PARKINSON'S DISEASE DETECTION

Submitted in partial fulfillment for the award of the degree of

Requirements Engineering Management (SWE2003) Project

by

TUNUGUNTLA SINDHU GAYATHRI 20MIS7007

P. SAI PIYUSHA 20MIS7009

G. PRAVEEN KUMAR 20MIS7021

Guided by

Dr. Arindam Dey

VIT-AP UNIVERSITY

VELLORE INSTITUTE OF TECHNOLOGY-AMARAVATI

SCOPE

NOVEMBER 2023

ABSTRACT

In our study we have aimed to detect the Parkinson's disease quick and efficiently predicting using machine learning approaches. Our study consists of algorithms like Logistic Regression, Support Vector Machine, Artificial Neural Network. As Parkinson's disease is a deadly disease earlier detection helps to cure the person suffering from disease.

This detection of Parkinson's disease is a binary classification problem; the problem classifies whether the person is suffering or not with disease. We must find the most efficient algorithm among the algorithms considered to find the best approach to earlier detection using comparative analysis between the algorithms considered.

INDEX

TABLE OF CONTENTS	PAGE NO.
INTRODUCTION	04
LITERATURE REVIEW	05
OBJECTIVE	06
METHODOLOGY	07-08
ALGORITHMS	09-10
RESULT Features Accuracy Confusion Matrix Scores Learning curves Comparative Analysis Flow Diagram	11-19
CONCLUSION	19
REFERENCES	20

INTRODUCTION

Parkinson's disease is described as a neurodegenerative condition caused by the loss of dopa stat producing cells. When dopaminergic neurons in the brain are lost, the achievable pace of transmission is reduced. Parkinson's disease affects the central nervous system, which influences the motor system. Tremors, stiffness, and movement problems are the most common symptoms.

Of all those suffering from Parkinson's disease, ninetieths have a speech impairment; only three to four percent of mental patients receive therapy; age is also one of the most important factors for mental patients; the majority of mental patients are between the ages of 45 and 60.

As the speech of mental patients have modification within the frequency spectrum in their voice as a result of the loss of the management of the limb, that decreases the frequency of the audio. So, the low- frequency region offers necessary information to differentiate the speech impairments in mental.

Because mental patients are unable to manipulate their limbs, their speech has altered within the frequency spectrum, resulting in a drop in audio frequency. Therefore, the low frequency area provides the information needed to distinguish between mental-related speech problems.

The Unified Parkinson Illness Rating Scale (UPDRS) facilitates clinical experience to determine the illness's severity. Parkinson's disease cannot be treated when it is discovered too late, and patients eventually die. Its early detection is therefore important. To determine if a certain individual has Parkinson's disease or not, we are using machine learning algorithms on the Parkinson's dataset.

LITERATURE REVIEW

PREVIOUS WORK

Mehrbakhsh Nilashi along with his team of 6 researchers named [1] Rabab Ali Abumalloh, [2] Behrouz Minaei-Bidgoli, [3] Sarminah Samad, [4] Muhammed Yousoof Ismail, [5] Ashwaq Alharagan, [6] Waleed Abdu Zogaan presented a paper for Parkinson's disease detection.

The scientists have done some research with a goal to present a comparison of machine learning approaches for remote tracking of Parkinson's disease progression.

The study completely depends on clustering and prediction learning approaches. The scientists used ensembles of support vector regression and different clustering techniques for Parkinson's disease data clustering.

OUR WORK

The results are then compared with other prediction learning approaches, deep belief network (DBN), support vector regression, linear regression, and neuro fuzzy techniques. We are going to test Parkinson's dataset with an algorithm having more accuracy by constructing a model to detect the person is suffering from the disease or not. These are the algorithms that we will be implementing on the dataset for building a model:

➢ Logistic Regression

➢ Support vector machine

➢ Artificial neural Network

After implementing the three models stated above, we will display the best accurate model and perform comparison analysis.

OBJECTIVE

Collecting data from various patients who were prone to disease and who are having symptoms and who got rid of the disease and comparing them properly. Preparing a project that helps to detect the Parkinson's disease.

Comparing our new algorithms with the existing work done. Predicting whether the person with the symptoms of his disease is affected to it or not with the help of our analysis.

