A PROJECT REPORT ON

Parkinson's Disease Prediction using Machine Learning

Submitted in partial fulfilment of the requirements for the award of the degree of

Bachelor of Technology

In

ELECTRONICS & COMMUNICATION ENGINEERING

Submitted By

<<1>ABIR LAL MANNA>	University Roll No. 12021002002035
<<2>RUPAYAN SAHA>	University Roll No. 12021002002037
<<3>SARTHAK CHAKRABORTY>	University Roll No. 12021002002027
<<4>DEBASMITA DAS>	University Roll No. 12021002002064

Under the guidance of

PROF. DEBANJANA GHOSH

Department of Electronics & Communication Engineering

UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA

University Area, Plot No. III – B/5, New Town, Action Area – III, Kolkata – 700160

CERTIFICATE

This is to certify that the project titled **anti-sleeping alarm mechanism for drivers** submitted by Abir Lal Manna(1), Rupayan Saha(31), Sarthak Chakraborty(37),Debasmita Das(12) Students of UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA, as submitted for the partial fulfilment of requirements for the degree of Bachelor of Technology in Electronics & Communication Engineering, is a bona fide work carried out by them under the supervision and guidance of **Prof. DEBANJANA GHOSH** during 7th Semester of academic session of 2024. The content of this report has not been submitted to any other university or institute for the award of any other degree.

I am glad to inform that the work is entirely original, and its performance is found to be quite satisfactory.

Prof. Debanjana Ghosh	Prof. Abir Chatterjee
Assistant Professor Department of ECE	Head of the Department Department of ECE
UEM, Kolkata	UEM, Kolkata

ACKNOWLEDGEMENT

We would like to take this opportunity to thank and acknowledge with due courtesy everyone whose cooperation and encouragement throughout the ongoing course of this project remains invaluable to us.

We are sincerely grateful to our guide Prof. Debanjana Ghosh of the Department of Electronics & Communication Engineering, UEM, Kolkata, for her wisdom, guidance and inspiration that helped us to go through with this project and take it to where it stands now.

We would also like to express our sincere gratitude to Prof. Abir Chatterjee, HOD, Electronics & Communication Engineering, UEM, Kolkata and all other departmental faculties for their ever-present assistant and encouragement.

Last but not the least, we would like to extend our warm regards to our families and peers who have kept supporting us and always had faith in our work.

Abir Lal Manna

Rupayan Saha

Sarthak Chakraborty

Debasmita Das

TABLE OF CONTENTS

ABSTRACT

Parkinson disease (PD) is a universal public health problem of massive measurement. Machine learning based method is used to classify between healthy people and people with Parkinson's disease (PD). This project presents a comprehensive review for the prediction of Parkinson disease by using **machine learning** based algorithms like **random forest** , **svm** etc. The brief introduction of various computational intelligence techniques based approaches used for the prediction of Parkinson diseases are presented . The dataset is preprocessed using **normalization** techniques to enhance the performance of the models. Feature importance is analyzed to identify the most significant predictors of the disease. Hyperparameter tuning is performed to optimize model performance, achieving improved precision, recall, and F1 scores. This project also presents the summary of results obtained by various researchers available in literature to predict the Parkinson diseases.

Keywords— Parkinson's disease , random forest, support vector machine, machine learning, normalization.

INTRODUCTION

Parkinson's disease (PD) is a chronic and progressive neurodegenerative disorder that affects millions worldwide, characterized by symptoms such as tremors, bradykinesia (slowness of movement), rigidity, and impaired balance. The disease primarily results from the degeneration of dopaminergic neurons in the substantia nigra region of the brain. While the exact cause remains unclear, early detection and diagnosis play a pivotal role in managing symptoms and slowing disease progression. Traditional diagnostic methods rely heavily on clinical observations and subjective assessments, which may lead to delays in diagnosis.

Machine learning (ML) has emerged as a powerful tool in the healthcare domain, enabling the analysis of complex datasets to detect patterns and make predictions with high accuracy. In the context of Parkinson's disease, ML techniques can analyze biomedical and voice-based features to identify subtle changes associated with the disorder. Features such as vocal frequency variations, jitter, shimmer, and nonlinear energy ratios have shown potential as indicators of PD, offering a non-invasive, efficient, and objective diagnostic approach.

This study focuses on leveraging ML algorithms to predict Parkinson's disease using clinical and voice datasets. The aim is to develop a predictive model capable of identifying PD cases accurately and to deploy this model as an accessible web application. By integrating feature selection, model optimization, and real-time prediction capabilities, this research seeks to demonstrate the feasibility and efficacy of ML in advancing early diagnosis and management of Parkinson's disease.

Machine Learning based approach

Machine learning (ML) is a branch of artificial intelligence (AI) that enables computers to “self-learn” from training data and improve over time, without being explicitly programmed. Machine learning algorithms are able to detect patterns in data and learn from them, in order to make their own predictions. In short, machine learning algorithms and models learn through experience. 7 Machine learning plays a crucial role in the medical field. Using machine learning, we can diagnose, detect, and predict various diseases. Recently, there has been a growing interest in using data mining and machine learning techniques to predict the likelihood of developing certain diseases. The already-existing work contains applications of data mining techniques for predicting the disease. Although some studies have attempted to predict the future risk of the progression of the disease, they have yet to find accurate results.

The scope of studying Parkinson's Disease (PD) prediction using Machine Learning (ML) is vast and promising. This research area can help in early detection, accurate diagnosis, and personalized treatment plans for patients. Below are the primary aspects and potential outcomes of such a study:

1. Problem Domain

Early Diagnosis: Parkinson's is often diagnosed in advanced stages due to subtle early symptoms. ML can help identify early biomarkers or patterns in data to aid in early detection.

2.Machine Learning Techniques

Feature Selection: Identify relevant features such as specific gait parameters, voice characteristics, or genetic markers.

Classification Models: Use algorithms like Random Forest, Support Vector Machines (SVM), Neural Networks, or Gradient Boosting for binary or multiclass classification (e.g., PD vs. healthy or disease stages).

3.Applications

Clinical Support Tools: Assist doctors in decision-making with predictive insights.

Remote Monitoring: Use wearable devices and IoT for continuous monitoring and symptom detection.

Personalized Medicine: Predict individual patient responses to therapies based on data.

LITERATURE REVIEW

The application of machine learning (ML) for Parkinson's disease (PD) prediction has gained significant traction in recent years, driven by the increasing availability of biomedical and voice datasets. This section reviews existing studies to understand the progress, challenges, and opportunities in this domain.

