1. **INTRODUCTION**
   1. **Problem Definition**

Technological advancements lead to smart objects being capable of identifying, locating, sensing, Predicting and connecting and thus leading to new forms of communication between people and things using Modern Technology. It’s a revolutionary change in the field of medicine and global technology to incorporating the life science with Technological innovations. Data Analytics is playing an major role in predicting and analyzing diseases and it is helping us to identify the best treatment for patients suffering from diseases; Neurological disorders are diseases of the brain, spine and the nerves that connect them. There are more than 600 diseases of the nervous system. This work promotes the development in the field of medicine for some of the neural diseases like Alzheimer, Parkinson’s and others. In monitoring them challenge is to find the best treatment. To analyze and predict best treatment for patients suffering from neurological diseases using Data analytics.Worldwide, one billion people are affected by neurological disorders, including 50 million who have epilepsy and 24 million with Alzheimer disease and other neurological diseases. After many new discoveries and research about the biology of the Parkinson’s disease and after almost 200 years since Parkinson’s disease was discovered, a diagnosis still depends on identifying the main features described by James Parkinson. Worldwide, it is estimated that four to six million people suffer from the Parkinson’s disease and in the USA complications from Parkinson’s disease are the 14th leading cause of death .The research in the field of Machine learning and Data analytics has led to the development of a large series of new tools and intelligent devices that can be used in the field of health services.It is important to note that, although there are common symptoms of PD, they can vary greatly from person to person. Moreover, how these symptoms change over time and whether other symptoms of PD emerge differ from person to person. Most people who develop the symptoms of PD do so sometime after the age of 50, but PD can affect younger persons as well. There are an estimated 1 million Americans living with PD and more than 10 million people worldwide. PD is a neurodegenerative disease. There is a loss of neurons (nerve cells) in certain areas of the brain, including a region called the substantia nigra (sub-STAN-she-uh NYE-gruh), Latin for “black substance.” The neurons in this region (which appear dark under a microscope) produce a neurotransmitter (a chemical messenger that allows neurons to communicate) called dopamine. Dopamine helps to regulate movement. As the number of cells in the substantia nigra decreases, there is less dopamine available in the brain. Dopamine is important to maintain normal movement patterns. This loss of dopamine is the reason that many treatments for PD are intended to increase dopamine levels in the brain.

* 1. **Project overview and specifications.**

Parkinson’s disease is an illness that affects the part of your brain that controls how you move your body. It can come on so slowly that you don't even notice it at first. But over time, what starts as a little shakiness in your hand can have an impact on how you walk, talk, sleep, and think. You're more likely to get it when you're 60 and older. It's also possible for it to start when you're younger, but that doesn't happen nearly as often. There's no cure for Parkinson's disease, but you can get treatment and support to help manage the symptoms.

Parkinson disease (PD) is one of the most common neurologic disorders, affecting approximately 1% of individuals older than 60 years and causing progressive disability that can be slowed, but not halted, by treatment. The 2 major neuro pathologic findings in Parkinson disease are loss of pigmented dopaminergic neurons of the substantia nigra pars compacta and the presence of Lewy bodies and Lewy neurites.

When that system is working well, your body moves smoothly and evenly. But when you have Parkinson's, the cells of your substantia nigra start to die. There's no replacing them, so your dopamine levels drop and you can't fire off as many messages to control your body.

Early on, you won't notice anything different. But as more and more cells die, you reach a tipping point where you start to have symptoms.

That may not be until 80% of the cells are gone, which is why you can have Parkinson's for quite a while before you realize it.

**How Does Parkinson's Affect the Body?**

The telltale symptoms all have to do with the way you move. You usually notice problems like:

Rigid muscles. It can happen on just about any part of your body. Doctors sometimes mistake early Parkinson's for arthritis

Slow movements. You may find that even simple acts, like buttoning a shirt, take much longer than usual.

Tremors. Your hands, arms, legs, lips, jaw, or tongue are shaky when you're not using them.

Parkinson’s disease is a nervous system disorder that affects movement. The disease is progressive and is marked by five different stages.

1. **Stage 1:** Mild symptoms that do not typically interfere with daily life, including tremors and movement issues on only one side of the body.
2. **Stage 2:** Symptoms continue to become worse with both tremors and rigidity now affecting both sides of the body. Daily tasks become challenging.
3. **Stage 3:** Loss of balance and movements with falls becoming frequent and common. The patient is still capable of (typically) living independently.
4. **Stage 4:** Symptoms become severe and constraining. The patient is unable to live alone and requires help to perform daily activities

.

1. **Stage 5:** Likely impossible to walk or stand. The patient is most likely wheelchair bound and may even experience hallucinations.
   * 1. **Objectives**

The main objective is to predict whether the patient is suffering from parkinson’s disease or not.

Parkinson is caused when certain nerve cells that produce dopamine begin to die.

This study aims to address four complementary challenges. To address the need for effective data management and reliable data accumulation,

**Challenge 1** involves gathering the harmonized and aggregating complex, multisource, and multisite Parkinson’s disease data. Then we apply machine learning techniques and controlled variable selection, e.g., knock off filtering.

**Challenge 2** is to identify salient predictive features associated with specific clinical traits, e.g., patient falls. (i.e using various symptoms)

**Challenge 3** involves forecasting patient symptoms using alternative techniques based on the selected features and evaluating the classification performance using internal (statistical) and external (prospective data) algorithms.

**Challenge 4**, is to provide the best treatment based on the result which occurred by applying algorithms and to provide better diagnosis by individual analysis of the data.