Showing which algorithm has the greater accuracy in predicting the disease appropriately.

So we are making a project for detection of Parkinson disease that is dangerous to human kind and in will be affecting the humans mental health and our neural system functioning.

We will be having chances to stop and so that we need to detect the disease and should take the proper precautions and so that we can avoid the disease and for that we use three algorithms in our project.

And each of them have their own individual purpose of using one will be detecting the final output.

One will be detecting in the form of yes or no and the other one will be performing based on the classifications and gather the requirements and analyze the issues.

And neural networks show how the whole input and inner mechanisms works and output for the inputs we have given for the system.

METHODOLOGY

In this project, we will be identifying whether a person is affected with the Parkinson's disease or not. For this purpose, we have taken a pre-existing project where they have done this method by using deep belief network, support vector regression, linear regression, and neuro fuzzy techniques and they have mentioned all these algorithms.

In our project keeping all those algorithms aside and we were using logistic regression, support vector machine, and artificial neural network and at last we will be training the algorithms and see that which among these have high accuracy of predicting the disease correctly and will suggest that algorithm to use in context to detect the disease accurately. So, all this depends upon the algorithms we use, and we compare the data of different gatherings and form a confusion matrix and test the data with it.

Logistic regression sees the probability of a particular event that occurs in the given scenario. And with the help of this algorithm, we can be able to predict whether we are prone to the disease or not prone to the disease and so that it will be helpful for us to know what we need to take care of after the detection of the disease in our body. This algorithm scans and sees about all the characteristic events and parameters of a particular event and sees how to analyze the particular situation and then we can be able to know how to further analyze the regression model with the help of the parameters that we have with us and it will be giving output in the form of yes or no like simply we can just understand whether the thing we predicted is yes or no.

The **support vector machine** is one of the most important and capable algorithms that can solve the various machine learning problems and its main functionality is to solve the several types of regression tasks by using classification process. The process in this is like we will be having a dataset and in this we do have the data related to patient health details like their voice analysis and speech testing and their blood report samples and all those content and Xray reports will be there with us and with the help of all those , we will start the classification testing and in that process we will see that the patients who are affected with the disease and take their health affected symptoms and compare them with the other patients and we will be then identifying how much related symptoms ate matched with the affected ones and then we will come to a conclusion that the person symptoms are related to the disease or

not and it will give the percentage of the person has the disease based on the above classifications.

Artificial neural networks are the almost useful algorithm in many of the situations and it functions based on the brain functions and it is useful for some advanced and level and complex level of patterns or issues and it is somehow related to how brain neurons work and relevant to some biological brain data like it thinks in various perspectives and it just simply do not give the result based on the Simple data sets of any of the project as we already said it can be described as an inspiration model from the human brain structure like we give some inputs like we in our project provide the data values of the patients and their reports and based on the process that is done in the internal activity we will be able to solve the problems and provide the results like the probability of the affected persons and the chances of getting the disease to them and with the help of this we can be exactly calculating the accuracy of the disease that is affected or been there to a person and so that it will be having the more accuracy compared to all the other mentioned algorithms.

ALOGORITHMS

PSEDUDO CODE FOR LOGISTIC REGRESSION

```
Input: Training data (X, y)
Output: Model parameters (theta)

1. Initialize model parameters (theta) to zeros
2. Repeat until convergence:
    2.1 For each training example (x, y):
        2.1.1 Calculate the predicted probability:
            z = theta^T * x
            prediction = sigmoid(z)

        2.1.2 Calculate the loss:
            loss = -y * log(prediction) - (1 - y) * log(1 - prediction)

        2.1.3 Calculate the gradient of the loss:
            grad_theta = x * (prediction - y)

        2.1.4 Update the model parameters using gradient descent:
            theta = theta - alpha * grad_theta

    2.2 Check for convergence:
        2.2.1 Calculate the change in loss:
            delta_loss = abs(previous_loss - loss)

        2.2.2 Check if the change in loss is less than a tolerance threshold:
            if delta_loss < tolerance:
                break

3. Return theta
```