TABLE 1: Summary of machine learning based methods for Parkinson disease prediction

Authors name	Machine learning	citation	Performance
Indira R, Almeida, Jefferson S.	fuzzy C- means	"Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques." *Pattern Recognition Letters* 125 (2019): 55-62.	68.04% accuracy, 75.34% sensitivity and 45.83% specificity
Indira R. (2014) Templeton, John Michael.	ANN	Templeton, John Michael, Christian Poellabauer, and Sandra Schneider. "Classification of Parkinson's disease and its stages using machine learning." *Scientific reports* 12.1 (2022): 14036.. "Classification	Recognition rate of 92 %.
R. Geeta (2012)	Classification	"Machine learning for the diagnosis of Parkinson's disease: a review of literature." *Frontiers in aging neuroscience* 13 (2021): 633752.	Random tree classification 100% accuracy
Betalu E., Pahuja, Gunjan, and T. N. Nagabhushan (2014)	SVM	"A comparative study of existing machine learning approaches for Parkinson's disease detection." *IETE Journal of Research* 67.1 (2021): 4-14.	76%accuracy 34% sensitivity

METHODOLOGY

CHARACTERISTICS OF PARKINSON'S DISEASE

Parkinson's disease (PD) is a progressive neurodegenerative disorder that primarily affects movement control. Its main characteristics include:

Motor Symptoms

1) **Tremor**:
   - Often starts in one hand (resting tremor).
   - "Pill-rolling" motion is a hallmark.

Non-Motor Symptoms

1) **Cognitive Impairment**:
   - Memory problems or difficulty concentrating.
   - Dementia may develop in advanced stages.

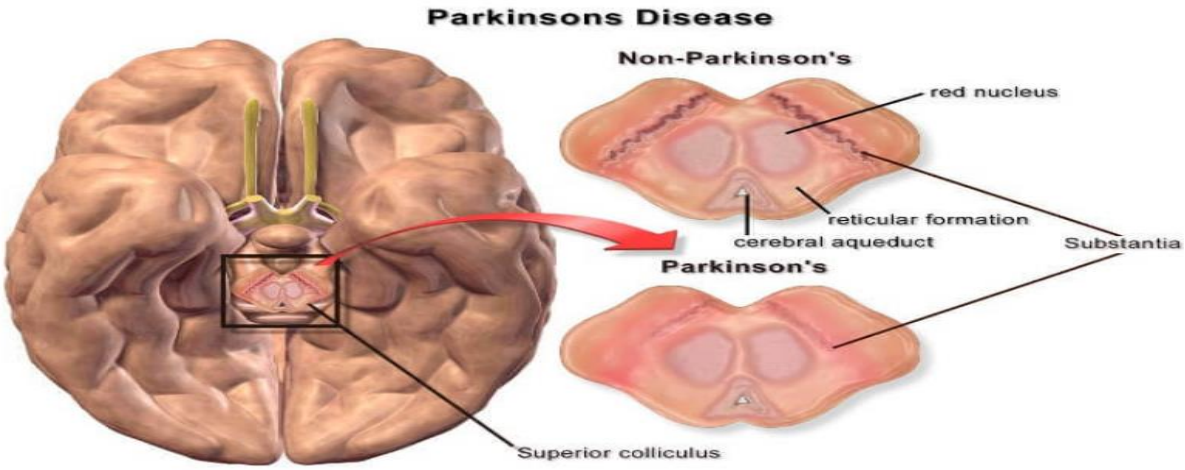

Fig 1 : Normal brain v/s heart affected by Parkinson's disease

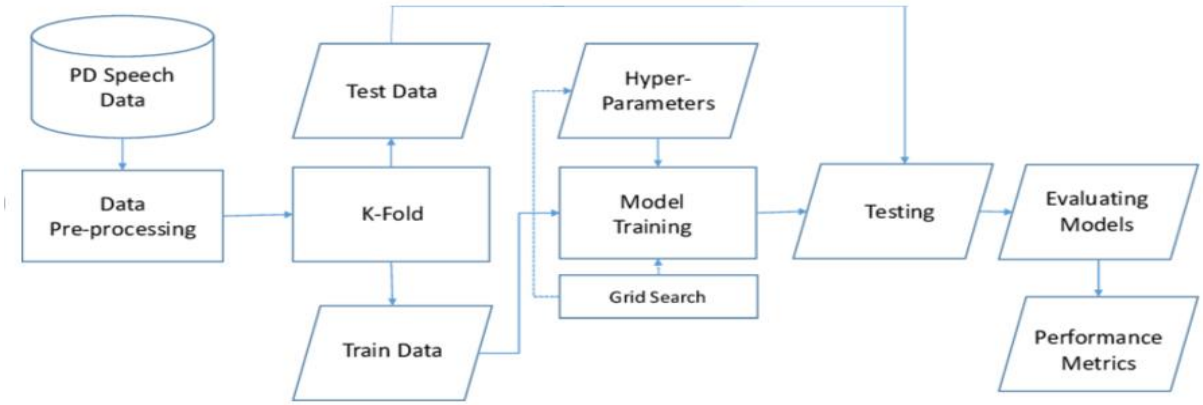

Fig 2 : Block Diagram

1. Data Collection

Data for this study was collected from reliable datasets, such as the UCI Machine Learning Repository or other Parkinson's-specific datasets. These datasets typically include voice recordings and clinical parameters. Key features used in the prediction model include:

Voice Features: Jitter, shimmer, harmonic-to-noise ratio (HNR), and noise-to-harmonic ratio (NHR).

2. Data Preprocessing

Preprocessing was essential to prepare the raw data for machine learning. The following steps were performed:

Handling Missing Values: Rows with missing or erroneous values were removed or imputed using mean or median imputation.

Feature Scaling: Continuous features were scaled using MinMaxScaler to normalize values between 0 and 1, which is crucial for algorithms sensitive to feature scaling.

3. Feature Selection and Engineering

To enhance model efficiency and reduce computation time, the following feature selection techniques were applied:

Recursive Feature Elimination (RFE): Identified the most relevant features by recursively removing less important ones.

Principal Component Analysis (PCA): Reduced dimensionality while retaining key information, ensuring the removal of redundant and highly correlated features.

