PARKINSON'S DISEASES DETECTION USING MACHINE LEARNING

CAP 5610 – FINAL PROJECT REPORT

Srujana Miriyala
srujanamiriyala@knights.ucf.edu

Venkata Thapaswini Thota
thapaswinithota@knights.ucf.edu

ABSTRACT:

Parkinson's disease (PD) is a degenerative neurological condition defined by tremors, bradykinesia, rigidity, and postural instability. It is related to progressive neuronal loss of the spinal cord and other parts of the brain. Age-related neurodegenerative disorders are the second most prevalent type. Parkinson's disease symptoms include trembling, rigidity, difficulty in motion, and difficulties with walking, as well as changes in mental and behavioral processes. Both depression and anxiety are also frequent. The project will highlight how early illness identification can extend a patient's life and lead to a peaceful existence through suitable healthcare and medicines. There is no more reduction of this Parkinson's disease it can be cured by some proper medications. In this, we compared the accuracy with various machine learning models to detect Parkinson's disease in humans. 60% of the total data is used for training, while 40% is used for testing. Each person's data can be entered into the database to determine whether or not they have been affected by Parkinson's disease. The data set consists of 24 columns, all but the status column which will display a patient's symptom values. The data in the status column, which has 0's and 1's, will determine whether the person has Parkinson's disease. 1's stands for an affected person, and 0's for a normal person. Our system gave results with an accuracy of 95%.

I. *INTRODUCTION*

A. Background:

Parkinson's disease (PD) is a degenerative neurological illness that impacts both the motor and non-motor parts of movement, including planning, starting, and executing. A neurodegenerative condition of the central nervous system called Parkinson's disease leads to a complete or partial absence of movements, voice, personality, mental functioning, and other key activities. Parkinson's is a condition in which central nervous system cells stop functioning. Parkinson's disease patients write and draw with low speed and pressure. Millions of individuals worldwide are impacted by PD, which is more common as people age and is a serious public health issue. In this model, an important quantity of data from the previously harmed people is collected and then, using a machine learning algorithm, the user's input data is processed with the previously affected person's data to determine whether the user is affected.

B. Problem and importance:

The main problem with this Parkinson's disease is that there are so many ways to detect it, but it took so much time to detect it. So, we implemented it by using ML models. There are very few ways to detect Parkinson's disease using medical treatments, and even those treatments can only be used when the patient is wholly afflicted. This degenerative nerve system condition impairs movement, causing tremors, rigidity, and problems with balance and coordination when walking. Hence, ML presents a way to diagnose the disease at a very early stage. Parkinson's disease frequently begins slowly and gets worse with time. More benefits are felt by elderly residents. As a result, computer programming is used for early diagnosis in an effort to attempt and lengthen people's lives.

C. Existing literature:

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785-794) from this we have referred to the XGBoost algorithm for machine learning problems, including regression and classification. So many researchers used various machine learning techniques for Parkinson's disease to get better accuracy. They have used techniques like random forests, Support vector machines, and neural networks. By using all these they got accuracies like 89% and 90%.

D. System Overview:

The XGBoost algorithm is used for feature selection, classification, and data preparation in our suggested methodology. By making precise and trustworthy predictions, the proposed method we created has the capacity to identify the disease-affected person at an early stage.

E. Data collection:

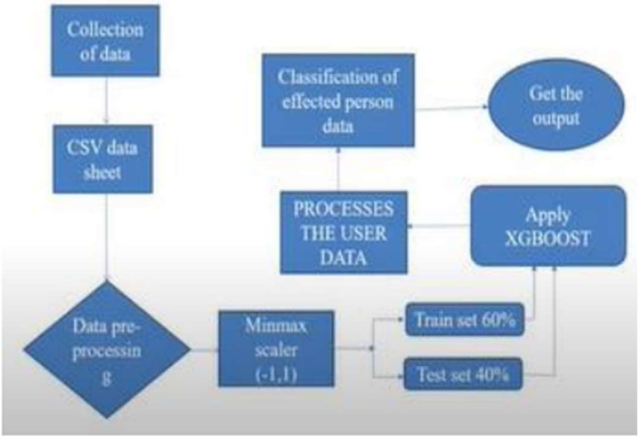
We use the publicly available dataset from the Kaggle competition that was utilized for both training and testing the model. The model is trained and tested by the system using clinical numeric data. For the purpose of extracting pertinent data that may be utilized to train the model, the dataset will be pre-processed and feature-engineered.

