Machine Learning Based Early Detection System for Parkinson's Disease

Moulieswaran R[1], Revanth M J[2], Gokul S[3], Karthikeyan B[4]
[1,2,3]*Department of Information Technology, Nehru Arts and Science College (Autonomous) Thirumalayampalayam, Coimbatore, (641105)*
[4]*Assistant Professor, Department of Information Technology, Nehru Arts and Science College (Autonomous), Thirumalayampalayam, Coimbatore, (641105)*

***Abstract*—Parkinson's Disease (PD) is a neurodegenerative disorder that causes the deterioration of motor coordination, speech, and other cognitive functions. Early detection is vital to reduce the progression rate and enhance the quality of life for patients suffering from Parkinson's Disease. The traditional methods of detecting Parkinson's Disease are based on observation and expertise, which may lead to delayed detection in the early stages. This study proposes the development of a machine learning-based intelligent system to detect Parkinson's Disease in its early stages using voice-derived biomedical features. The intelligent system will be based on the classification algorithm XGBoost, which is highly efficient and has the capability to produce high accuracy in the classification process using structured medical datasets. The biomedical features will be collected from the patient's voice, which will be affected in the case of Parkinson's Disease, and the XGBoost algorithm will be implemented to classify the patient's voice as normal or affected with Parkinson's Disease. The proposed intelligent system will be implemented using the Python Flask framework to create a web-based application that will be used to support the healthcare professionals in the classification process.**

***Index Terms*—Parkinson's Disease, Machine Learning, XGBoost, Voice Analysis, Biomedical Signal Processing, Early Detection, Data Science, Predictive Modeling, Healthcare Informatics, Clinical Decision Support System.**

I. INTRODUCTION

Parkinson's Disease is one of the most common neurodegenerative disorders in the world, and it is characterized by the progressive degeneration of dopamine-producing cells in the brain, resulting in motor disorders such as tremors, rigidity, and slowness of movement, and speech abnormalities. Among these motor disorders, speech abnormalities are often present in the early stages but remain subtle and hard to detect using physical examination. The detection of Parkinson's Disease is often based on the expertise and knowledge of neurologists, and this may cause delays in diagnosis. With the emergence of Artificial Intelligence and Machine Learning, disease detection tools have been developed, which can detect complex hidden patterns in data. Voice signals can be used as a medium for disease detection, as even the slightest change in the speech signal can indicate neurological damage. The aim of this research is to develop a reliable, scalable, and user-friendly web-based tool using voice signals and the XGBoost machine learning algorithm to detect Parkinson's Disease in the early stages.

II. LITERATURE REVIEW

Several research studies have been carried out to investigate the machine learning models that can be implemented in the detection of Parkinson's disease using biomedical and speech data sets. It has been shown that machine learning models can be used to detect Parkinson's disease using speech data sets as reliable biomarkers.

Md Abu Sayed et al. (2023) carried out a study on the analysis of vocal biomarkers using different machine learning models and reported that ensemble models can be used to achieve high accuracy in detecting Parkinson's disease. It was shown that preprocessing and feature extraction are important in improving the performance of machine learning models.

Govindu and Palwe (2023) carried out a study on the comparison of different machine learning models, including Random Forest, SVM, and Logistic Regression, and reported that ensemble models can be used to achieve more stable predictions compared to other models. It was shown that AI models can be used in the screening of patients with Parkinson's disease using remote healthcare.

Alshammri et al. (2023) carried out a study on the analysis of speech data sets to assess the performance of machine learning models in detecting Parkinson's disease and reported that speech features can be used to improve the accuracy of machine learning models in detecting Parkinson's disease.

Several recent research studies have shown that deep learning models can be used to detect Parkinson's disease using speech data sets with the aid of CNN and RNN models.

Although deep learning models can be used to achieve high accuracy in detecting Parkinson's disease, optimized machine learning models can be used to achieve a balance between accuracy and computational efficiency using models such as XGBoost.

