Prediction of Parkinson Disease Using Machine Learning Algorithm Models

Abstract

Parkinson's disease (PD), the second most widespread neuro-degenerative condition after Alzheimer's disease, is a neurological ailment that is most frequently observed in those over 60. Patients from all over the world are affected by the condition, which is generally identified by clinical evaluations of motor symptoms. However, due to the subjective nature of conventional diagnostic techniques and the subtlety of some of the disease's symptoms, they could misclassify the condition. Early diagnosis is difficult since early non-motor signs of PD may be subtle and frequently ignored.

To address this issue, a new Machine Learning technique has been developed to identify PD based on premotor features, allowing for early detection of the disease. The study utilized a relatively small dataset consisting of 183 healthy individuals and 401 early PD patients to evaluate the performance of two machine learning algorithms, SVM and XGBoost, in detecting PD.

The project's objective is to develop a reliable and accurate predictive model for Parkinson's disease, utilizing machine learning algorithms. The algorithms chosen for this project are SVM and XGBoost, with Jupyter Notebook serving as the backend tool and Anaconda as the front-end tool. The input data consists of various features, including age, gender, and medical history, which the algorithms analyze and process to predict the probability of developing Parkinson's disease. With an outstanding accuracy rate of 100% for both SVM and XGBoost algorithms, this model can help with the early and precise diagnosis of Parkinson's disease, leading to better treatment outcomes and

quality of life for patients. Moreover, the project can also identify individuals at high risk of the disease, enabling preventative measures. The successful implementation of this project can pave the way for similar predictive models for other neurodegenerative diseases, contributing significantly to the advancement of medical science.

1. Introduction

Nearly one million Americans today experience Parkinson's Disease (PD), a degenerative condition that impairs mobility throughout the body. Despite its widespread occurrence, PD has no established diagnostic procedure. It is challenging to identify at an early stage since current diagnostic procedures are restricted to a clinician's judgement in conjunction with a neurological exam.

As a result, machine learning algorithms are required to forecast PD symptoms without using invasive clinical testing. The impact of PD on voice patterns is one area that hasn't received much attention. Researchers believe that a brief speech recording, made even on a typical smartphone, can be used to identify the illness. Many additional disciplines might profit from the use of this technology, including biometrics

In this research, our goal is to develop a classification model that can reliably identify Parkinson's disease (PD) using voice patterns with at least a 95% accuracy rate. We will categorise speech recordings from PD patients and healthy test individuals using machine learning methods like support vector machines (SVM) and XGBoost. Based on how well the model performed on the test data, the most accurate model will be chosen. Our objective is to develop a small model that is accurate enough to recognise voice patterns from a little quantity of data and is usable in a wide range of applications. We want to contribute to the early diagnosis and treatment of PD by creating a reliable and usable diagnostic tool.

1.2 Background

One of the most important jobs for medical professionals and institutions is disease prediction because it helps them make decisions about patient treatment that are well-informed. Making the wrong choice may result in treatment delays or even death. The business component of healthcare is one more factor that cannot be disregarded. Patients are constantly searching for the finest treatment alternatives, which is why the healthcare sector is a market that is expanding quickly. Unfortunately, patients frequently find themselves in a tough situation due to the high expense of therapy.

Therefore, researchers are working tirelessly to find ways to reduce the cost of healthcare while still maintaining its quality. One approach is the development of an umbrella platform that can help to solve some of the challenges faced by patients, healthcare institutions, and medical professionals. The goal is to make healthcare more affordable and accessible to all patients.

In the case of Parkinson's disease, early detection is crucial to help patients lead healthier lives and reduce healthcare costs. Researchers are exploring innovative methods such as the use of artificial intelligence and wearable technology to aid in disease detection and tracking. By leveraging technology, medical professionals can analyze patient data, identify patterns, and make more accurate predictions about disease progression and treatment options.

Disease prediction and treatment are essential components of healthcare. Researchers are continually seeking new ways to make healthcare more affordable and accessible to all patients, while maintaining high-quality care. By embracing technology and developing innovative solutions, we can improve patient outcomes and make significant strides towards a healthier future for all.

1.3 Motivation

Worldwide, Parkinson's disease is a major cause of mortality and disability. The Parkinson's Disease Foundation estimates that 1 million persons in the United States will have Parkinson's disease by 2020 (Marras et al., 2018). The foundation for medical therapy of Parkinson's disease is based on neuropathologic and histopathologic investigations (Gelb, Oliver, neurology, & 1999, n.d.). Parkinson's disease is often diagnosed by assessing the sensitivity and specificity of the illness's defining symptoms. To better understand Parkinson's disease, including its prevalence and risk factors, in-depth research of clinical, pathologic, and nosological features are required (Aarsland, Andersen, neurology, & 2003, n.d.).

Parkinson's disease, which typically affects adults over 50, presently has no recognised aetiology. Parkinson's disease has no known treatment, however it is feasible to lessen symptoms, especially in the beginning stages (Singh, Pillay, neurobiology, & 2007, n.d.). "Speaking dysfunction in an extensive number of persons with Parkinson's disease," n.d., states that speech impairment affects almost 90% of those with the condition.

Many persons with Parkinson's disease struggle to pay for the required medical care since it can be highly expensive to treat the condition. Parkinson's disease must be identified early in order to lower healthcare expenditures and enhance patient outcomes. Machine learning-based platforms have recently demonstrated promising results in the detection of Parkinson's disease. Support Vector Machines (SVM) and XGBoost are two well-liked algorithms that are employed in the prediction of Parkinson's disease. These algorithms can analyse huge databases, find patterns and abnormalities, and spot Parkinson's disease early warning symptoms.

Early identification of Parkinson's disease is crucial for improving patient outcomes and lowering healthcare costs. Parkinson's disease is a major global health problem. SVM and XGBoost are two potential examples of recent developments in machine learning and data analysis techniques that open up new possibilities for the early recognition and diagnosis of Parkinson's

disease. We can significantly advance the goal of a healthy future for everyone by continuing to investigate cutting-edge technology and methods.

