I. PARKINSON'S DISEASE DETECTION USING MACHINE LEARNING ALGORITHMS

Name : AbhinshyamK
Rollno: 54
Registration no: 12110384

Abstract

As the second most prevalent neurological ailment, Parkinson's disease (PD) is incurable and results in severe impairment as well as a diminished quality of life. It has an impact on the brain's nerve cells that generate dopamine, a substance that aids in coordinating movement. Individuals may find it difficult to move, talk, write, or finish basic tasks as their dopamine levels drop. Speech problems affect about 90% of individuals with Parkinson's disease. The incidence increases dramatically with age, with the average age of onset being approximately 70 years. But a tiny proportion have early-onset Parkinson's disease (PD) before turning 50. PD affects around 10 million people globally. Even though there is no known cure, motor symptoms can frequently be significantly improved by continuous research, drugs, or surgery.

If Parkinson's disease (PD) is to be prevented or treated, early detection is essential. PD is predicted using machine learning classification algorithms, such as ensemble learning approaches that integrate many models for increased accuracy, k-nearest neighbors, logistic regression, and decision trees. Using machine learning models could greatly improve Parkinson's disease diagnosis. The XGBoost classification algorithm outperformed other algorithms with a high test accuracy rate of 95% in a research that used data from the UCI Machine Learning Repository.

II. INTRODUCTION

Parkinson's Disease (PD) is a progressive neurodegenerative condition leading to the death of the dopamine containing cells.(DeMaagd & Philip, 2015). The diagnosis is primarily a clinical one based on a history and examination.(Parkinson's Disease in over 20s: Diagnosis and Management | Guidance | NICE, n.d.). Parkinson's disease was first described by Dr. James Parkinson in 1817 as a "shaking palsy." Since then, our understanding of the disease has evolved significantly. Researchers now believe the pathophysiological changes associated with Parkinson's may begin before the onset of motor symptoms, during a preclinical phase. This has driven interest in developing protective or preventive therapies. The loss of striatal dopaminergic neurons is the cause of Parkinson's disease (PD) motor symptoms, while neuronal loss in non-dopaminergic areas is also supported by the existence of non-motor symptoms. "Looks like Parkinson's disease" is what the word "parkinsonism" refers to. This indicates to neurologists that the patient is stiff, has a slightly flexed posture, moves slowly, and typically walks slowly with little to no arm motion. The most frequent cause of parkinsonism is Parkinson's disease (PD), while there are a variety of secondary causes as well, such as illnesses that mimic PD and drug-induced causes.(Fahira et al., 2023; Little et al., 2007; *Parkinson's Disease in over 20s: Diagnosis and Management | Guidance | NICE*, n.d.; *[PDF] Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection | Semantic Scholar*, n.d.; Tiwari et al., 2021).

Doctors conduct a thorough clinical analysis of a patient's medical history, as this is crucial for accurate diagnosis. Unfortunately, this method alone is highly inaccurate. A study by the National Institute of Neurological Disorders found that early diagnosis (within five years of symptoms

appearing) is only 53% accurate, which is not much better than random guessing. However, early diagnosis is vital for effective treatment.

The global prevalence of Parkinson's disease (PD) has rapidly increased, with more than 10 million people worldwide living with the condition. PD progresses through five stages and affects over 1 million individuals annually in India. In the United States, it affects approximately 500,000 to 1 million people, or about 1% of those over 60 years old. While PD incidence rises with age, around 4% of PD cases are diagnosed before age 50. Men are 1.5 times more likely than women to have Parkinson's disease.(*Parkinson's Disease Statistics | Parkinson's News Today*, n.d.). While Parkinson's disease cannot be cured, medications can effectively manage its symptoms. These medications can help individuals with Parkinson's disease (PD) cope with issues related to movement, walking, and tremors by either increasing dopamine levels or acting as substitutes for dopamine. In certain cases, surgery may be recommended. Despite extensive research, the exact cause of PD remains unknown, and diagnosis can sometimes be challenging.(*(PDF) Performance Evaluation of Machine Learning Algorithms in the Classification of Parkinson Disease Using Voice Attributes*, n.d.).

