Machine Learning for Parkinson's Disease Prediction

Azizkhan F Pathan
Dept. of CSE
JIT Davanagere

Muskan M
Dept. of CSE
JIT Davanagere

Pooja K
Dept. of CSE
JIT Davanagere

Manu B M
Dept. of CSE
JIT Davanagere

Shreya B H
Dept. of CSE
JIT Davanagere

***Abstract*— Parkinson's disease (PD) is the second most prevalent neurological ailment that causes considerable impairment, has no cure, and causes reduction in quality of life of a patient. Dopamine, a neurotransmitter produced by nerve cells, is produced in this part of the brain. Dopamine is a neurotransmitter that helps regulate and coordinate body actions by acting as a messenger between brain and nervous system parts. As dopamine levels in the brain decline, it becomes more difficult to speak, write, walk, or complete other simple tasks. Approximately 90% of persons with Parkinson's disease experience speech problems. There is currently no cure for Parkinson's disease, but research is ongoing, and drugs or surgery can sometimes provide significant relief from movement symptoms. Even it causes loss of life. As a result, detecting it at an earlier stage may assist to prevent or mitigate the symptoms. The machine learning classification methods are used to determine if a person has Parkinson disease or not. The different machine learning techniques such as Support Vector Machine and XGBoost are utilized in this work. The experimental results show that the XGBoost algorithm outperforms the Support Vector Machine in terms of accuracy.**

Keywords: Parkinson disease, Machine Learning, Support Vector Machine, XGBoost.

I. INTRODUCTION

Parkinson's disease is a neurodegenerative disease in which the patient's motor abilities decrease over time as dopamine-producing brain cells are damaged. Tremors, stiffness, and difficulty walking, balancing, and coordinating are all symptoms of this neurological disorder [1]. Parkinson's disease symptoms typically begin gradually and worsen over time. As the disease progresses, people with Parkinson's disease may have difficulty walking and communicating. Some of the symptoms of this disorder include tremors, difficulty moving, behavioural difficulties, dementia, and depression. When the primary motor symptoms occur together, they are referred to as "Parkinsonism" or "Parkinsonian Syndrome."

In the current method, PD is only recognised in the secondary stage (Dopamine deficiency), posing medical problems. In addition, doctors must manually review and suggest medical diagnoses in which symptoms may differ from person to person, making pharmaceutical recommendations difficult [2]. As a result, mental diseases are poorly understood and have numerous health consequences. The following clinical procedures are commonly used to diagnose Parkinson's disease:

- Magnetic resonance imaging (MRI) or computed tomography (CT) scan - Conventional MRI cannot detect early indicators of Parkinson's disease.
- PET scan - this test is used to determine the activity and function of brain regions involved in movement.
- SPECT scan - can identify alterations in brain chemistry, such as dopamine deficiency.
- This leads to a high rate of misdiagnosis (up to 25% by non-specialists), and people can have the condition for many years before being diagnosed. As a result, the current system is ineffective in providing early warning and correct medical diagnosis to those who are afflicted.

Machine learning is the study of computer algorithms that can learn and develop on their own with experience and data. It is considered to be a component of artificial intelligence. Machine learning algorithms create a model based on training data to make predictions or judgments without having to be explicitly programmed to do so. Machine learning algorithms are utilised in a wide range of applications, including medicine, email filtering, speech recognition, and computer vision, where developing traditional algorithms to do the required tasks is difficult or impossible.

This research uses machine learning approaches to detect and identify Parkinson's disease in its early stages. Recurrent Neural Networks (RNN), which are employed by Apple's Siri and Google's voice search, are used to process sequential input. It's the first algorithm to recall its input using an internal memory, making it perfect for machine learning issues that require sequential data.

II. LITERATURE SURVEY

Caliskan et al. [3] suggested a DNN with stacked auto encoder and softmax classifier cascaded to one another for prediction on the Istanbul PD dataset and OPD dataset with 10 fold cross validation 30 times on the Istanbul PD dataset and OPD dataset. When compared to traditional classifier techniques, DNNs use auto encoders to reduce the dimension of features and softmax layers for classification. To demonstrate the efficiency of the deep neural network

classifier, several simulations are run on two datasets. On the PD and OPD datasets, the proposed DNN models outperform SVM, DT, and NB classification methods, with accuracy of 65.549 percent and 86.095 percent, respectively. The suggested DNN classifier has the capacity to extract hidden features, which improves the classifier's performance.

