Parkinson’s Disease Detection using Machine Learning

1. **Shriraj Vallamdeshi (Roll.no:14)**

ABSTRACT

Parkinson's disease (PD) is a neuro degenerative movement disorder in which the signs initially appear as a mild tremor in one hand and a general sense of stiffness, but gradually get worse. Over 6 million people are impacted globally. Currently, non-specialist clinicians have not been able to definitively diagnose this disease, especially in the early stages of the illness when it is very challenging to identify the symptoms. The issue can be resolved with a low mistake rate by utilizing machine learning techniques. As a result ,data mining offers a prediction method for systematically identifying Parkinson's disease. Parkinsonism cannot be diagnosed using a traditional test,; we suggest a statistical method based on the most prevalent PD symptoms,

Given that there is no established test to identify Parkinsonism, we suggest a statistical method based on the three most prevalent PD symptoms—gait, tremors, and micrographia. In order to determine the classification method that provides the maximum accuracy in diagnosing PD patients, this involves studying the correlation between the symptoms and classifying the obtained data using several classification algorithms. By combining inputs from Parkinson's sufferers and healthy individuals, our suggested method produces reliable results for the data sets obtained as the input. Our work will demonstrate how early disease identification can extend a patient's life and lead to a serene existence through appropriate medical care and medication.

INTRODUCTION

Parkinson's disease is characterized by the death of dopaminergic neurons in the substantia nigra pars compacta of the midbrain. Coordination problems, bradykinesia and voice alterations are among the signs of this neurodegenerative disease. Parkinson's disease (PD)patients can also develop dysarthria, an impairment of the motor-speech system that affects respiratory, phonatory, articulatory and prosodic functions. Parkinson's disease symptoms can be different for everyone. Early signs are mild that goes unnoticed .Symptoms usually begin on one side of your body and gets worsen on that side, afterwards it affects both the sides. Parkinson's symptoms may include:

• Tremor

• Slowed movement

• Rigid muscles.

• Impaired posture and balance.

• Loss of automatic movements

• Speech changes

• Writing changes

The Cause of Parkinson Disease is Still Unknown. But several factors appear to play a role, Including:

• Genes

• Environmental

• Triggers

Parkinson's disease, commonly known as Tremor, is affected by a reduction in dopamine levels in the brain which damages a person's motion functions, or physical functioning. It is one of the world's most common diseases. Intermittent neurological signs and symptoms result from these lesions, which get worse as the disease progresses .Because aging causes changes in our brains, such as loss of synaptic connections and changes in neurotransmitters and neurohormones, this condition is more frequent among the elderly. With the passage of time, the neurons in a person's body begin to die and become inimitable. The consequences of neurological problems and the falling dopamine levels in the patient's body show gradually, making them difficult to detect until the patient's condition requires medical treatment. However, the symptoms and severity levels are different for individuals. Major symptoms of this disease are deficiency in speech, short-term memory loss, loss of balance, and unbalanced posture.

Every year, 8.5 million individual cases of this disease are registered worldwide, as per the World Health Organization (WHO) report in 2019. The chance of developing this disease rises with age; currently, there are 4% of sufferers worldwide under 50 years of age. As invasive procedures are typically used for diagnosis and therapy, which are both expensive and demanding , a reasonably straightforward and accurate way to diagnose this disease looks very relevant.

1.1. Machine Learning-Based Detection of Parkinson's Disease

Over the past few decades, researchers have looked at a new way of detecting this disease through ML techniques, a subset of artificial intelligence (AI). Clinical personnel might better recognize these disease patients by combining traditional diagnostic indications with ML.

As walking is the most common activity in every person's day-to-day life, it has been linked to physical as well as neurological disorders. This disease, for example, has been identifiable using gait (mobility) data. Gait analysis approaches offer advantages such as being non-intrusive and having the potential to be extensively used in residential settings .Few researchers have attempted to combine ML methods to make the procedure autonomous and possible to do offline.

Furthermore, persons with the subject disease in its early stages might experience speech problems These include dysphonia (weak vocal fluency), repetitious echoes (a tiny assortment of audio variations), and hypophonia (vocal musculature disharmony). Information from human aural emissions might be detected and evaluated using a computing unit .

1.2. Research Problem and Motivation

Early PD detection in PD patients is a crucial challenge. Even if their health deteriorates, people can enhance their quality of life if they receive an early diagnosis. Another issue is that the diagnosis of PD requires a number of steps, including gathering a thorough neurological history from the patient and examining their motor abilities in various environments.

