Comparison of Machine Learning Methods for Breast Cancer Diagnosis

In this paper author is using SVM (Support Vector Machine) and ANN (Artificial Neural Network) to predict breast cancer diseases. First this algorithms will be trained using past disease dataset called ‘Wisconsin Breast Cancer’, this dataset contains 11 integer values and last value contains class label either 0 or 1, 0 means person is normal and 1 means person is infected with disease. Both algorithms will be trained on previous people’s dataset and new person test data will be applied on trained data to predict it class such as 0 or 1.

Both algorithms generate model from train dataset and new data will be applied on train model to predict it class. SVM algorithm is giving better prediction accuracy compare to ANN algorithm.

Machine learning involves predicting and classifying data and to do so we employ various machine learning algorithms according to the dataset. SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes. In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification. As a simple example, for a classification task with only two features (like the image above), you can think of a hyperplane as a line that linearly separates and classifies a set of data.

Intuitively, the further from the hyperplane our data points lie, the more confident we are that they have been correctly classified. We therefore want our data points to be as far away from the hyperplane as possible, while still being on the correct side of it.

So when new testing data is added, whatever side of the hyperplane it lands will decide the class that we assign to it.

How do we find the right hyperplane?

Or, in other words, how do we best segregate the two classes within the data?

The distance between the hyperplane and the nearest data point from either set is known as the margin. The goal is to choose a hyperplane with the greatest possible margin between the hyperplane and any point within the training set, giving a greater chance of new data being classified correctly.

An artificial neuron network (ANN) is a computational model based on the structure and functions of biological neural networks. Information that flows through the network affects the structure of the ANN because a neural network changes - or learns, in a sense - based on that input and output.

ANNs are considered nonlinear statistical data modelling tools where the complex relationships between inputs and outputs are modelled or patterns are found.

ANN is also known as a neural network.

An ANN has several advantages but one of the most recognized of these is the fact that it can actually learn from observing data sets. In this way, ANN is used as a random function approximation tool. These types of tools help estimate the most cost-effective and ideal methods for arriving at solutions while defining computing functions or distributions. ANN takes data samples rather than entire data sets to arrive at solutions, which saves both time and money. ANNs are considered fairly simple mathematical models to enhance existing data analysis technologies.

ANNs have three layers that are interconnected. The first layer consists of input neurons. Those neurons send data on to the second layer, which in turn sends the output neurons to the third layer.

Training an artificial neural network involves choosing from allowed models for which there are several associated algorithms.

To implement above two algorithms we have used python technology and ‘Wisconsin Breast Cancer’ dataset. This dataset available inside dataset folder which contains test dataset with dataset information file. Below are some dataset examples

**Sample\_code\_number,Clump\_Thickness,Uniformity\_of\_Cell\_Size,Uniformity\_of\_Cell\_Shape,Marginal\_Adhesion,Single Epithelial\_Cell\_Size,Bare\_Nuclei,Bland\_Chromatin,Normal\_Nucleoli,Mitoses,Class**

1000025,5,1,1,1,2,1,3,1,1,0

1002945,5,4,4,5,7,10,3,2,1,0

1015425,3,1,1,1,2,2,3,1,1,0

1016277,6,8,8,1,3,4,3,7,1,0

1017023,4,1,1,3,2,1,3,1,1,0

1017122,8,10,10,8,7,10,9,7,1,1

1099510,10,4,3,1,3,3,6,5,2,1

1100524,6,10,10,2,8,10,7,3,3,1

All above are the names of columns in bold format and then in below lines we can see dataset values. In last column we can find either 0 or 1 which means 0 indicate normal value and 1 indicate infected with disease. We will train both algorithms using above dataset values. Below are some test data values which will not contains either 0 or 1 and application will predict it from trained model data.