* + 1. **Parkinson’s disease symptoms**
* **Tremor:**-The characteristic tremor in PD is a slow, rhythmic tremor that typically starts in one hand, foot, or leg and eventually affects both sides of the body. Tremor may also occur in the jaw, chin, mouth, or tongue. The classic tremor of PD is a resting tremor, which is strongest when the affected limb is at rest, and may become less apparent or even disappear during a purposeful movement. An action tremor (a tremor that occurs with intentional movement) may also be a feature of PD.
* **Muscular:**-stiff muscles, difficulty standing, difficulty walking, difficulty with bodily movements, involuntary movements, muscle rigidity, problems with coordination, rhythmic muscle contractions, slow bodily movement, or slow
* **Mood:-**shuffling gait.
* **Sleep:-**People with Parkinson's disease often have sleep problems, including waking up frequently throughout the night, waking up early or falling asleep during the day.
* **Whole body**:-fatigue, dizziness, poor balance, or restlessness.
* **Cognitive**:- amnesia, confusion in the evening hours, dementia, or difficulty thinking and understanding.
* **Speech:**-difficulty speaking, soft speech, or voice box spasms,You may speak softly, quickly, slur or hesitate before talking. Your speech may be more of a monotone rather than with the usual inflections
* **Nasal:**-distorted sense of smell or loss of smell.
* **Urinary:**-dribbling of urine or leaking of urineUrinary frequency (the need to urinate often) and urinary urgency (the feeling that one must urinate right away, even if the bladder is not full) are other possible symptoms of PD.
* **Facial:**-jaw stiffness or reduced facial expression involve many subtle, complex muscle movements. People with Parkinson's disease often have a reduced ability to make facial expressions. This is called facial masking.Facial masking is related to bradykinesia. The facial muscles move more slowly or rigidly than usual.
* **Constipation:**-[Constipation](https://www.medicalnewstoday.com/articles/150322.php) is a common problem that can have a wide range of causes.Constipation is one of the most common non-motor symptoms associated with Parkinson's disease. Nearly [25 percent](https://www.nature.com/articles/s41598-017-16790-8) of people with the condition experience constipation before they develop motor symptoms.
* **Depression:**-Depression is a fairly common non-motor symptom of PD. It can range in severity and may improve with PD treatment, anti-depressant medications, or psychotherapy, such as cognitive behavioral therapy (CBT).
* **Fear of falling from heights:-** Poor balance (postural instability) and freezing are probably the most common causes but there may be many contributing factors. Stooped posture - the stooped posture that often occurs as Parkinson's progresses increases the risk of falling forwards.

**1.3 Hardware specifications**

1) Laptop with minimum RAM of 8 GB(preferred)

2) 1TB Hard disk storage space

3) Intel i5 processor

**1.4 Software specifications**

1) Python Programming

2) Machine Learning Algorithms

3) Jupyter Notebook/ Spyder

**2. LITERATURE SURVEY**

When we looked into existing systems we came to know that the Parkinson’s disease diagnosis were done using the concepts of DBN i.e. Deep Belief Neural network to recognize the voice of patients, so that Parkinson’s disease can be identified. .In another survey, they diagnosed Parkinson’s Disease using coarse audio recordings of test subjects’ voices. The best machine learning model as tested using Microsoft Azure Machine Learning Studio is Two-Class Boosted Decision Trees, an ensemble model of boosted regression trees made in a stepwise method. The next conclusion is that when evaluating Parkinson’s disease, the most important features to address are nonlinear measures of changes in fundamental frequency, as extrapolated from the voice recordings. Another method they proposed is by using the voice parameter such as pitch, formants, jitters, shimmer, glottal pulse and MFCC have been investigated for both normal and PD affected males and females. The observation, on the basis of analysis done, can be summarized as under conditions like PD patient male have higher pitch as compared to normal person. Formants have more variations as compared to normal person. In case of normal person glottal pulse, there emerged a similar pattern whereas it is not in case of PD affected persons. Jitter and Shimmer values are more in PD patients than normal persons.

**2.1 Existing System**

**Ali H. Al-Fatlawi [1]**  In this paper we collect dataset and we will apply DBN. Deep Belief Network (DBN) is a type of deep neural networks that uses multiple processing layers to model high- level abstraction in data with complex structure. These processing layers are connected to each other via connection weights but without any connection within the same layer. Therefore, it is a generative graphical model, consisting of multiple layers of hidden units. Two training stages are required to train this kind of networks: supervised and unsupervised learning.

DBN can be designed as a construction of Restricted Boltzmann Machines (RBM) that are stacked on of each other and trained separately. This stage is unsupervised learning and it called the pre-training stage. It divides the network into groups of stacked sub-networks, each of them includes two processing layers. Basically, this operation is performed to solve the problems that are associated with selecting random values in initializing the connection weights and providing the network with pre-trained weights. It can use the Greedy Layer- Wise unsupervised training algorithm for this function. RBM is a generative stochastic of neural network that can be learned based on a probability model by using unsupervised learning technique.

**R.K. Sharma [2]** In this paper, the analysis of voice a database of 21 persons (especially five vowels a, e, i, o, u) has been recorded using Sony IC Recorder (ICD-UX513F). As each person was asked to pronounce the vowels loudly two or three times. Thus the total voice samples in the study were 105. In the available database 14 were of PD affected persons and rest were of normal healthy persons. The male persons whose voices were recorded have an age between 50 to 74 years and for females it was 45 to 85 years.

Mel-frequency Cepstral Coefficients are very important and popular parameter of voice. MFCC are most acoustic and robust in nature, therefore popularly used in speech recognition systems. They are derived from a type of cepstral representation of voice signal.

According to Gunnar Fant, Formants are the spectral peaks of the sound spectrum of voice. Formants are the meaningful parameter of voice that can distinguish the vowels. This is due to the frequency components of voice.

Pitch depends upon the vocal folds vibrations, more vibrations of vocal folds will result in higher pitch values and vice versa. Pitch also depends on emotions, mood, swallowing on face etc. vocal folds contracts due to swallowing on face, neck (as in case of PD affected patients).

Jitter and Shimmer are measures of the cycle-to-cycle variations of fundamental frequency and amplitude, respectively, which have been largely used for the description of pathological voice quality.

**Jennifer He [3]** In this paper, the data from the voice recordings had to be pre-processed in a format that is easily analyzed by the machine learning algorithm. To do so, a dataset with processed voice recordings was utilized from the University of California - Irvine’s public library of datasets, as discussed.

In the dataset, the label called ‘status’ represented whether or not the patient has Parkinson’s Disease -- a status of 0 meant the person was healthy, and a status of 1 meant the person has been diagnosed with PD by a clinician. The ‘status’ column was used as the label for two-class classification in the tested machine learning models. All of the other columns, aside from ‘name,’ were considered as typical features in the model.

The first step in pre-processing the data was to clean the missing data, however, this was added only as a precautionary measure since the data had already been cleaned when collected. There were no missing values in any of the rows, so this step did not make a noticeable change. After that, we selected the columns that pertained to the problem. In this step, the column titled ‘name’ was removed, as it was not a relevant feature. Then, the data was put through a Filter-Based Feature Selection algorithm to determine the top 10 features to use out of 22 total columns. The specific algorithm used was the Pearson Correlation scoring method. This method is commonly used on numerical data with binary classification and provides weights for each feature to show how well they correlate to the label. The top 10 features were then used as features in the model, excluding all others. The most weighted feature was ‘spread1,’ with a weightage of 0.564838. Spread1 is one of three nonlinear measures of fundamental frequency variation, as according to the initial dataset. These three measures had the highest weights.

**Najmeh Fayyazifar [4]** In this paper, early Parkinson’s disease is detected using machine learning techniques. Firstly, Genetic algorithm is employed to reduce the size of feature vector and select the optimal set of features. Then two ensemble methods including AdaBoost and Bagging are used to classify input data into two healthy and patient categories.

