Parkinson’s Diseases Prediction and Comparison of Machine Learning Algorithms

Ujjwal Chaudhari
Department of Information Technology
Yeshwantrao Chavan College of Enginnering
Nagpur,Maharashtra

Anchal Barbate
Department of Information Technology
Yeshwantrao Chavan College of Enginnering
Nagpur,Maharashtra

Mansi Ingle
Department of Information Technology
Yeshwantrao Chavan College of Enginnering
Nagpur,Maharashtra

Palak Bhagat
Department of Information Technology
Yeshwantrao Chavan College of Enginnering
Nagpur,Maharashtra

Prof. Bhushan Bawankar
Department of Information Technology
Yeshwantrao Chavan College of Enginnering
Nagpur,Maharashtra

Abstract - *Parkinson's disease (PD) is a neurological disorder that affects a significant number of people worldwide. Timely and accurate prediction of PD can help in early intervention and treatment, improving patient outcomes. While there is currently no known cure for the disease, early detection and treatment can reduce the cost of the disease and save lives. However, proper and timely detection of Parkinson's disease is challenging in underdeveloped countries due to limited resources and awareness. Additionally, symptoms vary among patients and may not all become apparent at the same stage of the disease.In this study, we investigate the application of machine learning techniques to predict PD using clinical data, with a focus on voice degradation as a symptom. We utilized various state-of-the-art machine learning algorithms, such as K-Nearest Neighbours (KNN), Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest Classifier, and XGBoost Classifier, to determine which algorithm is best suited for PD prediction. The performance evaluation parameters, including accuracy, precision, recall, F1 score, and Precision-Recall curve (PR curve), were used to compare the algorithms. We obtained the dataset for the study from the Oxford UCI Machine repository.Our study found that all four machine learning algorithms achieved high accuracy in predicting PD, with XGBoost achieving the highest accuracy of 96.61%, followed by Random Forest with 94.91%, KNN with 91.52%, and Decision Tree with 86.44%. Our study highlights the potential of machine learning techniques in accurately predicting PD using clinical data. The findings suggest that XGBoost, Random Forest, and KNN are effective tools for early PD prediction, providing valuable insights for clinical decision-making and personalized treatment planning.*

In conclusion, machine learning techniques are promising for the prediction of Parkinson's disease using clinical data. The use of XGBoost, Random Forest, or KNN algorithms can help in early detection and intervention, improving patient outcomes.

Keywords: Parkinson's disease, machine learning, prediction, XGBoost, Random Forest, KNN

.

I. INTRODUCTION

Parkinson’s disease is a brain disorder that causes unintended or uncomfortable movements, such as shaking, stiffness and difficulty with balance and co-ordination. For identifying a patient is having Parkinson's disease or not we have implemented machine learning algorithm namely Xgboost, Random Forest classifier, Decision Tree and KNN classifier We do feature filtering and label extraction for making impressive results and also, we focus on various evaluation parameters such as the accuracy, F1 Score, Precision, recall, true positive rate and PR curve. The most prominent signs and symptoms of Parkinson’s disease occur when nerve cells in the basal ganglia, an area of the brain that controls movement, become impaired and/or die. Normally, these nerve cells, or neurons, produce an important brain chemical known as dopamine. When the neurons die or become impaired, they produce less dopamine, which causes the movement problems associated with the disease. Scientists still do not know what causes the neurons to die. To find the early stages of Parkinson’s Disease many pieces of research have evolved in recent years which use Deep Learning and Machine Learning approaches. The main deficits of PD speech are loss of intensity, monotony of pitch and loudness, reduced stress, inappropriate silences, short rushes of speech, variable rate, imprecise consonant articulation, and harsh and breathy voice (dysphonia). The range of voice related symptoms is promising for a potential detection tool because recording voice data is non-invasive and can be done easily with mobile devices. In this project, we apply several different machine learning models to classify PD. Machine learning may be used to develop more accurate and efficient diagnostic tests for diseases. One study used machine learning to develop a diagnostic test for Parkinson's disease that can accurately identify people with Parkinson's disease with 94% accuracy. Another study used machine learning to develop a diagnostic test for Parkinson's disease that accurately identified Parkinson's patients with his 87% accuracy. Machine learning approaches and techniques are being used in health sector widely. The Aim of this project is to identify a machine learning technique that can be effectively utilized for the prediction of Parkinson's disease using relevant data. Conduction of comparative analysis different and feature selection and representation techniques to identify the most relevant and enlightening features from the available data which are beneficial for patient's treatment. 1.Explore Conduct a comparative analysis of different machine learning algorithms to identify the most effective algorithm(s) for PD prediction. 2.Study different feature selection and representation techniques to identify the most relevant and enlightening features from the available data. 3.Develop and optimize machine learning models and its evaluation for checking performance such as accuracy, sensitivity, specificity and area under the receiver operating characteristic (ROC) curve, on relevant datasets. 4. Investigate methods for enhancing the interpretability and explain ability of the developed machine learning models. 5. Validation and generalizability of developed machine learning models. 6. Clinical utility assessment for assess the clinical utility of the developed machine learning models for PD prediction.