The majority of recent studies deal with the homo dataset (text, speech, video, or image). Problems with dataset modification and multi-data handling procedures have been highlighted in the suggested study. The effectiveness of disease prediction is regulated as a result of the examination of a particular dataset. More real-time solutions are made possible by the use of machine learning-based techniques for multivariate data processing. The multi-variate vocal data analysis (MVDA) is driven to provide multiple dataset attribute-based Parkinson's disease identification utilizing machine learning approaches. This study examines the potential for improving multi-variate and multimodal data processing, which aids in raising the disease detection rate. The existing research simultaneously concentrates on various ML-based techniques such as support vector machines, naïve Bayes, K-NN, and artificial neural network evaluations of Parkinson's data based on voice features. The MVDA employs extensive datasets and machine learning approaches to improve disease identification based on these works. The incorporation of numerous patients' multivariate acoustic characteristics in the proposed MVDA is encouraged. The subjective disease has been diagnosed with the help of proposed machine learning techniques under the MVDA system.

LITERATURE SURVEY

Previous studies to predict PD have been implemented on MRI scans, gait and genetic data, but research on audio impairment for early detection is minimal. For instance, Bilal et. al. studied genetic data to predict the onset of PD in senior patients with SVM model. They trained an SVM model to reach an accuracy of 0.889, while this research paper describes an improved SVM model with an accuracy of 0.9183. These results also corroborate the merits of classification of PD based on audio data, over genetic data. Raundale, Thosar and Rane used keystroke data from UCI telemonitoring dataset to train a Random Forest classifier to predict the severity of PD in older patients. Cordella et. al. use audio data to classify PWP, however their models are heavily reliant on MATLAB. Our research uses open-source models trained in Python, that are faster and memory efficient. Majority of research done emphasizes the use of deep learning in PD detection, such as, Ali et. al.who explain the use of ensemble deep learning models applied to phonation data, to predict the progress of Parkinson's disease. Their work lacked the use of feature selection that would improve Deep learning model (DNN) performance. Hence, this paper implements PCA on 22 attributes to select 7 major voice modalities in PD detection. Huang et. al. aim to reduce PD diagnosis dependence on wearable equipment by training a traditional decision tree on 12 complex speech features of the MDVR-KCL dataset. Wodzinski et. Al. trained a ResNet model on images of audio data, instead of training the model on the nuances of the frequency of audio. aimed to remove subjectivity of doctors in prediction of PD using an unbiased ML model, however their results achieved peak accuracy of 85% only. Wang et. al implemented 12 machine learning models on 401 voice biomarkers dataset to classify patients as PD or not. They built a custom deep learning model (DEEP) with a classification accuracy of 96.45%, however the model was expensive due to large memory requirements. Alkhatib et. al. implemented a linear classification model with 95% accuracy to characterize shuffling movement of PD patients. Their study focused on gait of patient and future work encouraged the use of audio and sleep data to improve the results. performed spatial-temporal analysis of brain MRI scans. They implemented decision trees, random forest and KNN to detect Mild Cognitive Impairment (MCI) in PWP. However, dataset was small and artificial data augmentation was needed. A. U. Haq and without considering healthy individuals in a lower age bracket. explain the importance of ML to detect PD, as subtle non-motor symptoms can be missed during subjective evaluation by a doctor. Their work reviews 209 studies based on dataset, ML methods and outcomes achieved. Based on our literary review, we have implemented a PD classification model on audio data. Through our findings, we aim to contribute to the advancement of detection of PD through telemedicine. Keeping in mind, past research on biomarker data and models implemented, our research aims to explore KNN, logistic regression, random forest regression and SVM models to classify Parkinson's patient audio data. Our preliminary findings show that K nearest neighbor model is the best performing model with an accuracy of 91.83% and sensitivity of 0.95 colleagues implemented L1-support SVM, without feature identification on vowel phonation dataset for neurological disorder patients. Their paper focused on patient age group of 46-85 years,

PROPOSED METHODOLOGY

Artificial Neural Network Algorithm

ANN is a subfield of deep neural networks that predict how the human brain works. In general, there is a significant distinction between the human brain and ANN. The brain has 'n' number of parallel neurons, whereas the machine only has a finite sum of processors. Additionally, neurons are meeker and more relaxed than computer processors. Another major disparity between computer systems and the brain is the ability to process information on a larger scale. Neurons are made up of synapses or networks that operate together [64,68]. In this article, the main aim is to classify the functionality of ANN techniques in the early detection of this disease which is built on the subsequent phases:

1. Identifying the responsibility and function of ANN in the detection of this disease.
2. Making observations on labels and features of datasets.
3. Grouping the types of the studied disease centered on their symptoms.
4. Examining the accurate outcomes.

