Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques

In this paper author is evaluating performance of various classification/prediction algorithms such as SVM, Naïve Bayes, and Logistic Regression etc to predict heart disease. All this algorithms are good in prediction but accuracy is not good enough. To get better prediction accuracy author is combining two classification algorithms such as Linear Model and Random Forest to build new algorithm called Hybrid Machine Learning to get better prediction accuracy of heart dataset. Hybrid algorithm will form up by using Voting classifier, Internally Voting classifier will build up using Linear Model and Random Forest and while classification voting algorithm will evaluate prediction accuracy of both algorithms and vote for that algorithm which gives better accuracy. So by using hybrid model always we will have better prediction accuracy algorithm which helps in better prediction of heart disease.

To implement above project we are using Cleveland Heart disease dataset and this dataset saved inside ‘heart\_dataset’ folder.

This project consist of 9 modules

1. Upload Module: using this module we will upload heart disease dataset of previous patients
2. Pre-process Module: Using this module we will remove all those records which contains missing values. Dataset will be splitted to two parts called training and testing, all classifier will build train model using training data and then test train model by applying test data on that train model to get classification accuracy.
3. SVM Module: Using this module we will build train model using SVM algorithm and then apply test data on that SVM model to get classification accuracy.
4. Naïve Bayes: Using this module we will build train model by using Naïve Bayes algorithm and apply test data to get Naïve Bayes classification accuracy.
5. Logistic Regression: Here train model accuracy will be check with Logistic Regression algorithm
6. ANN Module: Deep Learning Artificial Neural Network train model will be generated and its accuracy can be calculated using test data.
7. HRFLM: Propose Hybrid Algorithm which is combination of Linear model and Random Forest algorithm. Hybrid model will be generated by using both algorithms and then Voting classifier will be used to choose best performing algorithm.
8. Extension Extreme Machine Learning Module: This is an extra module which is built for extension purpose and this module is based on advance Extreme Machine Learning algorithm which can get better prediction accuracy compare to all algorithms. Extreme Learning Machine (ELM) is a novel method for pattern classification as well as function approximation. This method is essentially a single feed forward neural network; its structure consists of a single layer of hidden nodes, where the weights between inputs and hidden nodes are randomly assigned and remain constant during training and predicting phases. On the contrary, the weights that connect hidden nodes to outputs can be trained very fast. Experimental studies in the literature showed that ELMs can produce acceptable predictive performance and their computational cost is much lower than networks trained by the back-propagation algorithm.
9. Graph: This module display accuracy of all algorithms in graph format as comparison

Algorithm Details

**SVM 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 hyper plane 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 hyper plane as a line that linearly separates and classifies a set of data.

Intuitively, the further from the hyper plane 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 hyper plane as possible, while still being on the correct side of it.

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

**Random Forest Algorithm:** it’s an ensemble algorithm which means internally it will use multiple classifier algorithms to build accurate classifier model. Internally this algorithm will use decision tree algorithm to generate it train model for classification.

**Decision Tree Algorithm:** This algorithm will build training model by arranging all similar records in the same branch of tree and continue till all records arrange in entire tree. The complete tree will be referred as classification train model.

**Gradient Boosting Algorithm**:Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting. Gradient boosting models are becoming popular because of their effectiveness at classifying complex datasets, and have recently been used to win many Kaggle data science competitions.