1.4 Challenges

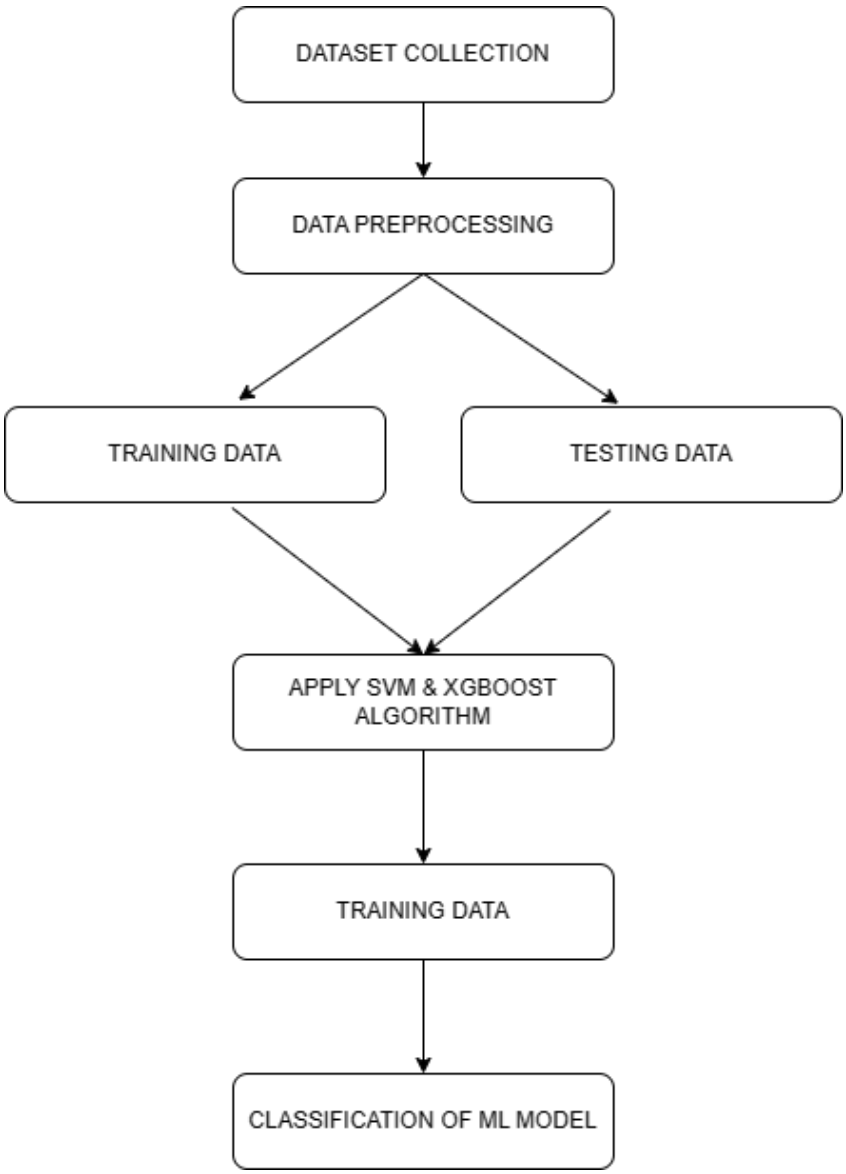
Parkinson's syndrome is a widespread neurological disorder that is predicted to increase in prevalence over the coming decades. Even though Parkinson's disease diagnosis can be difficult, new developments have enhanced our knowledge of the illness's early phases. In addition to the testing of preliminary Parkinson's disorder test results and the finding of genetic subtypes and multiple genetic variations linked to the development of the disease, these breakthroughs also involve the validation of clinical diagnostic criteria.

The development of diagnostic biomarkers has also advanced significantly, with novel tissue and fluid indicators being researched and genetic and imaging tests already being included into standard clinical regimens. Parkinson's disease is progressing towards a diagnostic strategy backed by biomarkers, enabling early detection and classification of several subtypes with differing prognoses. Additionally, potential therapies that can change the course of illness are now being created.

Over the past few years, machine learning algorithms have been successfully used to identify patterns in medical data that can predict the onset of Parkinson's disease. These algorithms are capable of analyzing large amounts of data, such as genetic and imaging tests, to identify potential biomarkers for the disease.

2.Planning & Requirements Specification

2.1 System Planning :



MODULE DESCRIPTION

DATA COLLECTION:

Moving information or datasets via an initial file, folder, of app to a library or comparable programme is known as data collecting or loading. Data loading's primary objective is to transmit data that is digital from an original location and load it into a tool for data processing or storage. This stage, which enables the loading of datasets with pertinent characteristics that may be utilised to train a predictive model, is crucial to the data analysis process. These characteristics may include the demographics, medical history, physical exam, and neurological exam outcomes of the patient. Once the required data has been entered, it may be processed and analysed to yield valuable insights that can be used to spot trends and make predictions regarding Parkinson's disease.
Typically, database-based extraction and loading strategies involve data loading. In this procedure, data is taken from an external source and converted into a format that is compatible with the target application. Typically, this transformation entails converting the data's original source location's format to another one, like.CSV or DAT.

Data loading is a crucial stage in the data analysis pipeline while trying to anticipate Parkinson's disease. Datasets containing pertinent information that might be utilised to develop a predictive model are loaded throughout this phase. These attributes may include the patient's demographics, medical background, and the outcomes of their physical and neurological exams.

The pertinent data may be further processed and analysed to derive useful insights once it has been imported into a database or equivalent programme. On the basis of this data, machine learning algorithms may be used to find patterns and forecast the progression of Parkinson's disease.

In conclusion, data loading is an essential stage in Parkinson's disease prediction. It entails moving information from a source file, folder, or programme to a database or comparable programme so that it may be be further analyzed and processed to extract meaningful insights.