These outcomes can be further used in the medical sector as direction for developers considering ANN deployment to enhance the civic health potential as a reaction to the studied disease .

In the experiment of an artificial neural network, the dataset was split into two parts i.e., the training dataset (80%) and the test dataset (20%). The classification results of the artificial neural network were found to be very high in the form of the average accuracy score which was the highest among all the classification methods, i.e., 96.7% shown in Table 5 and graphical representation is shown in

1.Dataset

The dataset of recorded speech signals was obtained from Max Little of the University of Oxford contains the details of the dataset. This dataset has an assortment of acoustic speech measures from 195 persons, where 147 persons have Parkinson's disease. All the attributes in the dataset characterize an individual voice measure, and each tuple represents a total number of voice recordings made by these people. The objective of the dataset is to differentiate fit persons compared to the unhealthy using the "status" column, which is set to negative for fit persons and positive for those having the disease.

2.Parkinson's Disease Diagnosis Based on Voice Analysis and Machine Learning

Some studies have concentrated on the acoustic level or the fluctuations in fundamental frequency (F0) caused by vocal activities. The effects of power spectral analysis of F0 phonation in persons with sensorineural audibility loss and the disease have been examined in. F0's rhythm was unique in the incidence and amplitude of the diseases. Further, the study demonstrated that the F0 analysis can be a useful tool for neurological diseases under investigation. The autocorrelation function approach was used to find the basic frequencies of speech transmissions. According to the concept, Parkinson disorder is frequently described as a simple neuro-motor disorder.

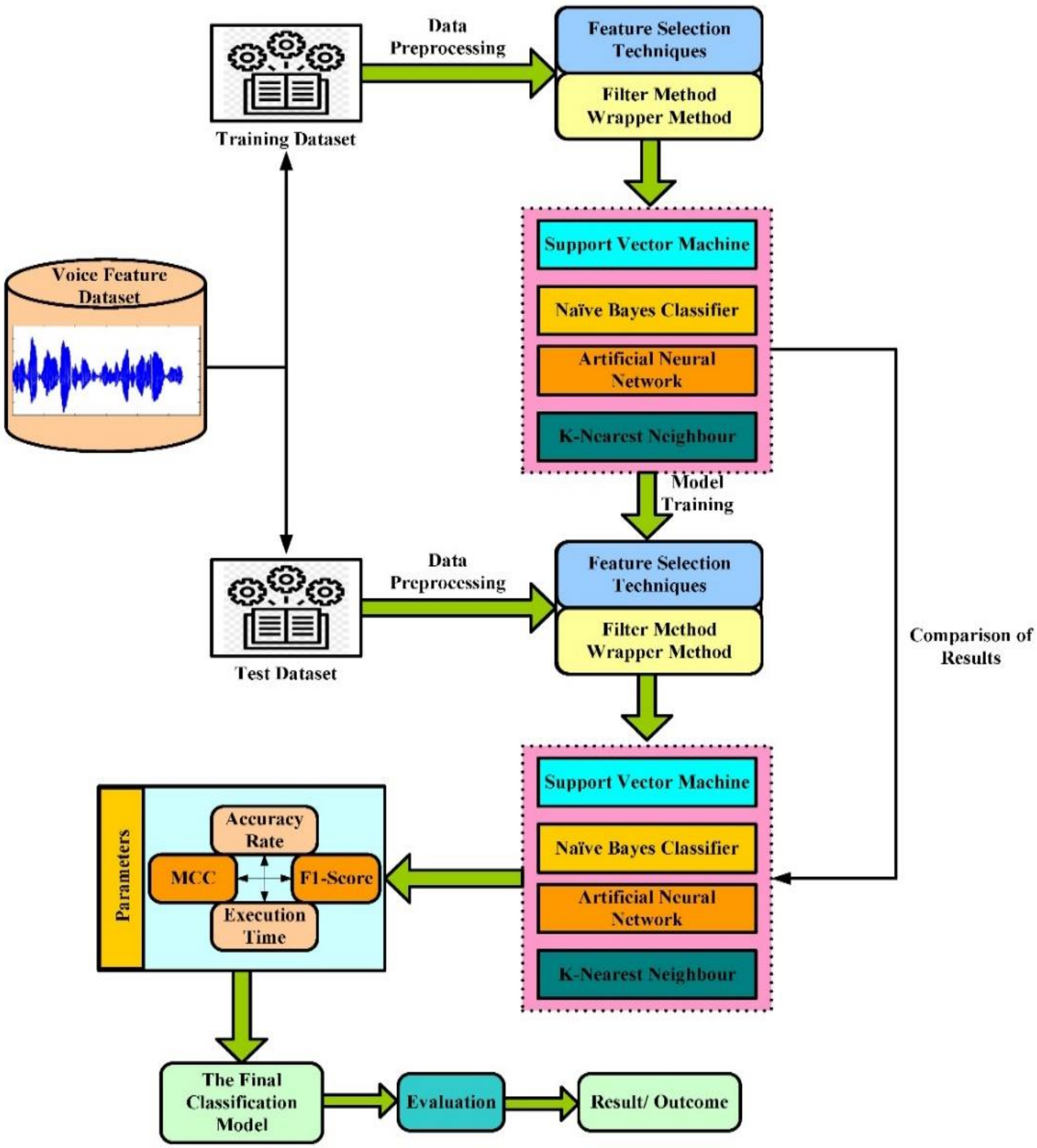
The understanding and generation of pitch characteristics in a group of patients were examined to confirm the idea. Conventional medications, such as LDOPA, define that in the early stages of PD, LDOPA is a very effective treatment of subjective disease. In the authors use deep learning to categorize the patient's speech data as "severe" and "not severe". The evaluation measures employed in this study were the unified Parkinson's disease ranking scale (UPDRS). The motor UPDRS examines the patient's motor ability on a 0–108 scale, while the entire UPDRS provides a range of scores from 0 to 1766.

