Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection

In this paper author is evaluating performance of two supervised machine learning algorithms such as SVM (Support Vector Machine) and ANN (Artificial Neural Networks). Machine learning algorithms will be used to detect whether request data contains normal or attack (anomaly) signatures. Now-a-days all services are available on internet and malicious users can attack client or server machines through this internet and to avoid such attack request IDS (Network Intrusion Detection System) will be used, IDS will monitor request data and then check if its contains normal or attack signatures, if contains attack signatures then request will be dropped.

IDS will be trained with all possible attacks signatures with machine learning algorithms and then generate train model, whenever new request signatures arrived then this model applied on new request to determine whether it contains normal or attack signatures. In this paper we are evaluating performance of two machine learning algorithms such as SVM and ANN and through experiment we conclude that ANN outperform existing SVM in terms of accuracy.

To avoid all attacks IDS systems has developed which process each incoming request to detect such attacks and if request is coming from genuine users then only it will forward to server for processing, if request contains attack signatures then IDS will drop that request and log such request data into dataset for future detection purpose.

To detect such attacks IDS will be prior train with all possible attacks signatures coming from malicious user’s request and then generate a training model. Upon receiving new request IDS will apply that request on that train model to predict it class whether request belongs to normal class or attack class. To train such models and prediction various data mining classification or prediction algorithms will be used.

In this paper author is evaluating performance of SVM and ANN.

In this algorithms author has applied Correlation Based and Chi-Square Based feature selection algorithms to reduce dataset size, this feature selection algorithms removed irrelevant data from dataset and then used model with important features, due to this features selection algorithms dataset size will reduce and accuracy of prediction will increase.

To conduct experiment author has used NSL KDD Dataset and below is some example records of that dataset which contains request signatures. I have also used same dataset and this dataset is available inside ‘dataset’ folder.

Dataset example

**duration,protocol\_type,service,flag,src\_bytes,dst\_bytes,land,wrong\_fragment,urgent,hot,num\_failed\_logins,logged\_in,num\_compromised,root\_shell,su\_attempted,num\_root,num\_file\_creations,num\_shells,num\_access\_files,num\_outbound\_cmds,is\_host\_login,is\_guest\_login,count,srv\_count,serror\_rate,srv\_serror\_rate,rerror\_rate,srv\_rerror\_rate,same\_srv\_rate,diff\_srv\_rate,srv\_diff\_host\_rate,dst\_host\_count,dst\_host\_srv\_count,dst\_host\_same\_srv\_rate,dst\_host\_diff\_srv\_rate,dst\_host\_same\_src\_port\_rate,dst\_host\_srv\_diff\_host\_rate,dst\_host\_serror\_rate,dst\_host\_srv\_serror\_rate,dst\_host\_rerror\_rate,dst\_host\_srv\_rerror\_rate,label**

All above comma separated names in bold format are the names of request signature

0,tcp,ftp\_data,SF,491,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0,0,0,0,1,0,0,150,25,0.17,0.03,0.17,0,0,0,0.05,0,normal

0,tcp,private,S0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,166,9,1,1,0,0,0.05,0.06,0,255,9,0.04,0.05,0,0,1,1,0,0,anamoly

Above two records are the signature values and last value contains class label such as normal request signature or attack signature. In second record ‘Neptune’ is a name of attack. Similarly in dataset you can find nearly 30 different names of attacks.

In above dataset records we can see some values are in string format such as tcp, ftp\_data and these values are not important for prediction and these values will be remove out by applying PREPROCESSING Concept. All attack names will not be identified by algorithm if it’s given in string format so we need to assign numeric value for each attack. All this will be done in PREPROCESS steps and then new file will be generated called ‘clean.txt’ which will use to generate training model.