4. Model Development

Several machine learning algorithms were trained and compared for performance. Each model's hyperparameters were fine-tuned using grid search or random search for optimal results:

Logistic Regression: A baseline model to evaluate separability between classes.

Support Vector Machine (SVM): Implemented with radial basis function (RBF) kernel for non-linear separation.

Random Forest: An ensemble learning method chosen for its robustness and ability to handle high-dimensional data.

Gradient Boosting (XGBoost): Used for its efficiency in minimizing error on imbalanced datasets.

K-Nearest Neighbors (KNN): Implemented for comparisons based on proximity-based classifications.

5. Model Evaluation

Each model was evaluated using a separate testing dataset. The following metrics were computed to assess performance:

Accuracy: Proportion of correctly predicted samples.

Precision and Recall: Used to evaluate class-wise prediction reliability.

F1-Score: A harmonic mean of precision and recall, crucial for imbalanced datasets.

Confusion Matrix: Visualized to analyze true positive, false positive, true negative, and false negative rates.

\

PROPOSED SOLUTION

USE OF ALGORITHMS:

- Decision tree regression
- Random forest regression
- Support Vector Machine
- k-Nearest Neighbors algorithm (KNN)
- Logistic Regression
- Naïve Bayes
- XG Boost

	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	MDVP:Jitter(%)	MDVP:Jitter(Abs)	MDVP:RAP	MDVP:PPQ	Jitter:DDP	MDVP:Shimmer	MDVP:Shimmer(dB
count	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	195.00000
mean	154.228641	197.104918	116.324631	0.006220	0.000044	0.003306	0.003446	0.009920	0.029709	0.28225
std	41.390065	91.491548	43.521413	0.004848	0.000035	0.002968	0.002759	0.008903	0.018857	0.19487
min	88.333000	102.145000	65.476000	0.001680	0.000007	0.000680	0.000920	0.002040	0.009540	0.08500
25%	117.572000	134.862500	84.291000	0.003460	0.000020	0.001660	0.001860	0.004985	0.016505	0.14850
50%	148.790000	175.829000	104.315000	0.004940	0.000030	0.002500	0.002690	0.007490	0.022970	0.22100
75%	182.769000	224.205500	140.018500	0.007365	0.000060	0.003835	0.003955	0.011505	0.037885	0.35000
max	260.105000	592.030000	239.170000	0.033160	0.000260	0.021440	0.019580	0.064330	0.119080	1.30200

Fig 3 : . Dataset, variable name, type and description

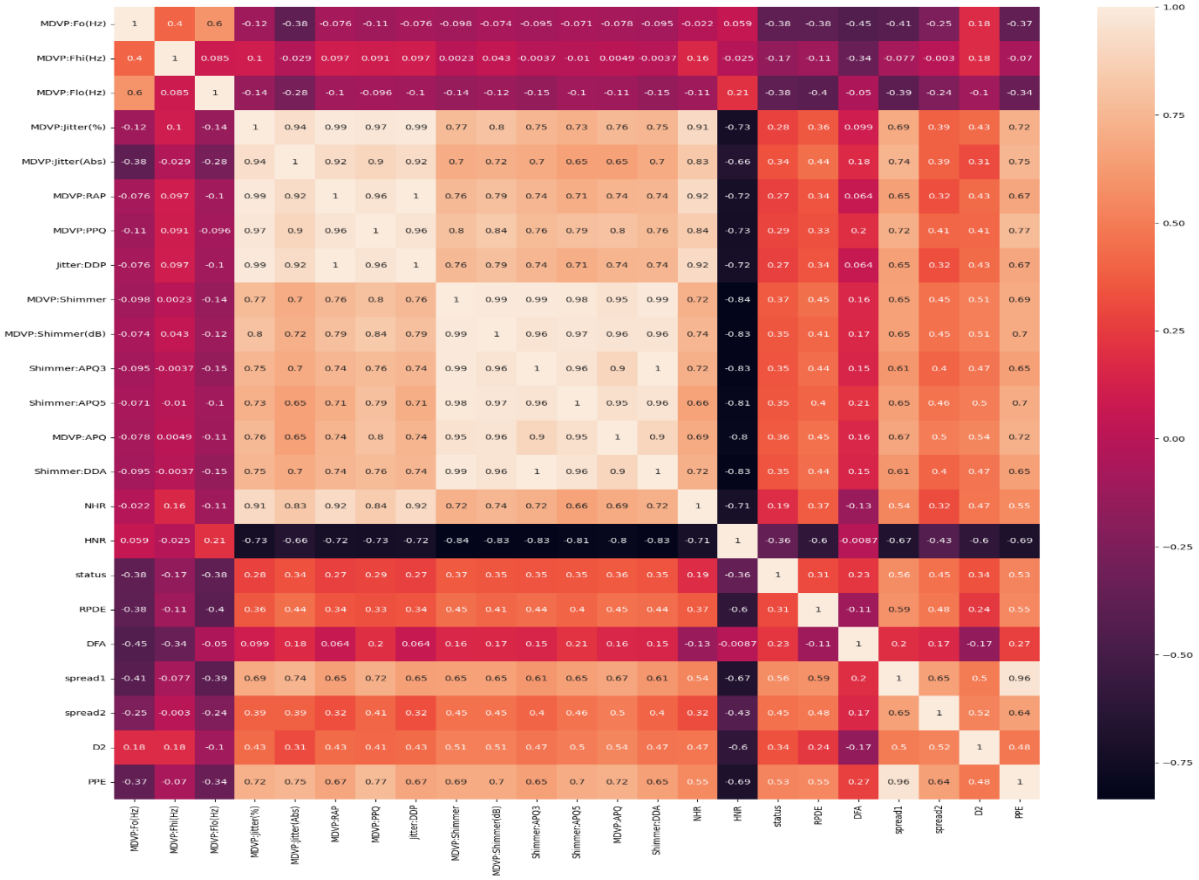

Fig 4 : Correlation Matrix using a Heat Map

Observations from the Correlation Matrix:

1. Highly Correlated Features:
    - Many shimmer-related variables (e.g., MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5) show strong positive correlations with each other (values close to 1), suggesting redundancy in these features.
2. Features Related to Status:
    - The variable status (likely representing Parkinson's disease diagnosis: 1 = PD, 0 = healthy) has moderate to strong positive and negative correlations with
3. Outliers or Weak Correlations:
    - Certain features, such as spread1 and spread2, seem to have weaker correlations with most other features, suggesting they may provide unique information about the dataset.
4. Potential Redundancy:
    - Features with very high correlations (e.g., shimmer or jitter groups) might be redundant. Dimensionality reduction techniques (like PCA) or feature selection might help eliminate redundant variables while retaining relevant information.