PSEDUDO CODE FOR SUPPORT VECTOR MACHINE

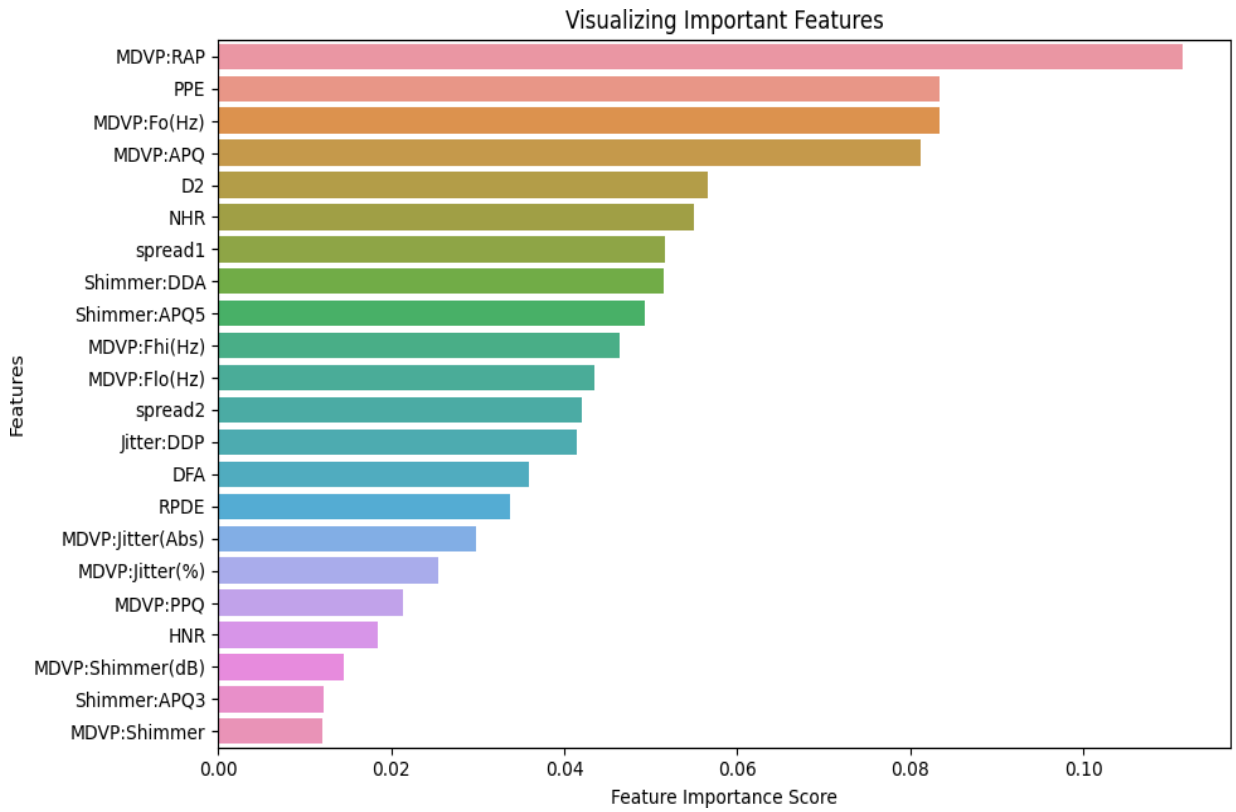
```
Input: Training data (X, y)
Output: Support vectors (sv), hyperplane parameters (w, b)

1. Initialize the weights (w) and bias (b) to random values
2. Repeat until convergence:
    2.1 For each training example (x, y):
        2.1.1 Calculate the hinge loss for the current example:
                    loss = max(0, 1 - y * (w^T * x + b))
        2.1.2 Calculate the gradient of the hinge loss:
                    grad_w = y * x
                    grad_b = y
        2.1.3 Update the weights and bias using gradient descent:
                    w = w - alpha * grad_w
                    b = b - alpha * grad_b
    2.2 Calculate the margin:
                    margin = 1 / ||w||
3. Identify the support vectors:
    3.1 sv = []
    3.2 For each training example (x, y):
        3.2.1 if y * (w^T * x + b) < 1:
            3.2.2 append x to sv
4. Return sv, w, b
```