The Genetic algorithm is an efficient optimization technique which can aims to the selection of optimal features in classification problems. In our research, we have applied the Genetic algorithm to reduce the size of the feature vector. The number of extracted feature in UCI Parkinson database is 22 features for each recording. In our proposed method, every chromosome has 22 genes. The genes are the binary representation of primary features. We initiate the algorithm with 20 random- generated chromosomes. The fitness function used in this paper is the classification rate of AdaBoost as a classifier which diagnoses the PD. In each iteration of this algorithm, four of the best chromosomes, chromosomes with the higher classification rate, are chosen for the production and mutation of the next generation. In the next step, we have employed the Bagging algorithm as a classifier in order to evaluate the performance of this classifier in the detection of PD. In both experiments, the algorithm terminates when a fixed number of iteration has carried out which is 15 and 5 for AdaBoost and Bagging, respectively. The value of parameters is obtained by trial and error.

**Timothy J.[5]** In this paper, the raw audio was cleaned with VoiceBox’s Voice Activation Detection (VAD) algorithm, activlev, to extract and remove background noise of the audio. This prepossessing step was required in order to pass only raw voice into the audio feature extraction algorithms. This cleaned audio was then passed through two separate algorithms for feature extraction before being input into the machine learning models.

Methods drawn from Audio-Visual Emotion recognition Challenge (AVEC) from 2013 were used for preliminary audio analysis and the method of Minimum Redun-dancy Maximum Relevance (mRMR) was applied to these AVEC 2013 audio features. mRMR extracts the most relevant features of a given dataset with respect to an output

class, while minimizing the redundancy. The mRMR technique yielded an array of ranked features indexed from highest to lowest predictive correlation on the labeled data. The ranked feature indexes were then used to further pre-process the data before being fed into the machine learning models (e.g. random forest, support vector machine etc.). The raw audio was also passed into the algorithm that extracted The Geneva Minimalistic Acoustic Parameter Set

(GeMaps) using the openSMILE toolkit for feature ex- traction before being sent to the machine learning models.

The GeMaps feature algorithm extracts a number of lower level features such as pitch, jitter, shimmer, loudness, and harmonics-to-noise ratio, in addition to temporal features, such as rate of loudness and number of continuous voiced regions per second. In total, this analysis yielded 62 features per audio sample.

An open-source tool for feature extraction, OpenSmile, was used to extract AVEC and GeMaps features. The AVEC feature set, as well as the GeMaps feature set, included mel ceptrum frequency coefficients (s) (MFCCs) as features. MFCCs are helpful to represent sound as perceived by the human ear, which interprets audio fre- quencies in a non-linear logarithmic fashion. PD is well known to cause a decrease in pitch variation and loudness. MFCCs provide information regarding the frequencies produced by the vocal tract without requiring pitch detection and incorporating the contribution of anatomy including the effects of the vocal chords, tongue, jaw, lips, on voice. The anatomy of the tract and functioning of voice articulators determines the resonant frequencies which are altered in PD. MFCCs offer a means to detect these effects quantitatively. A diverse range of machine learning classifiers were ex- amined to find the highest categorical accuracy for PD diagnosis. The decision tree and support vector machine classifiers were developed with the help of the Scikit-Learn machine learning library as well as the TensorFlow and Keras Deep Learning Libraries. Models were optimized through stratified cross validation with accuracy, F-1, recall and precision as metrics.

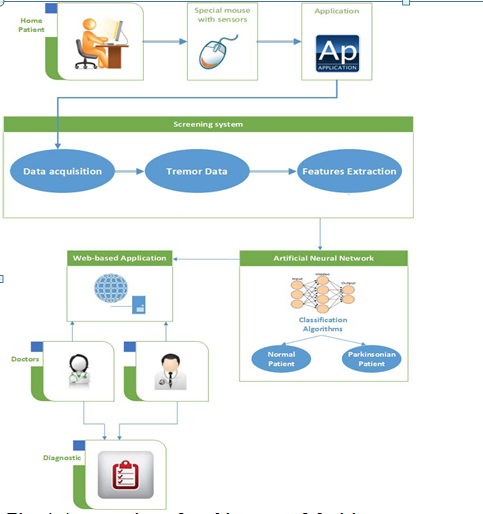
* **Advantages**
* We are able to detect the disease by using the concepts of DBN i.e. Deep Belief Neural network to recognize voice of patients.
* It is able to diagnose the disease using coarse audio recordings of test subject’s voices.
* The best machine learning model as tested using Microsoft Azure Machine Learning Studio is Two- Class Boosted Decision Trees
* We are using various data mining methods and technologies to detect the parkinson’ disease
* **Disadvantages**
* Even though we are able to detect the disease its been not able to provide an specific treatment to specific individual because of more general analysis. It requires individual care and attention in medical treatments to get rid of the disease

**2.2 Proposed System**

In the below figure 2.2.1 we have described how this architecture works by this particular project our software requirements consists of Patients Datasets, python programming, machine learning algorithms and its concepts and finally to implement all this we will be using either spyder or Jupyter Notebook.

Firstly, we will detect the disease by using the dataset collected. The dataset gathered is trained by using various algorithms. Here, we will check the algorithms with the best accuracy.

Once we get the accurate algorithm next we are using that particular algorithm to perform few operations on our dataset to know whether the patient is having parkinson’s disease or not.

****

**Fig 2.2.1: An overview of architecture system of patients suffering from Parkinson’s disease.**

**2.3 Feasibility study**

Parkinson’s disease (PD) results primarily from dying of dopaminergic neurons in the Substantia Nigra, a part of the Mesencephalon (midbrain), which is not curable to date. PD medications treat symptoms only, none halt or retard dopaminergic neuron degeneration

"Big data" is a field that treats ways to analyze, systematically extract information from, or otherwise deal with [data sets](https://en.wikipedia.org/wiki/Data_set) that are too large or complex to be dealt with by traditional [data-processing](https://en.wikipedia.org/wiki/Data_processing) [application software](https://en.wikipedia.org/wiki/Application_software).

Quantitative analysis (QA) is a technique that seeks to understand behavior by using mathematical and statistical modeling, measurement, and research. Quantitative analysts aim to represent a given reality in terms of a numerical value.

Machine Learning (ML) is the [scientific study](https://en.wikipedia.org/wiki/Branches_of_science) of [algorithms](https://en.wikipedia.org/wiki/Algorithm) and [statistical models](https://en.wikipedia.org/wiki/Statistical_model) that [computer systems](https://en.wikipedia.org/wiki/Computer_systems) use to perform a specific task without using explicit instructions, relying on patterns and [inference](https://en.wikipedia.org/wiki/Inference) instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a [mathematical model](https://en.wikipedia.org/wiki/Mathematical_model) based on sample data, known as "[training data](https://en.wikipedia.org/wiki/Training_data)", in order to make predictions or decisions without being explicitly programmed to perform the task.