Bereus.et.al [4] used the LOSO technique to apply multiple ANNs to the PD Dataset. LOSO has a lot less bias and can anticipate nearly anything. The method involves picking features and then using a majority vote methodology to select the best result from multiple ANN classifiers. Pearson and Kendall's correlation coefficients, pca, and self-organizing maps were used to choose features. According to Pearson and Kendall's correlation coefficients, the number 4 and short sentence 4 from the voice samples had better significance in identifying PD. NN has been fine-tuned, and it now has a test accuracy of 86.47 percent. The main flaw is that the performance of ANNs could be improved by employing other feature selection processes and doing more fine-tuning work.Several voice tests in other languages were not included in the study, and using those datasets to classify the models would assist expand their reach in predicting PD.

Ali et al. [5] presented automatic identification of Parkinson's disease based on several types of extended phonation using linear discriminant analysis and a genetically tailored neural network. The PD in this situation lacked generality, had low prediction accuracy, and had issues like subject overlap. To predict PD, a hybrid intelligent system is developed that uses LDA and genetic algorithms to reduce dimensionality and optimises hyperparamters of neural networks. After balancing the gender unbalanced dataset, the proposed model had a low complexity and delivered accuracy of 82.14 percent on the test dataset. This results in a generalised model that highlights the gender disparity in the Istanbul PD dataset and how these associated variables could be removed to achieve high accuracy. The method was not employed for complex tasks such as prodromal and differential diagnosis. As test data, an independent dataset was employed that was only obtained from PD patients and is severely skewed. There was no consideration for missing information concerning the feature extraction procedure, such as the extraction of features corrected for pitch.

On the UCI Parkinson speech sample dataset, Raza Rizval et al. [6] employed DNN and LSTM to detect and forecast Parkinson disease. The sizes of the first, second, and third hidden layers for DNN were found to be 256, 128 and 64. To avoid overfitting, a 0.5 ADAM optimizer dropout was utilised, as well as categorical cross entropy loss. Dropout also ensured that the model is not reliant on a single node. In the hidden and output layers, the activation functions relu and softmax were utilised. A batch size of 16 is recommended for LSTM. When trained for 80 epochs, a hidden layer size of 32 was obtained. These proposed models outperformed methodologies that had previously been used and studied.

ErdogduSakar.et.al [7] examined the collecting and analysis of a Parkinson Speech Dataset with Multiple Types of Sound Recordings. They presented a dataset that consisted of voice samples from 40 people, 20 of whom were healthy and 20 of whom had Parkinson's disease. On the classification algorithms KNN and SVM, we compared and summarised leave one subject out validation approaches. The s-loso approach outperformed the loo method significantly. Rather of considering each and every sample of a subject, mean and standard deviation were shown to be preferable metrics for summarising the information gathered from the speech samples.. The model's main flaw is that it performed poorly with KNN utilising the loso and s-loso schemes, because vowels were given greater weight in PD classification than sentences or words.

Eskodere et al. [8] proposed the use of a random subspace classifier ensemble to detect Parkinson's disease using vocal data. To increase the performance of individual classifiers, use the random subspace ensemble approach using knn. KNN, LDA, and QDA were employed as the basis classifiers with the help of k-fold cross validation. These were applied to the Istanbul PD dataset's normalised features. They were put to the test on a variety of features and with different numbers of KNN, LDA, and QDA learners, with the ensemble of KNN learners outperforming the LDA and QDA. On 7 dimensional subspaces, it was discovered that using k = 10 and 144 KNN classifiers resulted in the lowest classification error.The random subspace classifier ensemble method's performance was also demonstrated to be affected by variance and the type of base classifier used. The main disadvantage is the random selection of feature subspaces, as some of the subsets chosen at random may have poor discrimination capacity.

For Parkinson disease categorization, Li.et.al [9] described simultaneously learning of speech feature and segment. Feature selection algorithm that generates new hybrid features for classification without transforming existing features. After constructing hybrid features by merging features and segments, the Istanbul Dataset was separated into training and test datasets. After normalising with loso, hybrid voice features were chosen for generating a new training set. On datasets, the SVM classifier was used to classify them. In comparison to corrcoef, the results showed that the linear kernel outperformed the rbf kernel, and that p value and sdc were mean of 82.5% accuracy, 85% sensitivity and 80% specificity. Main drawback is very few samples were considered for feature selection. New samples could be obtained for additional testing and modification.