PSEDUDO CODE FOR ARTIFICIAL NEURAL NETWORKS

```
Initialize network weights and biases
Repeat until convergence:
  For each training example:
    Calculate the network outputs
    Calculate the error between the actual and desired outputs
    Update the network weights and biases using backpropagation
End
```

RESULT

VISUALIZING IMPORTANT FEATURES



LOGISTIC REGRESSION

- ACCURACY

```
Training Accuracy =  76.282
Testing Accuracy =  71.795
0.717948717948718
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        11
           1       0.72      1.00      0.84        28

    accuracy                           0.72        39
   macro avg       0.36      0.50      0.42        39
weighted avg       0.52      0.72      0.60        39
```

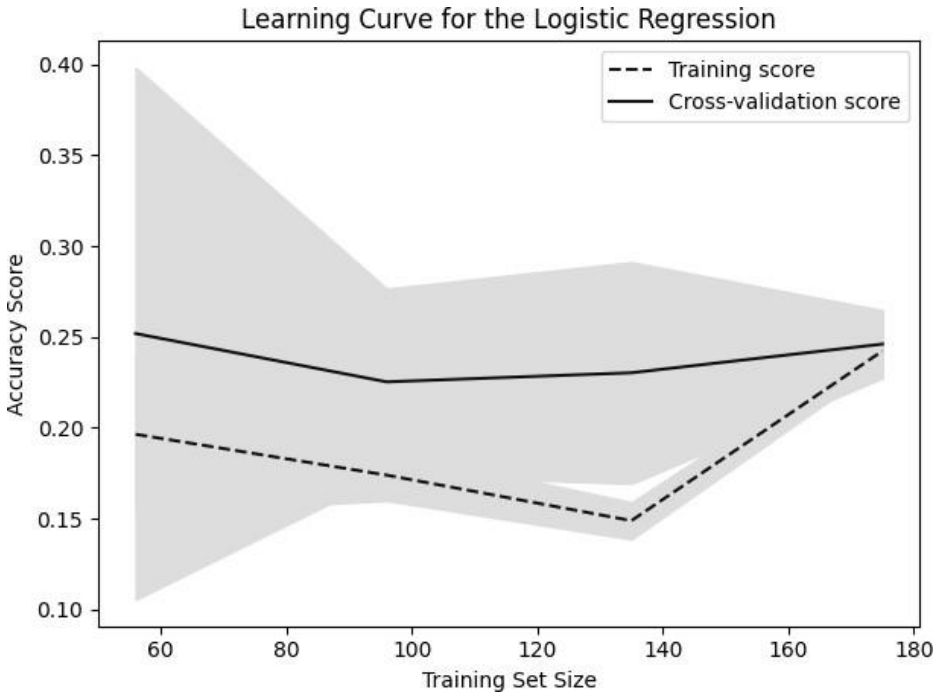
- CONFUSION MATRIX

```
[[ 0 11]
 [ 0 28]]
```

- SCORES

```
Mean training scores

 17          NaN
56      0.196429
96      0.173958
135     0.148889
175     0.242286
dtype: float64

 --------------------

Mean validation scores

 17          NaN
56      0.251842
96      0.225263
135     0.230263
175     0.246053
dtype: float64
```

- LEARNING CURVE



SUPPORT VECTOR MACHINE

- ACCURACY

```
Accuracy: 0.8205128205128205

Precision: 0.8262108262108262

Recall: 0.8205128205128205

              precision    recall  f1-score   support

           0       0.67      0.73      0.70        11
           1       0.89      0.86      0.87        28

    accuracy                           0.82        39
   macro avg       0.78      0.79      0.78        39
weighted avg       0.83      0.82      0.82        39
```