Data Preprocessing

Any missing data were imputed in order to guarantee that all the algorithms employed in the study could manage missing values. Some algorithms, like XGBoost, can manage missing data without imputation, though. All missing values were imputed depending on their data type to make the comparison process easier. The median of the entire entries was used to replace missing values in numbers, and the mode of the complete entries was used to replace missing values in categories. This method made it possible for all algorithms employed in the study to handle missing data consistently.

Data Cleaning

His module involves cleaning the data, followed by data clustering, which groups the data according to the project's specifications. To guarantee completeness, any blank values in the dataset are found and, if required, replaced with default values. During this stage, data that has to be formatted differently is changed. Data pre-processing refers to the full procedure of preparing and cleaning the data before the prediction stage. Following pre-processing, the data is used for the prediction and forecasting phases.

Splitting of data

In advance of training and testing the model, normalising the data comes after data cleansing. In order to do this, the data must be divided into training and testing sets, with the training set being used to fine-tune the algorithm's parameters. To make sure that all values are placed under the same scale, feature extraction is utilised.

A dataset should be divided into three subsets for machine learning: training, test, and validation sets. The model is trained using the training set, which also helps identify the best parameters for the model to learn from the data.

The trained model's generalisation capability is assessed using the test set. This refers to the model's capacity to find patterns in fresh, unexplored data following training over a training set. To prevent overfitting, or the model's failure to generalise, which can happen when the model is trained on the same data used for testing, it is crucial to utilise distinct subsets for training and testing.

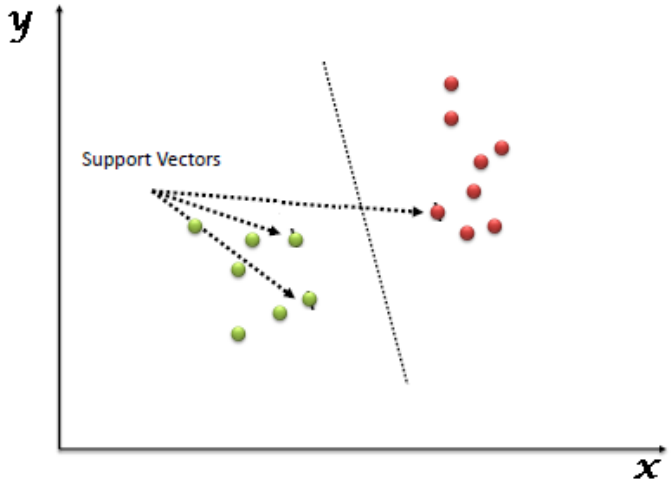
DATA TRAINING

Algorithms rely on data to learn, and they analyze relationships, make decisions, and develop understanding based on the training data they receive. The quality and quantity of training data are as important as the algorithms themselves in determining the success of a data project. However, even if a large amount of well-structured data is available, it may not be labeled in a way that is suitable for training a model. For instance, autonomous vehicles require labeled images of cars, pedestrians, and street signs, while sentiment analysis projects require labels that help an algorithm identify slang or sarcasm. In order to use data for training, it needs to be enriched or labeled, or more data needs to be collected. However, the data that is stored may not be prepared for training classifiers, which is where professional help can come in handy. With years of experience, they can help create a training set for any type of data that will make the models more successful.

SUPPORT VECTOR MACHINE:

Although it is primarily utilised for classification tasks, the "Support Vector Machine" (SVM) is a form of supervised machine learning strategy that may be used to address regression or classification problems. Each piece of data is represented as a point in a space of n dimensions by the SVM, where n is the total number of features. Each feature's value is allocated to a certain coordinate as its value. The algorithm then performs classification by locating the hyperplane that is capable of distinguishing between the two classes (see the graphic below for an example).

Individual observers' coordinates inside a dataset are referred to as support vectors. In a classification issue, a Support Vector Machine, or SVM, method is created to discover the hyper-plane or plane that best divides the two classes, given the Support Vectors lying closest to the hyper-plane



Classification

The preparation of the data is the initial step in the detection of Parkinson's disease. We may use a variety of machine learning algorithms to categorise and ensemble the data after it has been cleaned and organised. Then, we evaluate the effectiveness of various approaches, ascertain their accuracy, and pinpoint the key characteristics that are crucial to the prediction process. On the Pima Indians Parkinson dataset, we use several classification and ensemble strategies to do this.

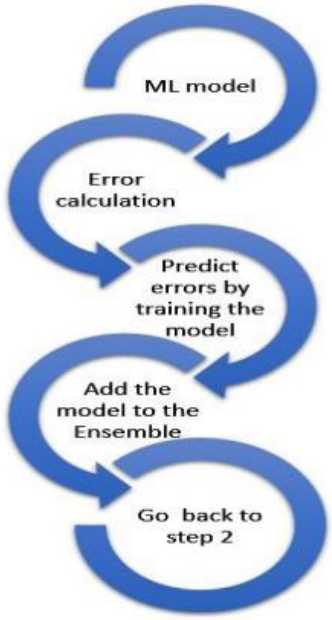
SUPPORT VECTOR MACHINE

The well-known supervised machine learning technique supported vector machine (SVM) is frequently utilised for categorization issues. In high-dimensional space, it divides two classes by forming a hyperplane. Regression analysis may also make use of the hyperplane. SVM has the ability to categorise entities that lack data support as well as instances in certain classes. The hyperplane nearest to any training point of any class is used to execute the separation. The algorithm functions as follows:

• It chooses the hyperplane that divides the classes most effectively.
• The method determines the Margin, or the distance between the planes and the data, in order to choose a better hyperplane.
• The likelihood of miss conception increases with decreasing class distance and vice versa.
• The algorithm chooses the class with a high margin, which is calculated as the sum of the distances to the positive and negative points.