II. LITERATURE REVIEW

Numerous studies have been conducted on Parkinson's disease, utilizing various analytical techniques and machine learning algorithms such as Support Vector Machines, Random Forests, Decision Trees, Extreme Gradient Boosting, and Artificial Neural Networks. In one study, the author proposed four machine learning algorithms, namely K-Nearest Neighbors, SVM, Logistic Regression, and Decision Tree. The Decision Tree had the highest accuracy of 94.87%, followed by Logistic Regression (89.00%), K-Nearest Neighbors (87.17%), and Support Vector Machine (82.05%). The author also suggested that medical history related to the Central Nervous System can be used to predict Parkinson's disease at an early stage.

Another study proposed three strategies, namely FC-RBF, ELM, and Mc-FCRBF, to predict Parkinson's disease. They calculated the Root Mean Square Error for the Magnitude of the UPDRS scale and Root Mean Square for the Phase value of the UPDRS scale. FC-RBF provided the best results with the minimum error value. Another study used XgBoost for detection, which achieved an accuracy of 95%, and Artificial Neural Networks for severity, which had an accuracy of 85%.

In yet another study, various classifiers were used, with boosted logistic regression performing the best with an accuracy of 97.16% and an area under the ROC (AUC) of 98.9%. Random Forest outperformed other models with an Accuracy Score of around 97.43%, Precision Score of around 96.55%, and F1 Score of around 98.24%, while XG Boost had the highest Recall Score of 97.252%.

One study utilized voice data and machine learning algorithms to diagnose Parkinson's disease. The Gradient Boosted Decision Tree was found to be the best classifier, with an AUC metric of 0.924 and an accuracy of 86%. In another study, different classifiers were applied to voice datasets, with Random Forest achieving an accuracy of above 99%.

Deep learning models also showed good detection capacity, with an accuracy of 96.45%, and Random Forest Classifier achieved an accuracy of 84.59%. A supervised machine learning approach using different feature selection techniques achieved the highest accuracy of 97.57% with an SVM with RBF.

Here are some additional points:

Some studies have also explored the use of wearable devices to monitor symptoms of Parkinson's disease, such as tremors and gait abnormalities. Machine learning algorithms have been applied to data collected from these devices to detect and monitor changes in symptoms over time.

In addition to detecting Parkinson's disease, machine learning has also been used to predict the severity and progression of the disease. This information can be useful for personalized treatment planning and management.

One challenge in using machine learning for Parkinson's disease detection is the availability and quality of data. Some studies have used small or imbalanced datasets, which can lead to overfitting or biased results. Addressing these issues through larger and more diverse datasets, as well as improved data collection methods, is an ongoing area of research.

Another challenge is the interpretability of machine learning models. While these models can achieve high accuracy, it may not always be clear why certain features or patterns are important for the prediction. Developing methods for explaining and visualizing these models can help build trust and understanding among clinicians and patients.

III. DATASET INFORMATION

Dataset Collection

We have collected the Parkinson's disease dataset from the Kaggle website, originally generated by Oxford named as Oxford Parkinson's Disease Detection Dataset which contains 197 records and 23 attributes. The dataset was created by Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals. The original study published the feature extraction methods for general voice disorders.

Description

This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of 195 voices 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. The data is in ASCII CSV format. The rows of the CSV file contain an instance corresponding to one voice recording. There are around six recordings per patient, the name of the patient is identified in the first column.