1000025,5,1,1,1,2,1,3,1,1

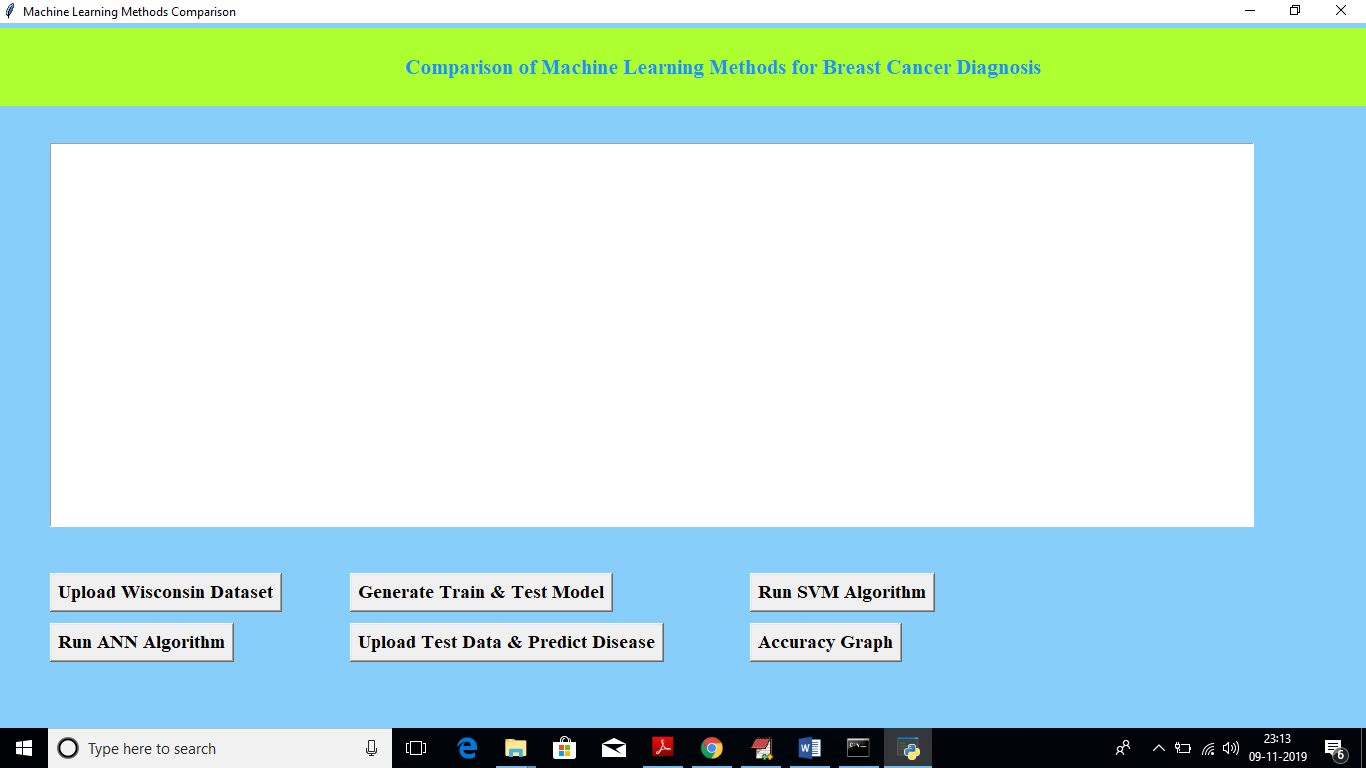
1002945,5,4,4,5,7,10,3,2,1

1017122,8,10,10,8,7,10,9,7,1

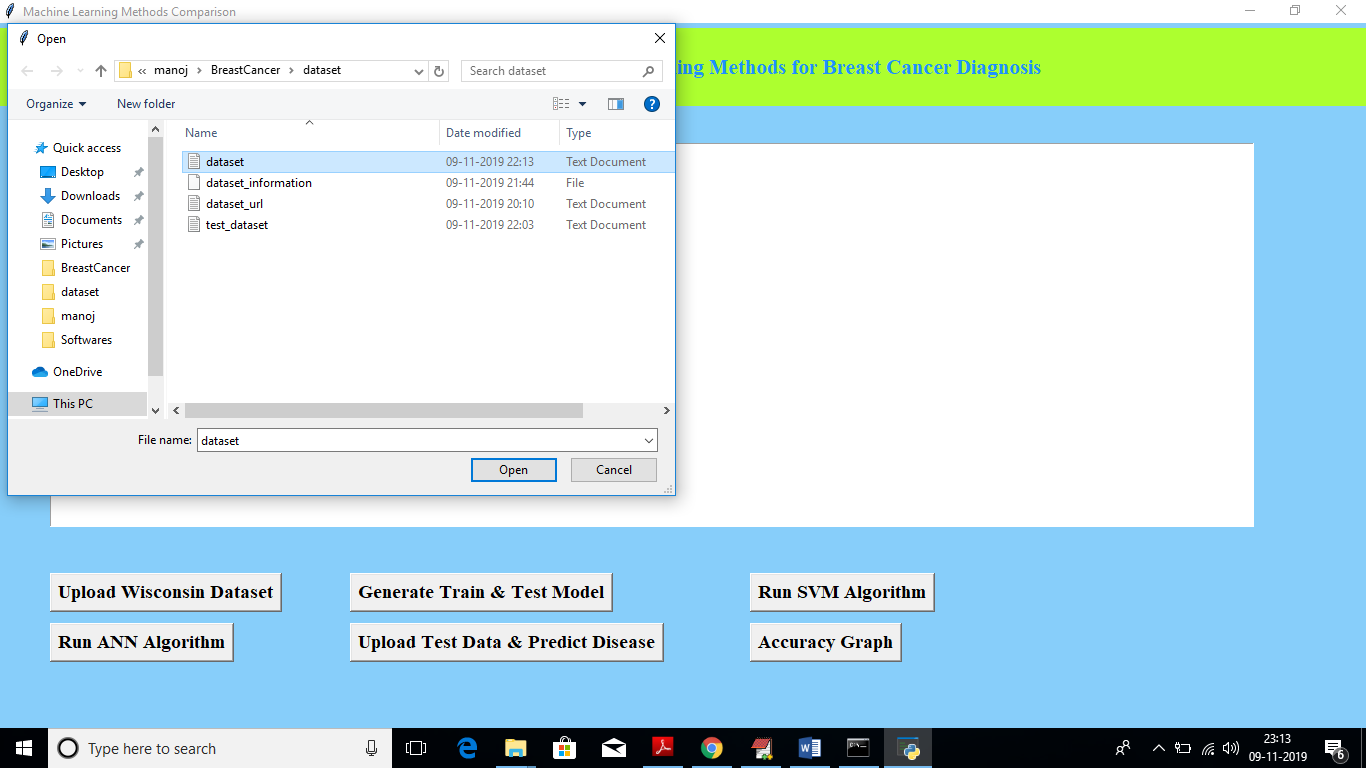
In above test data contains only 10 values as last value 0 or 1 missing