Here, in this paper various machine learning based methods such as linear and logistic regression, random forest, apriori algorithm, support vector machine, and decision tree based methods are used with minimum redundancy maximum relevance feature selection algorithms to select the most important feature among all the features from the speech articulation difficulty symptoms of Parkinson’s disease affected person to predict the Parkinson disease. Random forest classifier used an ensemble of random trees. Each of the random trees is generated by using a bootstrap sample data. At each node of the tree a subset of feature with highest information gain is selected from a random subset of entire features. Thus random forest used bagging as well as feature selection to generate the trees. The final classification is based on the majority voting of a particular class. Random forest performs better in comparison with single tree classifiers.This project will help doctors to provide the specific treatment for specific person who is suffering from parkinson’s disease.

**3. SYSTEM ANALYSIS AND DESIGN**

**3.1 Requirement Specification**

Data Source Healthcare databases have collected a significant amount of patient's records. The term Parkinson’s disease circulates on various conditions which are harmful to the human heart. Parkinson’s disease is one of the most harmful in nature. The term "Parkinson’s" deals with the situation by which a patient tends to show multiple abnormal behaviors which intendly affecting the patients physical and mental state of patient body leading to an unimaginable dangerous situation. Records were obtained from the medical report analysis collected from kaggle website. Datasets segregate the patterns related to the disease. The records classify into two datasets: training dataset and testing dataset. The dataset consists of 193 rows and 23 columns following table. shows the 23 list of attributes on which the system is working. Analysis of Data This phase has the major jobs of performing data pre-processing such as data cleaning, data integration, filling of missing values, removing redundant data as the dataset contains missing values and redundant data. It leads to fault prediction.

Operating Environment. The language Python provides a stage for performing statistical computation and graphical representation, especially for data analysis. Due to the collection of packages which helps in statistical computation and graphical representation; a user can make the quick analysis of data and graphical representation which leads to the development of prediction system for given application. Python provides an open source software which makes the best compatibility in UNIX and Windows For prediction results, Python offers a better outcome compared to other languages. Although heart disease can happen in various structures, there is a typical arrangement of center hazard factors that impact whether somebody will, at last, be in danger of heart disease or not. These are the basic characteristics that ought to be checked to know whether the heart disease will come or not.

Attributes to be collected:

MDVP:Fo(Hz) - Average vocal fundamental frequency

MDVP:Fhi(Hz) - Maximum vocal fundamental frequency

MDVP:Flo(Hz) - Minimum vocal fundamental frequency

MDVP:Jitter(%),MDVP:Jitter(Abs),MDVP:RAP,MDVP:PPQ,Jitter:DDP - Several measures of variation in fundamental frequency

MDVP:Shimmer,MDVP:Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,MDVP:APQ,Shimmer:DDA - Several measures of variation in amplitude

NHR,HNR - Two measures of ratio of noise to tonal components in the voice

RPDE,D2 - Two nonlinear dynamical complexity measures

DFA - Signal fractal scaling exponent

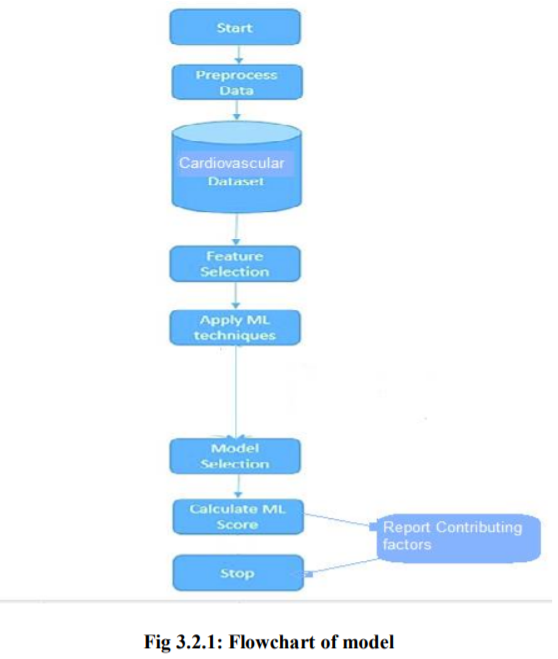
spread1,spread2,PPE - Three nonlinear measures of fundamental frequency variation

status - Health status of the subject (one) - Parkinson's, (zero) – healthy

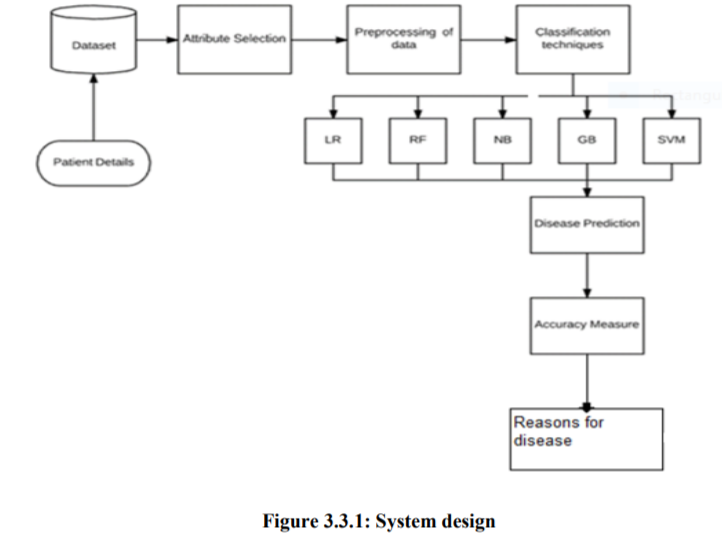
**3.2 Flowchart:**

Any flowchart might be of an iterative process, but this one involves testing the same program with different set of input instances of several patients. The general process of prediction is involved and the algorithm is written accordingly with respect to the cardiovascular disease dataset used in the project.

In the below fig 3.21. we have described the entire process .Generally, the prediction process in machine learning involves data selection, and data pre-processing. It considers the patient’s dataset with and data cleaning is done. Data cleaning is the process of detecting and correcting corrupt or inaccurate records from a record set, it may include fixing null values, noisy data and so on. Further, the data is considered for attribute selection, this involves considering the necessary column attributes for building prediction model in machine learning. Then various machine learning algorithms are applied with the dataset and most accurately predicting model is considered which is trained and tested at best score. At last, prediction is done and calculation of the selected model’s accuracy after which, the contributing factors of cardiovascular disease are reported which involve the major attributes which are responsible for the disease.