The above research studies have shown that machine learning models can be used to detect Parkinson's disease using speech data sets as biomarkers.

III. EXISTING SYSTEM

The existing system of diagnosing Parkinson's disease is based on clinical evaluation and neurological examination of patients carried out by highly qualified specialists. Doctors examine patients for motor symptoms of Parkinson's disease, such as tremors, muscle stiffness, bradykinesia (slowness of movement), and postural instability. Besides motor symptoms, other symptoms of Parkinson's disease, such as medical history of patients and observations of patients' behavior, are considered during the diagnosis process. In some cases, MRI and CT scans of patients' brains are carried out to rule out other neurological disorders. However, there is no single definitive laboratory test for diagnosing Parkinson's disease in its early stages.

One of the major disadvantages of the existing system of diagnosing Parkinson's disease is its dependency on highly qualified specialists. Diagnosis of Parkinson's disease is often based on the judgment and expertise of neurologists. Parkinson's disease symptoms appear very mild in the early stages and often resemble other normal aging and movement disorders. Such mild symptoms of Parkinson's disease are very difficult to identify and diagnose. For instance, minor speech and motor impairments may not be clearly noticeable during brief periods of observation and monitoring of patients' behavior. As a result, patients are often diagnosed with Parkinson's disease only after its symptoms have become very pronounced. The process of diagnosing Parkinson's disease is often very time-consuming and expensive. Patients often need to visit hospitals frequently and undergo continued monitoring and examination before they are diagnosed with Parkinson's disease. Such prolonged periods of observation and monitoring of patients' behavior often put patients under a lot of physical, emotional, and financial strain, especially for those living far from hospitals and clinics where highly qualified neurologists are based. Besides, the process of diagnosing Parkinson's disease is often subjective, and sometimes patients may not receive proper and timely medical attention.

The other major disadvantage of the existing system of diagnosing Parkinson's disease is its failure to use biomedical voice data for diagnosing Parkinson's disease. Although research evidence suggests that abnormalities of speech are symptoms of Parkinson's disease, the existing system of diagnosing Parkinson's disease is not based on the use of automated voice analysis software for diagnosing Parkinson's disease

IV. PROPOSED SYSTEM

The system, as proposed, is based on machine learning principles, aimed at early detection of Parkinson's disease through voice-based biomedical analysis. Instead of relying on subjective judgment, this data-driven approach promises greater accuracy, consistency, and early detection of the disease. It is based on the concept of developing an intelligent decision-support tool, assisting doctors in making better-informed decisions.

As proposed, the system involves the collection of voice-based biomedical parameters from patients, such as jitter, shimmer, pitch, harmonic-to-noise ratio, and other nonlinear features. These parameters are likely to reveal subtle vocal impairments, which are likely to occur during the early stages of Parkinson's disease. Once the data is collected, preprocessing occurs, including noise removal, normalization, handling missing values, and scaling features. After data preprocessing, feature selection techniques are applied, which identify the most important features used in the prediction process. This improves the overall efficiency of the system. Finally, the XGBoost classifier is used, known for its high prediction accuracy and robust performance with structured data. After training the model on labeled data, metrics are used to ensure the reliability of the classifier. Once the model is trained, the system is integrated into a web-based application using the Python-Flask framework. It features an interactive user interface using HTML, CSS, and JavaScript, enabling doctors and hospital staff to easily input patient voice parameters. Once the data is input, the system processes the data in real time and generates the output, indicating whether Parkinson's disease is detected, along with the scores. This method is advantageous in several ways, including early detection, non-invasive, less reliance on subjective judgment, faster results, consistency, and overall decision-support, helping doctors make better-informed decisions.

V. METHODOLOGY

The project presents a step-by-step guide to catching Parkinson's disease at an early stage using machine learning. The process is divided into steps, including data collection, data cleaning and preparation, selection of features, training the model, evaluation, and finally, usage.