CLASSIFICATION

1.DECISION TREE CLASSIFIER:

A **decision tree** is a supervised machine learning algorithm commonly used for classification and regression tasks. It is a tree-like model of decisions and their possible consequences, where the data is split into subsets based on feature values. Each node in the tree represents a decision based on a feature, and the branches
represent the outcomes of that decision.

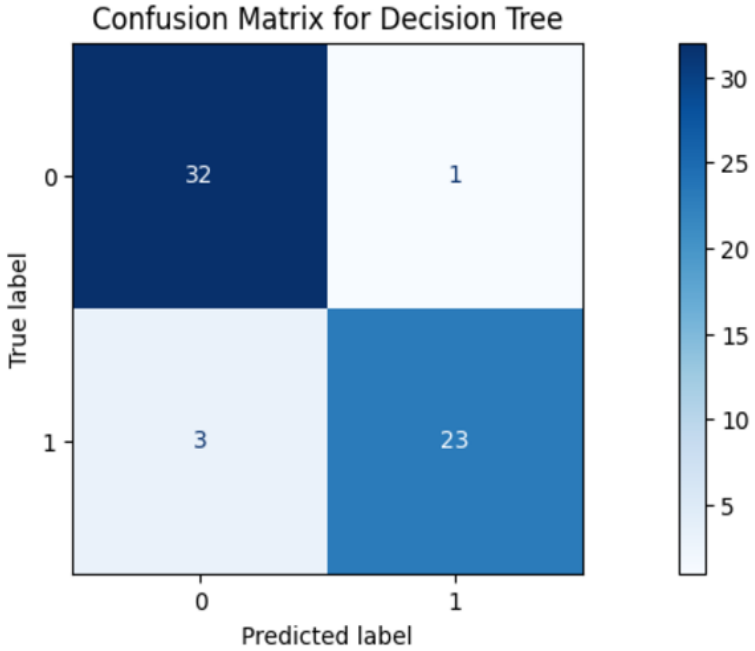

Fig 5 : Confusion Matrix for Decision Tree

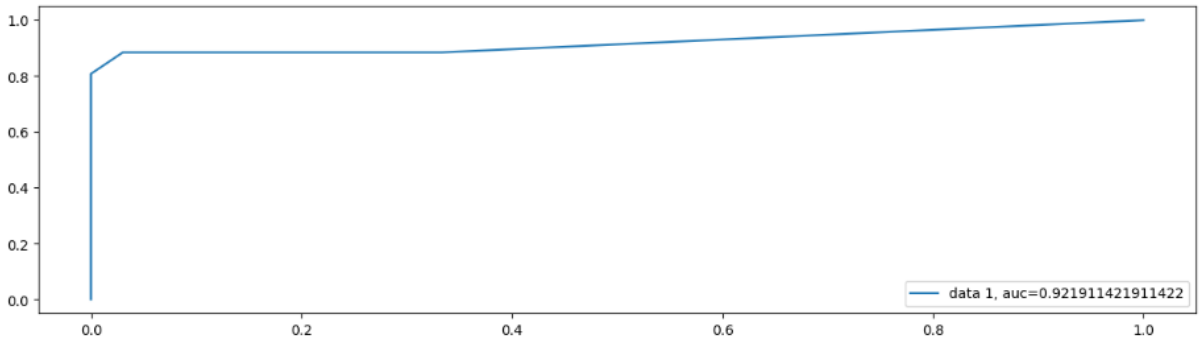

Fig 6 : Receiver Operating Characteristic (ROC) curve of DT

2.RANDOM FOREST CLASSIFIER:

- As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.
- The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.
- Random Forest is capable of performing both Classification and Regression tasks.
- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the overfitting issue

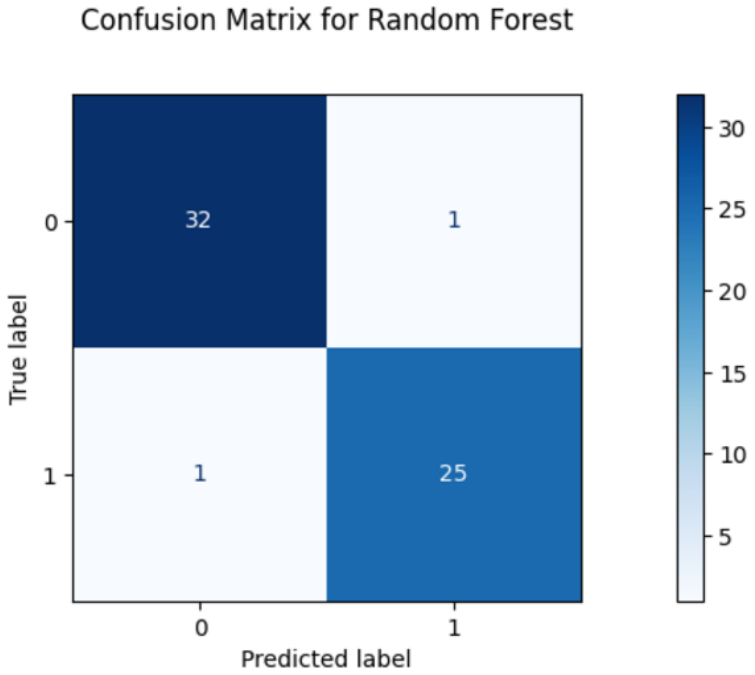

Fig 7 : Confusion Matrix for Random Forest

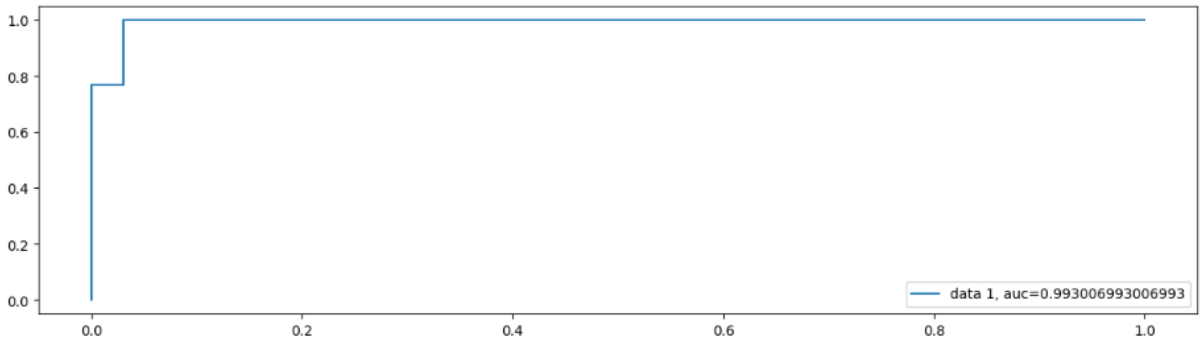

Fig 8 : Receiver Operating Characteristic (ROC) curve of RF

3.SUPPORT VECTOR MACHINE (SVM):

