# PREDICTION OF PARKINSON DISEASE WITH CO-RELATION OF MULTIPLE SYMPTOMS THROUGH MACHINE LEARNING

Dr.T. Venkat Narayana Rao,
Professor, Department of C.S.E,
Sreenidhi Institute of Science and Technology,
Yamnampet, Hyderabad, T.S, India.

Kapa Vinutha , Malka Alekya & Neerudu Sai Divya
Student(s), Department of C.S.E,
Sreenidhi Institute of Science and Technology,
Yamnampet, Hyderabad, T.S, India.

**ABSTRACT**:

Parkinson's disease is a neural disorder that causes shaking, stiffness, and difficulty with walking and balance. Parkinson's symptoms for the most part start steadily and deteriorate after some time. As the sickness advances, individuals may experience issues like strolling and talking. They may likewise have mental and social changes, rest issues, gloom, memory troubles, and exhaustion. Parkinson's infection signs and indications can be distinctive for everybody. Early signs might be mellow and go unnoticed. The symptoms of Parkinson's disease are tremors, small handwriting, trouble moving or walking, constipation, low voice, masked face, dizziness, stooping or hunching over. Some of these symptoms are common in other diseases as well and therefore it is kind of complicate to conclude infection of Parkinson's disease and it is highly impossible drawing results based on few symptoms. Therefore, it can be said that in order to detect Parkinsons's disease it requires accuracy. Furnished conclusions can be drawn correlating all the symptoms and generating a function that can analyze the data and then generate results by detecting the Parkinson's disease. With advanced technology like machine learning we can find a way detecting Parkinson's disease by correlating the symptoms. There are several machine learning algorithms which help in classifying the input depending upon certain factors. Support Vector Machine(SVM) and Logistic Regression are two machine learning algorithms which can be used to classify the data. This paper implements the above two algorithms we train and test data with high accuracy which can help to detect whether the person is infected with Parkinsons disease or not. To predict the results accurately multiple symptoms are analyzed.

**Keywords**: Termor ,Gait, Logistic Regression, Support VectorMachine.

# I.INTRODUCTION

Machine analysis (ML) is the analysis of mathematical models that gain knowledge automatically. It seems like an artificial intelligence ( AI) sub-set. In order to construct a method for analyzing or making choices whilst being fully integrated by a web developer, machine learning models create a statistical equation sample statistic destroyed in software. The research of data analysis automating the predictive analytics is often seen as machine learning. This is founded on the assumption that computers will take negligible manual involvement from given information, analyze and take choices and assumptions.

A series of commands used to address issues are algorithms. In order to recruit machines to do new duties for advanced digital environment that we're seeing nowadays those are equations created by programmers. Considering certain specifications and regulations, algorithms handle enormous volumes of data for info and organizational programs. This is an important concept to know because training opcodes-not software developers-create the regulations in computer science[1].

The main process of machine learning is to get more software data sets. A new line of guidelines is then generated dependent on dataset assumptions by studying algorithms. This method provides the machine with specifications to understand through details without the designer's recent phase-by-step specifications instead of every move in the ways the machine is programmed. This ensures that innovative and complex activities which cannot be physically coded can be used with machines. This creates essentially a new method, officially called the machine learning model. A common classification technique may be used to produce multiple predictions by using various training data samples[3] as shown in figure 1.

Machine learning is used to construct algorithms to admit input information and to test hypothesis-using statistics, dependent on the specific of statistics compiled. These ml techniques are classified as supervised, semi-conducted, uncontrolled and enhanced teaching, with a range of unlimited apps in all of these techniques. The methods for learning can commonly be classified as three general types as follows:

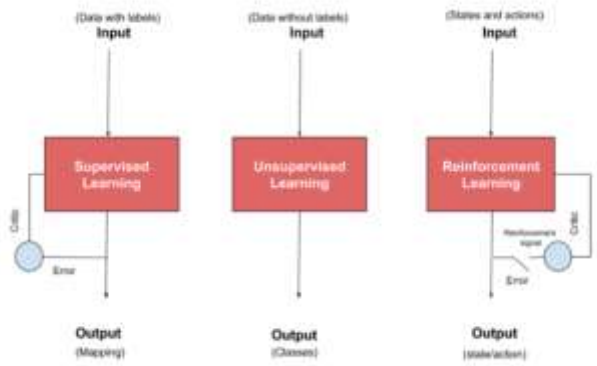

**Fig.1:Categories of Learning Process**

**A. Supervised learning:** The training samples should provide sources coupled with appropriate results in the practice of a supervised learning algorithm. The algorithms searches for data correlations which correspond to the required results while processing. Following the practice of the template, we will pick up a supervised learned algorithms (testing data), that mark would identify new items depending on first trained data. The algorithm will be used for the analysis. A supervised learning model is comprised of objective is to foretell the precise mark of freshly given inputs.

