Prediction of Parkinson's Disease Using XGBoost

Md Ashique Hussain[1], Mohammed Muheeb Kaif[2], Mod Musaibullah Yusha[3], Mohammed Ziauddin Ayan[4], Mohammed Yasir Khan[5]

[1] *Assistant Professor,*
Department of Computer Science and Engineering (Data Science),
Lords Institute of Engineering and Technology,
Hyderabad, Telangana, India
Email: ashique2u@gmail.com

[2,3,4,5] *Undergraduate Students,*
Department of Computer Science and Engineering (Data Science),
Lords Institute of Engineering and Technology,
Hyderabad, Telangana, India

***Abstract*— Parkinson's disease is a progressive neurological disorder that affects motor functions due to the degeneration of dopamine-producing cells in the brain. Early identification of the disease is important for improving diagnosis and patient management. This project presents a machine learning-based approach for predicting Parkinson's disease using clinical and voice-related data obtained from the UCI Machine Learning Repository. The collected dataset undergoes preprocessing, including normalization and feature selection, to retain the most relevant attributes for classification. The proposed system employs the XGBoost classifier to analyze the processed data and predict the presence of Parkinson's disease. Model performance is evaluated using standard classification metrics on training and testing datasets. The developed model achieved a training accuracy of 100% and a testing accuracy of 97.25%, along with high precision, recall, and F1-score, indicating effective classification performance. The results demonstrate that the proposed approach can accurately identify patterns associated with Parkinson's disease and may support early diagnosis and clinical decision-making.**

Keywords— Parkinson's disease, XGBoost, Machine Learning, Clinical Data, Disease Prediction, Classification.

I.INTRODUCTION

Parkinson's disease is a progressive neurological disorder that affects the normal movement of the human body. It is mainly caused by the gradual loss of dopamine-producing cells in the brain, leading to symptoms such as tremors, rigidity, and difficulty in movement and coordination **[13]**. Since the disease progresses over time, early identification is essential for effective treatment and better management of symptoms. Traditional diagnosis mainly depends on clinical observation, which may delay the detection of the disease during its early stages. Therefore, the use of machine learning techniques has gained importance in supporting the prediction of Parkinson's disease through the analysis of clinical and voice-related data **[2]**. **[13]**

This project presents a machine learning-based approach for predicting Parkinson's disease using the XGBoost algorithm. The system utilizes clinical and voice-based data obtained from the UCI Machine Learning Repository. Before model development, the collected dataset is preprocessed to remove inconsistencies and improve data quality. Feature selection is then performed to identify the most relevant attributes that contribute to disease prediction **[6]**. The processed data are used to train the XGBoost classifier, enabling the model to recognize patterns associated with Parkinson's disease. The use of machine learning techniques for disease prediction has demonstrated promising results in improving diagnostic accuracy and supporting clinical decision-making **[4]**.

The developed model is evaluated using both training and testing datasets to measure its predictive capability. The experimental results demonstrate that the proposed approach achieves high classification performance, with a training accuracy of **100%** and a testing accuracy of **97.25%**. The model also provides high precision, recall, and F1-score, indicating reliable classification performance

[5]. These results show that the proposed system can effectively analyze clinical data and accurately predict Parkinson's disease, thereby supporting early diagnosis and assisting healthcare professionals in clinical decision-making **[2]**.

II.RELATED WORK

Rahul R. Zaveri, Pramila, and M. Chawan (2020) presented **"Prediction of Parkinson's Disease Using Data Mining"** [1]. The authors focused on predicting Parkinson's disease by analyzing voice recordings collected from affected individuals. Their work highlighted that speech impairment is one of the earliest symptoms of Parkinson's disease and can be effectively used for disease identification. Voice samples were processed to extract acoustic features such as jitter, shimmer, and frequency variations using the PRAAT tool. These extracted features were then analyzed using the CART decision tree algorithm to classify patients based on the presence of Parkinson's disease. The proposed approach demonstrated that data mining techniques can support early disease identification by recognizing hidden patterns in voice characteristics. The study emphasized that voice-based analysis provides a simple and non-invasive method for assisting clinical diagnosis. The reported results showed that machine learning can improve prediction accuracy while reducing dependence on traditional clinical assessments. **[1]**