**3.3 System Design:**



In the Figure 3.3.1 we have described the process by which we are predicting the disease by explaining each and every individual methods.

**1 - Data Collection**

• The quantity & quality of your data dictate how accurate our model is

• The outcome of this step is generally a representation of data (Guo simplifies to specifying a table) which we will use for training

• Using pre-collected data, by way of datasets from Kaggle, UCI, etc., still fits into this step

**2 - Data Preparation**

• Wrangle data and prepare it for training

• Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.)

• Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data

• Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis

• Split into training and evaluation sets

**3 - Choose a Model**

• Different algorithms are for different tasks; choose the right one

**4 - Train the Model**

• The goal of training is to answer a question or make a prediction correctly as often as possible

• Linear regression example: algorithm would need to learn values for m (or W) and b (x is input, y is output)

•Each iteration of process is a training step

**5 - Evaluate the Model**

• Uses some metric or combination of metrics to "measure" objective performance of model

• Test the model against previously unseen data

• This unseen data is meant to be somewhat representative of model performance in the real world, but still helps tune the model (as opposed to test data, which does not)

• Good train/eval split? 80/20, 70/30, or similar, depending on domain, data availability, dataset particulars, etc.

**6 - Parameter Tuning**   
• This step refers to hyper parameter tuning, which is an "art form" as opposed to a science

• Tune model parameters for improved performance  
• Simple model hyper parameters may include: number of training steps, learning rate, initialization values and distribution, etc.

**7 - Make Predictions**

• Using further (test set) data which have, until this point, been withheld from the model (and for which class labels are known), are used to test the model; a better approximation of how the model will perform in the real world.

**3.4 Algorithms and Code Implementation**

**3.4.1 Logistic Regression:**

In [statistics](https://en.wikipedia.org/wiki/Statistics), the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc... Each object being detected in the image would be assigned a probability between 0 and 1 and the sum adding to one.

Logistic regression is a [statistical model](https://en.wikipedia.org/wiki/Statistical_model) that in its basic form uses a [logistic function](https://en.wikipedia.org/wiki/Logistic_function) to model a [binary](https://en.wikipedia.org/wiki/Binary_variable) [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable), although many more complex [extensions](https://en.wikipedia.org/wiki/Logistic_regression#Extensions) exist. In [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis), **logistic regression** (or **logit regression**) is [estimating](https://en.wikipedia.org/wiki/Estimation_theory) the parameters of a logistic model (a form of [binary regression](https://en.wikipedia.org/wiki/Binary_regression)). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an [indicator variable](https://en.wikipedia.org/wiki/Indicator_variable), where the two values are labeled "0" and "1". In the logistic model, the [log-odds](https://en.wikipedia.org/wiki/Log-odds) (the [logarithm](https://en.wikipedia.org/wiki/Logarithm) of the [odds](https://en.wikipedia.org/wiki/Odds)) for the value labeled "1" is a [linear combination](https://en.wikipedia.org/wiki/Linear_function_(calculus)) of one or more [independent variables](https://en.wikipedia.org/wiki/Independent_variable) ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a [continuous variable](https://en.wikipedia.org/wiki/Continuous_variable) (any real value). The corresponding [probability](https://en.wikipedia.org/wiki/Probability) of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The [unit of measurement](https://en.wikipedia.org/wiki/Unit_of_measurement) for the log-odds scale is called a [*logit*](https://en.wikipedia.org/wiki/Logit), from *logistic unit*, hence the alternative names. Analogous models with a different [sigmoid function](https://en.wikipedia.org/wiki/Sigmoid_function) instead of the logistic function can also be used, such as the [probit model](https://en.wikipedia.org/wiki/Probit_model); the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a *constant* rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the [odds ratio](https://en.wikipedia.org/wiki/Odds_ratio).

The binary logistic regression model has [extensions](https://en.wikipedia.org/wiki/Logistic_regression#Extensions) to more than two levels of the dependent variable: [categorical](https://en.wikipedia.org/wiki/Categorical_variable) outputs with more than two values are modeled by [multinomial logistic regression](https://en.wikipedia.org/wiki/Multinomial_logistic_regression), and if the multiple categories are [ordered](https://en.wikipedia.org/wiki/Level_of_measurement#Ordinal_type), by [ordinal logistic regression](https://en.wikipedia.org/wiki/Ordinal_logistic_regression), for example the proportional odds ordinal logistic model. The model itself simply models probability of output in terms of input, and does not perform [statistical classification](https://en.wikipedia.org/wiki/Statistical_classification) (it is not a classifier), though it can be used to make a classifier, for instance by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class, below the cutoff as the other; this is a common way to make a [binary classifier](https://en.wikipedia.org/wiki/Binary_classifier). The coefficients are generally not computed by a closed-form expression, unlike [linear least squares](https://en.wikipedia.org/wiki/Linear_least_squares_(mathematics));.

**3.4.2 Gaussian NB :**

The Gaussian Naive Bayes is one classifier model. Beside the Gaussian Naive Bayes there are also existing the Multinomial naive Bayes and the Bernoulli naive Bayes. I picked the Gaussian Naive Bayes because it is the simplest and the most popular one.

**3.4.3 Gradient Boosting:**

It is a method of converting weak learners into strong learners. In boosting, each new tree is a fit on a modified version of the original data set. The gradient boosting algorithm (gbm) can be most easily explained by first introducing the AdaBoost Algorithm. The AdaBoost Algorithm begins by training a decision tree in which each observation is assigned an equal weight. After evaluating the first tree, we increase the weights of those observations that are difficult to classify and lower the weights for those that are easy to classify. The second tree is therefore grown on this weighted data. Here, the idea is to improve upon the predictions of the first tree. Our new model is therefore Tree 1 + Tree 2. We then compute the classification error from this new 2-tree ensemble model and grow a third tree to predict the revised residuals. We repeat this process for a specified number of iterations. Subsequent trees help us to classify observations that are not well classified by the previous trees. Predictions of the final ensemble model is therefore the weighted sum of the predictions made by the previous tree models. Gradient Boosting trains many models in a gradual, additive and sequential manner. The major difference between AdaBoost and Gradient Boosting Algorithm is how the two algorithms identify the shortcomings of weak learners (eg. decision trees). While the AdaBoost model identifies the shortcomings by using high weight data points, gradient boosting performs the same by using gradients in the loss function (y=ax+b+e ,e needs a special mention as it is the error term). The loss function is a measure indicating how good are model’s coefficients are at fitting the underlying data. A logical understanding of loss function would depend on what we are trying to optimise. For example, if we are trying to predict the sales prices by using a regression, then the loss function would be based off the error between true and predicted house prices. Similarly, if our goal is to classify credit defaults, then the loss function would be a measure of how good our predictive model is at classifying bad loans. One of the biggest motivations of using gradient boosting is that it allows one to optimise a user specified cost function, instead of a loss function that usually offers less control and does not essentially correspond with real world applications

**3.4.4 Random Forest Classifier**

Random forests or random decision forests are an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and other tasks that operates by constructing a multitude of [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time and outputting the class that is the [mode](https://en.wikipedia.org/wiki/Mode_(statistics)) of the classes (classification) or mean prediction (regression) of the individual trees.Random decision forests correct for decision trees' habit of [overfitting](https://en.wikipedia.org/wiki/Overfitting) to their [training set](https://en.wikipedia.org/wiki/Test_set).