In below line i am assigning numeric id to each attack

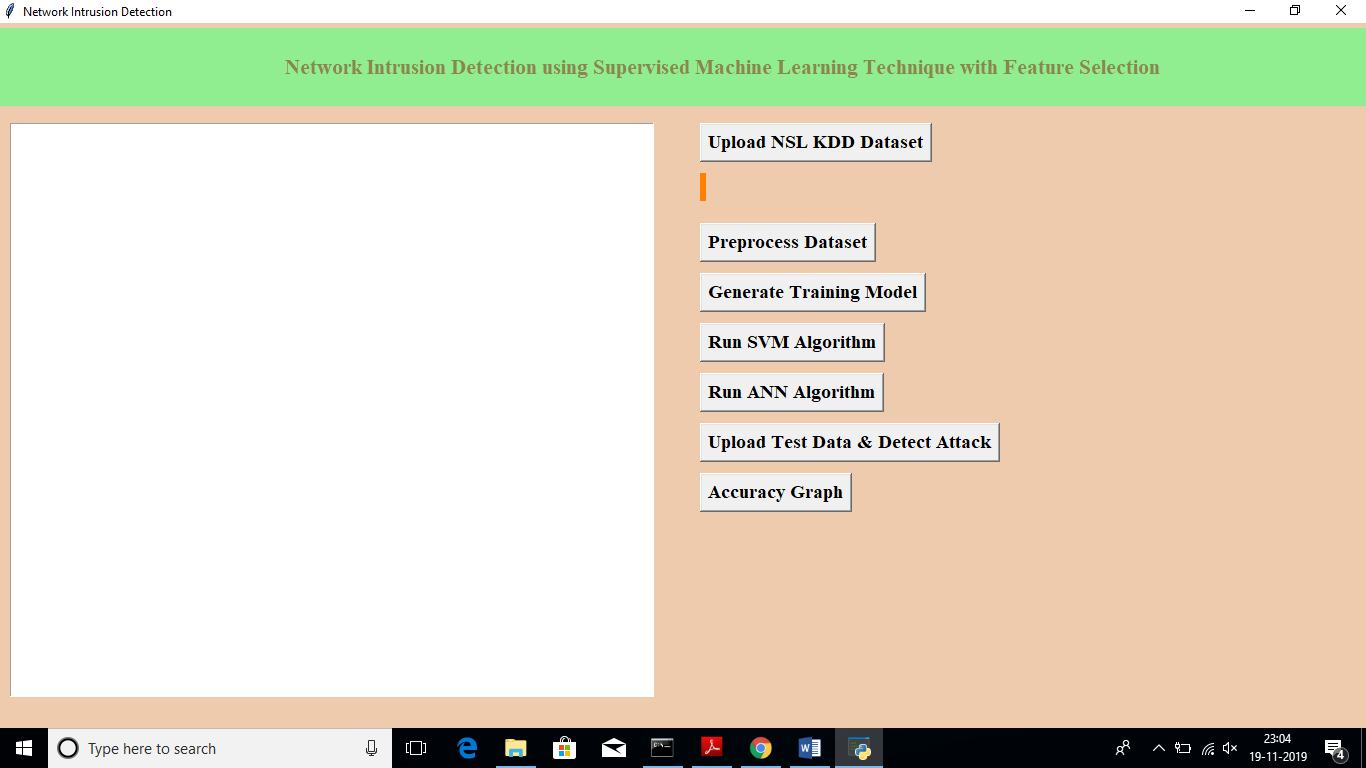
"normal":0,"anamoly":1

In above lines we can see normal is having id 0 and Anomaly has id 1 and goes on for all attacks.