which will predict by application. In train data you can see 11 column values are there but test data contains 10 values.

Screen shots

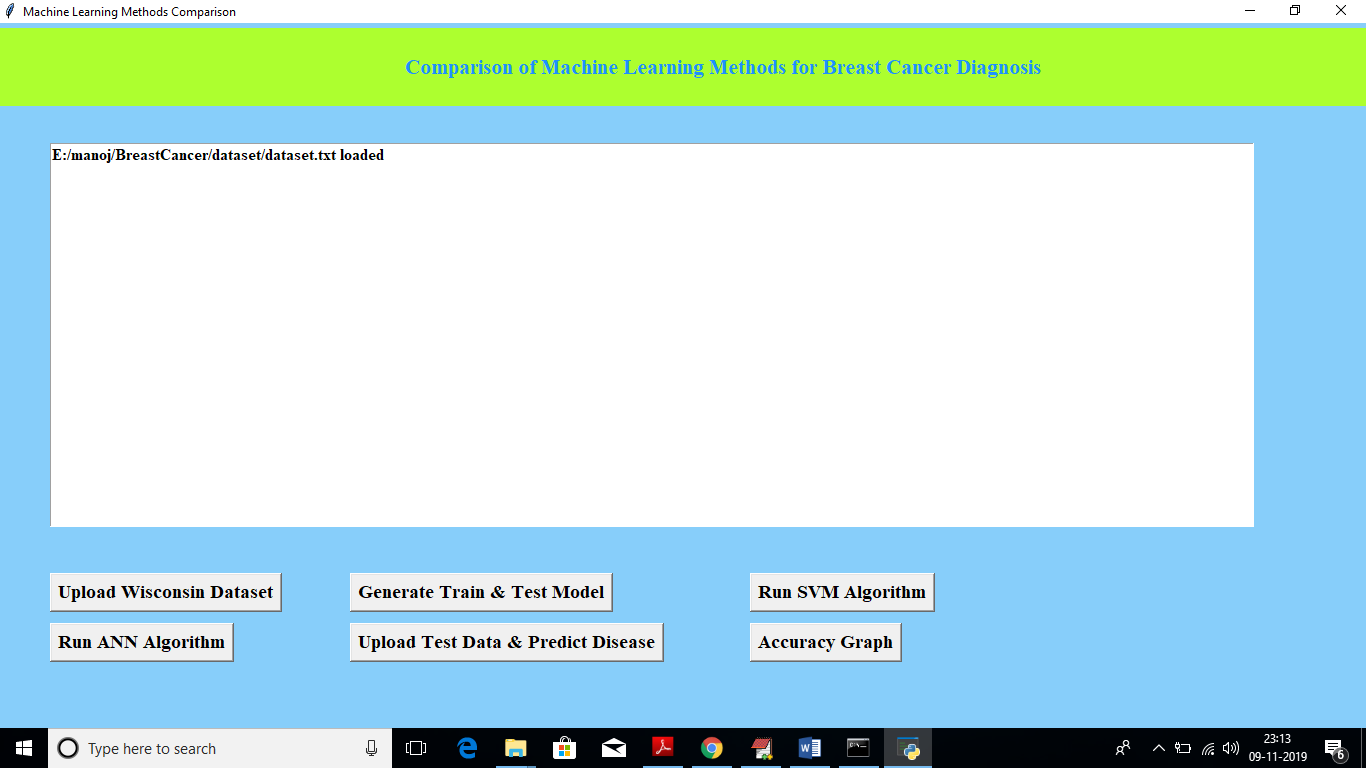
Double click on ‘run.bat’ file to get below screen