**B. Unsupervised learning:** This form of education is being used against records without past labeling. The design is not called the appropriate reaction. The system should decide what is displayed. The primary goal is to examine the details and to find a pattern. Uncontrolled learning excellently works with transactional datasets. It may, for instance, identify groups of consumers with same characteristics that could then be used in business models in a similar way. Perhaps it will find the basic characteristics that separate company sections from each other.

**C. Reinforcement learning**: The design is used for automation, simulation and mapping most effectively. The algorithm discovers by "truth and errors" that acts are better rewarded by enhanced training. There are multiple important elements of this sort of training: the person (the trainer or the choice-maker), the atmosphere (whatever the person interacts with) and intervention (what person may do). The key objective is for the client to select acts that increase the predicted rewards across period[2].
The grading method is used to grade the whole data collection in 2 phases, 1 and 0. Using the master classified function supporting vector machine and logistic regression models, the classified method is implemented to the sample. These designs are used to boost the degree of precision of the classified process. This model is graded and projected by both. These designs are carried out using coding from Python.

Parkinson’s disease in progressive neurological disorder. Gait issues are the primary symptoms. A material called dopamine in the mind makes soft and synchronized muscle contractions of the body. An area of the mind called the substanitanigra produces dopamine. It is the large nigra cells begin to die in Parkinson's. Dopamine levels are reduced as this occurs. After 60 to 80% decrease, parkinson's signs begin to show.

SYMPTOMS OF PARKINSONS

1. MAJOR SYMPTOMS
    - Termor
    - Slow movements
    - Stiffness of arms,legs and trunk
    - Problems with balance and tendency to fall

2. Secondary Symptoms
    - Muffled, low-volume speech
    - Reduced arm swinging when walking
    - A tendency to get struck when walking

Stages of parkinson's diseases

There are five Stages in parkinsons

1. Stage 1: It is mildest form where the symptoms are not experience , not noticeable and they don't exist in there in daily life and tasks.
2. Stage 2: This is the progression of stage 1 which can be seen in few months or a year , it is called as moderate stage
    - Termor
    - Muscle stiffness
    - Trembling a

    Are the symptoms seen in this moderate stage.

3. Stage3: this is the middle stage where symptoms reach to turning point. Symptoms are noticeable and can be experienced in daily life and tasks. Mainly balance issues become more significant in this stage.But still the patient can do his work by himself.
4. Stage 4: This is pre-advance stage where the patient experience great difficulty in standing without any support . Patients are not allowed to live alone as it leads to dangerous.
5. Stage 5: This is advance stage where all symptoms are clearly noticeable . The effected person suffer a lot from termors, gait, muscle stiffness ets . And other Symptoms the patient may experience are confusion, delusion and hallucinations.

Through this paper, we are trying to co-relate different symptoms like termor, gait, micrographia in order to increase the accuracy in diagnosing Parkinson. The dataset will include features such as jitters and stride. This data will be analyzed using different classification techniques thus providing a reliable and accurate approach to diagnose Parkinson's at an early stage [5][6].

# II.ARCHITECTURE

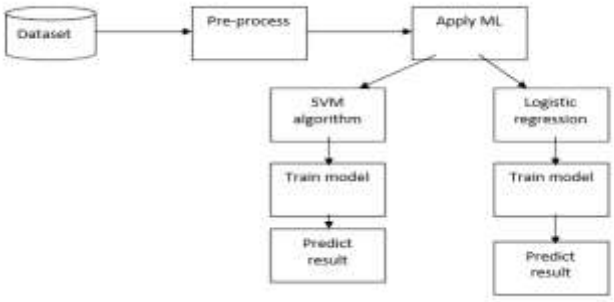


**Fig.2. System Architecture**

The above Figure 2 shows the architecture of how the data is processed using machine learning algorithms. It shows the flow among various elements throughout the process of preprocessing the data retrieved from the database and how the flow of training the data is followed by using Support vector machine algorithm and Logistic regression algorithm