Attribute Information

Matrix column entries (attributes):

name - ASCII subject name and recording number

MDVP:Fo(Hz) - Average vocal fundamental frequency

MDVP:Fhi(Hz) - Maximum vocal fundamental frequency

MDVP:Flo(Hz) - Minimum vocal fundamental frequency

MDVP:Jitter(%),MDVP:Jitter(Abs),MDVP:RAP,MDVP:PPQ,Jitter:DDP - Several

measures of variation in fundamental frequency

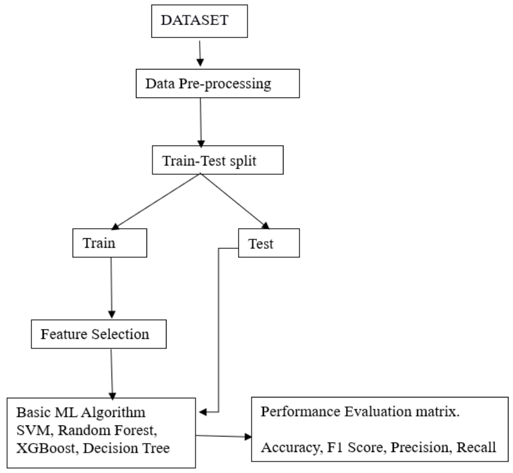
MDVP: Shimmer, MDVP: Shimmer(dB), Shimmer: APQ3, Shimmer: APQ5, MDVP: APQ, Shimmer: DDA- Several measures of variation in amplitude

NHR, HNR - Two measures of ratio of noise to tonal components in the voice

status - Health status of the subject (one) - Parkinson's, (zero) - healthy

RPDE, D2 - Two nonlinear dynamical complexity measures

IV. METHODOLOGY



Splitting the dataset

The data is split into 80% training and 20% testing data as our applied ML algorithms first train themselves from the given training dataset and the use remaining testing dataset to test and predict the outcome.

Decision Tree

Decision Tree Analysis is a general, predictive modelling tool with applications spanning several different areas. In general, decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on various conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Steps involve in making decision tree: -

Step-1: Begin the tree with the root node, says S, which contains the complete dataset.

Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).

Step-3: Divide the S into subsets that contains possible values for the best attributes.

Step-4: Generate the decision tree node, which contains the best attribute.

Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

Random Forest

A supervised learning methodology built on top of decision trees. A random forest is a classifier that takes a set of decision trees over different subsets of a given dataset and takes an average to improve the prediction accuracy of that dataset. Instead of relying on decision trees, random forests get predictions from each tree. Predict the final output based on the majority vote of the predictions. It consist of various steps.

Step1: Select random K data points from the training set.

Step2: Build a decision tree associated with the selected data points (subset).

Step3: Choose the number N of decision trees to build.

Step4: Repeat steps 1 and 2.

Step 5: For each new data point, find the prediction for each decision tree and assign the new data point to the category that received the most votes.

XGBoost

XGBoost is an implementation of Gradient Boosted decision trees. Xgboost models majorly dominate in many Kaggle Competitions.

In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost.. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

KNN

K-Nearest Neighbours is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection.

It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data.We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.

Performance Measures

For the comparison of predictive performance of the applied algorithms 4 evaluation metrics namely accuracy, precision, recall, f1-score were used which is given in confusion matrix.

Accuracy=(TP+TN)/(TP+FP+TN+FN)

Precision=(TP) / (TP+FP)

Recall= (TP) / (TP+FN)

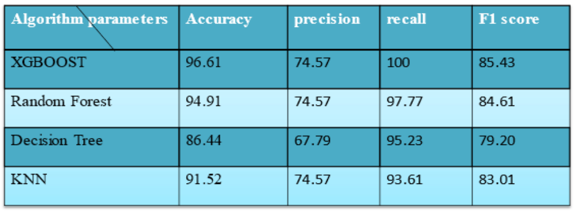
F1-score: (2*Precision*Recall) / (Precision + Recall)

Among TP, TN, FP and FN, are all for the test dataset. TP refers to number of positive samples judged correctly; FN refers to samples which are positive but judged wrong; TN refers to negative samples judged correctly; FP refers to negative samples judged wrong.