**Deep Learning ANN Algorithm**: 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.

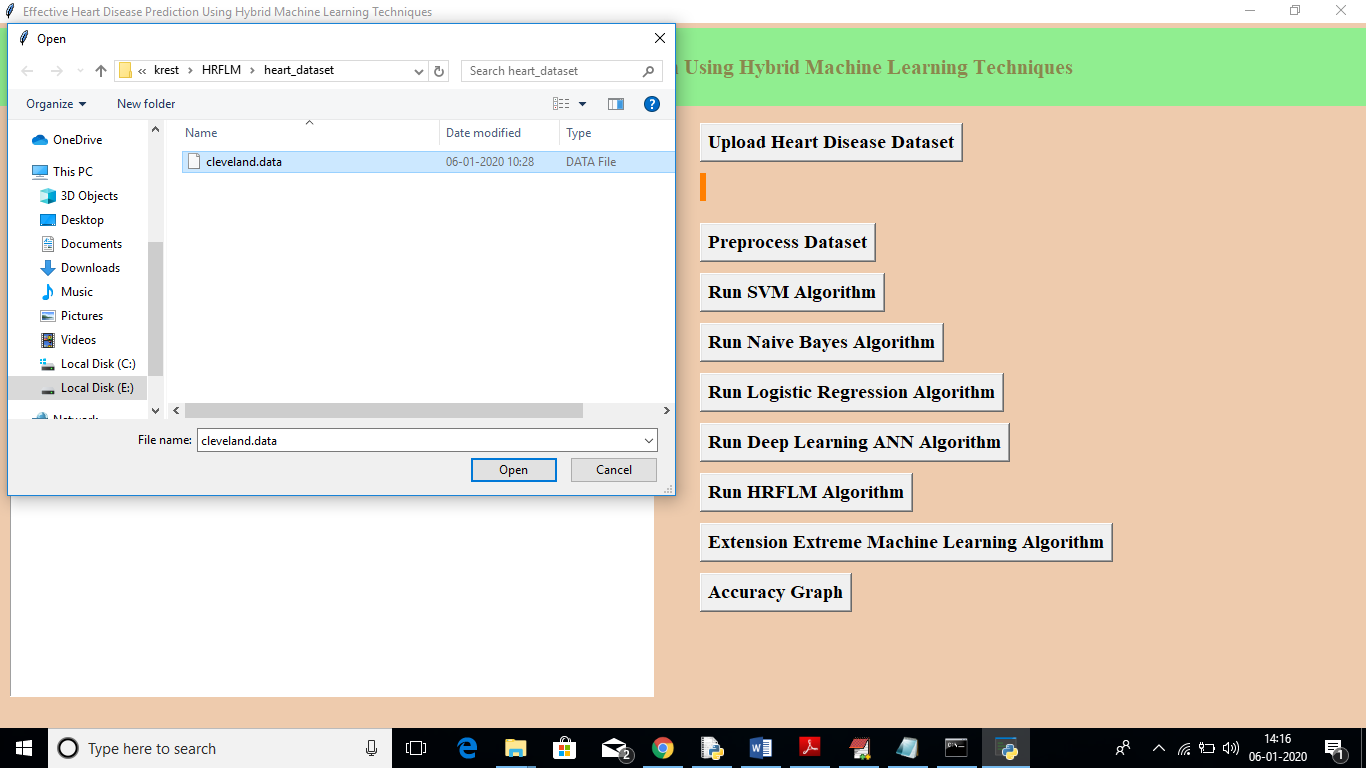
All above describe algorithms are to predict heart disease and using this paper we are comparing their performance.

Screen shots

To run this project double click on ‘run.bat’ file to get below screen



In above screen click on ‘Upload Heart Disease Dataset’ button to upload heart dataset



In above screen I am uploading ‘cleveland.data’ dataset, after uploading dataset will get below screen



In above screen we can see dataset contains total 303 records, now click on ‘Pre-process Dataset’ button to apply pre-processing technique to remove out all non-numeric data.



In above screen after applying pre-processing dataset size reduced to 297 records and we can see application randomly splitted complete dataset in to tow parts called train and test. For training application using 237 records and for testing application using 60 records. Application will choose random 60 records so always accuracy of same algorithm will be different as records for testing are randomly chooses.