An efficient diagnostic method for determining Parkinson's disease (PD)-related speech problems is to use speech signals to categorize those with PD and those who are healthy. For this goal, numerous machine learning-based classification techniques have been put out in the literature. Because machine learning (ML) is highly accurate and simple to use, it is being used more and more in the diagnosis of medical diseases. Furthermore, ML has been used to treat Parkinson's disease (PD), according to published reports.(Fahira et al., 2023). (Tsanas et al., 2012)used a data set consisting of 263 speech samples from 43 people and 76.7 % of dataset were PD, the leftover data set was healthy. (Little et al., 2009) carried out a test to identify dysphonia as a means of differentiating between people with Parkinson's disease (PD) and healthy people. Vowels were recorded in their dataset, which comprised 23 PD patients and 8 healthy individuals for the study. They classified data using a Support Vector Machine (SVM) and were 91.4% accurate. In addition, they used several classifiers to distinguish between PD and healthy participants, including AdaBoost, K-Nearest Neighbors (K-NN), Multilayer Perceptron (MLP), and Naïve Bayes (NB). According to the study, phonation is the best job for identifying Parkinson's disease. K-NN, MLP, Optimum Path Forest, and SVM were assessed as classifiers for the ML-based diagnosis of Parkinson's disease (PD) in a different study. In this study, artificial neural networks were used to minimize voice features.(Ioffe & Szegedy, 2015; Parisi et al., 2018).

III. MATERIALS AND METHODS

A. *Description of Dataset.*

ATTRIBUTE	DESCRIPTION
MDVP: Fo (Hz)	Average vocal fundamental frequency
MDVP: Fhi (Hz)	Maximum vocal fundamental frequency
MDVP: Flo (Hz)	Minimum vocal fundamental frequency
MDVP:Jitter(%) MDVP:RAP Jitter:DDP MDVP:Jitter(Abs) MDVP:PPQ	several measures of variation in fundamental frequency.
MDVP:Shimmer Shimmer:APQ3 MDVP:Shimmer(dB) Shimmer:APQ5 MDVP: APQ Shimmer: DDA	Several measures of variation in amplitude.
PDE, D2	Two nonlinear dynamical complexity measures.
NHR, HNR	Two measures of the ratio of noise to tonal components in the voice
DFA	Signal fractal scaling exponent
Spread1, spread2, PPE	Three nonlinear measures of fundamental frequency variation.
Status	Health status of the subject (one)- Parkinson's, (zero) healthy.

Table 1: Parkinson's Disease dataset

The dataset used for the analysis is taken from the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals.(Little et al., 2009). 31 people, including 8 in the control group and 23 with Parkinson's disease (PD), had their biological voice measurements included in the dataset. The ASCII CSV format of the data contains 195 feature samples. The patient's name, three different fundamental frequency types (high, low, and average), multiple fundamental frequency variation measures (jitter), multiple amplitude variation measures (shimmer), two measures of the noise-to-tonal components ratio in the voice (NHR and HNR), two nonlinear complexity measures (RPDE and D2),

a signal fractal scaling exponent (DFA), and three nonlinear fundamental frequency variation measures (spread1, spread2, PPE) are mostly included in these features.

The primary objective of the dataset is to differentiate between healthy individuals and those with PD, based on the "status" column which is assigned a value of 0 for healthy and 1 for PD. Each patient has approximately 6 recordings. The dataset consists of around 75% PD cases and 25% healthy cases.(*Parkinsons - UCI Machine Learning Repository*, n.d.).

B. Feature Importance Analysis.

A group of techniques known as "feature importance" rate input features in a predictive model to show how important they are in relation to the predictions they make. Through the identification of traits that are more pertinent to the goal variable, this study provides insight into the dataset. The model can be better understood by looking at feature importance. Usually, a prediction model trained on the dataset computes these ratings. The association between several variables in the Parkinson's disease dataset is depicted in Figure 1.