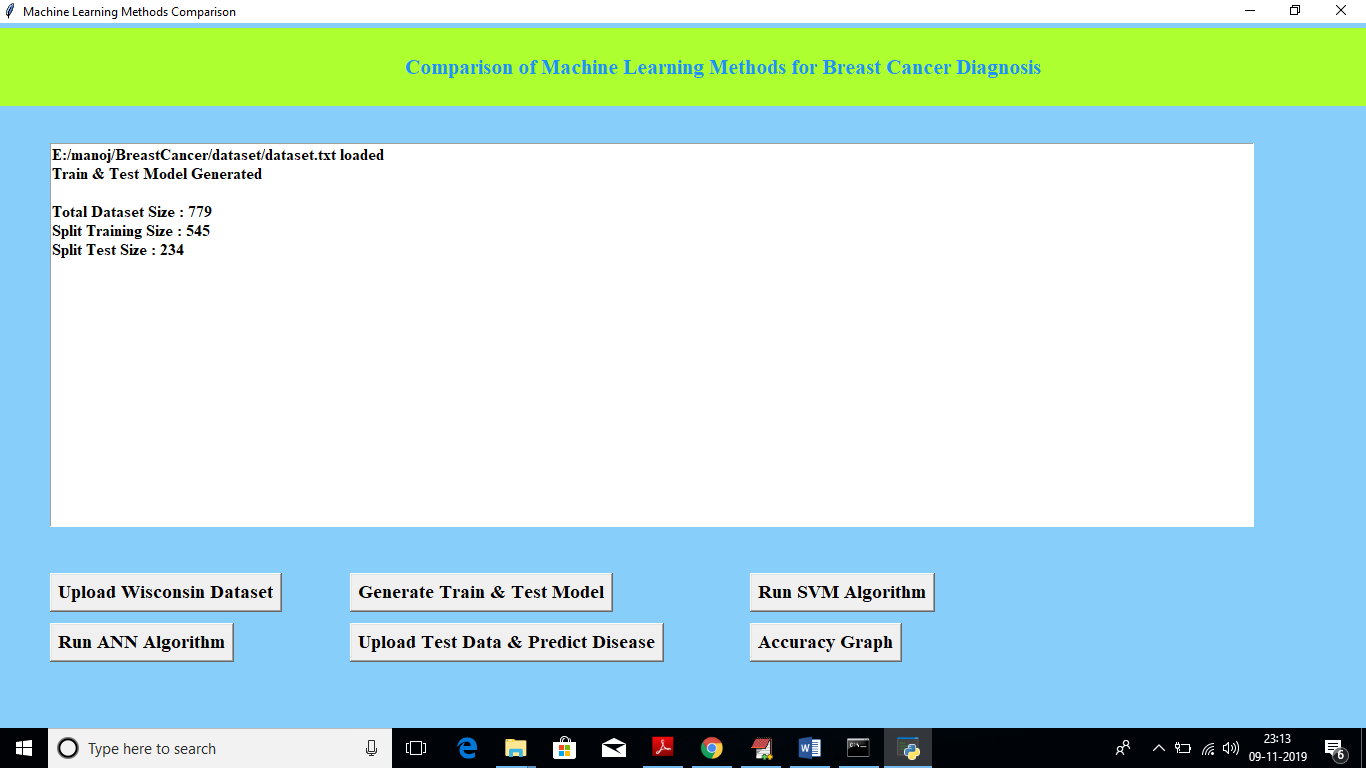
Now click on ‘Upload Wisconsin Dataset’ button to upload dataset