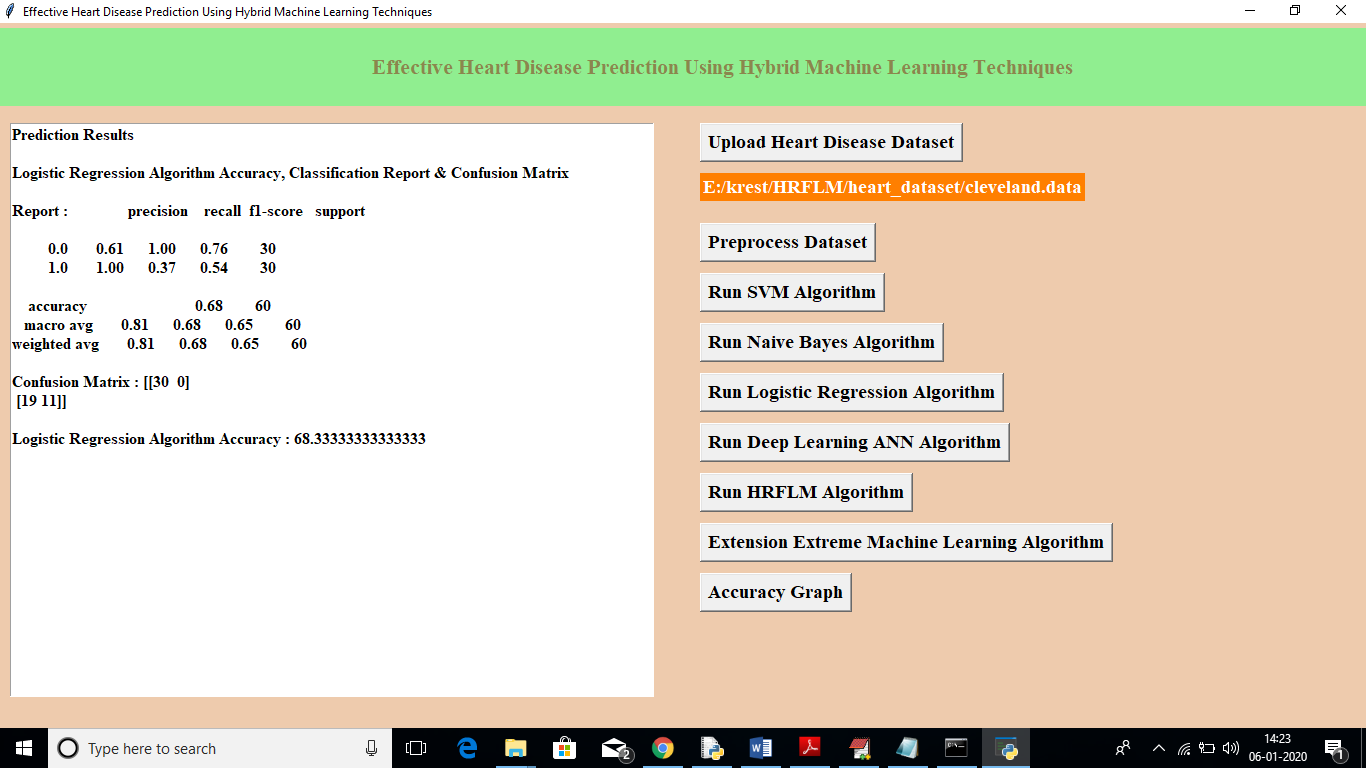
Now click on ‘Run SVM Algorithm’ button to generate SVM model on train dataset and to apply test data to get SVM classification accuracy.



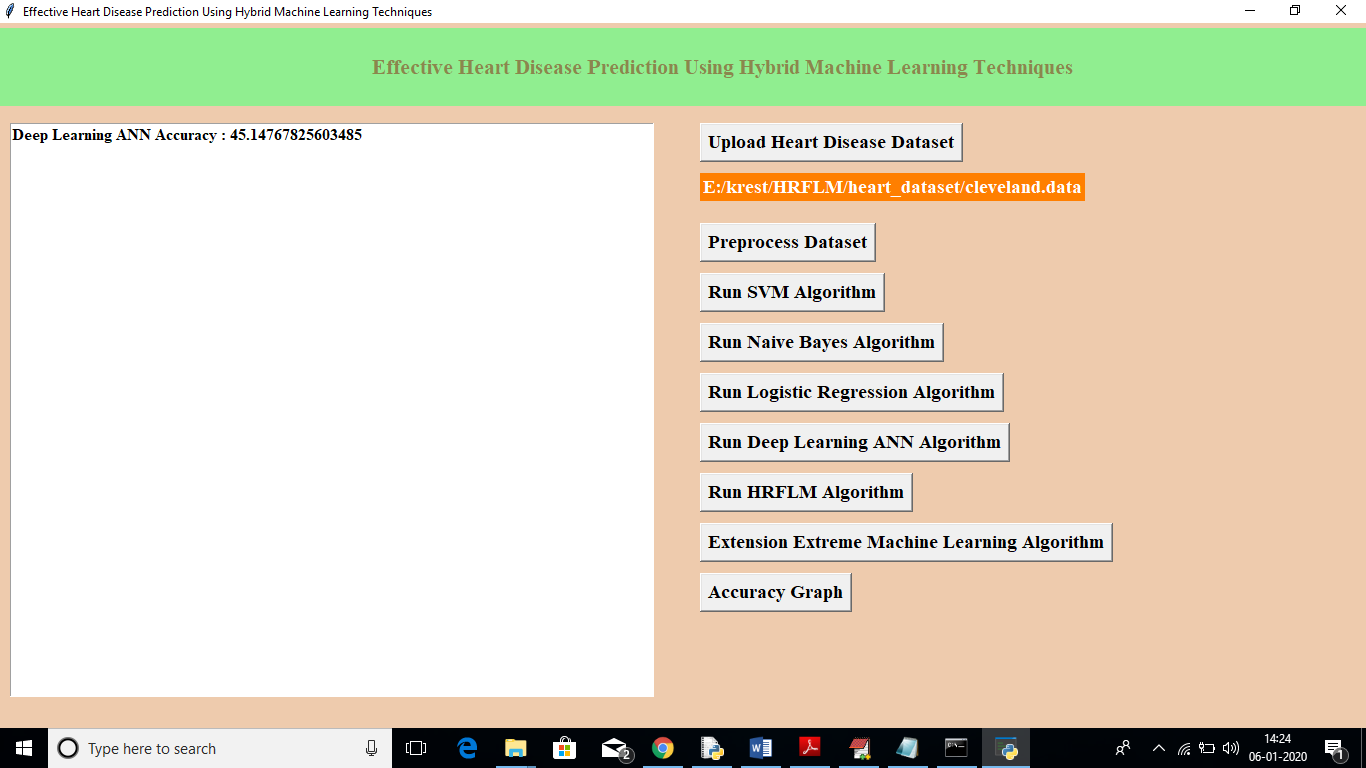
In above screen SVM got 62% accuracy, now click on ‘Run Naïve Bayes Algorithm’ button to get its accuracy