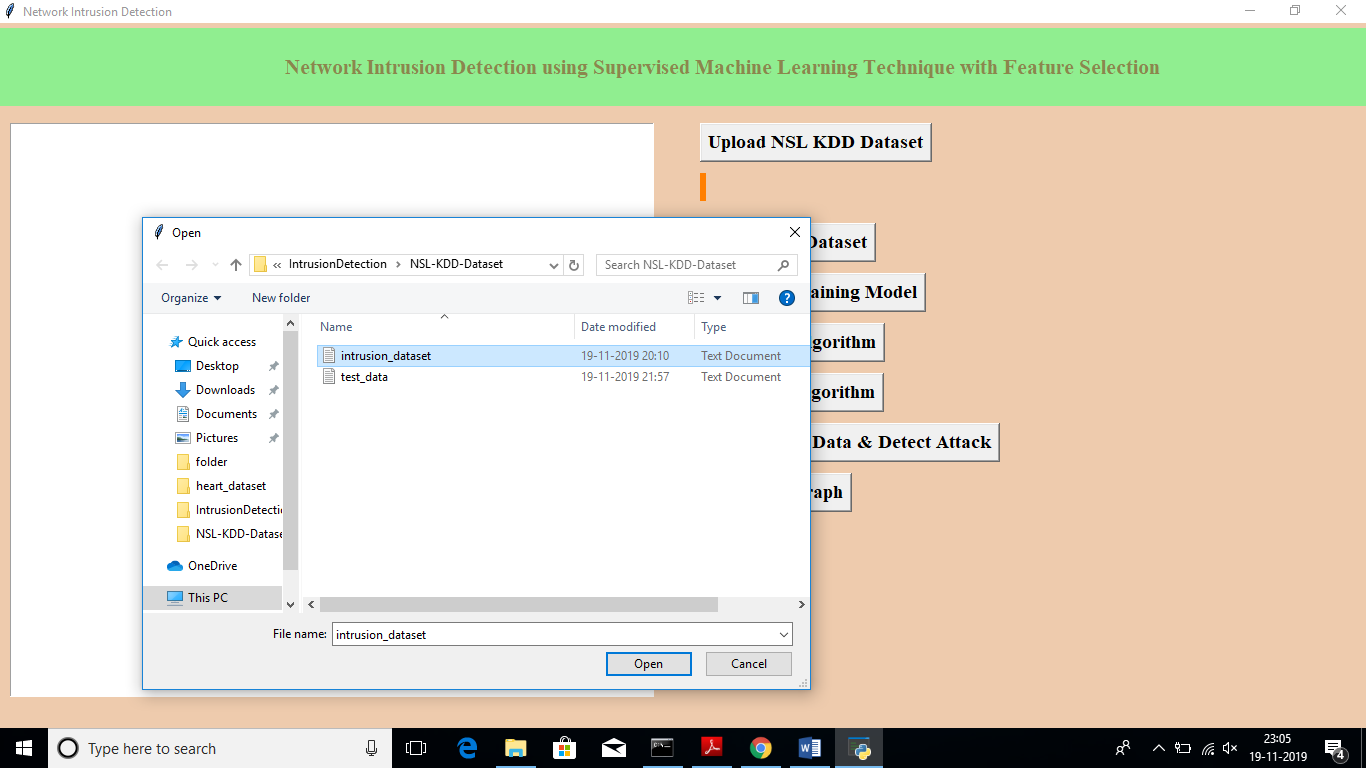
Before running code execute below two commands