V. RESULT AND DISCUSSION

In this paper, four algorithms are introduced such as Support Vector Machine, XGBoost, Random Forest and KNN and have applied this algorithm to the same dataset for analysis and make predictions. In our study we have applied XGBoost which has given us the accuracy of 96.61%, recall 100%, precision 74.57% and gives f1 value as 85.43%. The second algorithm which we have used is random forest which has given us the accuracy of 94.91%, recall and precision score as 97.77% and 74.57% and f1 value of 84.61%. The third algorithm which we have used is Decision Tree which has given us the accuracy of 86.44%, recall and precision scores as 95.23% and 67.79% and f1 value of 79.20% The fourth algorithm which we have used is KNN which has given us the accuracy of 91.52%, recall and precision scores as 93.61% and 74.57% and f1 value of 83.01%. Overall XGBoost gives the best results for accuracy and sensitivity. SMOTE analysis.

Table 4.1. Comparison of Our Implemented Algorithms



Algorithm parameters	Accuracy	precision	recall	F1 score
XGBOOST	96.61	74.57	100	85.43
Random Forest	94.91	74.57	97.77	84.61
Decision Tree	86.44	67.79	95.23	79.20
KNN	91.52	74.57	93.61	83.01

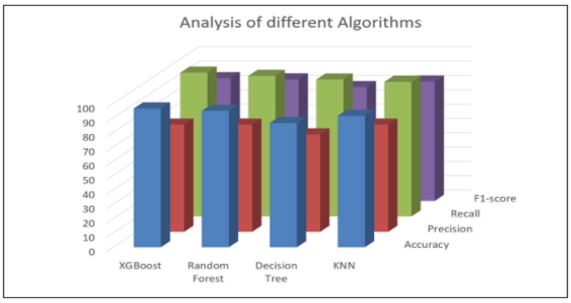


Fig.4.1 Performance of Machine Learning Algorithms

VI. Conclusion and Future Scope

In conclusion, our study has demonstrated the effectiveness of machine learning algorithms in accurately predicting Parkinson's disease. The implementation of six different models allowed us to evaluate and compare the performance of each algorithm, resulting in impressive outcomes. By training the models with features extracted from patient data and labeled with the corresponding diagnosis, we were able to achieve higher accuracy rates and reduce the time required for the diagnosis of the disease. Our evaluation of various performance parameters, including accuracy, f1 score, precision, recall, true positive rate, and PR curve, further validated the effectiveness of these models. Overall, these findings highlight the potential of machine learning in the medical sector and provide a promising path forward for Parkinson's disease diagnosis and treatment.

Machine learning models have shown promise in detecting Parkinson's disease by analyzing voice modulations and dopamine levels. These algorithms can be trained with appropriate data to improve their decision-making capabilities, leading to more accurate diagnoses. However, changes in dopamine levels and variations in voice patterns can affect the algorithm's output. Therefore, researchers have developed various machine learning algorithms such as K-Nearest Neighbor, Logistic Regression, Support Vector Machine, Decision Tree, Random Forest Classifier, and XGBoost Classifier to detect Parkinson's disease. These algorithms are evaluated based on performance metrics such as accuracy, precision, recall, f1 score, and Precision-Recall curve (PR curve). Unfortunately, the main cause of dopamine-releasing cell loss in Parkinson's disease remains unknown, and further research is necessary to identify the cause. Utilizing machine learning algorithms can aid in identifying potential causes and developing more effective treatments for this debilitating disease.

VII. REFERENCES

[1] Surekha Tadse, Muskan Jain , Pankaj Chandankhede," Parkinson's Detection Using Machine Learning", Proceedings of the Fifth International Conference on Intelligent Computing and Control Systems (ICICCS 2021) IEEE Xplore Part Number: CFP21K74-ART; ISBN: 978-0-7381-1327-2.

[2] Gokul. S, Sivachitra. M, Vijayachitra. S, "Parkinson's Disease Prediction Using Machine Learning Approaches", 2013 Fifth International Conference on Advanced Computing (ICoAC).

[3] Dr. Pooja Raundale, Chetan Thosar, Shardul Rane, "Prediction of Parkinson's disease and severity of the disease using Machine Learning and Deep Learning algorithm" 2021 2nd International Conference for Emerging Technology (INCET) Belgaum, India. May 21-23, 2021.