XGBoost Classifier:

A powerful machine learning model created for categorisation is the XGBoost algorithm. It is quick and uses a strengthened decision tree to perform better. The total process is improved by the employment of this classification model, both in terms of speed and efficiency. Extreme Gradient Boost, often known as XGBoost, is a group computation that is based on GBDT. The boosting method is predicated on the notion that many decision trees outperform a single one in terms of performance. Even though individual decision trees might not perform well, numerous decision trees working together produce superior outcomes.



The machine learning algorithm known as XGBoost makes use of decision tree-based grouping and gradient boosting. While decision tree-based algorithms are thought to be the best for small to medium organised or prepared data, artificial neural networks frequently beat other algorithms and architectures in prediction challenges involving unstructured input (such as photos and text).

.

2.2 Requirements

2.2.1user requirements

Functional and non-functional requirements:

Requirements analysis is a critical step in determining if a software or system project will be successful. Functional and non-functional needs are the two broad categories into which requirements may be divided.

Functional Requirements:

Functional requirements are the essential attributes and functions that the system must expressly provide for the end user. As a requirement of the agreement, all of these features must be incorporated into the system. Functional requirements are outlined as the system's inputs, processes, and anticipated outputs. Functional needs, as opposed to non-functional criteria, are visible in the finished product and are obvious to the user. Functional requirements can include things like requiring users to authenticate when they log in, shutting down the system in the event of a cyberattack, and sending a confirmation email to users who register for the first time on a software system.

Non-functional requirements

The quality criteria that the system must achieve in order to comply with the project agreement are known as non-functional requirements. The importance and extent of these aspects' application vary from project to project. These specifications, which are often referred to as non-behavioral requirements, cover topics like portability, security, maintainability,

dependability, scalability, performance, reusability, and flexibility. Examples of non-functional criteria include sending emails with a maximum 12-hour delay from a specific action, responding to each request in under ten seconds, and launching the website in under three seconds when there are more than 10,000 concurrent visitors.

2.3 System Requirements

2.3.1 Hardware Requirements

TThe hardware specifications for a system can serve as the basis for a contract for its implementation, therefore they should provide a thorough and cohesive description of the complete system. During the system design phase, these criteria serve as the software engineers' jumping-off point. They specify the intended functionality of the system rather than describing how it should be used.

OPERATING SYSTEM: Intel I5
RAM : 4GB
Hard Drive: 40 GB

2.3.2 SOFTWARE REQUIREMENTS:

The software requirements document serves as the specification of the system, encompassing both a definition and a detailed listing of requirements. It outlines the system's intended functionality, rather than prescribing how it should be implemented. The software requirements serve as the foundation for creating the software requirements specification, which is a more detailed and precise description of the software to be developed. This document is helpful in estimating costs, planning team activities, executing tasks, and monitoring the team's progress throughout the development process.

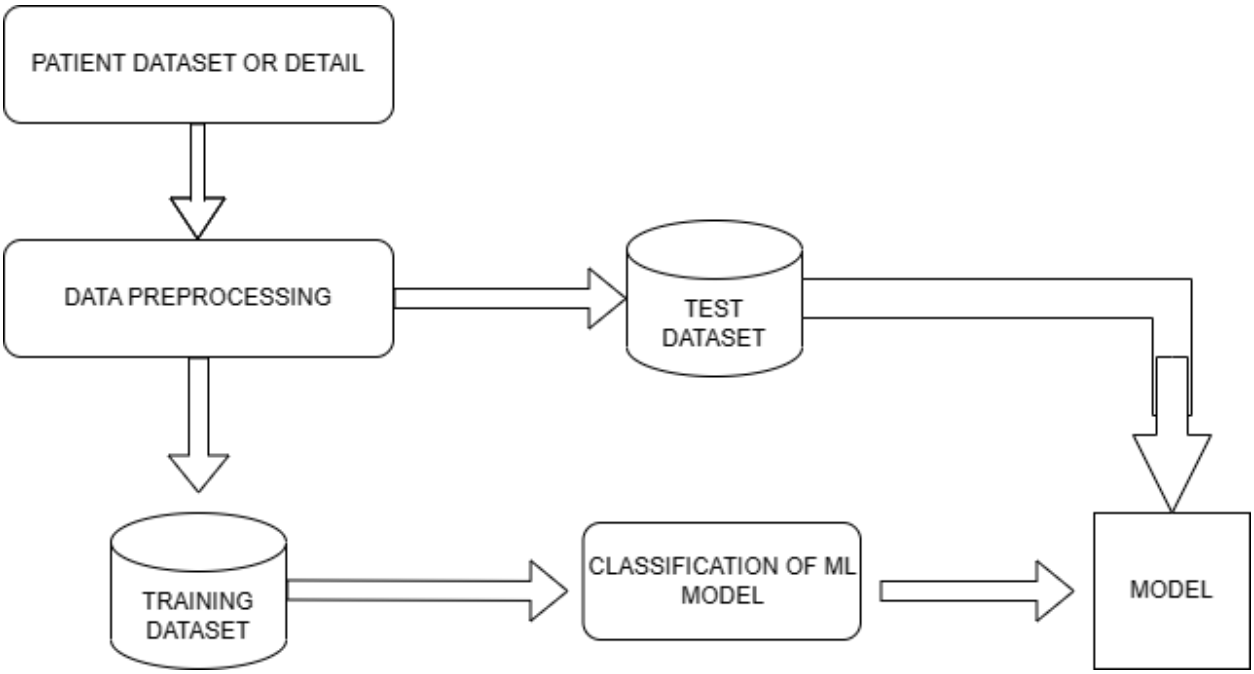
PYTHON IDE : Anaconda Jupyter Notebook
PROGRAMMING LANGUAGE : Python