1 DATAFLOW DIAGRAM

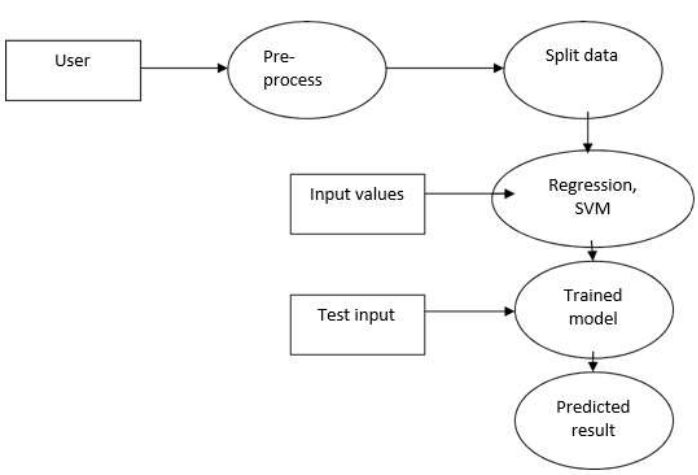


**Fig.3. Flow of Data in Different levels**

Figure 3 represents the flow of various data from different levels. The data is preprocessed and then it is split. Input values are given to the data with regressor or classifier methods applied. To this data test inputs are given to train the model to predict the results.

# III.PROPOSED SYSTEM

Parkinson's disease detection using gait, tremors and handwriting samples as the dataset, in order to increase the accuracy by finding the co-relation between these symptoms.

Since individual analysis of every symptom has some drawback attached to it, for example handwriting is a complex activity where other factors can influence motor movement, in speech recognition additional steps such as noise removal and speech segmentation are required, using breath samples has been proved to fail to meet clinically relevant results. Thus in order to avoid the above problems, we have included multiple symptoms rather than relying on one of them.

# IV.DATA PREPROCESSING AND PREDICTION

We have taken multiple symptoms in our case study, in which we combined the patient's dataset with speech and keystroke dataset. Pre-processing of dataset is done for converting the string attributes to numerals and missing data records are dropped. The pre-processed data is stored in "newdata.csv" file, which is given as input for machine learning models[5].

Step 1 - Import libraries os, csv, pandas, numpy&matplot.pyplot .
Step 2 - Define a variable "newdata" for storing the data after preprocessing.
Step 3 - Open the .csv file in read mode with delimiter as "\t".
Step 4 - Normalize the values to the range [ 0:1 ].(For example : Male=1,Female=0)
Step 5 - Print the "newdata".
Step 6 - Declare a variable "df_main" and assign the data in the dataset to "df_main"
Step 7 - Convert the data in the dataset to "FLOAT" as type.
Step 8 - Normalize the values to the range [ 0:1 ]
Step 9 - Split the data into Dependent and Independent variables
Step 10 - Append the preprocessed data to the variable "newdata"..

## **A.** LOGISTIC REGRESSION ALGORITHM

Logistic Regression is a machine learning which uses gauge the connection between the needy variable(dependent variable) and the at least one autonomous factor(independent variable), here dependent variable is our mark that we want to predict and independent variables are the features which contribute to predict the categorization. All of the mentioned procedure is done by predicting probabilities using the logistic function (also called sigmoid function).

In order to create a precise calculation, Sigmoid converts these odds into boolean digits. The Sigmoid function is an S-shaped curve that accepts any actual value integer. It shows variables from the scale 0 to 1, but at certain cutoff points not precisely. This variables between 0 and 1 are translated using a thresholds classification into either 0 or 1.

$$g(x) = \frac{1}{1 + e^{-x}}$$

- g(x) = output between 0 and 1
- x = input to the function
- e = base of natural log

Step 1 - Import libraries os, csv, pandas, numpy&matplot.pyplot .

Step 2 - Define five lists MSE, MAE, RSQ, RMSE & ACY.

Step 3 - Declare a variable “df_main” and assign the data in the dataset to “df_main”

Step 4 - Convert the data in the dataset to “FLOAT” as type.

Step 5 - Normalize the values to the range [ 0 : 1 ]. (For example: Male=1, Female=0)

Step 6 - Split the data into Dependent and Independent variables.