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well it's best suited for classification. The main objective of the SVM algorithm is to find the optimal hyperplane in an N-dimensional space that can separate the data points in different classes in the feature space. The hyperplane tries that the margin between the closest points of different classes should be as maximum as possible.

1. The dimension of the hyperplane depends upon the number of features.
2. If the number of input features is two, then the hyperplane is just a line.
3. If the number of input features is three, then the hyperplane becomes a 2-D plane.
4. It becomes difficult to imagine when the number of features exceeds three.

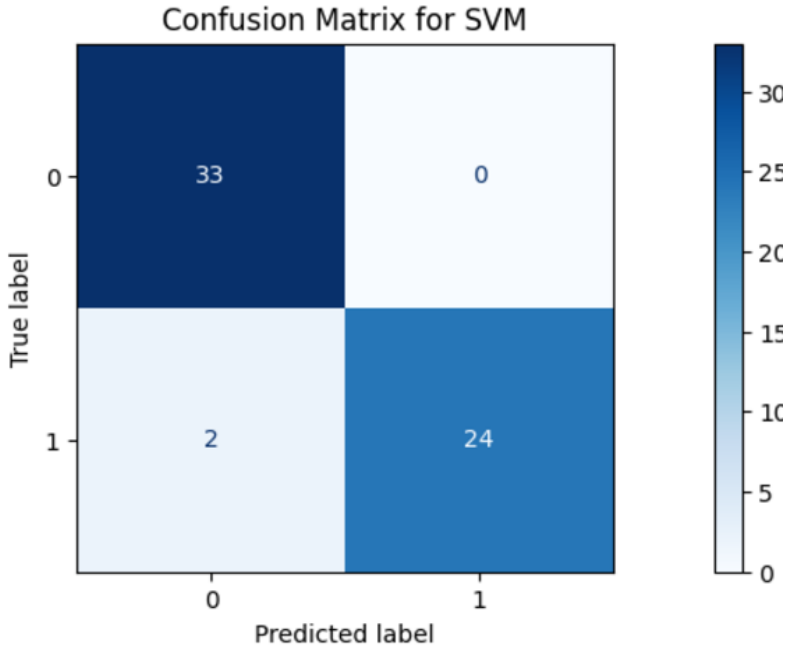

Fig 9 : Confusion Matrix for SVM

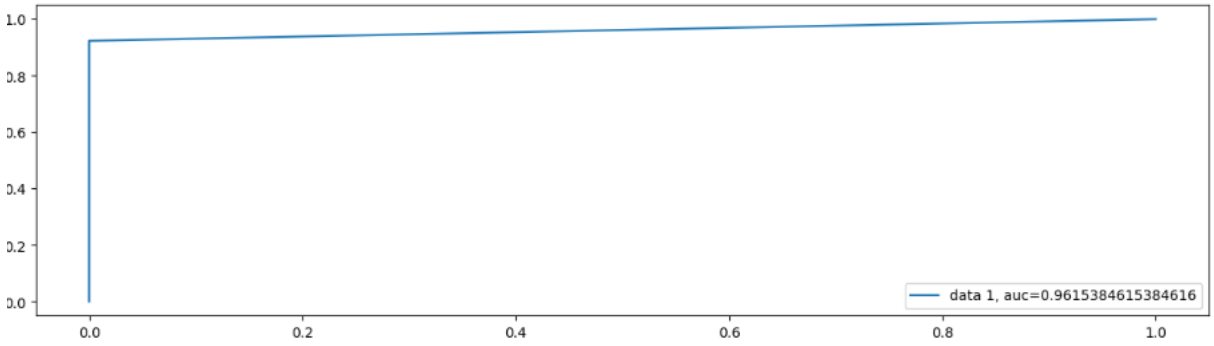

Fig 10 : Receiver Operating Characteristic (ROC) curve of SVM

4.K-NEAREST NEIGHBOUR (KNN):

- KNN is a non-parametric algorithm, which means it does not make any assumption on underlying data.
- It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.
- It is capable of Feature Transformation.