3. System Design

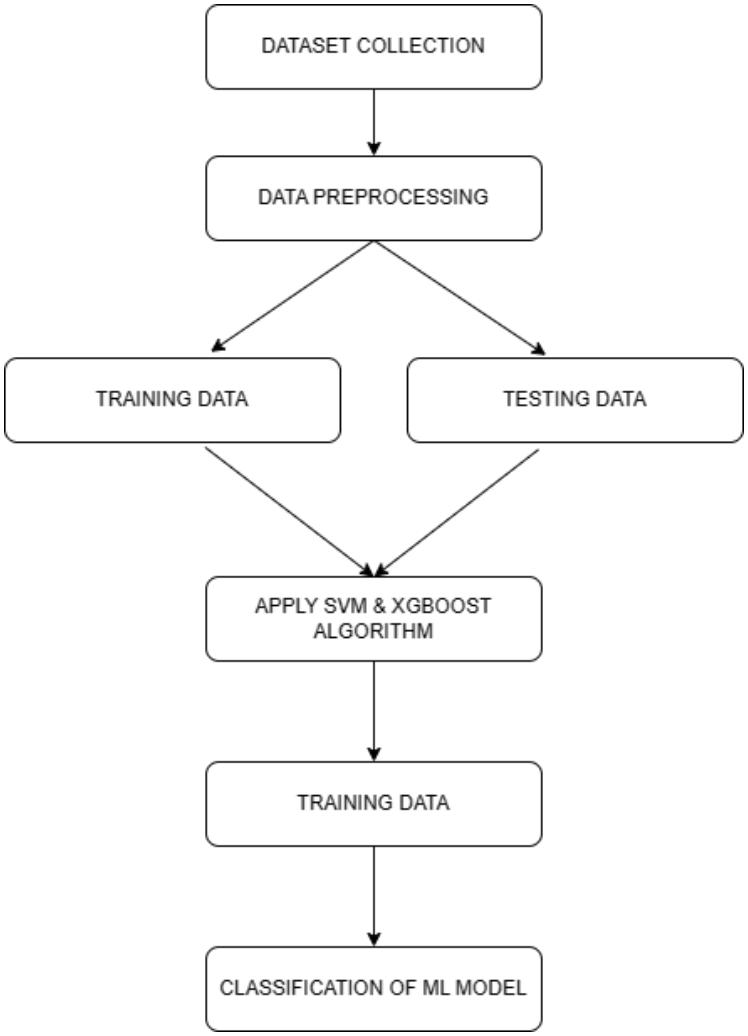
The interface, modules, and data required to satisfy certain criteria are defined as part of the system design process. The foundation of it is system theory. The creation of a system architecture that offers the information and data required for the system's execution is the main goal of system design. This entails thinking about the various system parts and how they will interact with one another to get the desired result.

.