The first algorithm for random decision forests was created by [Tin Kam Ho](https://en.wikipedia.org/wiki/Tin_Kam_Ho) using the [random subspace method](https://en.wikipedia.org/wiki/Random_subspace_method), which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg.

Linear discriminant analysis (LDA), normal discriminant analysis (NDA), or discriminant function analysis is a generalization of Fisher's linear discriminant, a method used in [statistics](https://en.wikipedia.org/wiki/Statistics), [pattern recognition](https://en.wikipedia.org/wiki/Pattern_recognition), and machine learning to find a [linear combination](https://en.wikipedia.org/wiki/Linear_combination) of [features](https://en.wikipedia.org/wiki/Features_(pattern_recognition)) that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a [linear classifier](https://en.wikipedia.org/wiki/Linear_classifier), or, more commonly, for [dimensionality reduction](https://en.wikipedia.org/wiki/Dimensionality_reduction) before later [classification](https://en.wikipedia.org/wiki/Statistical_classification).

LDA is closely related to [analysis of variance](https://en.wikipedia.org/wiki/Analysis_of_variance) (ANOVA) and [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis), which also attempt to express one [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable) as a linear combination of other features or measurements. However, ANOVA uses [categorical](https://en.wikipedia.org/wiki/Categorical_variable) [independent variables](https://en.wikipedia.org/wiki/Independent_variables) and a [continuous](https://en.wikipedia.org/wiki/Continuous_variable) [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable), whereas discriminant analysis has continuous [independent variables](https://en.wikipedia.org/wiki/Independent_variables) and a categorical dependent variable (*i.e.* the class label).[Logistic regression](https://en.wikipedia.org/wiki/Logistic_regression) and [probity regression](https://en.wikipedia.org/wiki/Probit_regression) are more similar to LDA than ANOVA is, as they also explain a categorical variable by the values of continuous independent variables. These other methods are preferable in applications where it is not reasonable to assume that the independent variables are normally distributed, which is a fundamental assumption of the LDA method.

LDA is also closely related to [principal component analysis](https://en.wikipedia.org/wiki/Principal_component_analysis) (PCA) and [factor analysis](https://en.wikipedia.org/wiki/Factor_analysis) in that they both look for linear combinations of variables which best explain the data.LDA explicitly attempts to model the difference between the classes of data. PCA, in contrast, does not take into account any difference in class, and factor analysis builds the feature combinations based on differences rather than similarities. Discriminant analysis is also different from factor analysis in that it is not an interdependence technique: a distinction between independent variables and dependent variables (also called criterion variables) must be made.

LDA works when the measurements made on independent variables for each observation are continuous quantities. When dealing with categorical independent variables, the equivalent technique is discriminant correspondence analysis.

Discriminant analysis is used when groups are known a priori (unlike in [cluster analysis](https://en.wikipedia.org/wiki/Cluster_analysis)). Each case must have a score on one or more quantitative predictor measures, and a score on a group measure. In simple terms, discriminant function analysis is classification - the act of distributing things into groups, classes or categories of the same type.

**3.4.5 Code Implementation**

The overall program code does the prediction of presence of cardiovascular disease for an instance data of a patient using the machine learning algorithms such as Logistic Regression, Random Forest Classifier, Gaussian NB, Gradient Boosting Classifier and Support vector Classifier. In addition to that of the prediction, this program code also generates a report representing all the necessary major contributing factors for the disease such as systolic BP, diastolic BP, glucose, cholesterol, smoking and alcohol. It reports the deviation of SBP and DBP from their standard values if they do not exit between the safety ranges, the glucose and cholesterol have 1-less, 2-considerable, 3-more contribution values which is taken care in the report. It also reports the binary values of smoking and alcohol 0-no contribution and 1-contributing.

PYTHON CODE:

# Importing necessary modules

importnumpy as np

import pandas as pd

importseaborn as sns

fromsklearn.model\_selection import train\_test\_split

fromsklearn.linear\_model import LogisticRegression

fromsklearn.ensemble import RandomForestClassifier

fromsklearn.naive\_bayes import GaussianNB

fromsklearn.discriminant\_analysis import LinearDiscriminantAnalysis

fromsklearn import metrics

fromsklearn import preprocessing

#Loading Dataset

data = pd.read\_csv('parkinsons2.csv')

data.head().transpose()

data.columns

data.shape

data.columns=['Fo','Fhi','Flo','Jitter(%)','Jitter(Abs)','RAP','PPQ','DDP','Shimmer','Shimmer(dB)','APQ3','APQ5','APQ','DDA','NHR','HNR','RPDE','DFA','spread1','spread2','D2','PPE','status']

data.describe().transpose()

data[data.isnull().any(axis=1)]

data['status'].value\_counts().sort\_index()

#splitting and training datasets:

array = data.values

X = array[:,0:22]

Y = array[:,22]

X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size=0.3, random\_state=2)

#implementing algorithms

LR = LogisticRegression()

LR.fit(X\_train, Y\_train)

Y1\_predict = LR.predict(X\_test)

Y\_acc = metrics.accuracy\_score(Y\_test,Y1\_predict)

print("Accuracy of the model is {0:2f}".format(Y\_acc\*100))

fromsklearn.ensemble import RandomForestClassifier

array = data.values

X = array[:,0:22]

Y = array[:,22]

X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size=0.3, random\_state=2)

R = RandomForestClassifier()

R.fit(X\_train, Y\_train)

Y1\_predict = R.predict(X\_test)

Y\_acc = metrics.accuracy\_score(Y\_test,Y1\_predict)

print("Accuracy of the model is {0:2f}".format(Y\_acc\*100))

fromsklearn.naive\_bayes import GaussianNB

array = data.values

X = array[:,0:22]

Y = array[:,22]

#Splitting the data into train and test in 70/30 ratio with random state as 2.