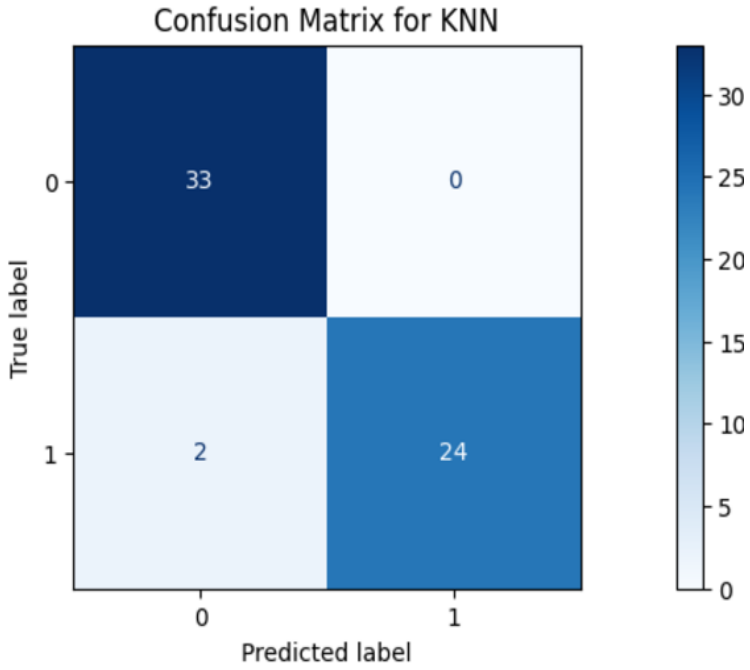

Fig 11 : Confusion Matrix for KNN

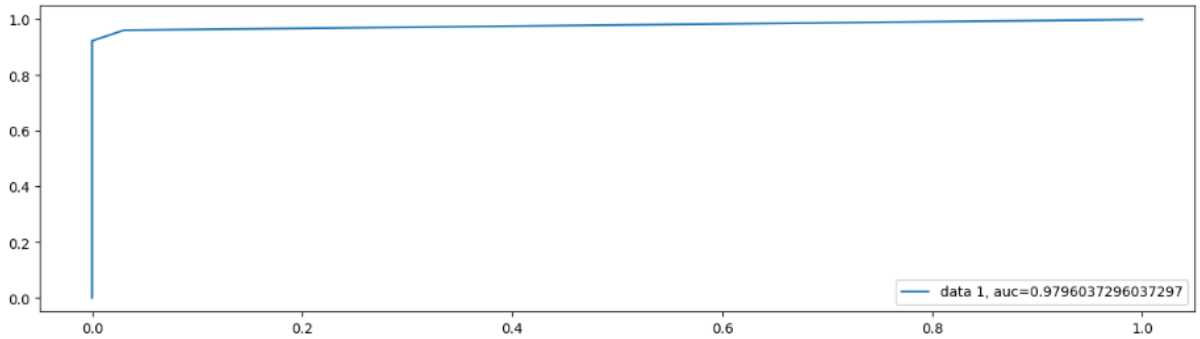

Fig 12 : Receiver Operating Characteristic (ROC) curve of KNN

5.LOGISTIC REGRESSION:

- Predicts the probability of an event occurring (e.g., yes/no, 1/0).
- Example: Predicting whether a patient has a disease (1) or not (0).
- The core of logistic regression is a linear equation that combines input features with coefficients to estimate an outcome.

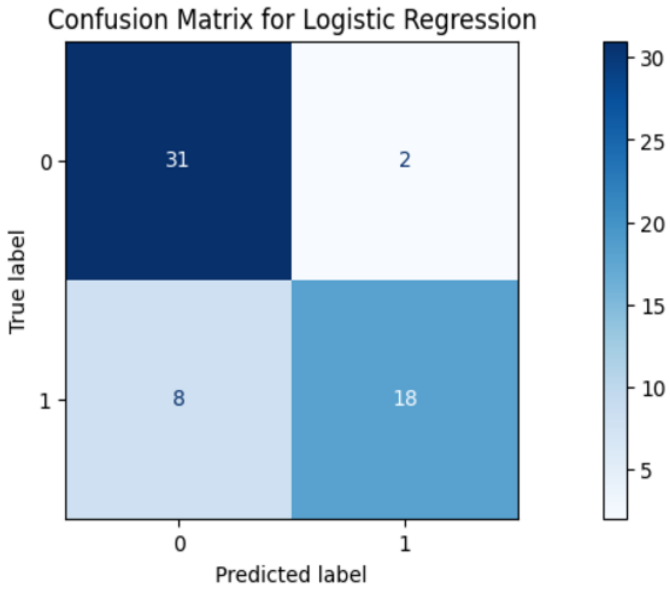

Fig 13 : Confusion Matrix for Logistic Regression

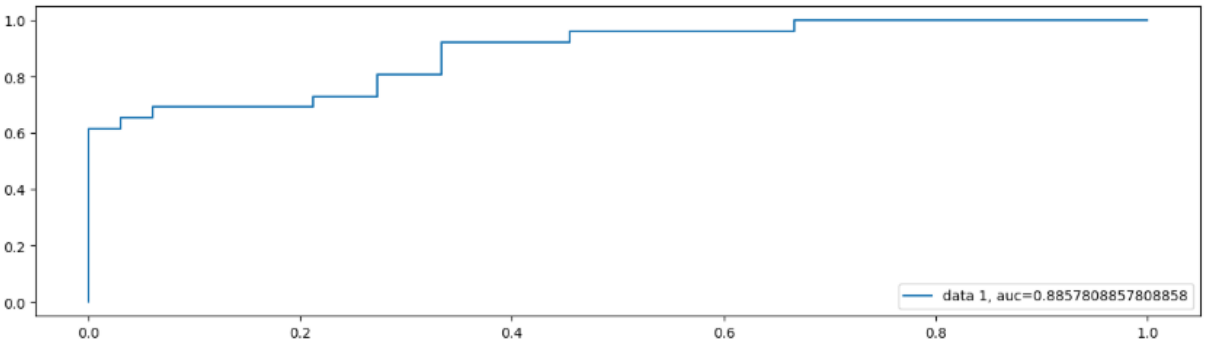

Fig 14 : Receiver Operating Characteristic (ROC) curve of LR

6.NAIVE BAYES:

- Provides the posterior probabilities for all classes, enabling confidence-based decisions.
- Requires calculating probabilities for each class and feature, making it computationally efficient.
- Handles large datasets effectively due to its simplicity and low computational complexity.

