scaling, which scales the data to a range between 0 and 1. The formula for min-max scaling is:

$$x_{norm} = \frac{x}{x}$$

Techniques can be applied to individual variables or entire datasets. Data normalization helps to ensure that all variables are given equal importance during machine learning analysis, leading to more accurate and reliable results.

3.4 Feature Generation

Feature generation is a technique of creating additional features or variables with the help of already-existing data which will result in machine learning models performance. This can be achieved by combining, transforming, or extracting features that may not have been originally present in the data. Feature generation is particularly useful when the original features are not sufficient to accurately represent the underlying patterns or relationships in the data.

One common technique for feature generation is polynomial feature expansion, which involves creating new features by raising the existing features to various powers. For example, if we have a single feature X, we can generate a new set of features by squaring, cubing, and so on, to create new features X^2, X^3and so on. This can be expressed mathematically as:

$$x_{new} = X^2 X^3 \ldots\ldots X^n$$

3.5 Feature Reduction

Feature reduction in machine learning is the process of lowering the amount of features or variables in a dataset in order to make data processing easy to compute and improve the efficacy of models. This can be achieved by eliminating irrelevant or redundant features that do not contribute to the accuracy of the model or may even negatively impact it.
One common technique for feature reduction is principal component analysis (PCA), It entails converting the data into a lower-dimensional space while keeping as much of the data's variance as possible. This can be expressed mathematically as:

$$x_{new} = X \times W$$

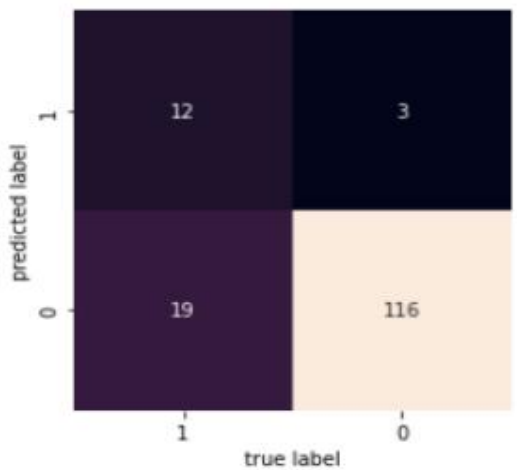
In equation (3) x_{new} is the new feature set, X is the original feature set, and W is the matrix of principal components that capturesthe maximum amount of variance in the data. Another common technique for feature reduction is feature selection, which involves selecting a subset of features that are most relevant to the model. This can be achieved through various methods such as correlation analysis, mutual information, or regularization.

3.6 Training Modal

Training a machine learning model involves using mathematical algorithms to adjust its parameters for accurate predictions on new data, based on concepts like linear algebra, calculus, probability, and statistics. The model is trained with labeled data to minimize differences between predicted and actual outputs. Once sufficiently accurate, the model can be used to make predictions on new data. The algorithms used to train the Model are:
1) Logistic Regression: It's a statistical technique that's applied in binary classification jobs with the objective of estimating the likelihood of an event happening based on input data. Logistic Regression is a statistical technique used in machine learning to predict the probability of a binary outcome. It represents the relationship between one or more independent variables and a binary dependent variable. The mathematical equation for logistic regression and this value represents the estimated probability of the outcome is:

$$\mathrm{p}\,(y = 1x) = 1 + \frac{1}{-(\mathrm{b0+b1\times1\ldots.+bk\times xk})(1+\mathrm{e})}$$

In equation (4) P(y=1|x) is the probability of the binary outcome being 1 given the input variables x, b0, b1, b2, ..., bk are the coefficients of the model learned during training, xk are the input variables, and e is the mathematical constant approximately equal to 2.71828. The output of a confusion matrix provides information about the true positive (12), true negative (116), false positive (3), and false negative (19) predictions made by the model. The confusion matrix aids in determining the model's accuracy, precision, sensitivity and specificity which are describes in Table 1