3.Classification of Parkinson's Disease with ML Classifier

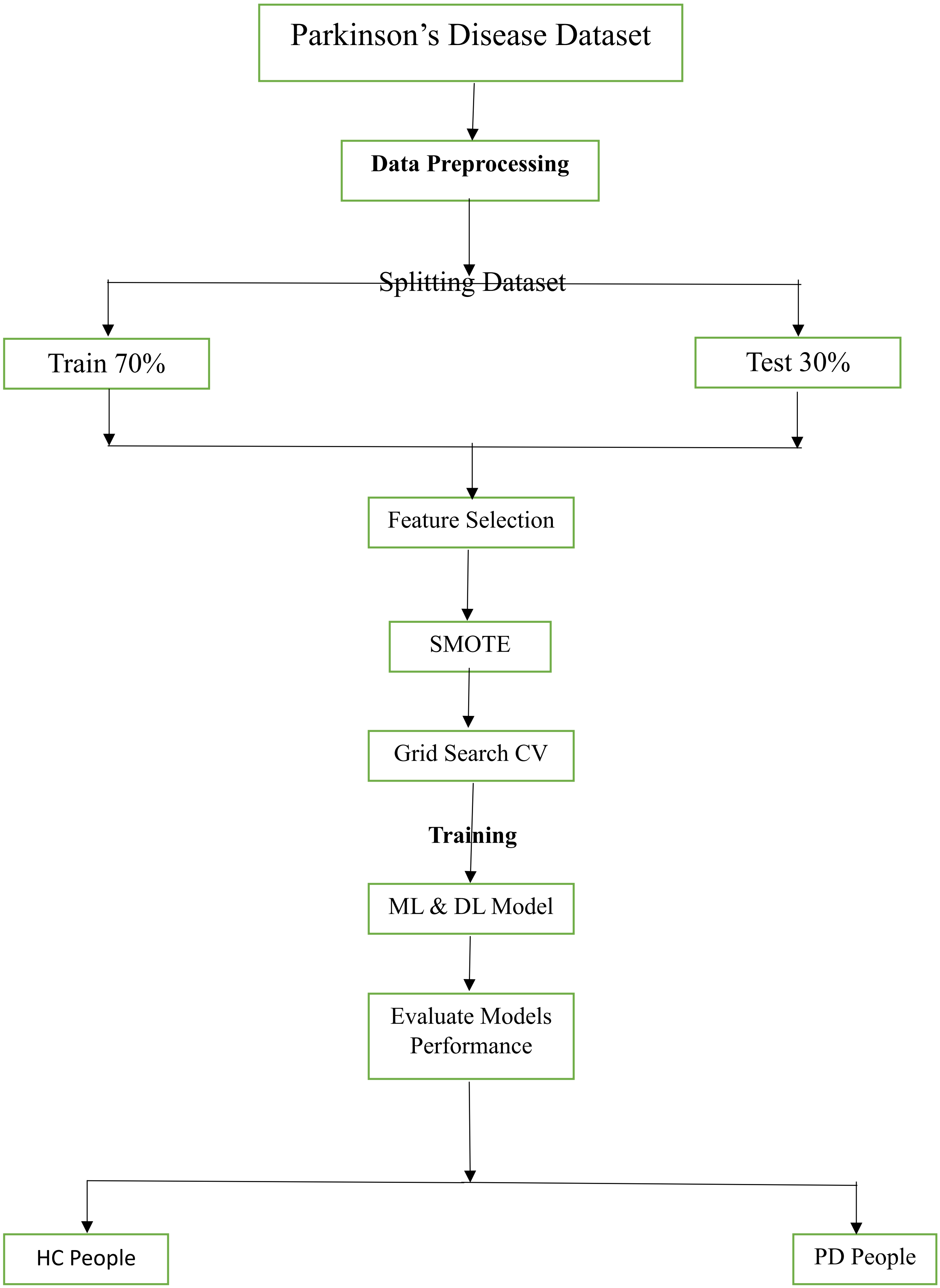
In this technique, we'll use an ML classifier to classify the disease. First, we select a target variable of patient health status and measure the number of patients in this report. We visualize the data graphically after assessing the health status of a patient. Two types of datasets were developed: 80% of the dataset was used for training and 20% for the testing dataset. In the following, the score of 0 represents the healthy persons in the sample, whose count is 48, and 1 represents the patients with Parkinson's disease, whose count is 147. The count of Parkinson's disease patients in the dataset: 147 out of 195 (75.38%). The count of healthy persons in the dataset: 48 out of 195 (24.62%).