3.1 ARCHITECTURE DIAGRAM

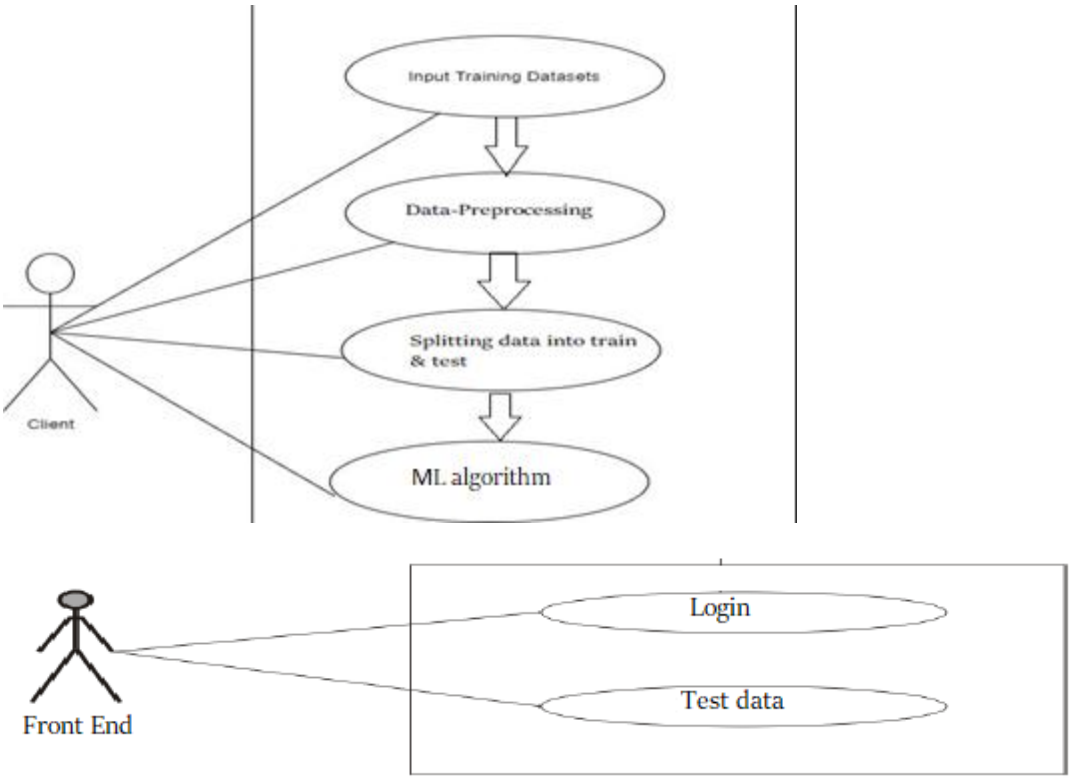


3.1.2DATA FLOW DIAGRAM



3.1.3 USECASE DIAGRAM:

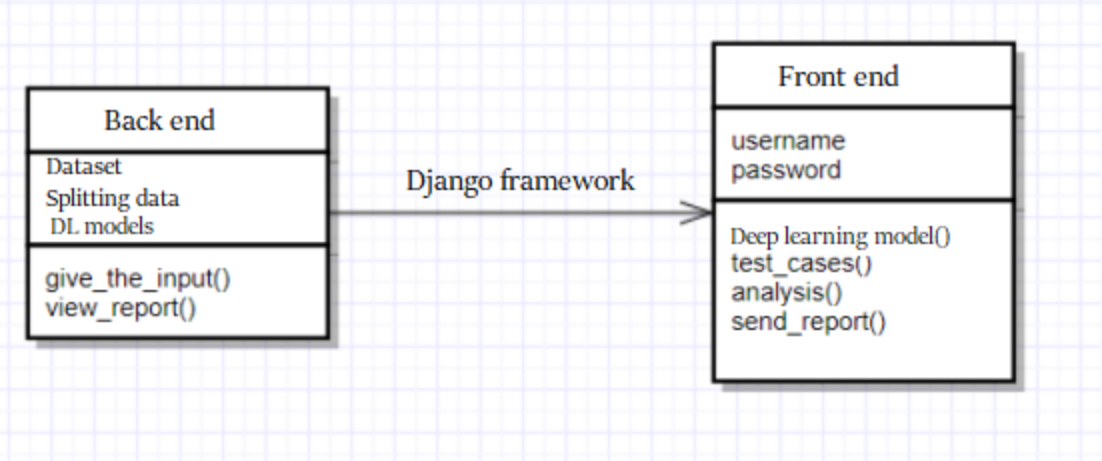
Use case diagrams are used to identify the functions that are provided by the use cases, the actors that interact with the system, and the relationships between the actors and the functions.



3.1.4 FRONT END MODULE DIAGRAMS:

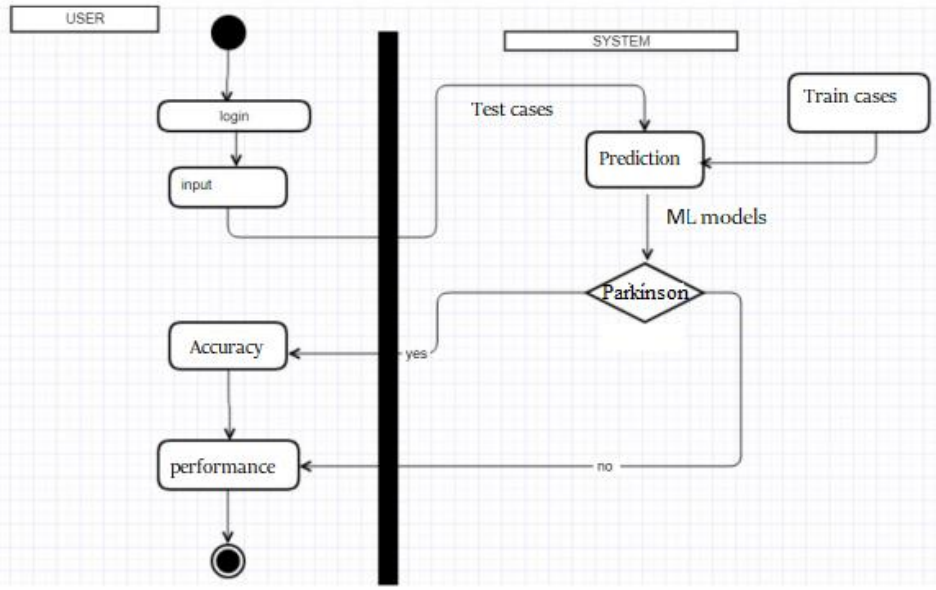
3.1.5 CLASS DIAGRAM:

A static diagram of this kind that displays a static representation of an application is called a class diagram. Class diagrams are helpful in creating executable code for the software programme in addition to helping with the visualisation, description, and documenting of many parts of a system.



3.1.6 ACTIVITY DIAGRAM:

The Activity Diagram is an effective tool for modeling the functionality of a system, as it portrays the activities, the types of flows between these activities, and the corresponding responses of objects to these activities.



4. Implementation of System

Method

Dataset

The dataset used in this study was comprised of voice recordings for the detection of Parkinson's illness and was obtained from the UCI Machine Learning Respirotary website. A total of 195 samples were included in the collection, including 48 voice data samples without Parkinson's disease and 147 voice data samples with Parkinson's disease.

	name	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	MDVP:Jitter(%)	MDVP:Jitt
0	phon_R01_S01_1	119.992	157.302	74.997	0.00784	
1	phon_R01_S01_2	122.400	148.650	113.819	0.00968	
2	phon_R01_S01_3	116.682	131.111	111.555	0.01050	
3	phon_R01_S01_4	116.676	137.871	111.366	0.00997	
4	phon_R01_S01_5	116.014	141.781	110.655	0.01284	
...	...	...	...	...	...	
190	phon_R01_S50_2	174.188	230.978	94.261	0.00459	
191	phon_R01_S50_3	209.516	253.017	89.488	0.00564	
192	phon_R01_S50_4	174.688	240.005	74.287	0.01360	
193	phon_R01_S50_5	198.764	396.961	74.904	0.00740	
194	phon_R01_S50_6	214.289	260.277	77.973	0.00567	

195 rows × 24 columns

Data Preprocessing

Imputation was carried out to make sure all algorithms could tolerate missing values. Some algorithms, such as XGBoost, can, however, automatically deal with missing variables without the requirement for imputation. The missing values were imputed depending on their data type to make the comparison easier. The median value of the entire values was used to replace

missing entries for numerical data types. For categorical data, the mode value of the entire items was used to replace any missing elements.

Data cleaning

In this module, the data is subjected to a cleaning process to remove any errors or inconsistencies. The cleaned data is then organized based on specific requirements, a process referred to as data clustering. Additionally, any missing values in the dataset are identified and replaced with default values. Any necessary changes to the data format are also made during this data pre-processing stage. Once the pre-processing is complete, the data is then used for prediction and forecasting purposes. It's important to note that this entire process is critical for ensuring accurate and reliable results.

- ### Splitting of data

Prior to training and testing the model, the data must first be normalised once it has been cleaned. The algorithm is then trained on the training set, with the testing set being set aside. The data is then divided into training and testing sets. A model is created throughout the training phase using the reasoning, algorithms, and values of the features in the training dataset. Bringing all feature values to the same scale is the aim of feature extraction.