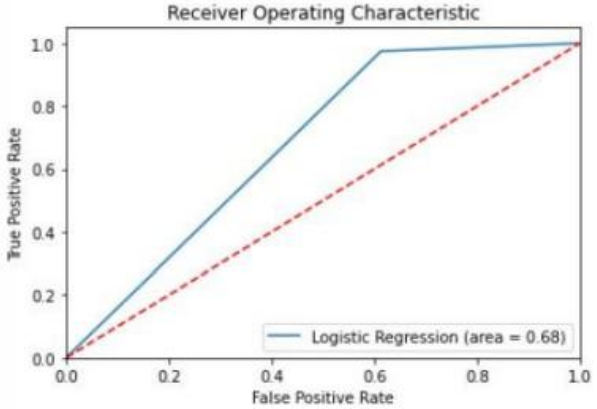
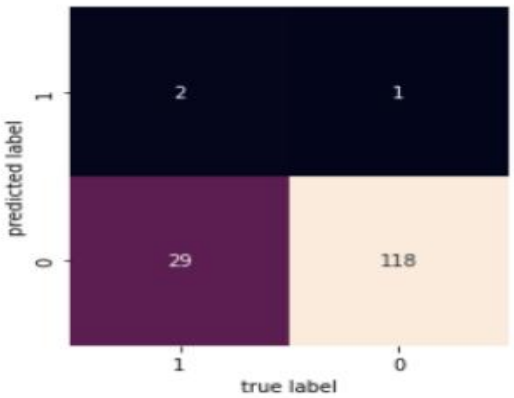


Fig. 3. Confusion Matrix of Logistic Regression **Fig. 4. ROC Curve of Logistic Regression**

2) Random Forest: A algorithm in machine learning called Random Forest is used for classification and regression tasks. To produce a final forecast, the system constructs a large number of decision trees and combines their results. The approach builds a forest of decision trees, where each tree is educated using a random subset of input features and training data. During prediction, the forest combines all of the individual trees' predictions to produce the ultimate outcome. Combining decision tree with ensemble learning equations yields the mathematical formula for Random Forest. By averaging the forecasts of each individual tree in the forest, the algorithm's output is produced. The output of a confusion matrix provides information about the true positive (2), true negative (118), false positive (1), and false negative (29) predictions made by the model. The confusion matrix aids in determining the model's accuracy, precision, sensitivity and specificity which are describes in Table 1.



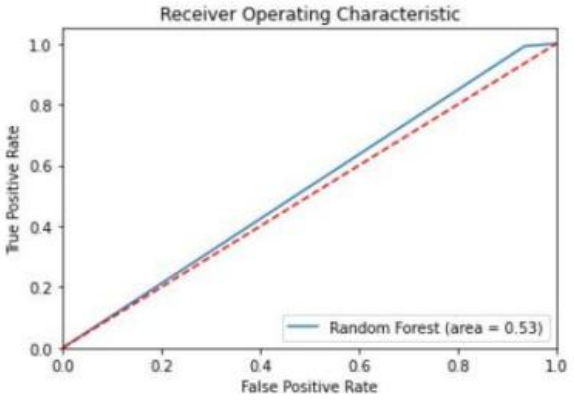
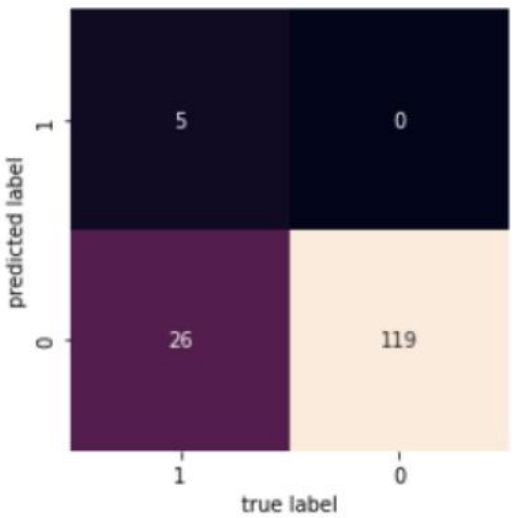


Fig. 5. Confusion Matrix of Random Forest **Fig.6. ROC Curve of Random Forest**

3) Gradient Boosted Tree: A machine learning algorithm called Gradient Boosted Trees is used for regression and classification tasks. Multiple decision trees are combined in this ensemble learning technique to produce predictions. The algorithm works by training decision trees in a sequential manner, where each subsequent tree learns from the errors ofthe previous tree. The algorithm minimizes the loss function by adding a new tree at each iteration, with the goal of reducing the residual errors. The mathematical equation for Gradient Boosted Trees involves calculating the sum of the output values of multiple decision trees. The output of each tree is weighted by a learning rate and added to the sum. In summary, Gradient Boosted Trees is a powerful algorithm that can handle both numerical and categorical data and can be used for a variety of machine learning tasks. Its strength lies in its ability to minimize the loss function by combining the output of multiple decision trees. The output of a confusion matrix provides information about the true positive (5), true negative (119), false positive (0), and false negative (26) predictions made by the model. It provides several key metrics that help assess the model's accuracy, precision, sensitivity, specificity and effectiveness in making predictions, which are describes in Table 1.