Screen shots

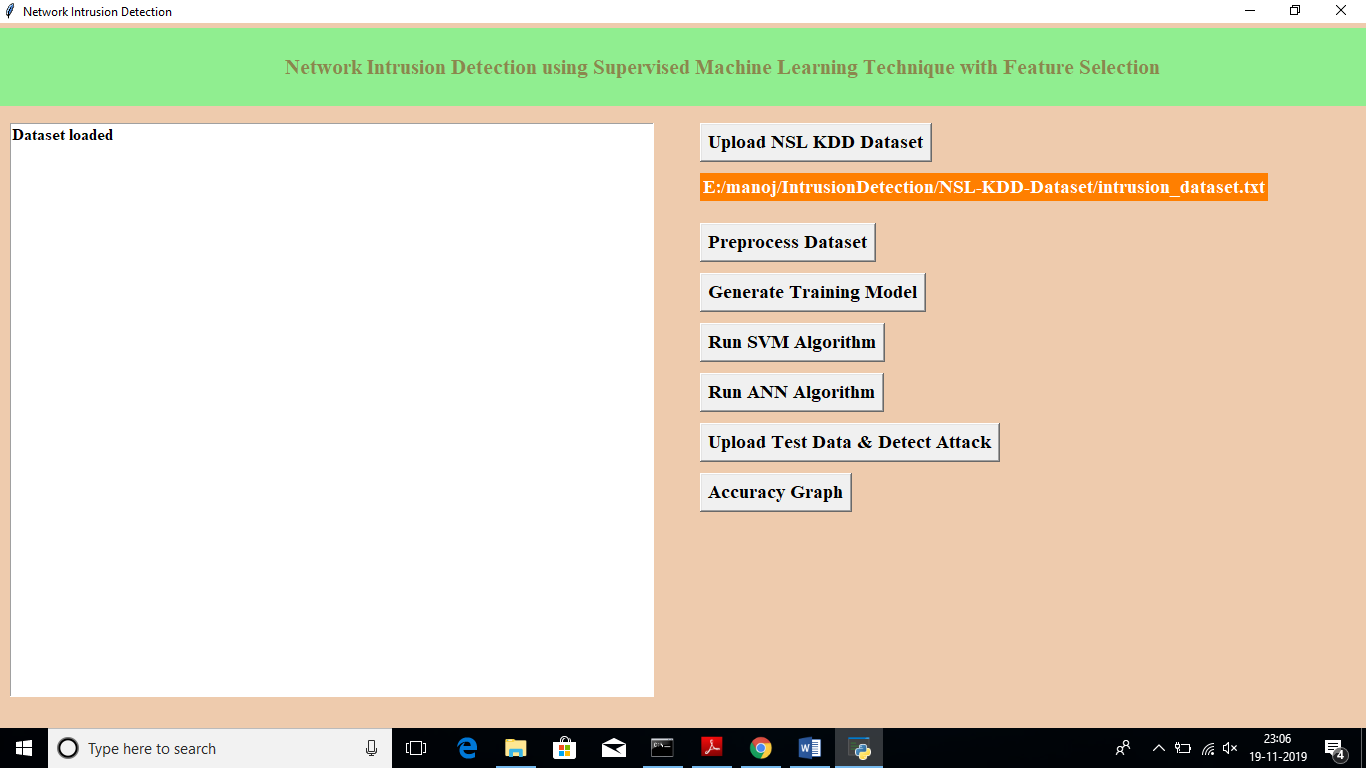
Double click on ‘run.bat’ file to get below screen



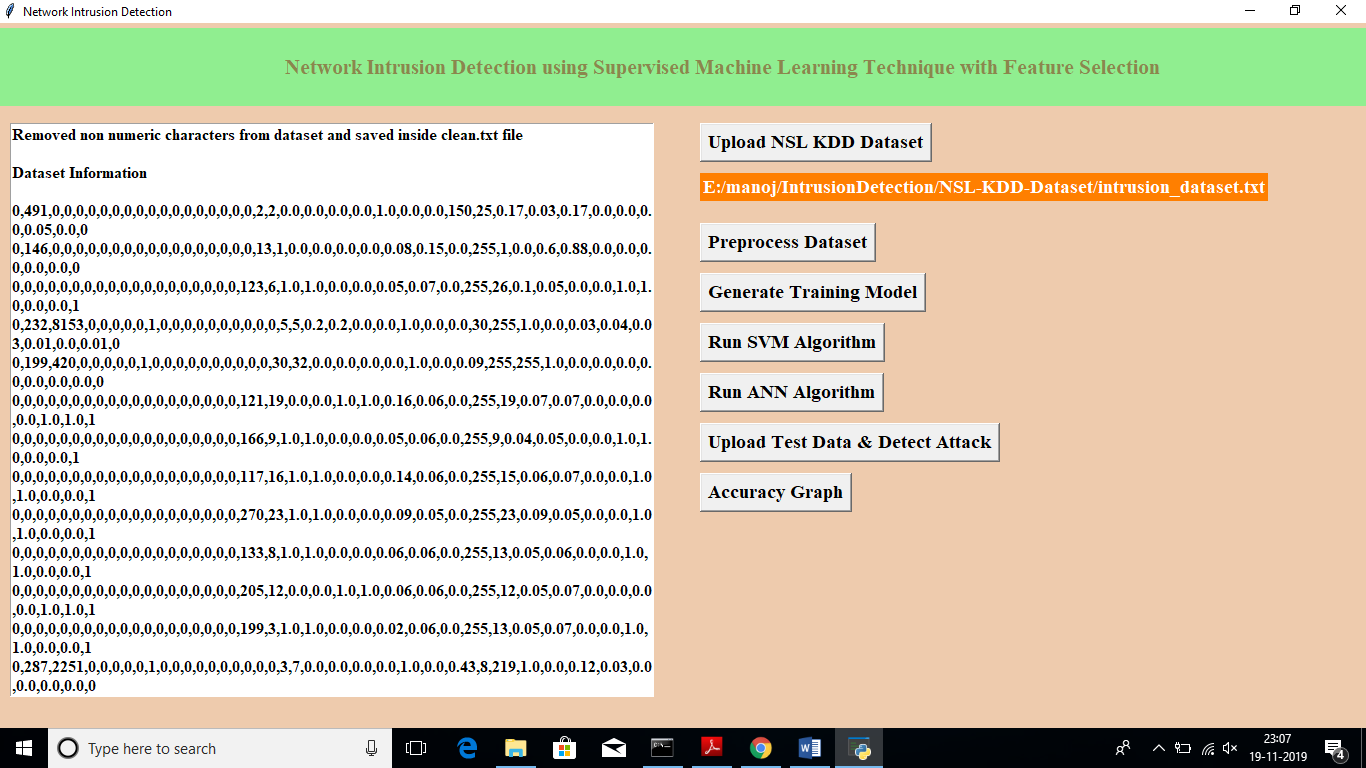
In above screen click on ‘Upload NSL KDD Dataset’ button and upload dataset