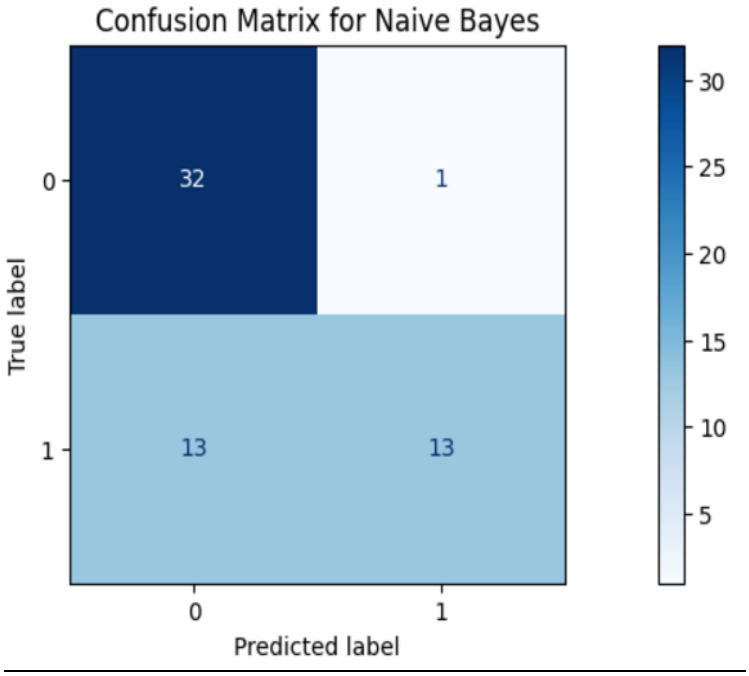

Fig 15 : Confusion Matrix for Naive Bayes

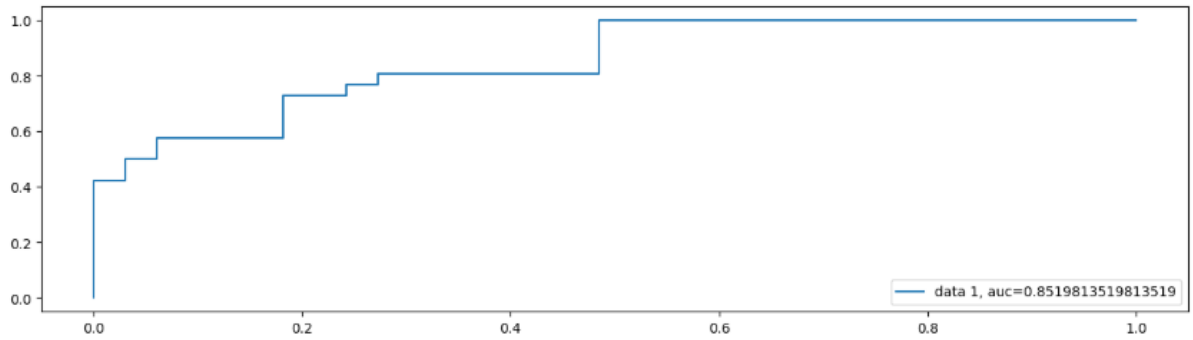

Fig 16 : Receiver Operating Characteristic (ROC) curve of NB

7.XG Boost Classifier:

- Based on boosting, where multiple weak learners (usually decision trees) are sequentially trained to correct the errors of prior models.
- Incorporates L1 (Lasso) and L2 (Ridge) regularization, reducing overfitting and improving generalization.
- Implements parallel processing and other optimizations for faster computation.
- Uses out-of-core computation for handling large datasets.

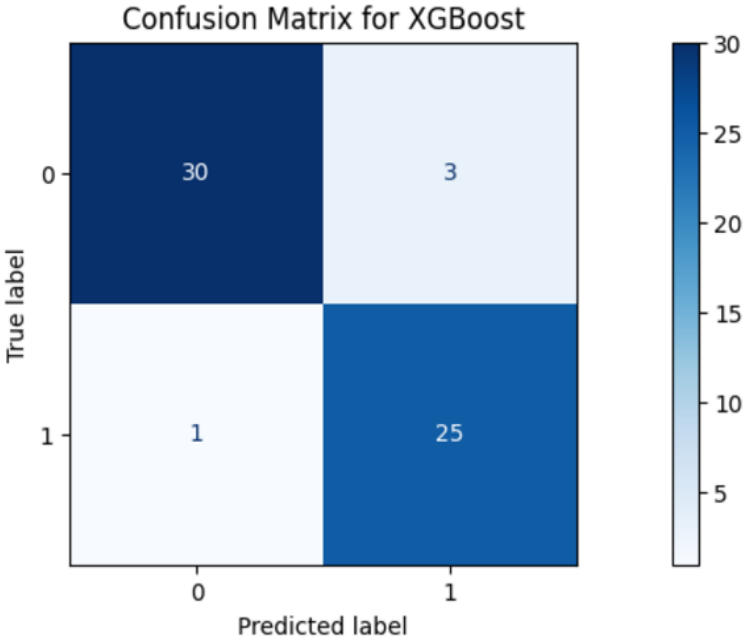

Fig 17 : Confusion Matrix for XGBoost

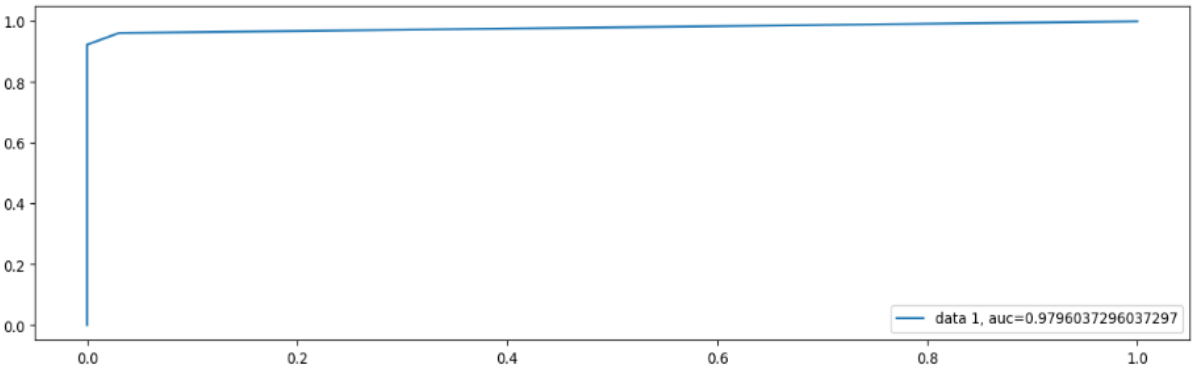

Fig 18 : Receiver Operating Characteristic (ROC) curve of XG Boost

PERFORMANCE EVALUATION:

Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate (TNR). The best specificity is 1.0, whereas the worst is 0.0.
Sensitivity (SN) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall (REC) or true positive rate (TPR). The best sensitivity is 1.0, whereas the worst is 0.0.
Accuracy (ACC) is calculated as the number of all correct predictions divided by the total number of the dataset. The best accuracy is 1.0, whereas the worst is 0.0. It can also be calculated by 1 – ERR.