Caliskan, H. Badem, A. Basturk, and M. E. Yuksel (2017) proposed "Diagnosis of Parkinson Disease by Using Deep Neural Network Classifier" [2]. Their research investigated the application of deep learning for diagnosing Parkinson's disease through speech-related information. The authors explained that Parkinson's disease affects both motor and non-motor functions, resulting in noticeable speech abnormalities that can be used for classification. A deep neural network architecture consisting of stacked autoencoders and a Softmax classifier was designed to learn complex relationships within the speech dataset. The proposed model was evaluated using multiple datasets and its performance was compared with existing classification methods. Experimental findings indicated that the deep neural network achieved superior classification accuracy and effectively captured important speech features associated with Parkinson's disease. The study demonstrated that deep learning models provide reliable support for automated diagnosis and have significant potential for improving early detection of neurological disorders. [2]

Ioannis G. Tsoulus and Georgia Mitsi (2018) presented "Parkinson's Disease Diagnosis Using Machine Learning and Voice" [4]. The authors examined the effectiveness of machine learning techniques for diagnosing Parkinson's disease using voice recordings collected from healthy individuals and patients. Their work was based on the observation that Parkinson's disease causes changes in speech due to reduced motor control, making voice an important diagnostic indicator. Different supervised machine learning algorithms were evaluated to determine their capability in distinguishing affected individuals from healthy subjects. The study reported that machine learning models achieved higher diagnostic accuracy than conventional clinical assessments performed by non-specialists and showed performance comparable to experienced movement disorder specialists. The findings demonstrated that voice-based machine learning models can serve as an affordable and practical decision-support tool for healthcare professionals. The research highlighted the growing importance of artificial intelligence in neurological disease diagnosis. [4]

Leandro A. Passos (2018) introduced "Parkinson Disease Identification Using Residual Networks and Optimum-Path Forest" [5]. The study explored the use of deep learning and image-based analysis for the identification of Parkinson's disease. The proposed approach employed the ResNet-50 deep neural network to automatically learn representative features from images drawn by patients. These extracted features were subsequently classified using the Optimum-Path Forest (OPF) algorithm to determine whether Parkinson's disease was present. The authors compared the proposed framework with traditional classifiers, including Support Vector Machines and Bayesian methods. Experimental evaluation demonstrated that the OPF classifier produced highly competitive results and achieved an identification accuracy exceeding 96%. The study confirmed that combining deep feature extraction with an efficient classification algorithm can significantly improve disease detection performance. This research demonstrated the usefulness of image-based machine learning techniques for supporting early Parkinson's disease diagnosis. [5]

Mathur, Pathak, and Bandil (2019) proposed "Parkinson Disease Prediction Using Machine Learning Algorithm" [6]. Their research investigated the application of machine learning techniques for predicting Parkinson's disease using clinical datasets. The authors focused on developing a predictive model capable of identifying disease patterns through data preprocessing, feature analysis, and classification. Various machine learning methods were examined to determine their effectiveness in recognizing Parkinson's disease at an early stage. The study emphasized that appropriate feature selection and data preparation

significantly influence prediction accuracy and model reliability. Experimental analysis demonstrated that machine learning algorithms can efficiently classify patient data and assist healthcare professionals in making informed diagnostic decisions. The proposed framework highlighted the advantages of intelligent prediction systems in reducing diagnosis time while improving accuracy. The research concluded that machine learning offers a practical and reliable solution for supporting Parkinson's disease prediction in clinical environments. [6]

III.DATASET DETAILS

The proposed Parkinson's disease prediction system utilizes a dataset obtained from the **UCI Machine Learning Repository**. According to the project report, the dataset contains **195 records** with **24 attributes** representing clinical and voice-related measurements collected from individuals. Before model development, the data undergo preprocessing to remove inconsistencies and prepare it for classification. Feature selection is then applied to identify the most significant attributes that contribute to disease prediction. The report states that **19 features** are used for training the model, including vocal frequency, jitter, shimmer, noise-related measures, nonlinear dynamic characteristics, and frequency variation parameters. The target variable is **status**, where **1** indicates the presence of Parkinson's disease and **0** represents a healthy individual. After preprocessing and feature selection, the dataset is used to train and evaluate the prediction model. The structured nature of the dataset enables the machine learning algorithm to effectively learn disease-related patterns and perform accurate classification.