The training, testing, and validation sets should be divided into three subsets when using a dataset for machine learning. This guarantees that the model is assessed on hypothetical data in order to avoid overfitting, eventually leading to more precise and trustworthy conclusions.

Training set:
A dataset is often divided into two subsets in machine learning: the training set and the test set. The training set is used to develop a model's optimum parameters, which are variables that the model must discover through data analysis.

Test set:

After the trained model has been trained on the training data, it is evaluated against the test set to determine how well it generalises, or how well it can spot patterns in fresh, unexplored data.

In order to prevent overfitting, which happens when the model gets too complicated and is unable to generalise to new data, it is crucial to employ separate subsets for training and testing. We can ensure that the model is tested on unknown data by employing distinct subsets, producing findings that are more accurate and trustworthy.

- ## Classification

Following data preparation, we use a variety of machine learning algorithms to forecast mental disease. These approaches are applied to the CSV dataset and include classification and ensemble methods. The major goal is to evaluate the effectiveness and precision of these procedures. Additionally, we want to pinpoint the crucial qualities or factors that contribute significantly to the prediction process.

By applying machine learning techniques, we can gain insights into mental health and predict the likelihood of mental illness in individuals. Additionally, identifying the important features can help us understand the key factors contributing to mental illness, ultimately leading to better prevention and treatment strategies.

Classifier Training:

A classifier in machine learning is a function that analyses input information and forecasts the appropriate class label. Based on the learning function and underlying assumptions, many types of classifiers may be created. Numerous classifiers have been used in neuroimaging research to forecast mental disease.

When using classification methods, it's crucial to take into account the dimensionality issue brought on by the relatively large number of characteristics and the few samples.

This means that the potential for overfitting must be taken into account, and steps should be taken to ensure that the model is not too complex and can generalize well to new data. Overall, selecting an appropriate classifier is critical for accurate and reliable predictions in mental illness studies.

Data splitting

We divided the full dataset into a 70% training set and a 30% test set for our studies. The test set was used to assess the performance of the trained model whereas the training set was utilised for resampling, hyperparameter adjustment, and model training.

To guarantee the same data split each time the programme was run, we supplied a random seed (i.e., any random number) when splitting the data. This is crucial for the consistency and reproducibility of our findings. We can be sure that any changes in the model's performance are brought on by adjustments to the method or parameters rather than variations in how the data was split by utilising the same random seed.

DATA TRAINING

Algorithms derive knowledge from data. They discover relationships, gain understanding, make decisions, and assess their confidence based on the training data they are fed. The efficacy of your data project is determined as much by the quality and quantity of the training data as by the algorithms themselves.

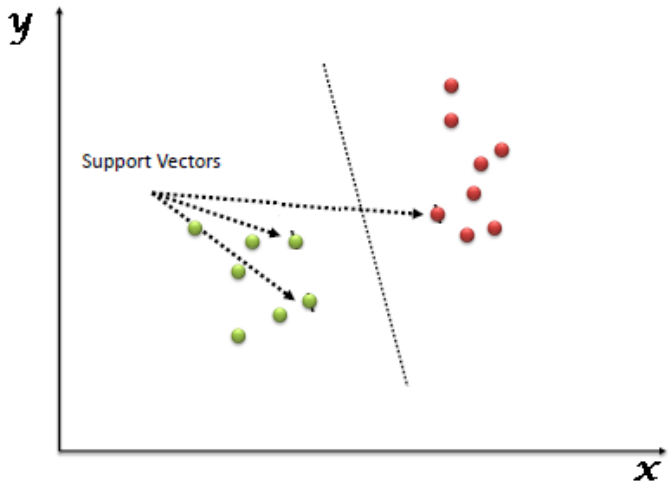
However, even if you have amassed a large amount of well-organized data, it may not be properly labeled to facilitate training your model. For instance, autonomous vehicles require labeled images, with each vehicle, pedestrian, and road sign annotated. Similarly, sentiment analysis projects necessitate labeled data to help an algorithm detect when someone is using slang or sarcasm. Chatbots require entity extraction and thorough syntactic analysis rather than raw language data.

Therefore, the data you intend to employ for training usually necessitates enrichment or labeling. Alternatively, you may need to collect additional data to enhance the performance of your algorithms. The data you've collected may not be ideal for training your classifiers if you are striving to develop a high-quality model. Fortunately, we specialize in creating exceptional training data. We have labeled over 5 billion rows of data for some of the most forward-thinking corporations worldwide. Regardless of whether the data in question is comprised of images, text, audio, or any other form of data, we can assist you in developing the training set that will make your models successful.

ALGORITHM IMPLEMENTAION

SUPPORT VECTOR MACHINE(SVM)

A The Support Vector Machine (SVM), a supervised machine learning technique, may be used to solve classification or regression problems. The bulk of its applications are for classification concerns, even though it may be used for both. n dimensions of space, where n is the number of features in the dataset, are used to represent each data point. This is the way SVM works. The approach is then used to identify the hyperplane that most effectively separates the two classes. Support vectors are the positions of certain observations that are closest to the hyperplane. In essence, the SVM model is the hyperplane or line that has the largest margin of separation between the two classes.



- **Classification**

We first prepare the data before using a variety of machine learning algorithms to look for Parkinson's disease. Applying several classification and ensemble algorithms to the Pima Indians Parkinson dataset is the main goal of this study. This analysis's main objectives are to assess the efficacy of machine learning methods, ascertain their accuracy, and pinpoint the most relevant elements that are essential to prediction. The following are some of the methods used for this purpose: [insert the particular methods here].