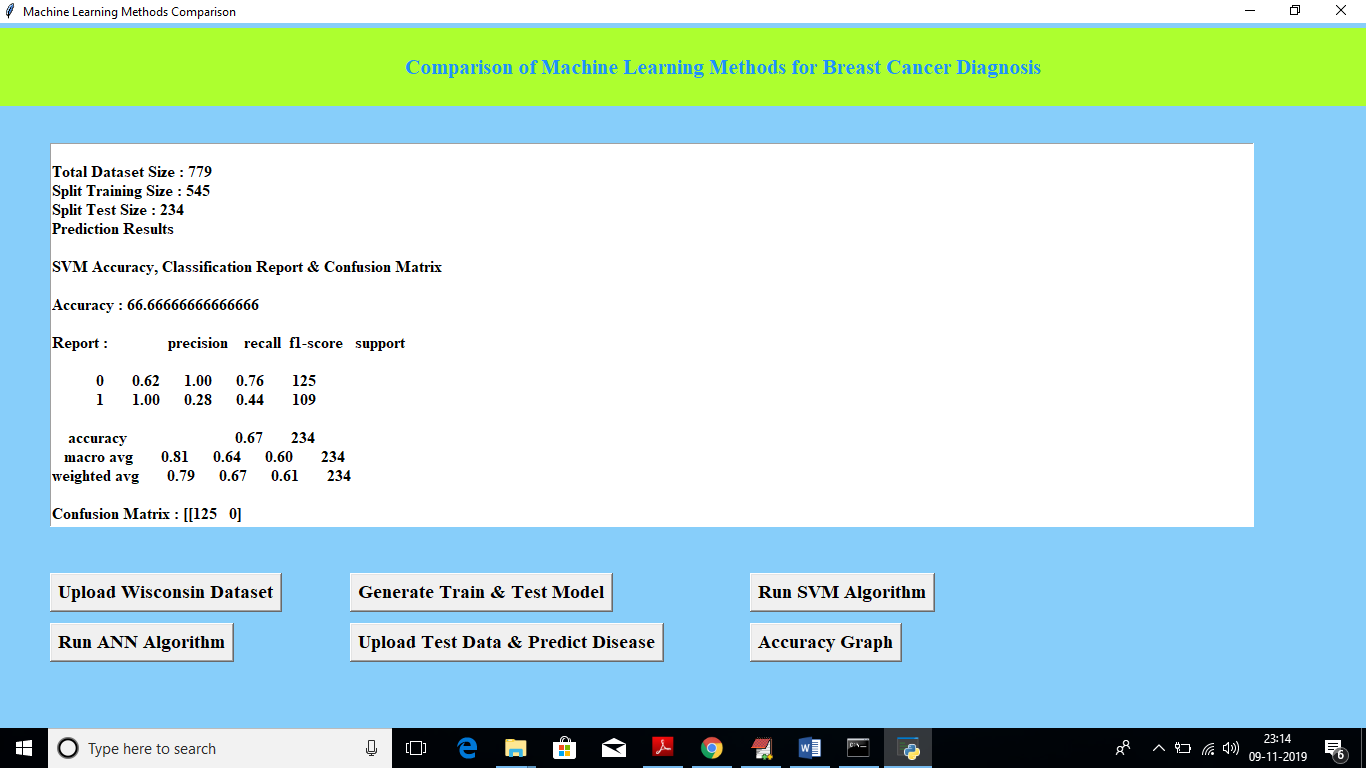
In above screen I am uploading ‘dataset.txt’ file and after upload will get below screen