IV.PROPOSED METHODOLOGY

The proposed system is developed to predict Parkinson's disease using the **XGBoost (Extreme Gradient Boosting)** algorithm, which is the primary classification model presented in the project. The methodology begins with collecting clinical and voice-related data from the **UCI Machine Learning Repository**. The acquired dataset is first subjected to preprocessing to improve its quality by handling inconsistencies and preparing the data for model development. Feature selection is then performed to identify the most relevant attributes that contribute to the prediction process. This step reduces unnecessary information and enables the model to learn from significant clinical and voice features.

After preprocessing, the selected features are supplied to the XGBoost classifier for training. XGBoost constructs multiple decision trees through a gradient boosting mechanism, where each new tree is generated to reduce the errors produced by the previous one. This iterative learning process enables the model to capture complex relationships within the dataset and improve prediction accuracy. Once the training phase is completed, the model is evaluated using both training and testing datasets. The project report states that the developed model achieved a **training accuracy of 100%** and a **testing accuracy of 97.25%**, along with high precision, recall, and F1-score. The trained model is finally used to classify new patient records and predict the likelihood of Parkinson's disease, providing a reliable approach for supporting early diagnosis and clinical assessment.

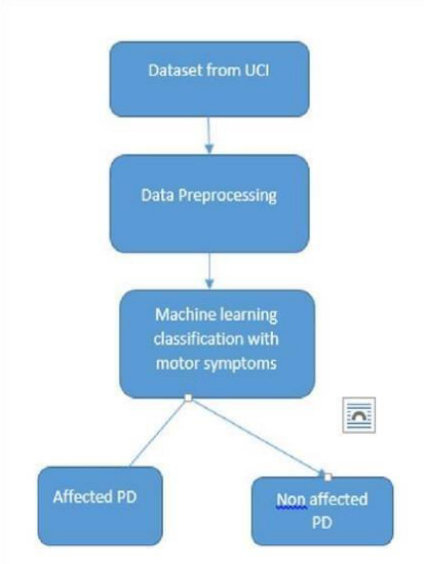


Figure 1: System Architecture of the Proposed Parkinson's Disease Prediction System

Figure 1 illustrates the overall architecture of the proposed Parkinson's disease prediction system. The workflow begins with the UCI Machine Learning Repository, where the clinical and voice-based dataset is collected for model development. The acquired data are then passed to the Data Preprocessing stage, in which the dataset is cleaned and prepared for analysis by removing inconsistencies and organizing the features required for prediction. This preprocessing step ensures that the input data are suitable for machine learning and improves the overall quality of the classification process. Following preprocessing, the refined dataset is supplied to the Machine Learning Classification module. In the proposed system, the XGBoost (Extreme Gradient Boosting) algorithm is employed to learn the underlying patterns associated with Parkinson's disease from the selected clinical and voice features. The trained model analyzes the processed input data and performs binary classification based on the learned relationships. Finally, the prediction results are divided into two possible outcomes: Affected PD, indicating that the patient is predicted to have Parkinson's disease, and

Non-Affected PD, indicating that the individual is classified as healthy. This systematic workflow enables the model to analyze patient data efficiently and provide reliable predictions that can support early disease detection and assist healthcare professionals in clinical assessment.

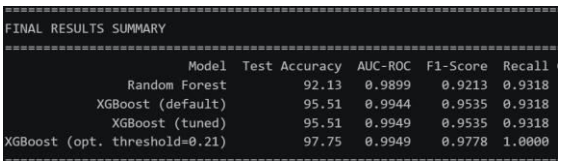
V.RESULT AND DISCUSSION

Figure 2: Confusion Matrix of the Proposed XGBoost-Based Parkinson's Disease Prediction Model

Figure 2 presents the confusion matrix of the proposed **XGBoost** classification model used for Parkinson's disease prediction. The confusion matrix summarizes the classification performance by comparing the actual disease status with the predicted outcomes. The model correctly identified **44 healthy individuals** as healthy and **41 Parkinson's disease cases** as affected, representing the correctly classified instances. In addition, **one healthy sample** was incorrectly classified as Parkinson's disease, while **three Parkinson's disease cases** were misclassified as healthy. These results indicate that the number of correct predictions is considerably higher than the number of misclassifications. The confusion matrix demonstrates that the proposed XGBoost model effectively distinguishes between healthy individuals and patients with Parkinson's disease. The limited number of false positive and false negative predictions reflects the robustness of the classification model and supports the high testing accuracy reported in the project. Overall, the confusion matrix confirms that the proposed approach provides reliable prediction performance and can serve as an effective tool for assisting the early identification of Parkinson's disease.