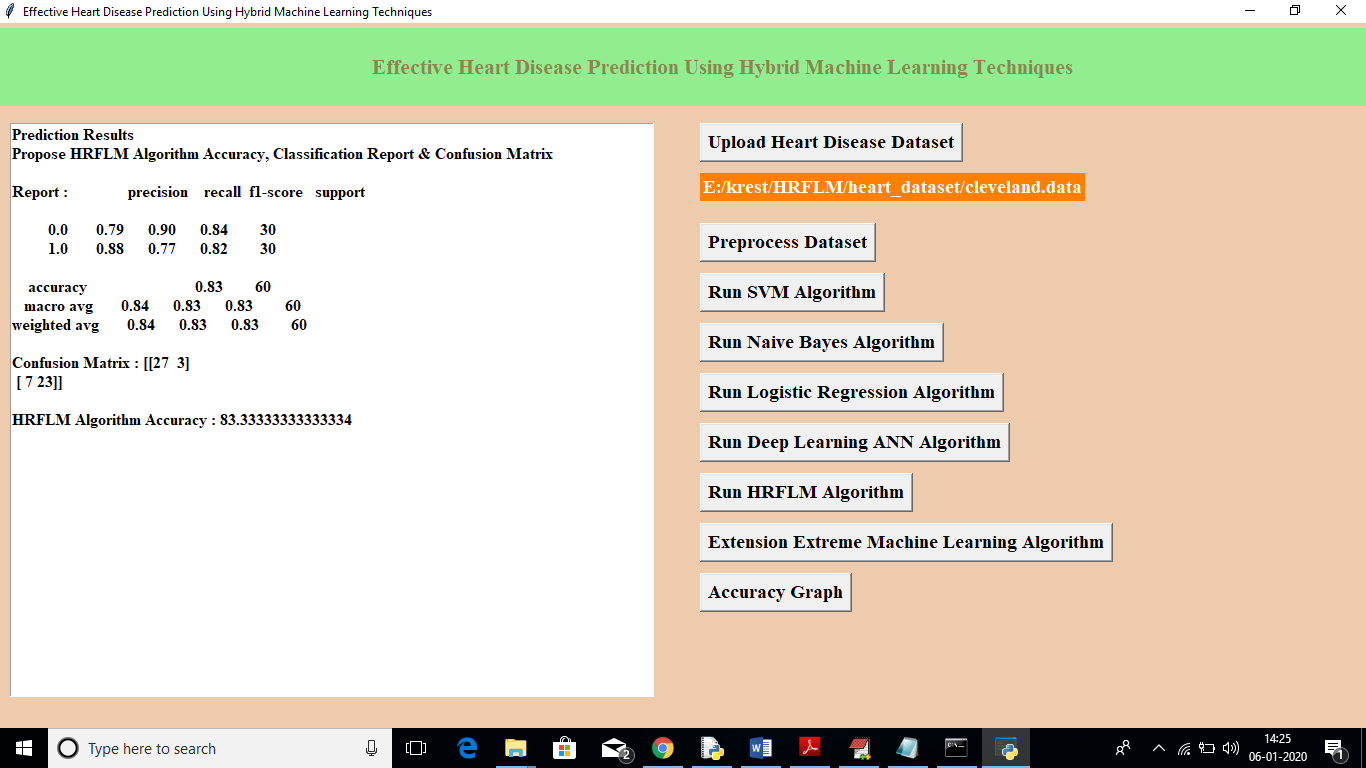
In above screen we can see Naïve Bayes got 72% accuracy, now click on ‘Run Logistic Regression Algorithm’ to get its accuracy