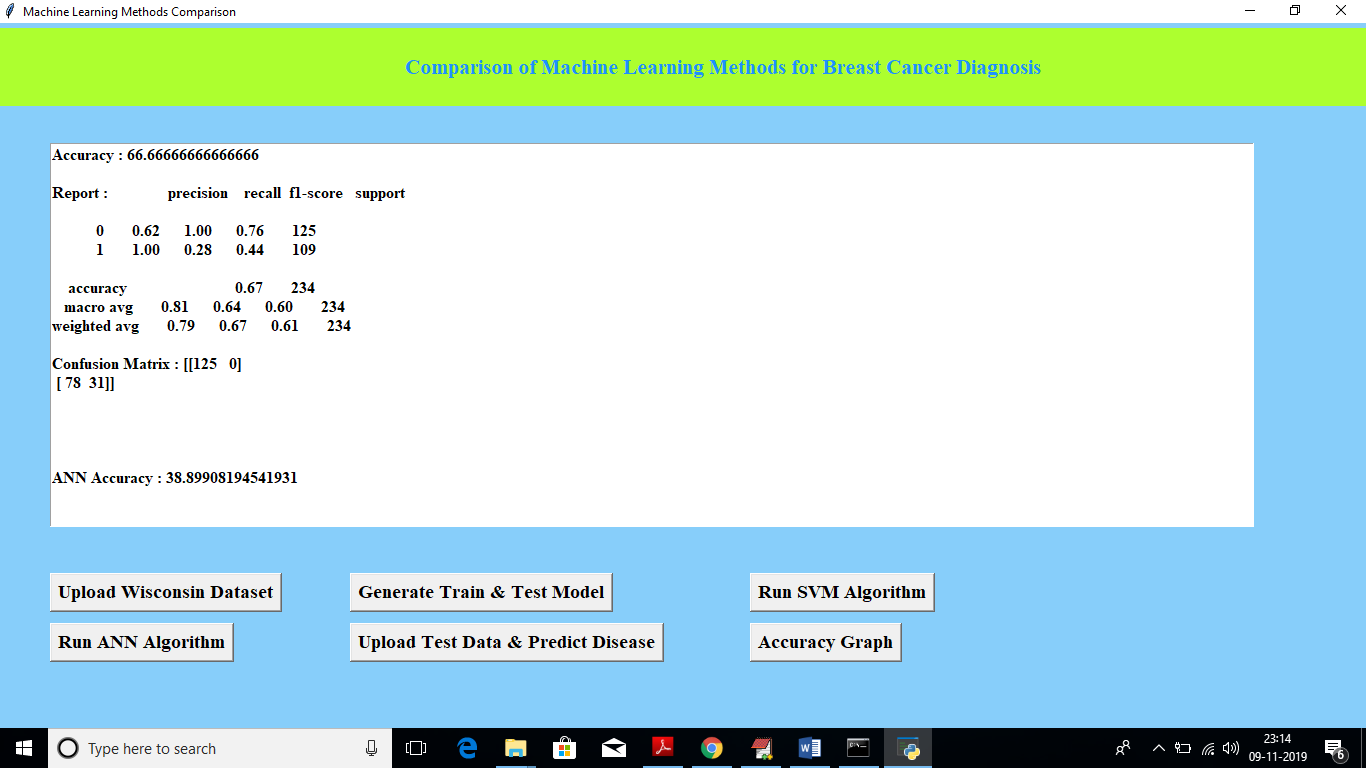
Now click on ‘Generate Train & Test Model’ button to split dataset into train and test



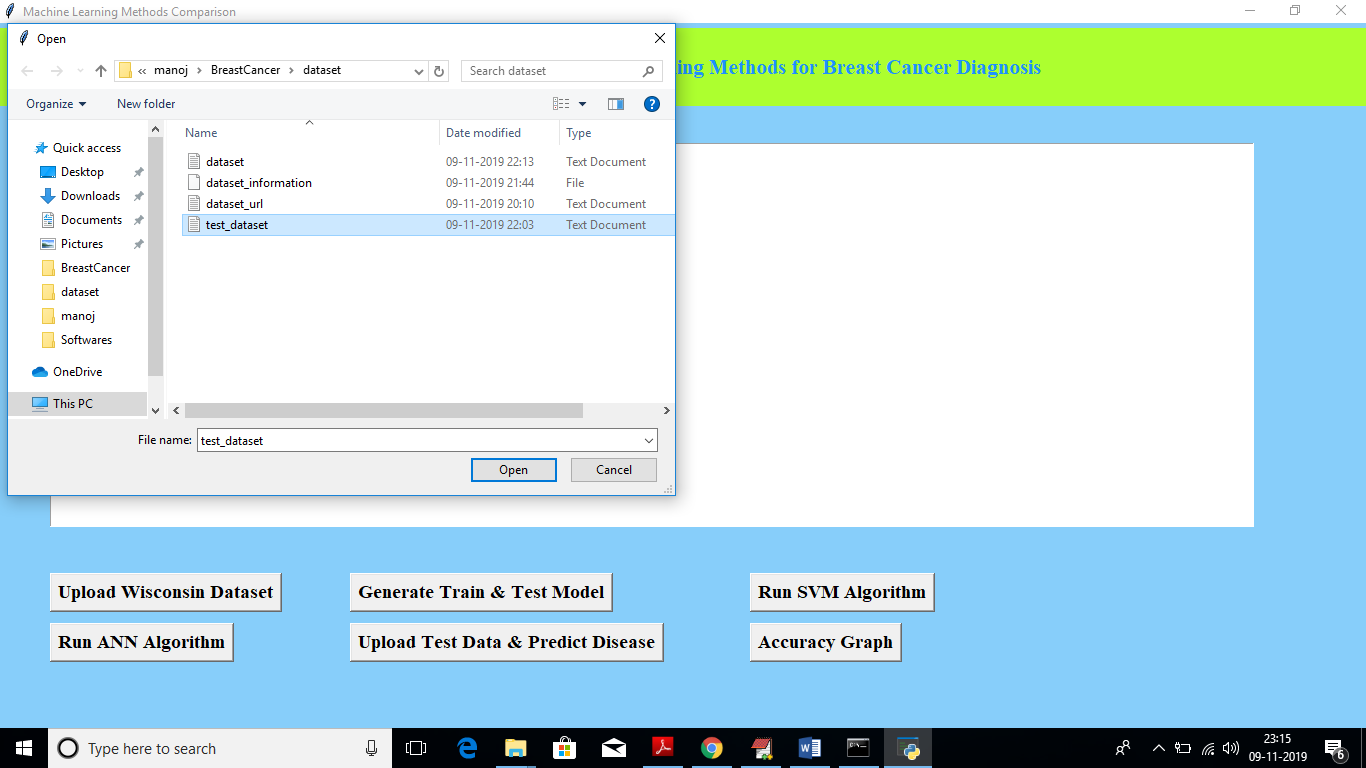
In above screen we can see dataset contains total 779 records and training size splits to 545 and test size splits to 234. Now click on ‘Run SVM Algorithm’ button to generate training model with SVM