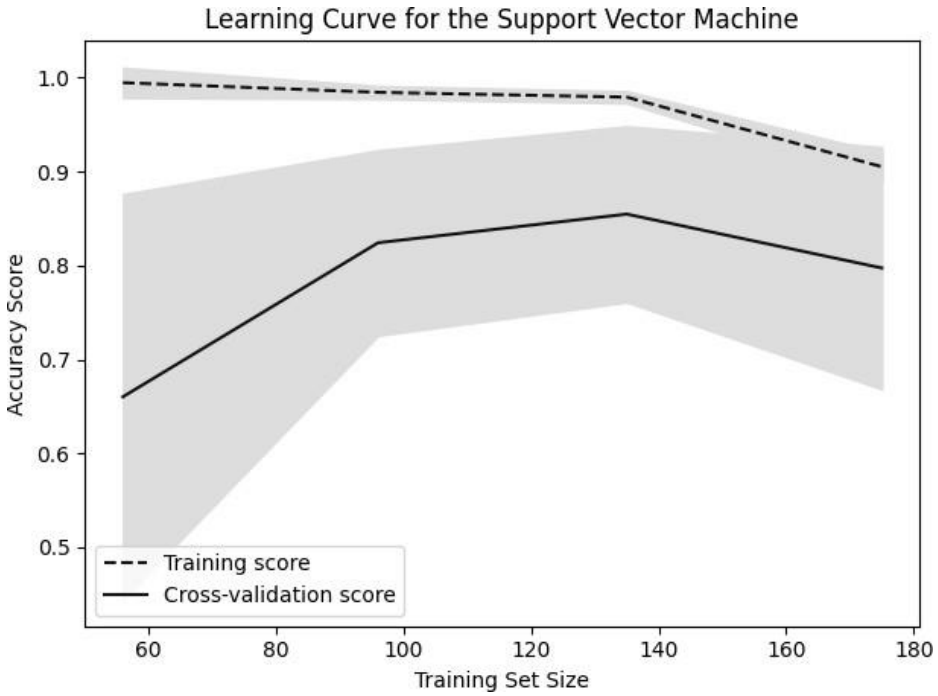
- CONFUSION MATRIX

```
[[ 8  3]
 [ 4 24]]
```

- SCORES

```
Mean training scores

 17          NaN
56     0.994643
96     0.984375
135    0.979259
175    0.905143
dtype: float64

 --------------------

Mean validation scores

 17          NaN
56     0.660263
96     0.824211
135    0.854737
175    0.797368
dtype: float64
```

- LEARNING CURVE



ARTIFICIAL NEURAL NETWORKS

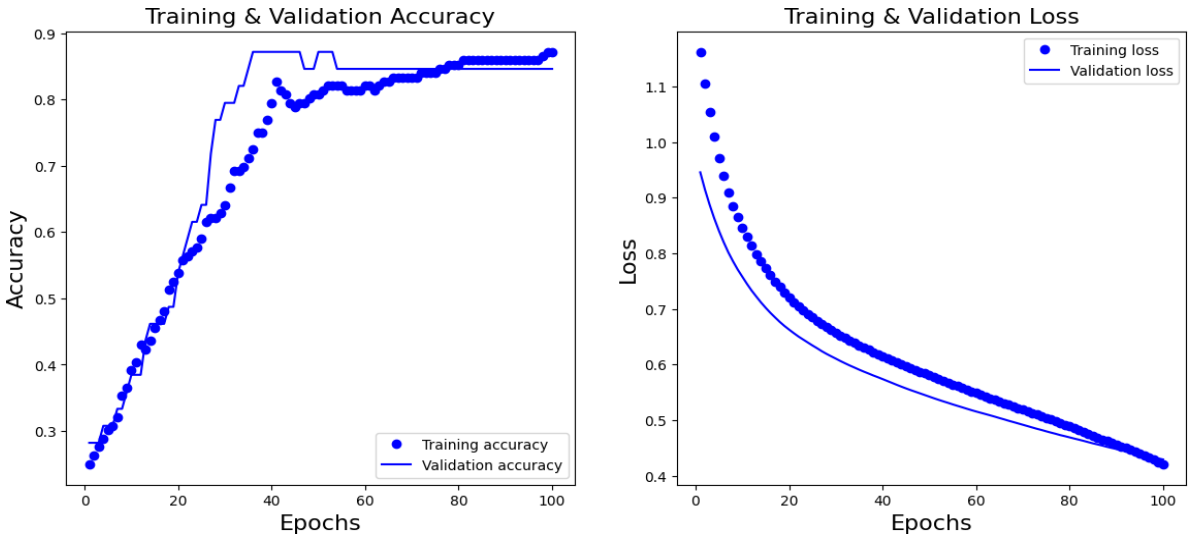
- ACCURACY

```
2/2 [==============================] - 0s 7ms/step
0.8461538461538461
              precision    recall  f1-score   support

           0       0.75      0.60      0.67        10
           1       0.87      0.93      0.90        29

    accuracy                           0.85        39
   macro avg       0.81      0.77      0.78        39
weighted avg       0.84      0.85      0.84        39
```

- CONFUSION MATRIX

```
[[ 6  4]
 [ 2 27]]
```

- LEARNING CURVES



→ So, we have trained and tested the 3 individual algorithms and got the results.