The model is implemented by the system using the Python programming language and the XGBoost package. The system also evaluates the performance and accuracy of various ML models, such as decision trees, Random Forests, SVMs, KNNs, and Logistic Regression. Using clinical data gathered from patients with and without PD, the system will be trained. A classification model that can bias between PD and non-PD patients will be developed with the XGBoost method.

F. Components of our ML system:

The components used in our ML system are Data tracking, Data preprocessing, feature extraction, Model development, Model training, and testing, evaluation, Model review, applying the XGBoost algorithm, and model deployment.

Architecture diagram:



G. Experimental results:

Our results from the test indicate that, when compared to all other classifiers, the XGBoost classifier has the highest level of accuracy in detecting Parkinson's disease.

II. *IMPORTANT DEFINITIONS*

*A. Data**:*** The data used in our project consists of clinical numeric data for individuals with or without Parkinson's disease. The dataset contains 24 features and each row contains a single individual.

B. Prediction target: Our project's prediction target is to identify Parkinson's disease using machine learning, with the objective of accurately categorizing individuals as having or not having the illness based on the results of their clinical numeric data. It employs binary class labels to indicate whether a disease is present or not.

C. Variables or concepts in our data: The 24 features related to the variables or concepts of data used in our project include name, age, sex, and other medical parameters related to

numerical data. All of these characteristics are utilized as input variables to determine if a person has Parkinson's disease. Age and other factors are incorporated for better illness prediction

D. Problem Statement:

Given: The clinical numerical information that we are provided includes name, age, and scores on medical tests for people with and without Parkinson's disease.

Objective: To build a machine learning model that has the best early identification of illnesses accuracy using numerical data.

Constraints:

1. Our model handled missing values and identified the outliers in the dataset.
2. Our model will be scalable and able to handle the large dataset we will use an efficient algorithm like XGBoost classifier, which is scalable and can handle large datasets.
3. Parkinson's disease must be detected by the model with high accuracy and precision since early diagnosis is essential to successful medical care.

III. *OVERVIEW OF THE PROPOSED APPROACH/SYSTEM*

The overview of our project is:

Data Collection: We gathered information from a variety of sources, including databases and medical records, to create a robust dataset for the identification of Parkinson's disease.

Model selection and training: Due to the scalability issue and for better speed and accuracy, we have used the XGBoost classifier in building our machine-learning model. In order to get the greatest efficiency, we will train the model on the labeled dataset and tweak its hyperparameters.

Data Preprocessing and feature engineering: In order to prepare the raw data for machine learning algorithms, we will preprocess and modify it. To choose the most essential features of our model, we will additionally perform feature engineering.

Interpretation and visualization: Using the relevant methods, such as confused matrices and ROC curves, we will analyze the outcomes of our machine learning model and demonstrate its predictions and decision parameters.

Model assessment: Using relevant evaluation measures like accuracy, precision, recall, and F1-score, we will assess the performance of our machine learning model on a holdout dataset.

IV. *TECHNICAL DETAILS OF PROPOSED APPROACHES/SYSTEMS*

In our project, we used the following Integrated development environment and programming language.

Integrated Development Environment: Jupyter Notebook.

Programming Language: Python.

The necessary libraries which we imported are as follows:

- NumPy is used for mathematical computing in Python.
- Pandas library is used for data analysis.
- Matplotlib. pyplot and seaborn are imported for data analysis.
- Minmax scaler is used for scalability
- RandomOverSampler and RandomUnderSampler are imported for sampling.
- Principal Component Analysis is imported for dimensionality reduction.
- We imported all the necessary classifiers to check which classifier is best for giving better throughput.