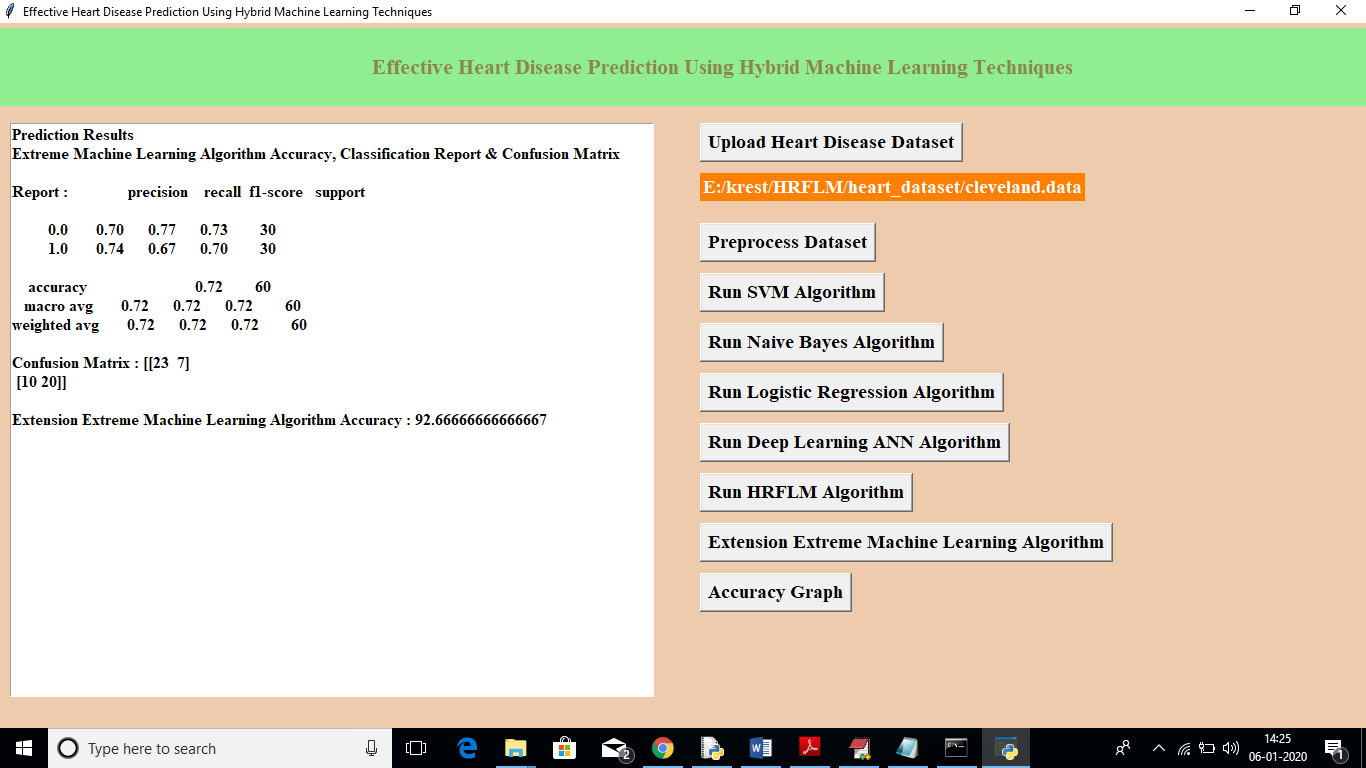
In above screen logistic regression got 69% accuracy, now click on ‘Run Deep Learning ANN Algorithm’ button to get its accuracy