In above screen I am uploading ‘intrusion\_dataset.txt’ file, after uploading dataset will get below screen

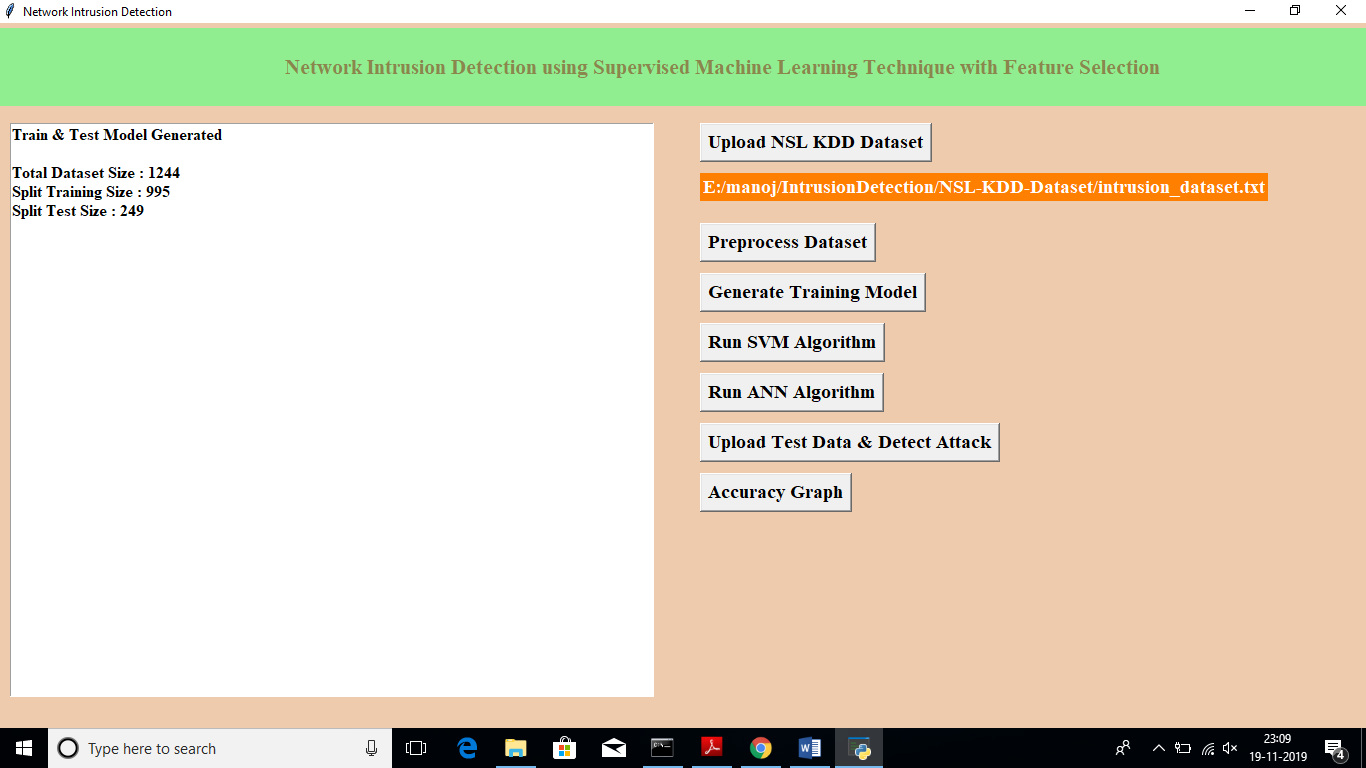


Now click on ‘Pre-process Dataset’ button to clean dataset to remove string values from dataset and to convert attack names to numeric values