Step 7 - Declare four variables X_train, X_test, Y_train, Y_test.

Step 8 - Divide the data into Training data and Testing data (by default 75% training data and 25% testing data).

Step 9 - Declare a variable “slf” for Logistic Regression algorithm.

Step 10 - Provide input training datasets ‘X_train’ and ‘Y_train’ to Logistic Regression algorithm using the function “slf.fit()”.

Step 11 - After training, test the model using “slf_predict()” function by providing “X_test” as input.

Step 12 - Declare a variable “result2” for the output of the Logistic Regression algorithm.

Step 13 - Open the .csv file for Logistic Regression algorithm in write mode and update the output values and close the .csv file.

Step 14 - Calculate MSE, MAE, RSQ, RMSE & ACY values for the Logistic Regression algorithm.

Step 15 - Print the MSE, MAE, RSQ, RMSE & ACY values for the Logistic Regression algorithm.

**B**. SVM ALGORITHM

SVM may be used to classify (difference from different classifications) and predict (acquiring the prediction statistical equation). It could be used for either linear or non - linear issues.

Support vector machines are profoundly preferred by many individuals as they produce effective accuracy with less computation power. SVM can be utilized for both regression and classification of the dataset. But, it is effectively used in classification objectives rather than regression. The important characteristic feature of the support vector machine algorithm is to discover a hyperplane, also called a decision boundary in an N-dimensional space (where N represents the number of features that are considered )that particularly classifies the data points into distinct classes[4].

Step 1 - Import libraries os, csv, pandas, numpy & matplot .pyplot .

Step 2 - Define five lists MSE, MAE, RSQ, RMSE & ACY.

Step 3 - Declare a variable “df_main” and assign the data in the dataset to “df_main”

Step 4 - Convert the data in the dataset to “FLOAT” as type.

Step 5 - Normalize the values to the range [ 0:1 ]. (For example : Male=1,Female=0)

Step 6 - Split the data into Dependent and Independent variables.

Step 7 - Declare four variables X_train, X_test, Y_train, Y_test.

Step 8 - Divide the data into Training data and Testing data (by default 75% training data and 25% testing data).

Step 9 - Declare a variable “clf” for SVM algorithm.

Step 10 - Provide input training datasets ‘X_train’ and ‘Y_train’ to svm algorithm using the function “clf.fit ()”.

Step 11 - After training, test the model using “clf_predict ()” function by providing “X_test” as input.

Step 12 - Declare a variable “result2” for the output of the SVM algorithm.

Step 13 - Open the .csv file for svm algorithm in write mode and update the output values and close the .csv file.

Step 14 - Calculate MSE, MAE, RSQ, RMSE & ACY values for the SVM algorithm.

Step 15 - Print the MSE, MAE, RSQ, RMSE & ACY values for the SVM algorithm.

## V. IMPLEMENTATION AND WORKING

In order to built the model, Firstly we need to take a raw dataset and perform Data Preprocessing by using classification techniques to remove noise and outliers. After completing this process the

data is now ready for Prediction where the data is trained and tested by Machine learning algorithms .Here we use two ML algorithms like Support Vector Machine and Logistic Regression algorithms . We find the co-relation between multiple attributes, in order to increase accuracy which is calculated by importing accuracy_core package in python . There are several attributes in our dataset but we consider only few attributes which are independent variables and others are dependent variables, this is done by using confusion matrix in python.

During prediction, the dataset is divided in to training data and testing data by using train_test_split method that allows to divide the dataset in specified ratio.The train dataset is used for training ,building the model with Machine learning algorithms i.e SVM , Logistic regression . The test data is used for testing the created model which is used to predict whether the person has Parkinson disease or not. We split the dataset in different ratios for training and testing in order to determine accuracy .Each time when we execute the model the accuracy is varying but it is always above 90% .

# VI. RESULTS AND DISCUSSION

Table 1: Inputs and the attributes

| Attribute Name | Attribute description |
|---|---|
| Name | ASCII subject name and recording number |
| 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 |
| Status | Health status of the subject (one) - Parkinson's, (zero) – healthy |
| RPDE,D2 | Two nonlinear dynamical complexity measures |
| DFA | Signal fractal scaling exponent |
| spread1,spread2,PPE | Three nonlinear measures of fundamental frequency variation |
| UserKey | 10 character code for that user |
| Date | YYMMDD |
| Timestamp | HH:MM:SS.SSS |
| Hand | L or R key pressed |
| Hold time | Time between press and release for current key milliseconds |
| Direction | Previous to current LL, LR, RL, RR (and S for a space key) |
| Latency time | Time between pressing the previous key and pressing current key. Milliseconds |
| Flight time | Time between release of previous key and press of current key. Milliseconds |

Table 1: It shows the Attributes names and their description which plays an major role in performing the study and to generate the outcomes. The data for this attributes is collected from the people.