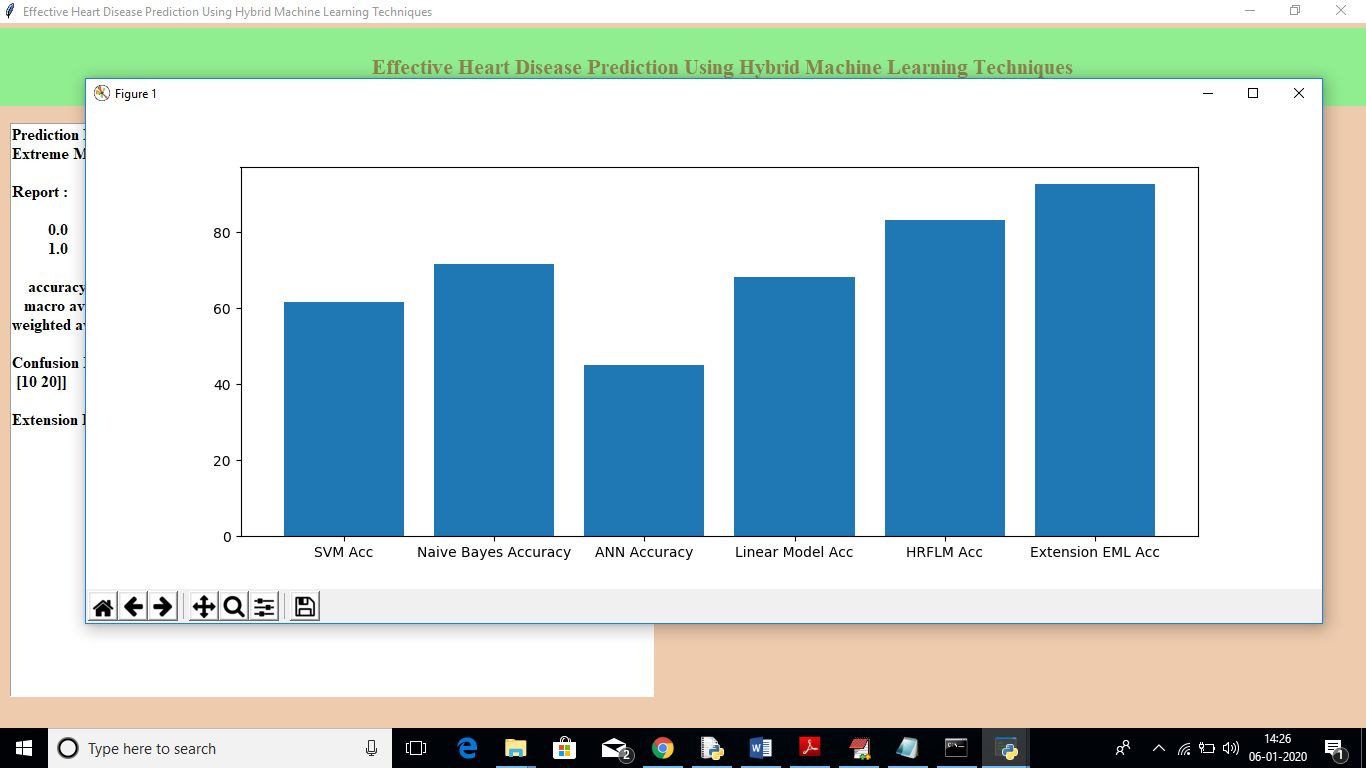
In above screen we can see ANN got 46% accuracy, now click on ‘Run HRFLM Algorithm’ button to get propose work accuracy



In above algorithm we can see HRFLM got 84% accuracy, now click on ‘Extension Extreme Machine Learning Algorithm’ button to check EML extension accuracy



In above screen we can see extension EML algorithm got 93% accuracy which is better than all algorithms. Now click on ‘Accuracy Graph’ button to get below graph



In above graph x-axis represents algorithm names and y-axis represents accuracy of that algorithm. In all algorithms propose HRFLM and extension algorithm got better accuracy