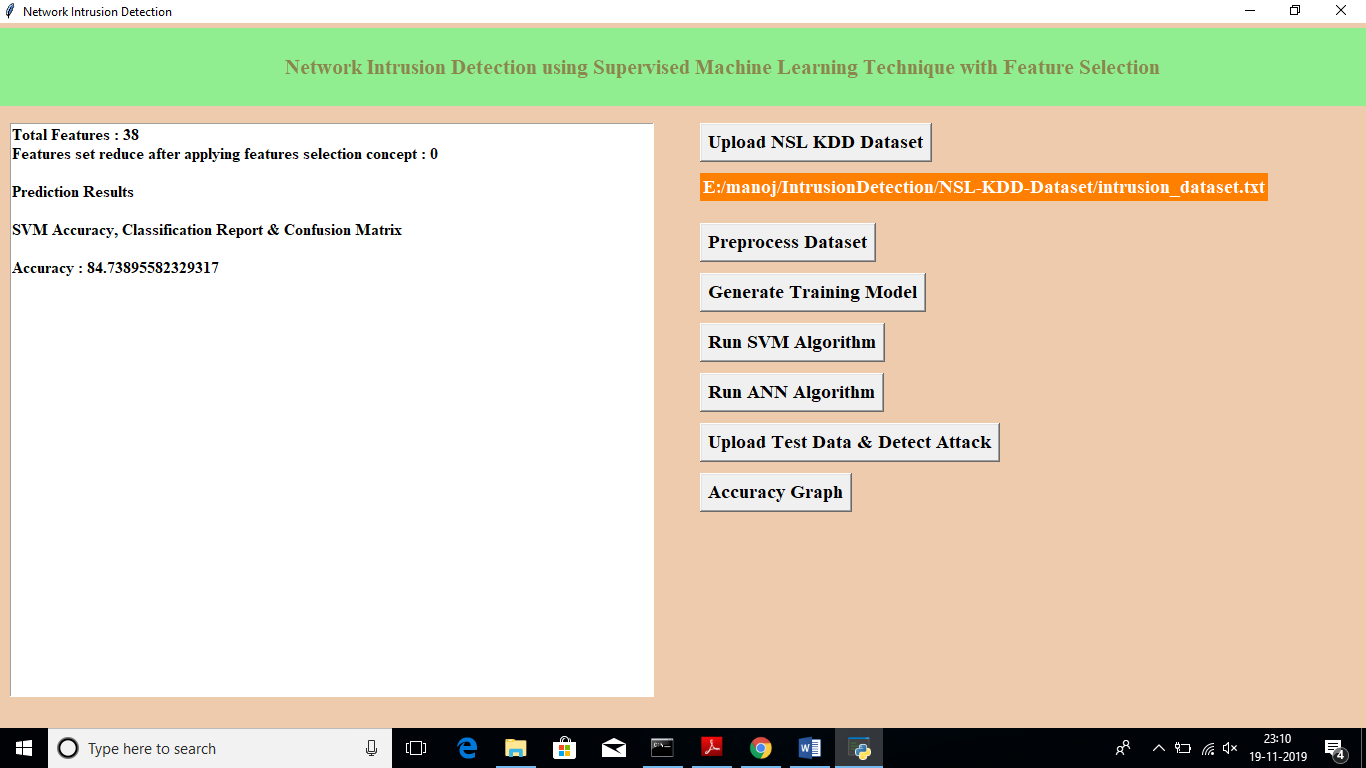


After pre-processing all string values removed and convert string attack names to numeric values such as normal signature contains id 0 and anomaly attack contains signature id 1.