Behroozi.et.al [10] developed a Multiple-Classifier system for Parkinson's disease detection based on numerous voice tests. Rather than looking at every single voice sample for Instanbul PD prediction, the authors separated each vocal sample from the PD datset and utilised classification algorithms. To choose features, the Pearson Correlation Coefficient was used. If the vocal tests lacked relevant features, MCFS and A-MCFS were used to select them based on common features and to eliminate unsuccessful voice samples. When the LOO CVV methodology was applied to KNN, SVM, Naive Bayes, and discriminant analysis classifiers, the A-MCFS method outperformed LOSO, s-LOSO, and MCFS techniques in terms of accuracy. A majority vote from all of the classifiers would determine the final classification result. The fundamental problem is that not all vocal words have the same discriminating ability; in fact, even vocal terms that have been

considered discriminating in the literature, such as vowel "a," have failed. Additional research on a variety of vocal terms from the proposed perspective, as well as a variety of vocal tests from various languages, can be looked into.

On the PD voice telemonitoring datasets, Behroozi.et.al [11] employed SAE for dimension reduction and applied several classification algorithms such as KNN, LDA, NB, LSBM, RSVM, CART, KELM, MSVM to predict Parkinson disease. The suggested SAE featured a batch size of 20 neurons and two hidden layers, each with 10,9,8 and 8,7,6 neurons. On the Istanbul PD dataset, SAE with KNN produced the highest accurate classification result. The proposed approach to remap temporal frequency characteristics has a small dimensional space.The main disadvantage is that a smartphone application was created to predict PD using this method, and it was discovered that noise must not be present when recording with a smartphone, and the microphone quality must be high for prediction using a smartphone application with the help of the web.

Guruler [12] suggested a complex-valued artificial neural network for Parkinson's disease diagnosis using a k-means clustering feature weighting strategy. A hybrid method for diagnosing Parkinson's disease has been developed using a mix of a k-means clustering-based feature weighting approach and a complex valued artificial neural network. The feature-based wights technique aids in high classification precision. The weighting approach collects similar data points and aids in the conversion of a nonlinear separable dataset into a linear separable dataset. The additional features were retrieved and transformed to a complicated number format, which was then fed into the neural network as input. With the tenfold CV approach, the proposed method produced a high accuracy of 99.52 percent, and with the 50-50 training testing data selection method, it achieved a high accuracy of 99.39 percent. Their method allowed for a quick and low-cost classification of Parkinson's disease.

III. PROPOSED METHODOLOGY

The proposed methodology for the early Detection and Prediction of Parkinson’s Disease is shown in Figure 1.

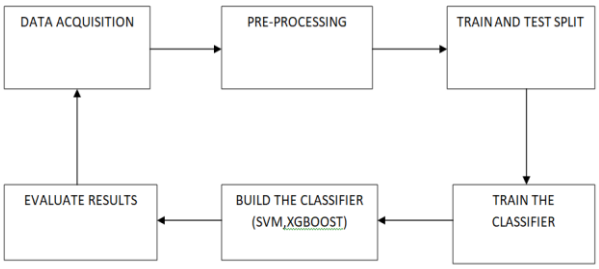

Figure 1: Parkinson's disease early detection and prediction methodology.

The Parkinson's dataset is the first item we need. It contains specific information on people who have Parkinson's disease and those who do not. The machine learning model can detect patterns in this data and determine what symptoms are frequent in people with Parkinson's disease and what symptoms are common in those who don't have Parkinson's disease. So, the machine learning algorithm can detect this ailment based on the data. This research uses the Voice Dataset for Parkinson's Telemonitoring from the UCI Machine Learning Repository [13]. A total of 31 patients' biomedical voice measures are included in the collection. Subject number, subject age, and vocal fundamental frequencies (MDVP) are among the data's many properties. There are 125 voice recordings of these people in the data collection. The information is stored in ASCII CSV format.

The input dataset needs to be pre-processed because it cannot be fed directly to machine learning models. The superfluous and missing values are removed from the dataset during processing. The dataset is then split into training and testing sets after it has been pre-processed. The training data is used to train the data, while the testing data is used to evaluate the model. After that, the training data is used to train the machine learning model, which is where the information is learned.

In this study, machine learning methods [14,15] like Support Vector Machine and XG Boost are utilised to detect and forecast Parkinson's illness.