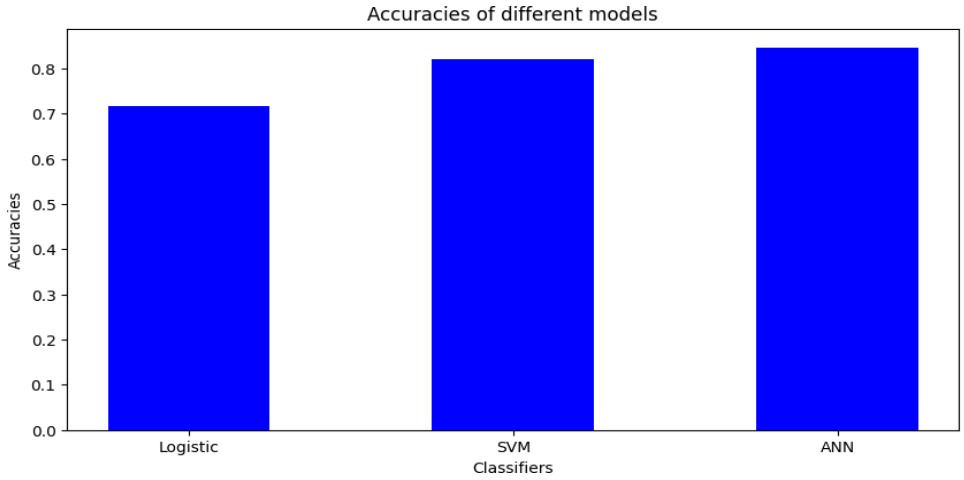
→ We obtained the accuracy levels and the import factor of the comparison, that is, confusion matrix, and the training and validation scores.

→ And then finally we plotted and got the learning curves suiting with the data given.