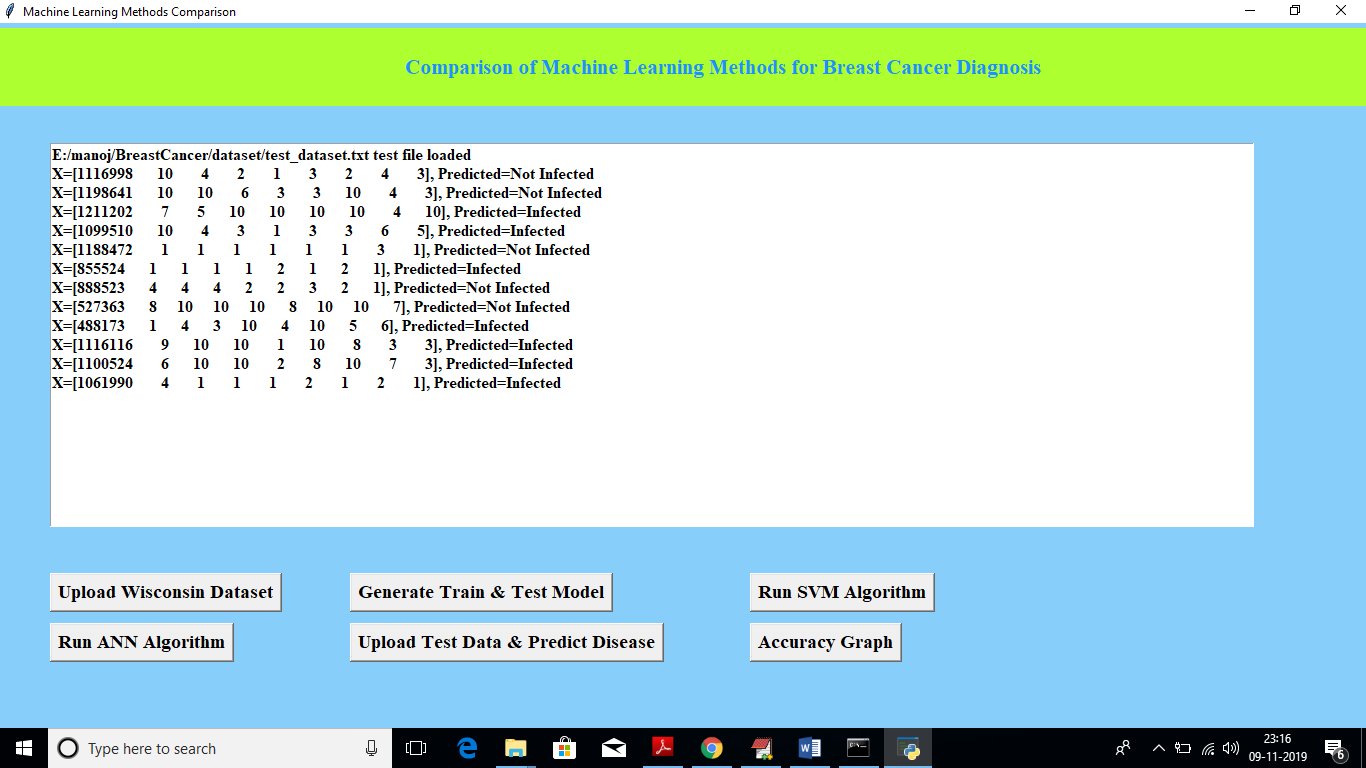
In above screen we can see Accuracy value, precision and recall obtained from SVM. SVM got 66.66% accuracy. Now click on ‘Run ANN Algorithm’ to get ANN Accuracy