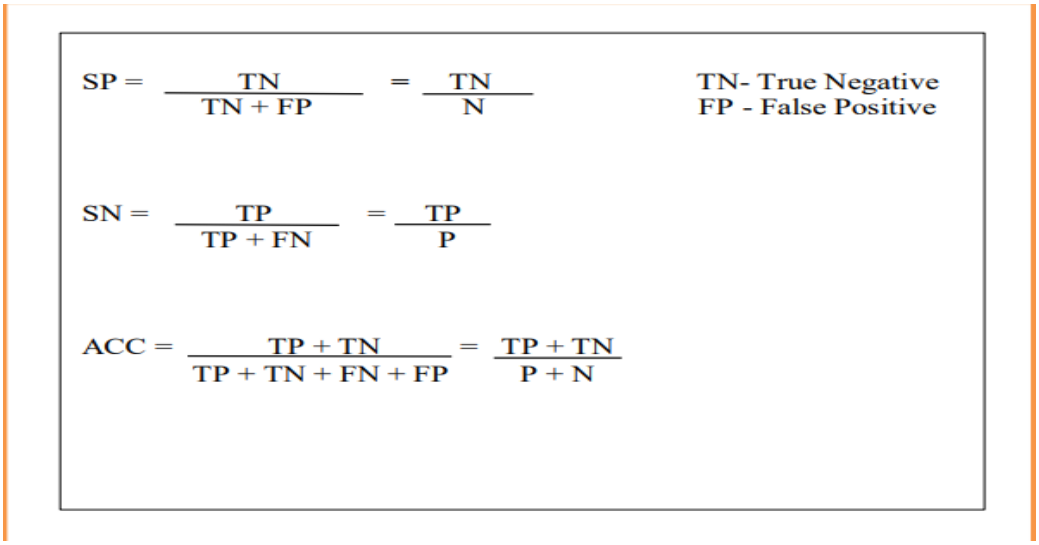

Fig 19 : Formulae for SP , SN , ACC

	Metric	DT	RF	LR	SVM	NB	KNN	XGB
0	Accuracy	0.932203	0.966102	0.830508	0.966102	0.762712	0.966102	0.932203
1	F1-Score	0.920000	0.961538	0.782609	0.960000	0.650000	0.960000	0.925926
2	Recall	0.884615	0.961538	0.692308	0.923077	0.500000	0.923077	0.961538
3	Precision	0.958333	0.961538	0.900000	1.000000	0.928571	1.000000	0.892857
4	R2-Score	0.724942	0.862471	0.312354	0.862471	0.037296	0.862471	0.724942

Fig 20 : Performance table of different machine learning models

Best Performing Machine Learning Model:

Random Forest Classifier was found to be the best performing Classifier with:

- Accuracy: 0.996102
- F1 Score : 0.961538
- R2 Score : 0.862471

FUTURE SCOPE

The future of Parkinson's disease prediction using machine learning holds immense promise, driven by advancements in data collection, algorithm design, and computational power. With wearable devices, medical imaging, and other sensor technologies generating vast amounts of data, ML models can analyze diverse data types—such as speech patterns, handwriting analysis, and neuroimaging data—to improve diagnostic accuracy and early detection. Integration of multi-modal data (e.g., genetic, clinical, and sensor data) offers the potential to provide personalized healthcare solutions, aiding clinicians in tailoring treatment plans to individual needs. Additionally, explainable AI (XAI) is emerging as a critical area to address the opacity of ML models, ensuring healthcare practitioners understand and trust the predictions.

However, challenges persist in achieving these goals. One major hurdle is the availability of high-quality, labeled datasets; obtaining large-scale, unbiased, and diverse datasets is difficult due to privacy concerns and the rarity of early-stage Parkinson's diagnoses. Another issue lies in managing imbalanced data, as Parkinson’s cases often constitute a minority in collected samples, potentially skewing predictions. Furthermore, the heterogeneity of Parkinson’s disease symptoms among patients complicates the development of generalized models. Addressing these challenges requires collaboration between technologists, healthcare professionals, and policymakers to ensure ethical data sharing, advanced algorithmic development, and clinical validation of ML models. Despite these obstacles, ongoing research and innovation promise a transformative impact on the detection and management of Parkinson’s disease in the years to come.

CONCLUSION

This project presented a comprehensive review for the prediction of Parkinson disease by using machine learning based approaches. The brief introduction of various computational intelligence techniques based approaches used for the prediction of Parkinson diseases are presented .The summary of results obtained by various researchers available in literature to predict the Parkinson diseases is also presented. The integration of machine learning into Parkinson's disease prediction marks a paradigm shift in healthcare. It empowers clinicians with data-driven tools, fosters earlier detection, and opens avenues for personalized treatment strategies. While challenges persist, the continued advancement of ML algorithms, along with improvements in data collection and processing, promises a future where Parkinson's Disease is detected earlier, managed more effectively, and its progression potentially slowed, ultimately improving patient outcomes and quality of life.

BIBLIOGRAPHY/REFERENCES

[1] Rustempasic, Indira, & Can, M. (2013). Diagnosis of Parkinson's Disease using Fuzzy C-Means Clustering and Pattern Recognition. SouthEast Europe Journal of Soft Computing, 2(1). Senturk, Zehra Karapinar. "Early diagnosis of Parkinson's disease using machine learning algorithms." *Medical hypotheses* 138 (2020): 109603.

[2] Rustempasic, I., & Can, M. (2013). Diagnosis of Parkinson's disease using principal component analysis and boosting committee machines. SouthEast Europe Journal of Soft Computing, 2(1). Senturk, Zehra Karapinar. "Early diagnosis of Parkinson's disease using machine learning algorithms." *Medical hypotheses* 138 (2020): 109603.

[3] Armañanzas, Ruben, Bielza, C., Chaudhuri, K. R., Martinez-Martin, P., & Larrañaga, P. (2013). Unveiling relevant non-motor Parkinson's disease severity symptoms using a machine learning approach. Artificial intelligence in medicine, 58(3), 195-202. Challa, Kamal Nayan Reddy, et al. "An improved approach for prediction of Parkinson's disease using machine learning techniques." *2016 international conference on signal processing, communication, power and embedded system (SCOPES)*. IEEE, 2016.

[4] Sakara, Batalu. E., & Kursunb, (2014)O. Telemonitoring of changes of unified Parkinson's disease rating scale using severity of voice symptoms. Wang, Wu, et al. "Early detection of Parkinson's disease using deep learning and machine learning." *IEEE Access* 8 (2020): 147635-147646.

[5] Tsanas, A., Little, M. A., McSharry, P. E., Spielman, J., & Ramig, L. O. (2012). Novel speech signal processing algorithms for highaccuracy classification of Parkinson's disease. Biomedical Engineering, IEEE Transactions on, 59(5), 1264-1271.

[6] Kotsavasiloglou, C., Tzallas, A. T., Tsipouras, M. G., Rigas, G., Bougia, P., & Fotiadis, D. I. (2017). Machine learning-based classification of simple drawing movements in Parkinson's disease. *Biomedical Signal Processing and Control, 31*, 174–180.