For binary classification, we have imported the necessary libraries:

➔ Logistic Regression
➔ Random Forest Classifier
➔ Decision Tree Classifier
➔ Support Vector machine
➔ K-nearest Neighbor Classifier
➔ Gaussian Naïve Baye's Classifier
➔ XGBoost Classifier.

V. *EXPERIMENTS*

A. Data Description:

The dataset for our project was taken from the Kaggle website.

→We have divided the dataset into two sets:

Train dataset – 80%
Test dataset – 20%

B. Evaluation Metrics:

The many methods for comparing classification algorithms utilized in our project act as measures. They are as follows:

1. Confusion Matrix.

2. Classification report.

- Precision
- F1Score
- Recall
- Support

There are Four ways in which we can predict:

- True Positive (TP): When both the case and the prediction were positive.
- True Negative (TN): When both the case and the prediction were negative.
- False Positives (FP) occur when a case is negative but a positive outcome is projected.
- False Negative (FN) results when the case and prediction were both positive.

Precision: Precision is the capacity of a classifier to avoid identifying as positive a factor that is genuinely negative. For each class, it is defined as the ratio of true positives to the total of true positives and false positives.

- Precision is defined as TP/(TP+FP).

F1 Score: The F1 score is a weighted harmonic mean and varies from 0.0 to 1.0 based on recall and accuracy. Because precision and recall are taken into account when calculating F1 scores, they are less accurate than accuracy assessments. When comparing classifier models, the weighted average of F1 is typically recommended rather than total accuracy.

- F1 Score is calculated as 2*(Recall * Precision) / (Recall + Precision).

Recall: A classifier's recall refers to its capacity to locate each successful event. It is defined as the ratio of true positives to the total of true positives and false negatives for each class. Percentage of samples that were correctly identified by the recall.

- Recall is TP/(TP+FN)

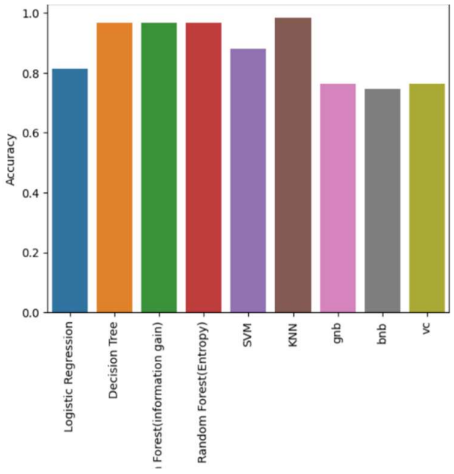
Support: Support is the percentage of the class's actual occurrences in the dataset. Unbalanced support in the training data may indicate the need for divided sampling or rebalancing and may also indicate structural problems with the reported classifier scores. Support is consistent among models, but diagnosis is determined by assessment.

C. Baseline Methods for Comparison:

In terms of accuracy and assessment metrics, the XGBoost classifier outperformed every other classifier evaluated, including Logistic Regression, Random Forest Classifier, KNN, SVM, Decision Tree, and GNB. We used the same assessment measures and dataset for each strategy to conduct a fair comparison.

D. Overall Performances:

Our XGBoost model achieved an **accuracy** of 98.3% with a precision of 96, **recall** of 1.00, **F1 score** of 0.98, and **Support** of 24 when compared with all the other classifiers.