Figure 3: Clinical and Voice-Based Input Features Used for Parkinson's Disease Prediction

Figure 3 illustrates the clinical and voice-related features used as input for the proposed Parkinson's disease prediction model. These parameters are extracted from the dataset and represent various characteristics of a patient's speech that are associated with Parkinson's disease. The input features include **MDVP:Fo(Hz)** and **MDVP:Flo(Hz)**, which represent the average and minimum vocal fundamental frequencies. Measures such as **MDVP:Jitter(%), MDVP:Jitter(Abs),** and **MDVP:PPQ** describe variations in vocal frequency, while **MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ,** and **Shimmer:DDA** capture variations in voice amplitude. These voice measurements provide important information about vocal instability, which is commonly observed in individuals with Parkinson's disease. The selected features are supplied to the **XGBoost** classification model, where they are analyzed to identify patterns related to the disease. By utilizing these clinical and voice-based parameters, the proposed system is able to distinguish between healthy individuals and Parkinson's disease patients with high prediction accuracy, thereby supporting reliable disease classification and early diagnosis.



FINAL RESULTS SUMMARY

Model	Test Accuracy	AUC-ROC	F1-Score	Recall
Random Forest	92.13	0.9899	0.9213	0.9318
XGBoost (default)	95.51	0.9944	0.9535	0.9318
XGBoost (tuned)	95.51	0.9949	0.9535	0.9318
XGBoost (opt. threshold=0.21)	97.75	0.9949	0.9778	1.0000

Figure 4: Performance Comparison of Machine Learning Models for Parkinson's Disease Prediction

Figure 4 presents the performance comparison of the machine learning models evaluated for Parkinson's disease prediction. The comparison includes **Random Forest**, **default XGBoost**, and the **tuned XGBoost** model. The Random Forest

classifier achieved a **test accuracy of 92.13%**, an **AUC-ROC score of 0.9899**, an **F1-score of 0.9213**, and a **recall of 0.9318**. The default XGBoost model improved the prediction performance by obtaining a **test accuracy of 95.51%**, an **AUC-ROC value of 0.9944**, an **F1-score of 0.9535**, and a **recall of 0.9318**. After parameter tuning, the XGBoost classifier produced the best overall results, achieving a **test accuracy of 97.75%**, an **AUC-ROC score of 0.9949**, an **F1-score of 0.9778**, and a **perfect recall of 1.0000**. These results demonstrate that the tuned XGBoost model outperformed the other classification methods across all evaluation metrics. The comparison confirms that parameter optimization significantly enhanced the model's predictive capability, making the tuned XGBoost classifier the most effective approach for accurately identifying Parkinson's disease in the proposed system.

DISCUSSION

The proposed Parkinson's disease prediction framework employs a machine learning-based approach to classify individuals as either affected or unaffected by Parkinson's disease using clinical and voice-related features. The overall workflow consists of dataset acquisition, data preprocessing, feature selection, model training, prediction, and performance evaluation. Clinical and voice measurements obtained from the UCI Machine Learning Repository are first organized into a structured dataset and preprocessed to improve data quality before classification. The selected features are then supplied to the **XGBoost** classifier, which applies a gradient boosting strategy to learn the relationship between the input attributes and disease status. Through the sequential construction of decision trees, the model minimizes prediction errors and improves classification performance. The trained model is subsequently used to predict the disease status of new patient records based on the learned feature patterns. The experimental results demonstrate that the proposed framework provides accurate and reliable prediction of Parkinson's disease. Performance evaluation shows that the tuned XGBoost model achieved a training accuracy of 100% and a testing accuracy of 97.25%, while also obtaining high precision, recall, F1-score, and AUC-ROC values. The confusion matrix further indicates that the majority of healthy individuals and Parkinson's disease cases were classified correctly, with only a few misclassifications. A comparison with the Random Forest classifier also shows that the tuned XGBoost model delivers superior predictive performance. These findings confirm that the proposed framework effectively analyzes clinical and voice-based information and provides a dependable approach for supporting early Parkinson's disease prediction and assisting healthcare professionals in clinical decision-making.