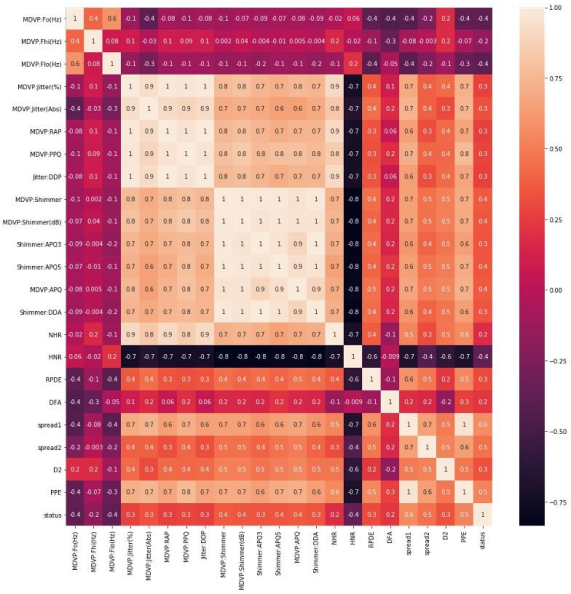

Fig 1: Correlation between features of Parkinson's disease.

C. Prediction Techniques

Logistic Regression:

In machine learning, the statistical technique known as logistic regression is frequently employed to address binary classification problems involving two class values. This prediction model works well with target variables that are categorical. When it comes to utilizing basic clinical variables to forecast the risk of major chronic diseases with low incidence rates, logistic regression can outperform other machine learning models. This approach requires the dependent variable to be binary, such as 0 or 1, and uses probability to evaluate the link between category dependent variables.

Decision Tree:

A supervised learning technique for both regression and classification problems is the decision tree algorithm. It is a predictive modelling technique used in machine learning, data mining, and statistics. A decision tree classifier uses conditional control statements to assist it reach a conclusion by partitioning the input data into sub-spaces according to specific functions. Decision nodes, which are used to make decisions and have several branches, and leaf nodes, which are the result of those decisions and have no further branches, make up the tree structure. Using basic decision rules that may be inferred from the data attributes, a decision tree aims to build a model that can forecast the value of the target variable.

Support Vector Machine (SVM):

In machine learning, the Support Vector Machine (SVM) is a widely used classification method. Although its main application is in classification, this supervised technique can also be utilized for regression applications. Because SVM can attain significant accuracy with little computer resources, it is often preferred.

Finding a hyperplane in an N-dimensional space—where N is the number of features—that successfully divides the data points into distinct classes is the main goal of support vector machines (SVM). This hyperplane, which serves as a border that effectively divides the two classes, is determined by the SVM classifier. SVM seeks to partition the dataset into classes while optimizing the maximum marginal hyperplane (MMH), or margin of separation. SVM's high generalization ability has made it successful in various classification tasks across different fields.

K-Nearest Neighbour (KNN):

Although it is generally used for classification, the K-Nearest Neighbors (KNN) algorithm is a straightforward machine learning method based on supervised learning. It may be applied to both regression and classification applications.

"K", which establishes the number of closest neighbors to consider when categorizing a new data point, is a crucial KNN parameter. The algorithm places the new data in the category that most closely resembles the existing categories based on the supposition that the new and current data are comparable.

KNN makes it simple to categorize new data into appropriate groups by storing all the existing data and classifying a new data point based on similarity. With bigger training datasets, this method can perform better and is resilient to noisy training data.

Steps include:

- Load the data.
- Initialize the value of K.
- Fitting the KNN algorithm to the training set.
- To predict a data, Calculate the distance between test data and each row of training data. Euclidean distance is used.
- Sort the calculated distance based on the distance value.
- Test the accuracy of the result.
- Visualizing the test-set result.

Random Forest Classifier:

A well-liked machine learning approach for classification and regression applications is the Random Forest classifier. As an ensemble learning technique, it improves accuracy by combining the predictions of several different individual models.

A group of decision trees is produced during training in a Random Forest. A random portion of the training data is used to train each tree, and at each node, a random selection of characteristics is used for splitting. This unpredictability enhances generalization and lessens overfitting.

Every tree in the forest makes an independent forecast about the outcome during prediction, and the final prediction is decided by a majority vote in the case of classification or by averaging the predictions of all the individual trees in the case of regression. Usually, this ensemble method produces a more reliable and accurate than a single decision tree.

Extreme Gradient Boosting Algorithm (XGBOOST):

In machine learning, XGBoost is a very well-liked technique that may be used for both regression and classification problems. It is well known for outperforming other algorithms in terms of performance. XGBoost, short for Extreme Gradient Boosting, uses decision trees as its foundation. It is a highly accurate option due to its great predictive potential. The fact that XGBoost is around ten times faster than conventional gradient boosting methods also plays a part in its popularity. Furthermore, XGBoost is called a Regularized Boosting method for the accuracy that it gives.