X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size=0.3, random\_state=2)

G = GaussianNB()

G.fit(X\_train, Y\_train)

Y1\_predict = G.predict(X\_test)

Y\_acc = metrics.accuracy\_score(Y\_test,Y1\_predict)

print("Accuracy of the model is {0:2f}".format(Y\_acc\*100))

fromsklearn.discriminant\_analysis import LinearDiscriminantAnalysis

array = data.values

X = array[:,0:22]

Y = array[:,22]

#Splitting the data into train and test in 70/30 ratio with random state as 2.

X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size=0.3, random\_state=2)

LD = LinearDiscriminantAnalysis()

LD.fit(X\_train, Y\_train)

Y1\_predict = LD.predict(X\_test)

Y\_acc = metrics.accuracy\_score(Y\_test,Y1\_predict)

print("Accuracy of the model is {0:2f}".format(Y\_acc\*100))

#predicting based of Accuratereults

pr=LD.predict([[198.764,396.961,74.904,0.0074,0.00004,0.0037,0.0039,0.01109,0.02296,0.241,0.01265,0.01321,0.01588,0.03794,0.07223,19.02,0.451221,0.643956,-6.744577,0.207454,2.138608,0.123306

]])

print(pr)

B=([['MDVP:Fo(Hz)','MDVP:Fhi(Hz)','MDVP:Flo(Hz)','MDVP:Jitter(%)','MDVP:Jitter(Abs)','MDVP:RAP','MDVP:PPQ','Jitter:DDP','MDVP:Shimmer','MDVP:Shimmer(dB)','Shimmer:APQ3','Shimmer:APQ5','MDVP:APQ','Shimmer:DDA','NHR','HNR','RPDE','DFA','spread1','spread2','D2','PPE'

]])

R=int(1)

C=int(22)

matrix=[]

for i in range(R):

d=[]

for j in range(C):

print(B[i][j],end=" ")

d.append(float(input()))

matrix.append(d)

matrix

a=LD.predict(matrix)

z=int(a[1])

if z==1 :

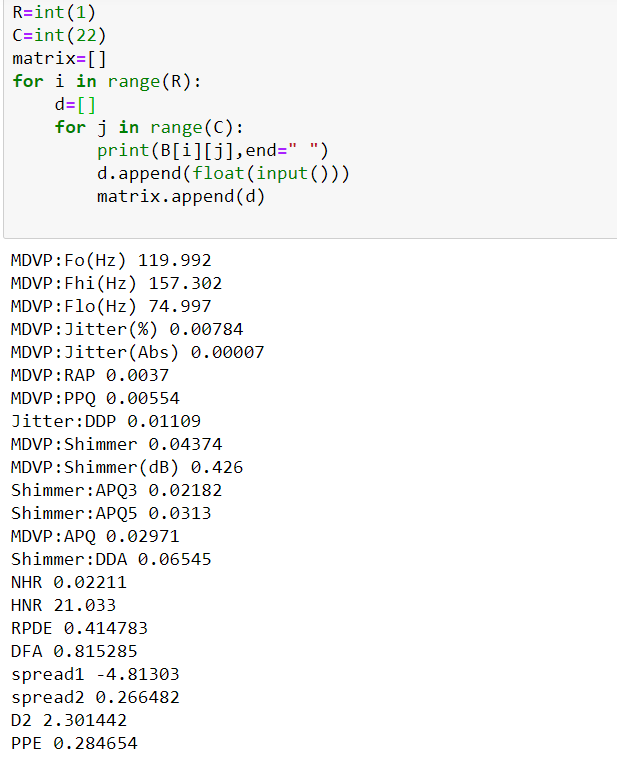
print("person as parkison disease")

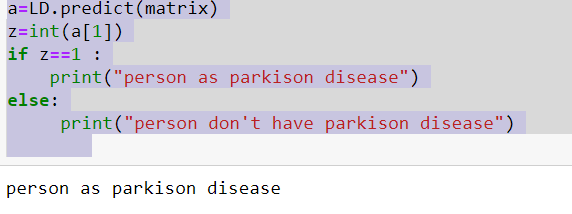
else:

print("person don't have parkison disease")

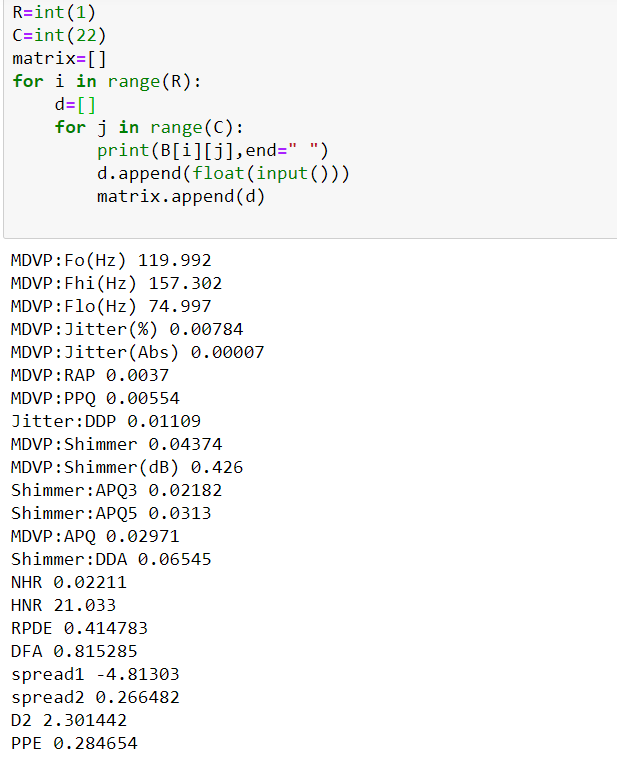
**3.4.6 Testing Process:**

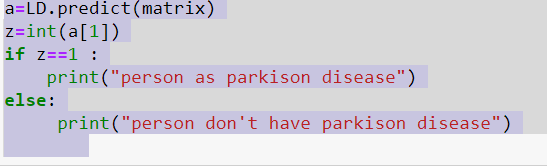
In case of a positive instance, the model should predict the output as 1i.e, the example patient data qualifies to have no Parkinson’s disease.





In case of a negative instance, the model should predict the output as 0i.e, the example patient data qualifies to have no Parkinson’s disease.