SVM is a supervised machine learning technique that can be used to classify and predict data. Despite the challenges with regression, classification is the best fit. The SVM algorithm seeks to locate a hyperplane in an N-dimensional space that clearly categorises data points. The number of features determines the hyperplane's size. The hyperplane is essentially a line if there are just two input features. The hyperplane converts into a two-dimensional plane when the number of input features approaches three. It gets impossible to imagine when the number of features exceeds three.

XGBoost is a decision tree-based gradient boosting-based ensemble Machine Learning algorithm. In prediction problems involving unstructured data, artificial neural networks outperform all existing algorithms or frameworks (pictures, text, etc.). However, for small-to-medium structured/tabular data, decision tree-based algorithms are now rated best-in-class.

Ⅳ. RESULT AND DISCUSSION

Exploratory graphs are frequently created rapidly, and many of them are created while checking out data. Exploratory graphs are typically used to establish a personal understanding of data and to prioritize actions for follow-up. Figure 2 shows the exploratory graphs for the Parkinson’s dataset that is considered for the work.

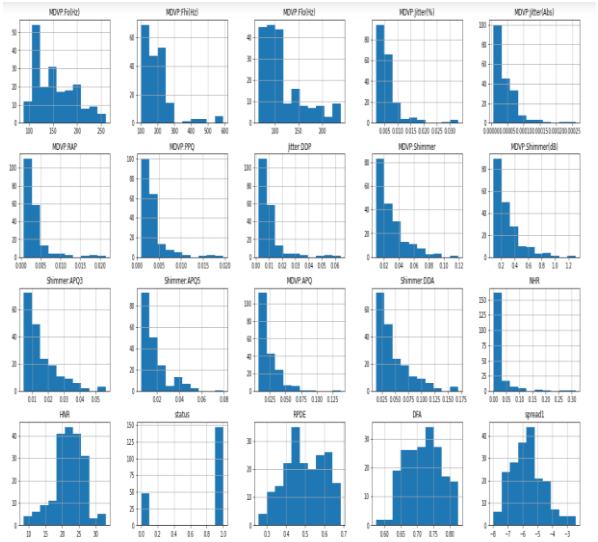

Figure 2. Exploratory graphs.

The location graph is only a visual depiction of the values of several attributes spaced out. Figure 2 shows the location graphs for the Parkinson's dataset that is considered for the work.

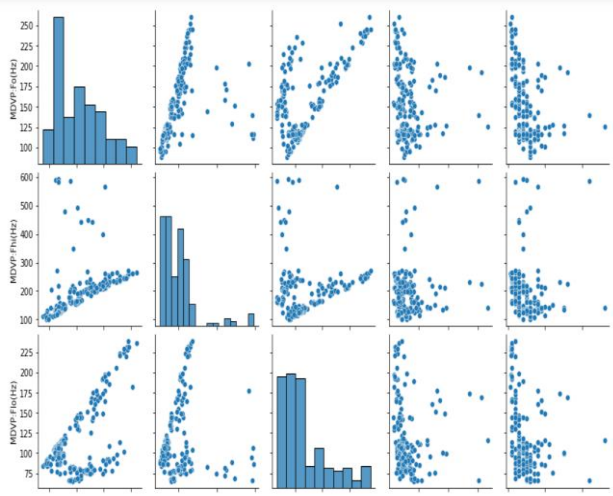

Figure 3. Location graphs.

A heat map is a graphical representation of data in which the value of the matrix is displayed using colours. Brighter colours, especially reddish colours, are used to symbolise more common values or higher activities, whilst darker colours are preferred to represent less common or activity values. The heatmap is defined by the name of the shading matrix. Figure 4 depicts the location graphs for the Parkinson's dataset used in this study.

Figure 4. Heat map.

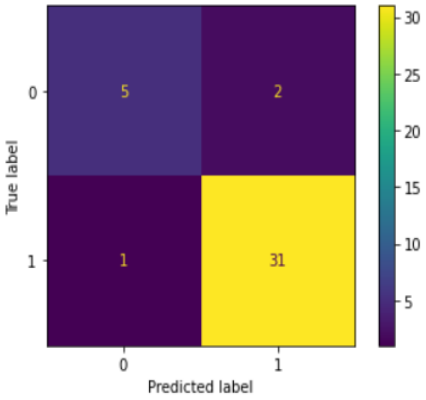

Figure 5: XGBoost's Confusion Matrix.