4.Building of Machine Learning Techniques with Classifier Evaluation Metrics

By using different types of classifiers, it becomes easy to detect the disease. Classification sensitivity, Matthews's correlation coefficient (MCC), accuracy, specificity, F-score (F-measure), and other measurement parameters are used to distinguish it. Each of these measurement criteria includes a formula for calculating it and determining which classifier is the most qualitatively appropriate for the analysis. It is requisite to focus on the confusion matrix before developing these criteria .



Training Dataset
Data Preprocessing
Feature Selection Techniques
Filter Method
Wrapper Method
Voice Feature Dataset
Support Vector Machine
Naïve Bayes Classifier
Artificial Neural Network
K-Nearest Neighbour
Model Training
Test Dataset
Data Preprocessing
Feature Selection Techniques
Filter Method
Wrapper Method
Comparison of Results
Support Vector Machine
Naïve Bayes Classifier
Artificial Neural Network
K-Nearest Neighbour
Parameters
Accuracy Rate
MCC
F1-Score
Execution Time
The Final Classification Model
Evaluation
Result/ Outcome



Parkinson’s Disease Dataset
Data Preprocessing
Splitting Dataset
Train 70%
Test 30%
Feature Selection
SMOTE
Grid Search CV
Training
ML & DL Model
Evaluate Models Performance
HC People
PD People

EXPECTED OUTCOME

1) Improved Early Diagnosis: The primary goal is to demonstrate that machine learning techniques can effectively diagnose Parkinson's disease in its early stages based on symptoms like gait, tremors, and micrographia. The outcome should be a more accurate and timely diagnosis, potentially allowing for better patient outcomes.

2) Classification Algorithms: The research should determine which classification algorithms are most effective in identifying Parkinson's disease based on the selected symptoms. This information will be crucial for implementing practical diagnostic tools.

3) High Accuracy: The paper mentions an accuracy score of 96.7% for the artificial neural network. Achieving high accuracy in disease diagnosis is a key outcome, as it demonstrates the reliability of the proposed approach.

4) Potential for Further Research: If the research is successful, it may open up avenues for further studies, including the exploration of other symptoms, data sources, or machine learning techniques to enhance early diagnosis.

5) Comparison with Existing Methods: The outcomes should include a comparison of the proposed machine learning-based approach with existing diagnostic methods, highlighting the potential advantages and limitations of the new approach.