IV. RESULT

In this study, I used a trustworthy dataset from the UCI Machine Learning Repository together with various prediction models to make predictions on Parkinson's illness. The dataset comprises 195 feature samples from 31 participants, including 8 in the control group and 23 with Parkinson's disease (PD). XGBoost attained the greatest test accuracy rate of 95% and the highest training accuracy rate of 100%, as Table 2 demonstrates.

Classification Techniques	Training Accuracy Rate	Test Accuracy Rate
LOGISTIC REGRESSION	0.88	0.79
DECISION TREE	1.00	0.90
SUPPORT VECTOR MACHINE (SVM)	0.89	0.92
K-NEAREST NEIGHBOUR (KNN)	0.94	0.95
RANDOM FOREST CLASSIFIER	0.99	0.92
XGBOOST	1.00	0.95

Table 2: Result

Three sorts of vocal basic frequencies are distinguished: maximum, minimum, and average. To determine how important shimmer, jitter, and fundamental frequencies are in

predicting Parkinson's disease, performance analysis was done.

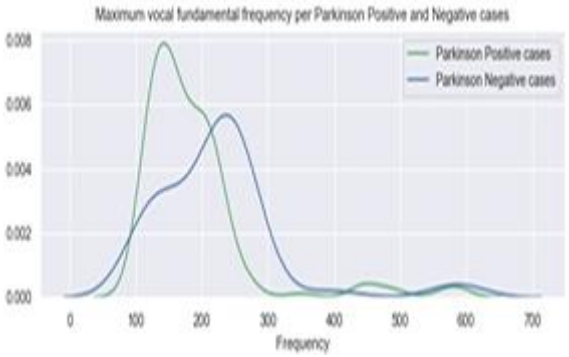

Fig 2: Distribution plot for Maximum vocal fundamental frequency

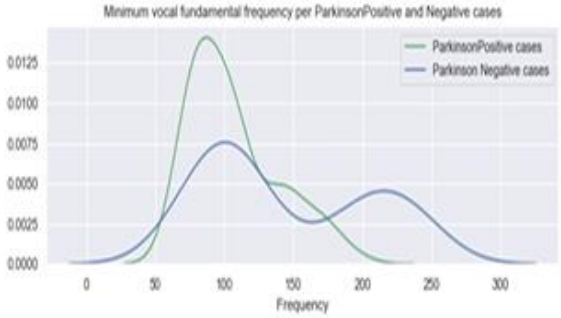

Fig 3: Distribution plot for Minimum vocal fundamental frequency

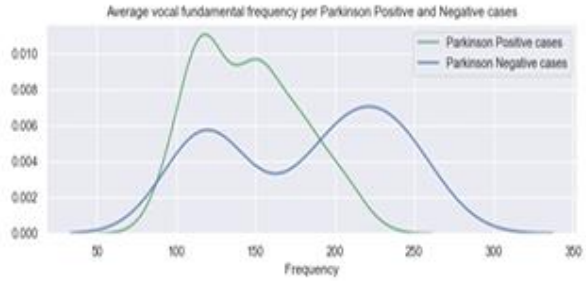

Fig 4: Distribution plot for Average vocal fundamental frequency

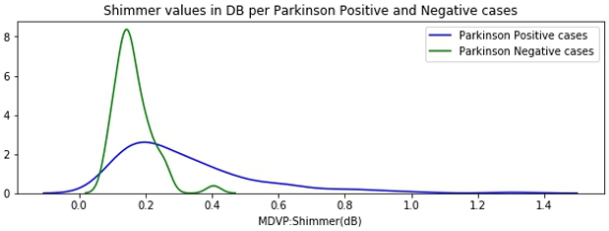

Fig 5: Distribution plot of shimmer values in DB Parkinson Positive and Negative cases.

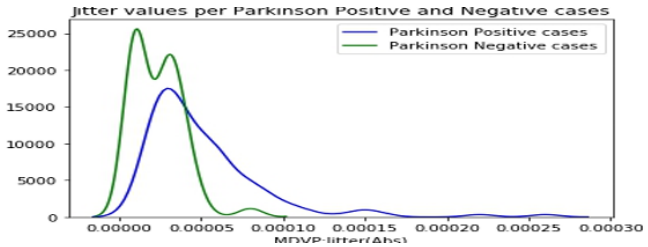

Fig 6: Distribution plot of Jitter values per Parkinson Positive and Negative cases.