1) Support Vector Machine-

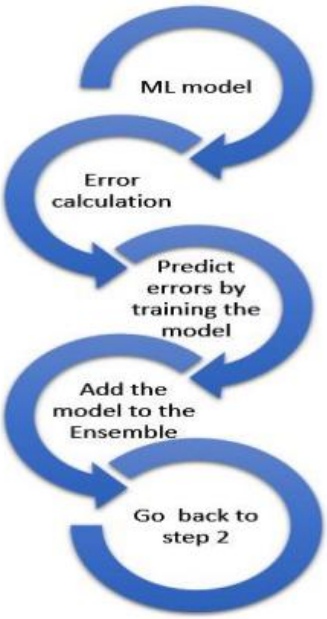
The support vector machine, or SVM, is a supervised machine learning technique. SVM is the most used classification technique. For the purpose of dividing two classes, SVM generates a hyperplane. SVM can differentiate between instances in certain classes in high-dimensional space and can classify objects for which there is no supporting data. Separation on a hyperplane is done at the nearest training point for each class. Algorithm: Pick the hyperplane that divides the class most effectively.

• To select the ideal hyperplane, you must first establish the Margin, or the separation between the planes and the data.

We must, then, choose the class with the greatest margin. The total of the distances to the positive and negative points is the margin.

XGBoost Classifier:

Researchers created the very successful machine learning classification system known as the XGBoost algorithm. It is a performance-enhancing, enhanced decision tree that is incredibly quick. This classification prototype's goal is to increase the model's effectiveness and speed. Extreme Gradient Boost, often known as XGBoost, is a group computation that is based on GBDT (Gradient Boosting Decision Tree). Since each decision tree may perform badly, the main principle behind the boosting technique is that several decision trees outperform a single one. The performance increases when several trees are used. While decision tree-based algorithms are thought to be the best for prediction issues requiring structured input, such as text and images, artificial neural networks frequently exceed all other algorithms or structures in such instances.



Data Prediction and forecasting:

The XGBoost algorithm, a very effective machine learning classification technique, was developed by researchers. It is a rapid, upgraded decision tree that improves performance. The purpose of this categorization prototype is to boost the model's efficiency and speed. GBDT (Gradient Boosting Decision Tree) is the foundation of the group computing known as Extreme Gradient Boost, or XGBoost. The primary idea behind the boosting strategy is that several decision trees outperform a single one since each decision tree may perform poorly. When several trees are employed, performance improves. Artificial neural networks routinely outperform all other algorithms or structures in prediction problems requiring structured input, such as text and images, even though decision tree-based techniques are regarded to be the best in these situations.

PERFORMANCE MATRICES:

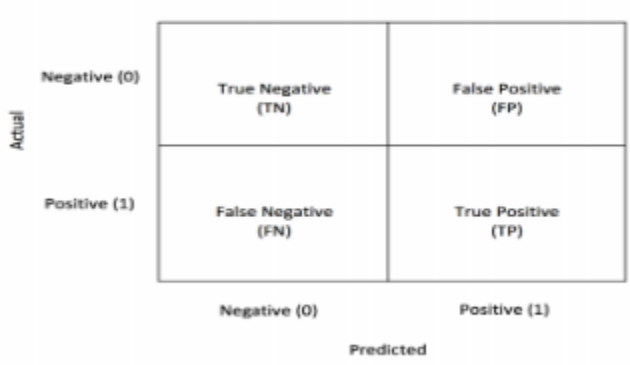
The data was divided into two groups, training data and testing data, containing 70% and 30% of the total data, respectively, to assess the performance of six different algorithms.

Using Enthought Canopy, these algorithms were run on the same dataset to produce the desired results. Predicting accuracy, or the percentage of accurate predictions, was the primary assessment criterion employed in this study to assess the algorithms' success rate. An equation may be used to determine the accuracy.

$$\text{Accuracy} = (TP+TN) / (P + N)$$

CONFUSION MATRIX:

Because it is simple to understand and can be used to determine other crucial metrics like recall, precision, etc., accuracy is the most often used assessment statistic in predictive analysis. The confusion matrix, which is a NxN matrix with N representing the number of class labels in the classification issue, summarises the overall performance of a model when applied to a dataset.



A table displays all algorithms' true positive (TP), true negative (TN), false negative (FN), and false positive (FP) predictions. A positive class is correctly predicted as positive by the model in true positive (TP), and the actual number of positive classes in the sample was predicted by the model. A positive class is incorrectly predicted as negative in false negative (FN), and the actual number of negative classes in the sample was predicted by the model. False positive (FP) predicts a negative class as positive, and the actual number of positive classes in the sample was predicted by the model. Finally, a negative class is correctly predicted as negative in true negative (TN), and the actual number of negative classes in the sample was predicted by the model.

Results, Conclusion and Future Work

Observations and Discussion

The experiment's findings demonstrated that the SVM and XGBoost algorithms may be used to build a machine learning model on the data to predict Parkinson's illness. Anaconda was used for the front end and Jupyter Notebook for the back end. Seventy percent of the dataset was utilised for training and thirty percent for testing. Both the SVM and the XGBoost algorithms exhibited 100% prediction accuracy, according to an evaluation of both algorithms' accuracy. Due to their excellent accuracy, these algorithms may one day help in the diagnosis of Parkinson's disease.

CONCLUSION AND REFERENCES

This study presents a detailed assessment of machine learning-based Parkinson's disease diagnostic tools. This project's goal was to present a comprehensive review of imaging and machine learning tools used in the diagnosis of mental diseases. In order to guarantee prompt treatment and care for patients, the significance of early identification and prediction of Parkinson's disease was also emphasised. While the majority of machine learning methods used by earlier writers yielded encouraging results, creating a quicker classifier employing a unique ML architecture and particular methodology may increase accuracy even more. Convolutional neural networks with various numbers of layers and nodes will be developed as a subsequent stage to compare accuracy.

Future Work

In the future, we may build on this research by investigating different methods for anticipating Parkinson's disease using a variety of datasets. Currently, we categorise patients using a binary property (1 for patients with diseases, 0 for patients without diseases). Future research will instead categorise individuals and identify various Parkinson's disease stages using a range of features.