Work	**Working Method**	**Advantages and Implementations**	**Scope for Future Work**
1. Data Collection and Preprocessing	- Collecting relevant data sources, such as patient health records and voice recordings. - Preprocessing data to clean and format it for machine learning.	Accurate data collection ensures reliable model training. - Data preprocessing is essential for feature engineering and model training.	| - Exploration of more diverse and extensive data sources. - Improved data preprocessing techniques for handling noise and outliers.
2. Feature Selection and Engineering*	Identifying relevant features or symptoms for Parkinson's disease prediction	Feature engineering can uncover hidden patterns and relationships in the data.	Incorporation of advanced feature selection techniques. - Development of novel features for early diagnosis.
3. Choice of Machine Learning Algorithms	Selection and implementation of machine learning algorithms, e.g., decision trees, SVM, neural networks. - Hyperparameter tuning and optimization.	Different algorithms can be compared for their performance. - Optimized hyperparameters can enhance model accuracy.	Exploration of ensemble methods for improved prediction. - Consideration of deep learning approaches.
*4. Model Training and Validation	Training models on the prepared dataset. - Cross-validation to assess model generalization. - Evaluation metrics such as accuracy, F1-score, ROC AUC	Model training allows for learning patterns from data. - Validation ensures the model's ability to generalize to new data. - Evaluation metrics provide quantitative measures of model performance.	Incorporation of larger and more diverse datasets. - Comparison of multiple evaluation metrics for comprehensive assessment.
5. Real-Time Monitoring and Telemedicine Integration	Development of a system for real-time patient monitoring and remote diagnosis. - Integration of machine learning models into telemedicine platforms	Allows for continuous monitoring and early detection of symptom changes. Extends the reach of healthcare to remote areas.	Enhancements in remote monitoring technology. - Privacy and security measures for telemedicine integration.
*6.Patient-Friendly	Creation of user-	Facilitates patient	Development of

Interfaces*	friendly interfaces for patients to report symptoms or provide data. - Integration with wearable devices and mobile applications.	involvement and data collection. - Wearable devices provide continuous data streams for analysis.	mobile apps with advanced reporting and feedback features. - Integration with smart wearable devices.
7. Ethics and Privacy Considerations	Ensuring data privacy and ethical use of patient information. - Complying with relevant regulations and guidelines	Builds trust with patients and stakeholders. - Adherence to regulations avoids legal and ethical issues	Advancements in privacy-preserving machine learning techniques. - Continuous monitoring of ethical and legal standards.
8. Interpretability and Explainability	Ensuring machine learning models provide interpretable results. - Implementing methods to explain model predictions.	Allows clinicians and patients to understand why a particular prediction was made. - Increases trust in AI-based diagnosis	Research on explainable AI (XAI) for complex machine learning models. - Development of user-friendly explanation interfaces.
9. Continuous Model Monitoring and Improvement	Developing mechanisms for continuous model monitoring. - Periodic retraining of models with new data	Ensures model performance remains reliable over time. - Incorporates new data trends and patterns for improved accuracy.	Automated model retraining based on dynamic data updates. - Adaptive learning approaches for continuous improvement
10. Collaboration with Healthcare Professionals*	Engaging medical experts in the model development process. - Collaborative decision-making based on model predictions.	Leverages clinical expertise for model refinement and validation. - Fosters a multidisciplinary approach to Parkinson's disease diagnosis.	Development of specialized training programs for healthcare professionals in machine learning. - Seamless integration with clinical workflows.

REFERENCES

1. https://chat.openai.com/
2. https://ieeexplore.ieee.org/document/9725409
3. https://www.analyticsvidhya.com/blog/2021/07/parkinson-disease-onset-detection-using-machine-learning/
4. https://www.bing.com/ck/a?!&&p=79bbfb227736a591JmltdHM9MTY5ODc5NjgwMCZpZ3VpZD0yMzkyOGRmNi04ZjZhLTZhMmQtMmJmZC05ZjBkOGVkODZiMmQmaW5zaWQ9NTIxNw&ptn=3&hsh=3&fclid=23928df6-8f6a-6a2d-2bfd-9f0d8ed86b2d&psq=parkinson%27s+disease+prediction+using+ml&u=a1aHR0cHM6Ly93d3cuc2NpZW5jZWRpcmVjdC5jb20vc2NpZW5jZS9hcnRpY2xlL3BpaS9TMTg3NzA1MDkyMzAwMDA3OA&ntb=1
5. https://app.litmaps.co/