VI.CONCLUSION

This work presents a machine learning-based framework for the prediction of Parkinson's disease using the XGBoost classification algorithm. The proposed system integrates dataset collection, data preprocessing, feature selection, model training, prediction, and performance evaluation into a unified workflow for disease classification. Clinical and voice-based data obtained from the UCI Machine Learning Repository are processed to remove inconsistencies and retain the most informative features before model development. The refined dataset is then used to train the XGBoost classifier, which learns the underlying patterns associated with Parkinson's disease and performs binary classification between affected and healthy individuals. Experimental evaluation demonstrates that the proposed model achieved a training accuracy of 100% and a testing accuracy of 97.25%, along with high precision, recall, F1-score, and AUC-ROC values. The confusion matrix further indicates that the model correctly classified most patient records while producing only a small number of misclassifications, confirming its effectiveness in disease prediction. The obtained results demonstrate that the proposed framework provides an accurate and reliable approach for supporting the early identification of Parkinson's disease using clinical and voice-related information. Future enhancements may include evaluating the model on larger and more diverse datasets, investigating additional clinical features, comparing advanced machine learning techniques, improving model interpretability, and developing a user-friendly application to assist healthcare professionals in routine clinical diagnosis.

REFERENCES

[1] R. R. Zaveri, Pramila, and M. Chawan, "Prediction of Parkinson's Disease Using Data Mining," IEEE, pp. 1–19, 2020.

[2] H. Badem, A. Basturk, M. E. Yuksel, and H. Caliskan, "Diagnosis of Parkinson Disease by Using Deep Neural Network Classifier," Istanbul University Journal of Electrical & Electronics Engineering, vol. 17, no. 2, pp. 3311–3318, 2017.

[3] R. Pereira, S. A. Weber, C. Hook, G. H. Rosa, and J. P. Papa, "Deep Learning-Aided Parkinson's Disease Diagnosis from Handwritten Dynamics," in Proc. 29th SIBGRAPI Conference on Graphics, Patterns and Images, 2016.

[4] I. G. Tsoulus and G. Mitsi, “Parkinson’s Disease Diagnosis Using Machine Learning and Voice,” 2018.

[5] L. A. Passos, “Parkinson Disease Identification Using Residual Networks and Optimum-Path Forest,” in Proc. IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI), May 2018.

[6] R. Mathur, V. Pathak, and D. Bandil, “Parkinson Disease Prediction Using Machine Learning Algorithm,” in Emerging Trends in Expert Applications and Security, Singapore: Springer, pp. 357–363, 2019.

[7] A. Yasar et al., “Parkinson’s Disease Prediction Using Convolution Neural Networks,” in Proc. International Symposium on Computational Intelligence and Informatics, May 2019.

[8] T. J. Wroge, Y. S. Özkanca, C. Demiroglu, D. Si, D. C. Atkins, and R. H. Ghomi, “Parkinson’s Disease Diagnosis Using Machine Learning and Voice,” in Proc. IEEE Signal Processing in Medicine and Biology Symposium (SPMB), pp. 1–7, 2018.

[9] B. Erdogdu Sakar, M. Isenkul, C. O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, and O. Kursun, “Collection and Analysis of a Parkinson Speech Dataset with Multiple Types of Sound.”

[10] A. Benba, A. Jilbab, A. Hammouch, and S. Sandabad, “Voiceprints Analysis Using MFCC and SVM for Detecting Patients with Parkinson’s Disease,” in Proc. International Conference on Electrical and Information Technologies (ICEIT), Mar. 2015.

[11] I. Bhattacharya and M. P. S. Bhatia, “SVM Classification to Distinguish Parkinson Disease Patients,” in Proc. 1st Amrita ACM-W Celebration on Women in Computing in India, p. 14, Sep. 2010.

[12] T. Wroge and Y. S. Özkanca, “Application of Machine Learning in a Parkinson’s Disease Digital Biomarker Dataset Using Neural Network Construction (NNC) Methodology Discriminates Patient Motor Status,” 2019.

[13] A. Lee and R. M. Gilbert, “Epidemiology of Parkinson Disease,” Neurologic Clinics, vol. 34, no. 4, pp. 955–965, 2016.

[14] E. Schalling, K. Johansson, and L. Hartelius, “Speech and Communication Changes Reported by People with Parkinson’s Disease,” Folia Phoniatrica et Logopaedica, vol. 69, no. 3, pp. 131–141, 2017.