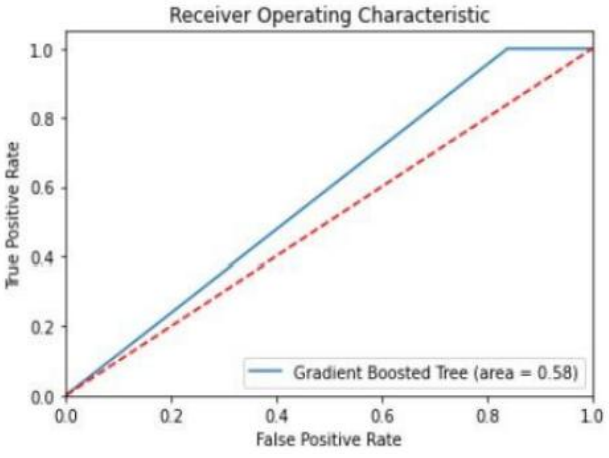


Fig. 7. Confusion Matrix of Gradient Boosted Tree **Fig. 8. ROC Curve of Gradient Boosted Tree**

Evaluating Performance: It's a statistical technique that's applied in binary classification jobs with the objective of estimating the likelihood of an event happening based on input data. Logistic Regression is a statistical technique used in machine learning to predict the probability of a binary outcome. It represents the relationship between one or more independent variables and a binary dependent variable. The mathematical equation for logistic regression is:

TABLE 1

PERFORMANCE METRICS

Model	Accuracy	Positive Precision	Negative Precison	Sensitivity	Specificity
Logistic regression	0.85	0.80	0.85	0.38	0.97
Gradient Boosted Tree	0.83	1.00	0.82	0.16	1.00
Random forest	0.80	0.67	0.80	0.06	0.99

Table showing five Performance Metrics with respective to algorithms used

4. Conclusion :

In conclusion, the project on Parkinson disease prediction system using machine learning has the ability to enhance illness identification and treatment dramatically. Machine learning algorithms can effectively predict the possibility of an individual having Parkinson's disease by analysing several traits and symptoms of the condition. The method can also assist in determining the stage of the disease and its severity, allowing doctors to give patients with personalised treatment options.

This project can be of great significance in the medical field, as it can help doctors to detect Parkinson's disease at an initial stage when it is most treatable. Additionally, the prediction system can be used to identify potential risk factors and develop preventative measures. The implementation of this system could lead to better patient outcomes, and ultimately, contribute to reducing the overall burden of Parkinson's disease on individuals and healthcare systems.

While this project is a significant step forward, there is still room for improvement. Further research and development could lead to an even more accurate prediction system by incorporating additional data sources and refining the machine learning algorithms used.

In conclusion, the machine learning-based Parkinson disease prediction system has the ability to have a big influence on Parkinson's disease identification and management, and it is an intriguing subject for future study.

5. REFERENCE :

1. Surekha Tadse, Muskan Jain, Pankaj Chandkheda "Parkinson's Detection Using Machine Learning" Published in: 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) by IEEE.
2. Haewon Byeon “Development of a depression in Parkinson's disease prediction model using machine learning” in World J Psychiatry. 2020 Oct 19; 10(10): 234–244. Published online 2020 Oct 19.
3. Alexandre Boutet, Radhika Madhavan, Gavin J. B. Elias, Suresh E. Joel, Robert Gramer, Manish Ranjan, Vijayashankar Paramanandam, David Xu, Jurgen Germann, Aaron Loh, Suneil K. Kalia, Mojgan Hodaie, Bryan Li, Sreeram Prasad, Ailish Coblentz, Renato P. Munhoz, Jeffrey Ashe, Walter Kucharczyk, Alfonso Fasano & Andres M. Lozano. Article number: 3043 (2021)” Predicting optimal deep brain stimulation parameters for Parkinson’s disease using functional MRI and machine learning” Open Access article Published: 24May 2021.
4. Diba Ahmadi Rastegar, Nicholas Ho, Glenda M.Halliday & Nicolas Dzamko « Parkinson’s progression prediction using machine learning and serum cytokines. Open Access article by Nature.com Published: 25 July 2019