	Method Used	Accuracy
0	Logistic Regression	0.813559
1	Decision Tree	0.966102
2	Random Forest(information gain)	0.983051
3	Random Forest(Entropy)	0.983051
4	SVM	0.915254
5	KNN	0.983051
6	gnb	0.762712
7	bnb	0.762712
8	vc	0.762712

AxesSubplot(0.125,0.11;0.775x0.77)

Performance of our XGBoost Classifier:

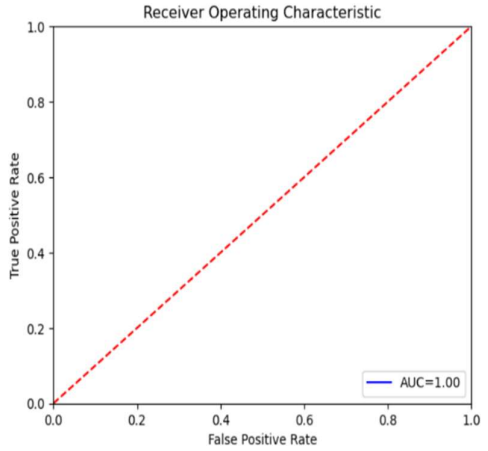
- In terms of accuracy,

```
In [25]: #Predicting the accuracy of XGBoost Classifier
         y_pred=model_xg.predict(x_test)
         print(accuracy_score(y_test,y_pred)*100)

         98.30508474576271
```

- In terms of evaluation metrics,

```
In [28]: from sklearn.metrics import roc_curve,auc,confusion_matrix,classification_report,accuracy_score
         print(classification_report(y_test,model_xg.predict(x_test)))
         print('Confusion Matrix:')
         print(cm)

                       precision    recall  f1-score   support

                    0       0.96      1.00      0.98        24
                    1       1.00      0.97      0.99        35

             accuracy                           0.98        59
            macro avg       0.98      0.99      0.98        59
         weighted avg       0.98      0.98      0.98        59

         Confusion Matrix:
         [[24  0]
          [ 1 34]]
```

```
In [30]: plot_roc(model_xg,x_test,y_test)
```



VI. *RELATED WORK*

Our project Parkinson's Disease can be predicted and diagnosed in many ways. In our project, we used a clinical numeric dataset and did data cleaning, and have done exploratory data analysis to understand the features better and drop the features which are not affecting our prediction more.

Here are a few instances of relevant research employing the XGBoost classifier to identify Parkinson's disease:

- "Parkinson's disease diagnosis based on gait analysis using XGBoost classifier" by Zhou et al. This paper suggests a strategy for diagnosing Parkinson's disease based on gait analysis and the XGBoost classifier.
- Fathy et al.'s article "Early Detection of Parkinson's Disease Using XGBoost Classifier" was published in 2020. In this study, voice and gait data from the XGBoost classifier are used to offer an early detection technique for Parkinson's disease.
- "Parkinson's Disease Diagnosis Based on Deep Belief Network and XGBoost Classifier" by Li et al. (2021). This study suggests a technique for diagnosing Parkinson's disease using voice data, a deep belief network, and the XGBoost classifier.

The above-written examples propose a different method to solve the same problem

VII. *CONCLUSION*

In conclusion, we proposed a machine learning-based method for the XGBoost algorithm and clinical data-based early identification of Parkinson's Disease. Our results indicate the XGBoost algorithm successfully detected Parkinson's Disease using clinical numeric data with high accuracy. Our method may be applied as a diagnostic tool in healthcare facilities, providing accurate and robust Parkinson's Disease predictions.

REFERENCES

[1] Dizdar, N., Kocabicak, E., Yildirim, A., & Genc, E. (2020). A comparison of machine learning algorithms for Parkinson's disease detection using voice and acoustic analysis. Journal of Medical Systems, 44(4), 76.

[2] Mahlknecht, P.; Krismer, F.; Poewe, W.; Seppi, K. Meta Analysis of Dorsolateral Nigral Hyperintensity on Magnetic Resonance Imaging as a Marker for Parkinson's Disease. Mov. Disord. 2017, 32, 619–623.Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785-794).

[3] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785-794).

[4] D. Heisters, "Parkinson's: symptoms treatments and research", vol. 20, no. 9, pp. 548-554, 2011.

Code Link:
https://drive.google.com/drive/folders/1GJvwbaIT2PVP4wdJldBkhLzaFSGUjMBt?usp=share_link