Conclusion

This work investigates the use of six classification algorithms on a dataset to infer, from speech input parameters, whether an individual is healthy or suffering from Parkinson's disease. Several techniques are employed, including Random Forest, K-Nearest Neighbors (KNN), Decision Tree, Support Vector Machine (SVM), Logistic Regression, and Extreme Gradient Boosting (XGBoost). To ascertain the efficacy of these algorithms, the study compares their output. The findings imply that when compared to basic classification algorithms, ensemble approaches like XGBoost and Random Forest yield predictions that are more accurate. Evaluations are also conducted on the basis classification algorithms, such as SVM, KNN, Decision Tree, and Logistic Regression. The classifiers attain a range of 70% to 100% accuracy. With a test accuracy rate of 100% and a training accuracy of 100%, and with a test accuracy of 95%.

References:

DeMaagd, G., & Philip, A. (2015). Parkinson's Disease and Its Management: Part 1: Disease Entity, Risk Factors, Pathophysiology, Clinical Presentation, and Diagnosis. *Pharmacy and Therapeutics*, *40*(8), 504. /pmc/articles/PMC4517533/

Fahira, N. R., Lawi, A., & Aqsha, M. (2023). Early Detection Model of Parkinson's Disease Using Random Forest Method on Voice Frequency Data. *Journal of Natural Sciences and Mathematics Research J. Nat. Scien. & Math. Res*, *9*(1), 30–38. http://journal.walisongo.ac.id/index.php/jnsmr

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *32nd International Conference on Machine Learning, ICML 2015*, *1*, 448–456.

Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., & Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Transactions on Bio-Medical*

Engineering, *56*(4), 1015. https://doi.org/10.1109/TBME.2008.2005954

Little, M. A., McSharry, P. E., Roberts, S. J., Costello, D. A. E., & Moroz, I. M. (2007). Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection. *BioMedical Engineering OnLine*, *6*. https://doi.org/10.1186/1475-925X-6-23

Parisi, L., RaviChandran, N., & Manaog, M. L. (2018). Feature-driven machine learning to improve early diagnosis of parKinson's disease. *Expert Systems with Applications*, *110*, 182–190. https://doi.org/10.1016/J.ESWA.2018.06.003

Parkinsons - UCI Machine Learning Repository. (n.d.). Retrieved April 26, 2024, from https://archive.ics.uci.edu/dataset/174/parkinsons

Parkinson's disease in over 20s: diagnosis and management | Guidance | NICE. (n.d.).

Parkinson's disease in over 20s: diagnosis and management | Guidance | NICE. (n.d.). Retrieved April 26, 2024, from https://www.nice.org.uk/guidance/cg35

Parkinson's Disease Statistics | Parkinson's News Today. (n.d.). Retrieved April 26, 2024, from https://parkinsonsnewstoday.com/parkinsons-disease-statistics/

[PDF] Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection | Semantic Scholar. (n.d.). Retrieved April 26, 2024, from https://www.semanticscholar.org/paper/Exploiting-Nonlinear-Recurrence-and-Fractal-Scaling-Little-McSharry/27e1dcd0d64bfc9d936e597d4f29b80c21571936

(PDF) Performance evaluation of machine learning algorithms in the classification of parkinson disease using voice attributes. (n.d.). Retrieved April 26, 2024, from https://www.researchgate.net/publication/322557855_Performance_evaluation_of_machine_learning_algorithms_in_the_classification_of_parkinson_disease_using_voice_attributes

Tiwari, H., Shridhar, S. K., Patil, P. V, Sinchana, K. R., & Aishwarya, G. (2021). *EasyChair Preprint Early Prediction of Parkinson Disease Using Machine Learning and Deep Learning Approaches*.

Tsanas, A., Little, M. A., McSharry, P. E., Spielman, J., & Ramig, L. O. (2012). Novel speech signal processing algorithms for high-accuracy classification of Parkinsons disease. *IEEE Transactions on Biomedical Engineering*, *59*(5), 1264–1271. https://doi.org/10.1109/TBME.2012.2183367