Data Collection

The process begins by collecting data from patients using voice-based biomedical data. The data collected consists of various features such as jitter, shimmer, pitch frequency, harmonic-to-noise ratio, and nonlinear speech parameters. These features are known to reflect the vocal changes in patients suffering from Parkinson's disease.

Data Preprocessing

After collecting the data, the next step is to clean and preprocess the data. This is done by filling in missing values, removing duplicates, and identifying outliers. The features are then normalized, and all data is placed in the same range, which is essential for the smooth functioning of the machine learning model.

Feature Extraction and Selection

After data preprocessing, the next step is to select the most important features in speech, which can help in the detection of Parkinson's disease. Feature selection is used to remove unwanted features, reducing the data and speeding up the process.

Dataset Splitting

The cleaned set of data is divided into two subsets, one for training and one for testing, usually 70/30 and 80/20, respectively. This is done for the model to learn and validate its accuracy without bias.

Model Training

The classification is done using the XGBoost algorithm. Using gradient boosting, the model identifies patterns between voice feature and disease status.

Model Evaluation

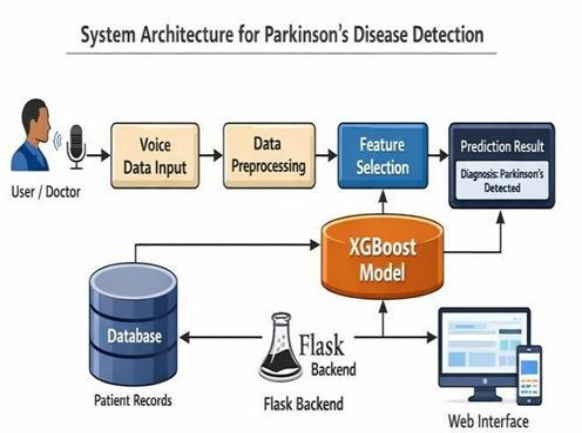
The model is evaluated based on accuracy, precision, recall, and F1-score. A confusion matrix is sometimes used for more detailed evaluation. Retraining is done if needed to obtain better results.

System Integration and Deployment

The model is integrated into a web-based application built using Python and Flask. Users input patient voice parameters, and the application processes them for real-time predictions, making it a convenient and reliable aid for screening.

In summary, this approach is accurate for prediction, efficient for processing, and convenient for deployment for early detection of Parkinson's disease.

VI. SYSTEM ARCHITECTURE



The Parkinson's Disease Detection is designed as an integrated system, and the voice data from patients is processed in a series of connected components to arrive at a prediction. It combines machine learning, database management, and web- based interface technologies to create a Parkinson's disease diagnosis tool that is both effective and easy to use. The major components involved in this process are voice data, preprocessing, feature selection, machine learning prediction, data management, and web interface communication.

Getting started with Machine Learning Based Early De…

The first step in this process is to input the voice data using the web interface, which can be done by patients or medical professionals. This voice data consists of parameters such as biomedical speech features, including jitter, shimmer, and frequency, which are directly linked to Parkinson's disease. After this, the process continues to the next step, which is data preprocessing.

In the data preprocessing step, the data is normalized and cleaned in such a way that it is ready for use in machine learning, which in turn increases the accuracy and reliability of the predictions.

Next in line is the feature selection module. In this module, the system focuses on the most relevant features in the speech that can be used for disease prediction. In this module, unnecessary features in the speech are eliminated. This results in improved performance of the system.

Next in the process is the XGBoost model. In this model, the features that have been selected in the previous module are used for the prediction of the disease. In this model, the features in the speech are used to predict whether a person has Parkinson's disease.

A database module has been added to the system. In this module, the results obtained from the system have been stored. In this module, the results obtained in the past have been stored.

A flask backend has been added to the system. In this module, the entire system has been connected. In this module, the results obtained from the system have been provided to the user in real time.