Figure 5 depicts the XGBoost classifier's confusion matrix. The outcomes of a prediction or estimate are displayed in this confusion matrix. The true label and the prediction label are used to create this matrix. This gives an accuracy of 94.87%. Table 1 shows the comparison of results of SVM and XGBoost algorithm. The XGBoost algorithm has better accuracy compared to SVM algorithm. Along with accuracy the Precision, Recall and F1-score results of both algorithms are also shown in the Table 1.

Table 1: Comparison of Results of SVM and XGBoost.

Model	Accuracy	Precision	Recall	F1-score
SVM	87.13	0.93	0.64	0.87
XGBoost	94.87	0.89	0.84	0.86

V. CONCLUSION

Parkinson's disease is a neurodegenerative disease in which the patient's motor abilities decrease over time as dopamine-producing brain cells are damaged. As a result, discovering it early on may help to prevent or lessen the symptoms. To identify whether a person has Parkinson's disease or not, machine learning classification algorithms such as Support Vector Machine and XGBoost are utilised. As a result, in terms of accuracy, the XGBoost algorithm outperforms the Support Vector Machine.

REFERENCES

[1] C. Blauwendraat, M.A. Nalls, and A.B. Singleton: Parkinson's disease genetic architecture. The Lancet Neurology, vol. 19, no. 2, pp. 170–178. (2020).

[2] Jayaprakash, S., Nagarajan, M.D., Prado, R.P.D., Subramanian, S. and Divakarachari, P.B., 2021. A systematic review of energy management strategies for resource allocation in the cloud:

Clustering, optimization and machine learning. Energies, 14(17), p.5322.

[3] Abdullah Caliskan, Hasan Badem, Alper Basturk, Mehmet Emin Yuksel. “Diagnosis of the Parkinson’s disease by using deep neural network classifier”.

[4] Lucijano Berus ,Simon Klancnik, Miran Brezocnik and Mirko Ficko. “Classifying Parkinson Disease based on measures using artificial neural networks”.2018

[5] Liaqat Ali, Ce Zhu, Zhonghao Zhang, Yipeng Liu. “Automated Detection of Parkinson's Disease Based on Multiple Types of Sustained Phonations”. 2019

[6] Danish Raza Rizvil , IqraNissar , Sarfaraz Masood Mumtaz Ahmed , Faiyaz Ahmad. “An LSTM based Deep Learning model for voice-based detection of Parkinson's disease”. 2020.

[7] Betul Erdogu Sarkar, M. Erdem Isenkul, C. Okan Sarkar, Ahmet sertbas, Fikret Gurgen, Sakir Delhi, Hulya Apaydin and Olcay Kursun. “Collection and Analysis of a Parkinson Speech Dataset with Multiple Types of sound Recordings”.2019.

[8] Omer Eskodere, Ali Karatutlu Cevet Unal.“Detection of Parkinson’s disease from vocal features using random subspace classifier ensemble”.2016.

[9] Youngming Li, Yunjian Jia, Xiaoheng Zhang, Cheng Zhang, Ping Wang. Tingjie Xie. “Simultaneously learning of speech feature and segment for classification of Parkinson disease”.2018.

[10] Mahnaz Behroozi and Ashkan Sami. “A Multiple-Classifier framework for Parkinson’s disease detection based on various vocal tests”.2019.

[11] Mahnaz Behroozi and Ashkan Sami. “Can a smartphone diagnose Parkinson disease,A deep neural network method and telediagnosis system implementation”.2018.

[12] "A novel Parkinson's disease detection system based on a complex-valued artificial neural network with k-means clustering feature weighting algorithm," says Huseyin Guruler. 2016.

[13] A. Benba, A. Jilbab, and A. Hammouch: Using pca and npca, voice assessments for detecting people with parkinson's illnesses. 743–754 in International Journal of Speech Technology, vol. 19, no. 4. (2016).

[14] Azizkhan F Pathan and Chetana Prakash. “Attention-based position-aware framework for aspect-based opinion mining using bidirectional long short-term memory”. Journal of King Saud University - Computer and Information Sciences. 2021. ISSN 1319-1578.

[15] Azizkhan F Pathan and Chetana Prakash. “Unsupervised Aspect Extraction Algorithm for opinion mining using topic modeling”. Global Transitions Proceedings. Volume 2. Issue 2. 2021. Pages 492-499. ISSN 2666-285X.