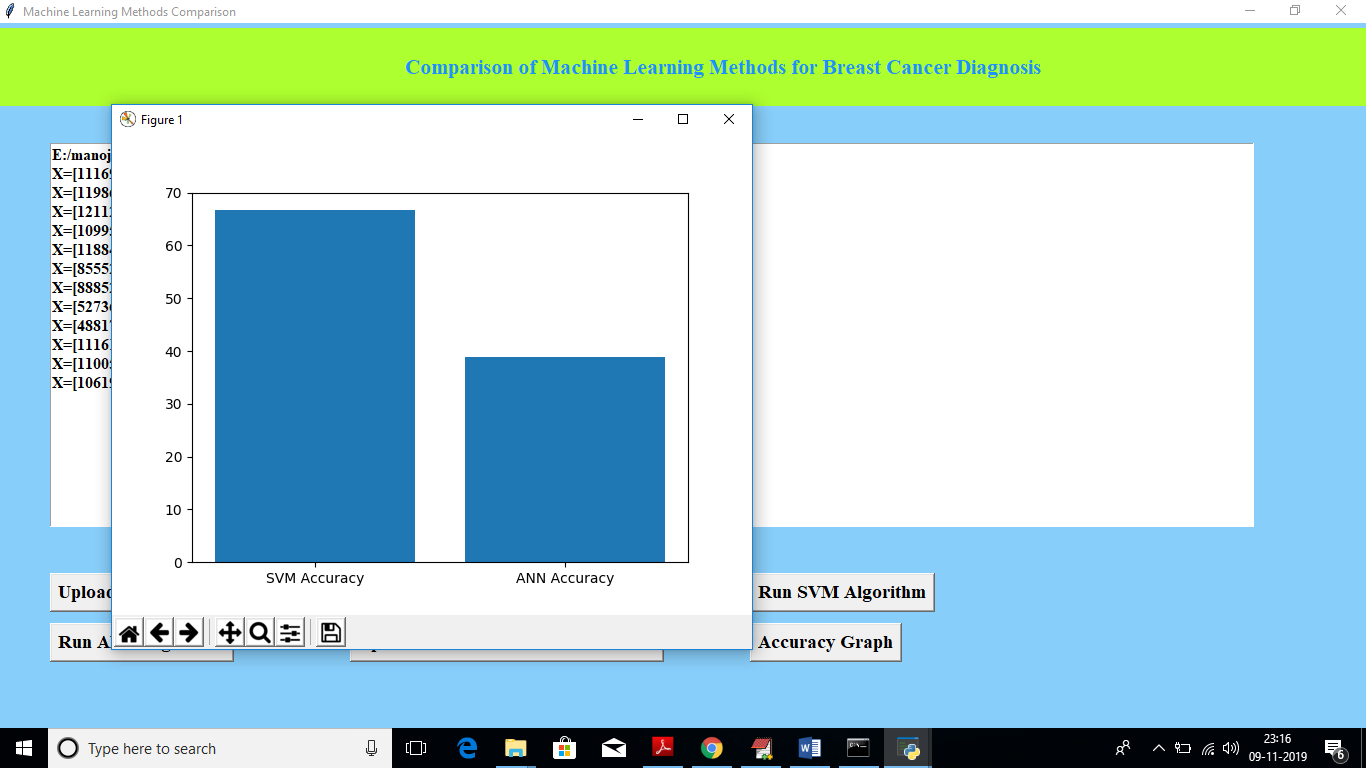
In above screen ANN got 38% accuracy and SVM got better accuracy than ANN. Now click on ‘Upload Test Data & Predict Disease’ button to upload test data and to predict disease



In above screen I am uploading test\_dataset file and below is predicted result



Now click on ‘Accuracy Graph’ button to get accuracy graph of both algorithms



In above graph x-axis represents algorithm name and y-axis represents accuracy of that algorithm. From above graph we can conclude that SVM is better than ANN