Table 2: Output of the system

| Patient ID | Original Value | Logistic regression output | SVM output |
|---|---|---|---|
| P1 | 1 | 1 | 1 |
| P2 | 1 | 1 | 0 |
| P3 | 1 | 1 | 1 |
| P4 | 1 | 1 | 1 |
| P5 | 1 | 1 | 1 |
| P6 | 1 | 1 | 1 |
| P7 | 1 | 1 | 1 |
| P8 | 0 | 0 | 1 |
| P9 | 0 | 0 | 0 |
| P10 | 1 | 1 | 1 |
| P11 | 0 | 0 | 1 |
| P12 | 1 | 1 | 1 |
| P13 | 1 | 1 | 0 |
| P14 | 1 | 1 | 0 |
| P15 | 1 | 1 | 1 |
| P16 | 1 | 1 | 1 |
| P17 | 0 | 0 | 0 |
| P18 | 1 | 1 | 0 |
| P19 | 1 | 1 | 0 |
| P20 | 1 | 1 | 1 |
| P21 | 0 | 0 | 0 |
| P22 | 1 | 1 | 0 |
| P23 | 1 | 1 | 0 |
| P24 | 1 | 1 | 1 |
| P25 | 1 | 1 | 1 |
| P26 | 1 | 1 | 0 |
| P27 | 0 | 0 | 1 |
| P28 | 1 | 1 | 1 |
| P29 | 0 | 0 | 1 |
| P30 | 0 | 0 | 0 |

Table 2: This table shows the predicted values generated by SVM and Logistic regression algorithms in two different columns .The status of person having Parkinson diseases is 0 &1 for No &Yes respectively. The difference between Original values and predicted values shows the accuracy of the model.

Table 3: N folds Training and Testing

| Train data | Test data | SVM accuracy | Logistic regression accuracy |
|---|---|---|---|
| 90% | 10% | 100 | 97.95 |
| 80% | 20% | 94.87 | 95.91 |
| 70% | 30% | 96.60 | 97.95 |
| 60% | 40% | 97.43 | 93.87 |
| 50% | 50% | 97.93 | 97.59 |
| 40% | 60% | 98.29 | 95.91 |
| 30% | 70% | 97.05 | 93.87 |
| 20% | 80% | 98.07 | 97.95 |
| 10% | 90% | 98.28 | 97.95 |

Table 3: The above table describe the accuracy of the model when the percentages of the train and test data is varied for each instance of execution.

## VII.CONCLUSION

The utilization of different occurrence learning for recognizing Parkinson sickness side effects is considered. The finding of the paper is co-relation of multiple symptoms of voice dataset with higher rate of accuracy as depicted in the results section. Moreover the dataset is divided in to different ratios of train and test data to prove that the created model is accurate.

## REFERENCES

[1] S.Andrews,I.Tsochantaridis,andT.Hofmann.Supportvectormachinesformultiple-instance learning. In Advances in Neural Information Processing Systems 15, pages 561–568. MIT Press, 2003.
[2] L. Bao and S. S. Intille. Activity recognition from user-annotated acceleration data. In Intl Conference on Pervasive Computing,2004.
[3] P. Bonato, D. M. Sherrill, D. G. Standaert, S. S. Salles, and M. Akay. Data mining techniques to detect motor fluctuations in Parkinson's disease .InProc.Conf.IEEEEng.Med.Biol.Soc,2004.
[4] C. Cortes and V. Vapnik. Support-vector networks. Machine Learning,1995.

[5] T. G. Dietterich, R. H. Lathrop, and T. Lozano-P´erez. Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell., 89(1-2):31–71, Jan.1997.
[6] N. L. Keijsers, M. W. Horstink, and S. C. Gielen. Automatic assessment of levodopa-induced dyskinesias in daily life by neural networks. Movement Disorder, 18:70–80,2003.