From the above explanation, it can be concluded that the system has been designed in a manner that there is efficient communication between all the modules. In the system, the process of Parkinson's disease prediction has been made efficient. In the system, the results have been provided to the user in real time.

VII. CONCLUSION

This system, as described, represents a smart and reliable way of identifying Parkinson's disease at an early stage through the use of machine learning. By analyzing various biomedical features like jitter, shimmer, and change in frequencies, the system can identify early signs of Parkinson's disease through the identification of speech patterns. This is a non-invasive method, which increases the chances of early diagnosis and timely intervention.

The use of XGBoost, as an algorithm, ensures high prediction accuracy and good performance in classifying data. When integrated into a web- based platform, the predictions are made in real- time, ensuring the system is user-friendly and can be used in hospitals. It is also useful in the sense that, unlike other systems, there is no need for deep technical knowledge, ensuring healthcare professionals can use the system.

This system, as described, is useful in the sense that, unlike other systems, there is no need for deep technical knowledge, ensuring healthcare professionals can use the system. It represents a useful tool in ensuring early intervention and reducing the need for subjective judgment. It is evident, through this project, the potential benefits of machine learning and health technologies in improving patient care, ensuring early intervention, and improving patient outcomes in the management of Parkinson's disease.

VIII. FUTURE WORK

The system has laid a good foundation for the intelligent screening of Parkinson's disease. At the same time, it is evident that there are various ways the system can be enhanced to improve its efficiency and effectiveness. For instance, the system can be enhanced to incorporate deep learning techniques such as the use of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) for the automatic extraction of features from the raw voice signal. The techniques have the capability to recognize complex patterns and can improve the efficiency of the system when more data is used.

Another way the system can be enhanced is by the incorporation of gait analysis, handwriting, wearable sensor data, and face expressions into the system. The use of a combination of biological signals will improve the efficiency and accuracy of the system while at the same time reducing the chances of false results.

Other ways the system can be enhanced include the incorporation of real-time voice capture and the automatic extraction of parameters from the signal. The system will be able to replace the current manual entry of parameters into the system. The system can also be enhanced to include a mobile application for the screening of the disease. The system will be able to offer telemedicine services, which will be very helpful for people living in remote areas.

Additionally, the system can be enhanced to incorporate the use of explainable AI techniques, which will be helpful in identifying the features used by the system for the prediction of the disease. The system can also be enhanced to incorporate the retraining of the model with the latest data to improve the efficiency of the system for the long term. The system can also be enhanced to incorporate the electronic health records system.

The system will be enhanced to make it comprehensive and efficient for the early detection of Parkinson's disease.

REFERENCE

[1] Little, M. A., McSharry, P. E., Roberts, S. J Costello, D. A., & Moroz, I. M. (2007). Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. Bio Medical Engineering Online.

[2] Sakar, C. O., Serbes, G., Gunduz, A., et al. (2013). A comparative analysis of speech signal processing algorithms for Parkinson's disease classification. Expert Systems with Applications.

[3] Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2012). Enhanced classical

dysphonia measures and machine learning. IEEE Transactions on Biomedical Engineering.
[4] Md Abu Sayed et al. (2023). Parkinson's disease detection using vocal biomarkers and ML. arXiv Preprint.
[5] Govindu, A., & Palwe, S. (2023). Early detection of Parkinson's disease using ML. Procedia Computer Science.
[6] Alshammri, R., et al. (2023). Machine learning-based Parkinson's diagnosis. Frontiers in Artificial Intelligence.
[7] Alalayah, K. M., et al. (2023). Feature selection for Parkinson's voice analysis. Diagnostics (MDPI).
[8] Tusar, M. T. H. K., et al. (2023). Multimodal Parkinson's detection using ML. arXiv Preprint.
[9] Swain, C., et al. (2023). Hybrid ML model for Parkinson's detection. Open Biomedical Engineering Journal.
[10] Bukhari, S. N. H., & Ogudo, K. A. (2024). Ensemble learning for Parkinson's detection. Mathematics (MDPI).