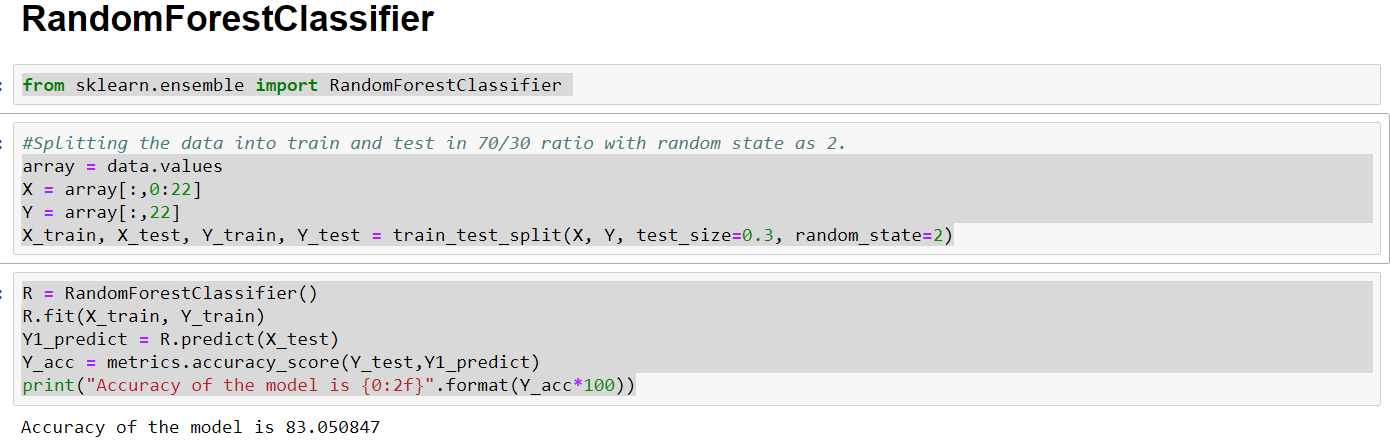


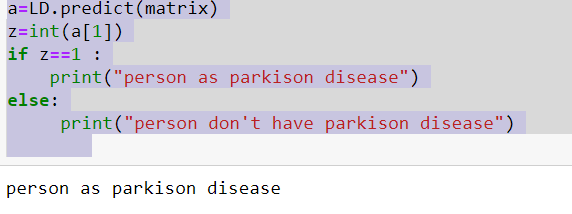


person don’t have Parkinson disease

**4. RESULTS AND ANALYSIS**

We used python programming for predicting our problem. Based on the results conducted we are able to detect whether the patient is having Parkinson or not and we were successful in predicting the above results. We collected patients datasets from medical report analysis and then we used them using some series of machine learning algorithms and then we trained and tested patients data by loading our data to those algorithms and used them to predict the accuracy based on the accuracy results obtained from the machine learning algorithm we implemented that particular algorithm for our code to predict whether the patient is suffering from Parkinson disease or not.





**5. CONCLUSION**

This project contributes the correlative application and analysis of distinct machine learning algorithms in the python software which gives an immediate mechanism for the user to use the machine learning algorithms in python software for forecasting the Parkinson’s diseases. This is non ethical study aims to use available machine learning techniques in python Jupiter notebook software. Future work includes different ensemble methods of these algorithms which can advance to better performance with more parameter settings for these algorithms. Identifying the processing of raw healthcare data of patient information will help in better treatment of patient lives and early detection of abnormalities in patient physical conditions. Machine learning techniques were used in this work to process raw data and provide a new and novel discernment towards Parkinson disease. Parkinson disease prediction is challenging and very important in the medical field. However, the Parkinson suffering rate can be drastically controlled if the disease is detected at the early stages and preventative measures are adopted as soon as possible. Further extension of this study is highly desirable to direct the investigations to real-world datasets instead of just theoretical approaches and simulations .The propose techniques use the series of machine learning algorithm and finding the best accurate results from the algorithm and we will predict the results obtained from the best accurate algorithm.

**FUTURE ENHANCEMENTS**:

The future course of this research can be performed with diverse mixtures of machine learning techniques to better prediction techniques. Furthermore, new feature selection methods can be developed to get a broader perception of the significant features to increase the performance of Parkinson disease prediction. This project’s main objective being achieved, it predicts the Parkinson disease in any instance patient’s data whether the patient has Parkinson disease or not. In addition, this program also reports major contributing factors for the disease listing certain attributes involved in contribution to the Parkinson disease and the deviation of their attribute values with the standard value. This can be mainly used to verify lab test reports or in general to predict the disease as early as possible.

**References**

[1] “Efficient Diagnosis System for Parkinson's Disease Using Deep Belief Network”. Ali H. Al-Fatlawi CNR Lab, Information Technology Research and Development Center, University of Kufa, Najaf, Iraq [Ali.H.Mahdi@uokufa.edu.iq](mailto:Ali.H.Mahdi@uokufa.edu.iq) Mohammed H. Jabardi CNR Lab, Information Technology Research and Development Center, University of Kufa, Najaf, Iraq [Mohammed.Naji@uokufa.edu.iq](mailto:Mohammed.Naji@uokufa.edu.iq) Sai Ho Ling, SMIEEE Centre for Health Technologies, University of Technology Sydney, Sydney, Australia Steve.Ling@uts.edu.au

[2] “Early Detection of Parkinson’s disease through Voice” Vikas School of VLSI Design & Embedded Systems. National Institute of Technology Kurukshetra ,India [vikas.rohilla11@gmail.com](mailto:vikas.rohilla11@gmail.com) R.K. Sharma,School of VLSI Design & Embedded Systems National Institute of Technology Kurukshetra, India [mail2drrks@gmail.com](mailto:mail2drrks@gmail.com)

[3] “Using​ ​Machine​ ​Learning​ ​to​ ​Diagnose Parkinson’s​ ​Disease​ ​from​ ​Voice​ ​Recordings” Akshaya Dinesh North Brunswick Township High School Rutgers School of Engineering Piscataway, New Jersey [akshayadinesh19@gmail.com](mailto:akshayadinesh19@gmail.com) Jennifer He Bergen County Technical High School Rutgers School of Engineering Piscataway, New Jersey [jenhe2000@gmail.com](mailto:jenhe2000@gmail.com)

[4] “Parkinson’s Disease Detection Using Ensemble Techniques and Genetic Algorithm” Najmeh Fayyazifar Information Technology and Computer Engineering Department Shahrood University of Technology Shahrood, Iran [najme.fayazi@shahroodut.ac.ir](mailto:najme.fayazi@shahroodut.ac.ir) Najmeh Samadiani Computer Engineering Department Kosar University of Bojnord Bojnord, Iran [nsamadiani@kub.ac.ir](mailto:nsamadiani@kub.ac.ir)

[5] Parkinson’s Disease Diagnosis Using Machine Learning and Voice” Timothy J. Wroge1, Yasin Ozkanca ̈ 2, Cenk Demiroglu2 , Dong Si3 , David C. Atkins4 and Reza Hosseini Ghomi4 1. Department of Bioengineering, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

[6] <https://ieeexplore.ieee.org/document/8284216/references>

[7] <https://ieeexplore.ieee.org/document/6887541/references>