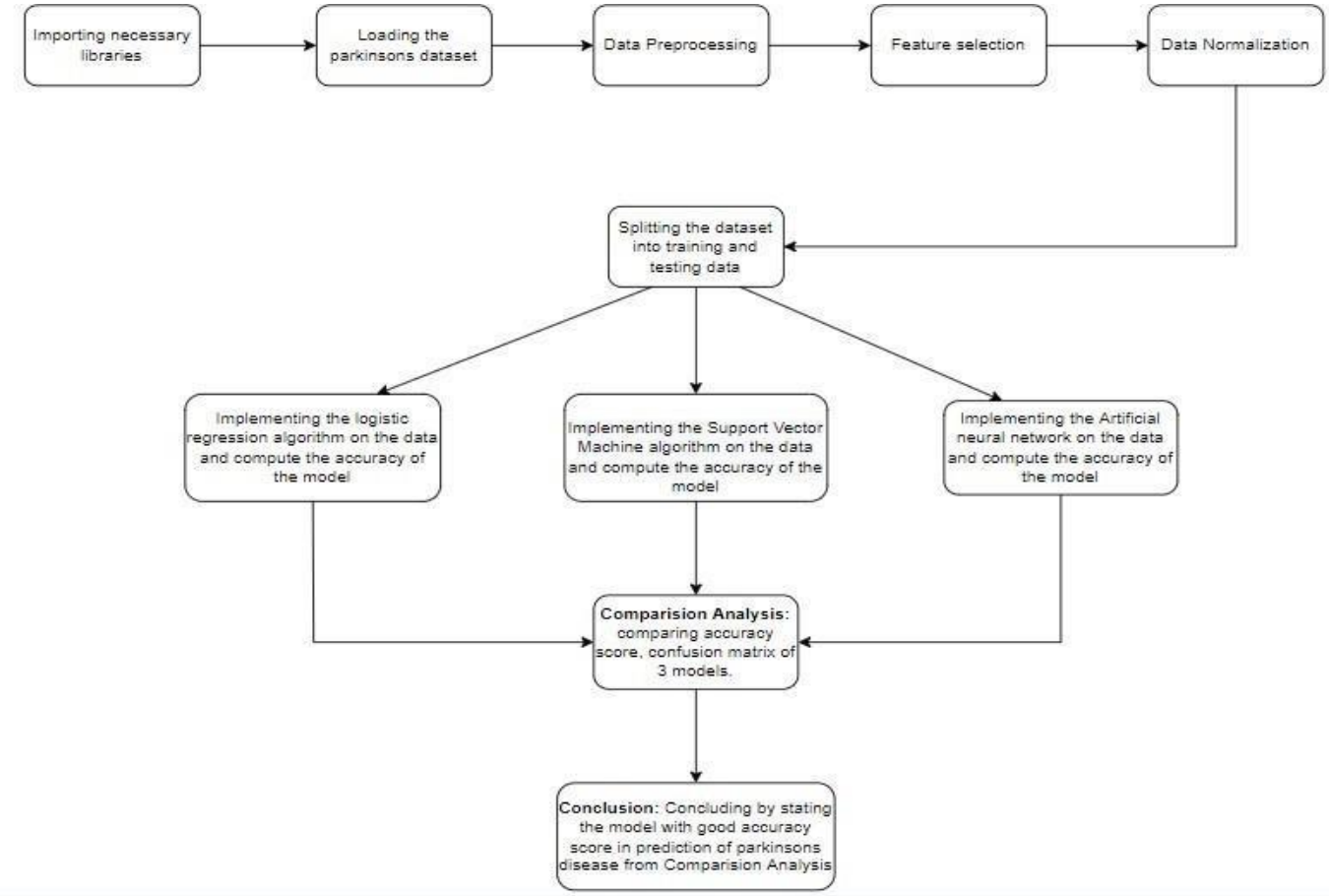
→ So, we will be comparing the three algorithms and see which is the better one and which gives the better accuracy in detecting the disease.

→ We have also drawn a flow diagram suiting our project.

COMPARITIVE ANALYSIS BETWEEN THREE MODELS



FLOW DIAGRRAM



CONCLUSION

By observing the graphs, the accuracy ranges of **logistic regression** are **76.2%** and **support vector machine** is **82.05%** and **artificial neural network** is **92.3%**. We conclude that the **Artificial Neural Network (ANN**) is the most effective algorithm for detection of the Parkinson's disease by observing the graph which is depicting the comparison between three algorithms considered in our study.

Therefore, **Artificial Neural Network** machine learning approach is a good model for detecting the Parkinson's disease quickly and efficiently.

REFERENCES

[1] IBM. Neural networks. IBM Topics.
https://www.ibm.com/topics/neural-networks.
[2] GeeksforGeeks. Understanding logistic regression.
https://www.geeksforgeeks.org/understanding-logistic-regression/.
[3] GeeksforGeeks. Support vector machine algorithm.
https://www.geeksforgeeks.org/support-vector-machine-algorithm/.
[4] ScienceDirect. Confusion matrix.
https://www.sciencedirect.com/topics/engineering/confusion-matrix.
[5] Journal of existing work.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8831050/
[6] M. Nilashi, H. Ahmadi, L. Shahmoradi, A. Mardani, O. Ibrahim, and E. Yadegaridehkordi, "Knowledge discovery and diseases prediction: a comparative study of machine learning techniques," Journal of Soft Computing and Decision Support Systems, vol. 4, pp. 8–16, 2017
[7] B. Sabeena, S. Sivakumari, and P. Amudha, "A technical survey on various machine learning approaches for Parkinson's disease classification," Materials Today Proceedings, 2020.
[8] S. Bhattacharya, P. K. R. Maddikunta, S. Hakak et al., "Antlion re-sampling based deep neural network model for classification of imbalanced multimodal stroke dataset," Multimedia Tools and Applications, 2020.
[9] B. M. Bot, C. Suver, E. C. Neto et al., "The mPower study, Parkinson disease mobile data collected using ResearchKit," Scientific Data, vol. 3, pp. 1–9, 2016.