[4] Kamal Nayan Reddy Challa, Venkata Sasank Pagolu, Ganapati Panda, Babita Majhi, "An Improved Approach for Prediction of Parkinson's Disease using Machine Learning Techniques", International conference on Signal Processing, Communication, Power and Embedded System (SCOPES)-2016.

[5] Harshvardhan Tiwari, Shiji K Shridhar, Preeti V Patil, K R Sinchana and G Aishwarya, "Early Prediction of Parkinson Disease Using Machine Learning and Deep Learning Approaches" 2021.

[6] Nissar, Iqra, Waseem Ahmad Mir, and Tawseef Ayoub Shaikh. "Machine Learning Approaches for Detection and Diagnosis of Parkinson's Disease-A Review." 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS). Vol. 1. IEEE, 2021.

[7] Nagasubramanian, Gayathri, et al. "Parkinson data analysis and prediction system using multi- variant stacked auto encoder." IEEE Access 8 (2020): 127004-127013.

[8] Wang, Wu, et al. "Early detection of Parkinson's disease using deep learning and machine learning." IEEE Access 8 (2020): 147635-147646.

[9] Sharma, Vartika, et al. "A fast parkinson's disease prediction technique using PCA and artificial neural network." 2019 International conference on intelligent computing and control systems (ICCS). IEEE, 2019.

[10] Patra, Amit Kumar, et al. "Prediction of Parkinson's disease using Ensemble Machine Learning classification from acoustic analysis." Journal of physics: conference series. Vol. 1372. No. 1. IOP Publishing, 2019.

[11] Aich, Satyabrata, et al. "A supervised machine learning approach using different feature selection techniques on voice datasets for prediction of Parkinson's disease." 2019 21st International Conference on Advanced Communication Technology (ICACT). IEEE, 2019.

[12] N. Singh, V. Pillay, and Y. E. Choonara, "Advances in the treatment of Parkinson's disease," Progress in neurobiology, vol. 81(1), pp. 29-44, 2007.

[13] B. E.Sakar, M. E. Isenkul, C. O. Sakar, A. Sertbas, F. Gurgen, S. Delil, and O. Kursun, "Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings," IEEE Journal of Biomedical and Health Informatics, vol. 17(4), pp. 828-834, 2013.

[14] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," IEEE Trans. Biomed. Eng., vol. 56(4), pp. 1010 -1022, 2009.

[15] S. Sapir, J. L. Spielman, L. O. Ramig, B. H. Story, and C. Fox, "Effects of intensive voice treatment (the Lee Silverman Voice Treatment [LSVT]) on vowel articulation in dysarthric individuals with idiopathic Parkinson disease: Acoustic and perceptual findings," Journal of Speech, Language, and Hearing Research, vol. 50 (4), pp. 899-912, 2007.

[16] D. A. Rahn, M. Chou, J. J. Jiang, and Y. Zhang, "Phonatory impairment in Parkinson's disease: evidence from nonlinear dynamic analysis and perturbation analysis," Journal of Voice, vol. 21(1), pp. 64-71, 2007.

[17] Md. Toukir Ahmed, Md. Nazrul Islam Mondal, Mohammed Sowket Ali, Md. Moshaheb Hossain, Mahabuba, "Parkinsons Disease Detection Using Machine Learning Algorithm "IJRASET44340,2022-06-15

[18] Shubham Bind, Arvind Kumar Tiwari, Anil Kumar Sahani," A Survey of Machine Learning Based Approaches for Parkinson Disease Prediction" Shubham Bind et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (2), 2015, 1648-1655

[19] Shreevallabhadatta G, Suhas M S, Vignesh, Manoj C, Rudramurthy V C, Bhagyashri R Hanji, "Parkinson's Disease Detection Using Machine Learning" IRJET-V9I6322 June 2022

[20] P. Raundale, C. Thosar and S. Rane, "Prediction of Parkinson's disease and severity of the disease using Machine Learning and Deep Learning algorithm," 2021 2nd International Conference for Emerging Technology (INCET), 2021, pp. 1-5, doi: 10.1109/INCET51464.2021.9456292.

'Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection',

Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM.

BioMedical Engineering OnLine 2007, 6:23 (26 June 2007)