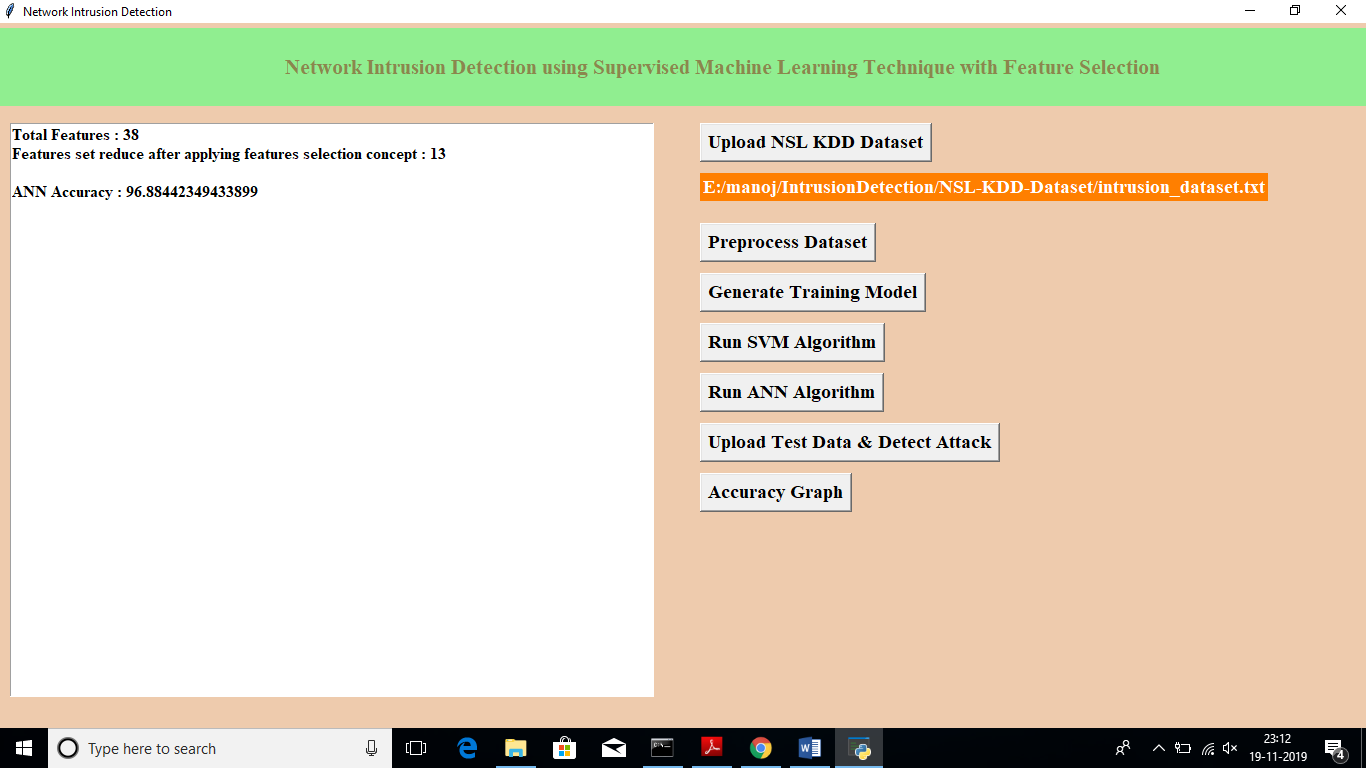
Now click on ‘Generate Training Model’ to split train and test data to generate model for prediction using SVM and ANN



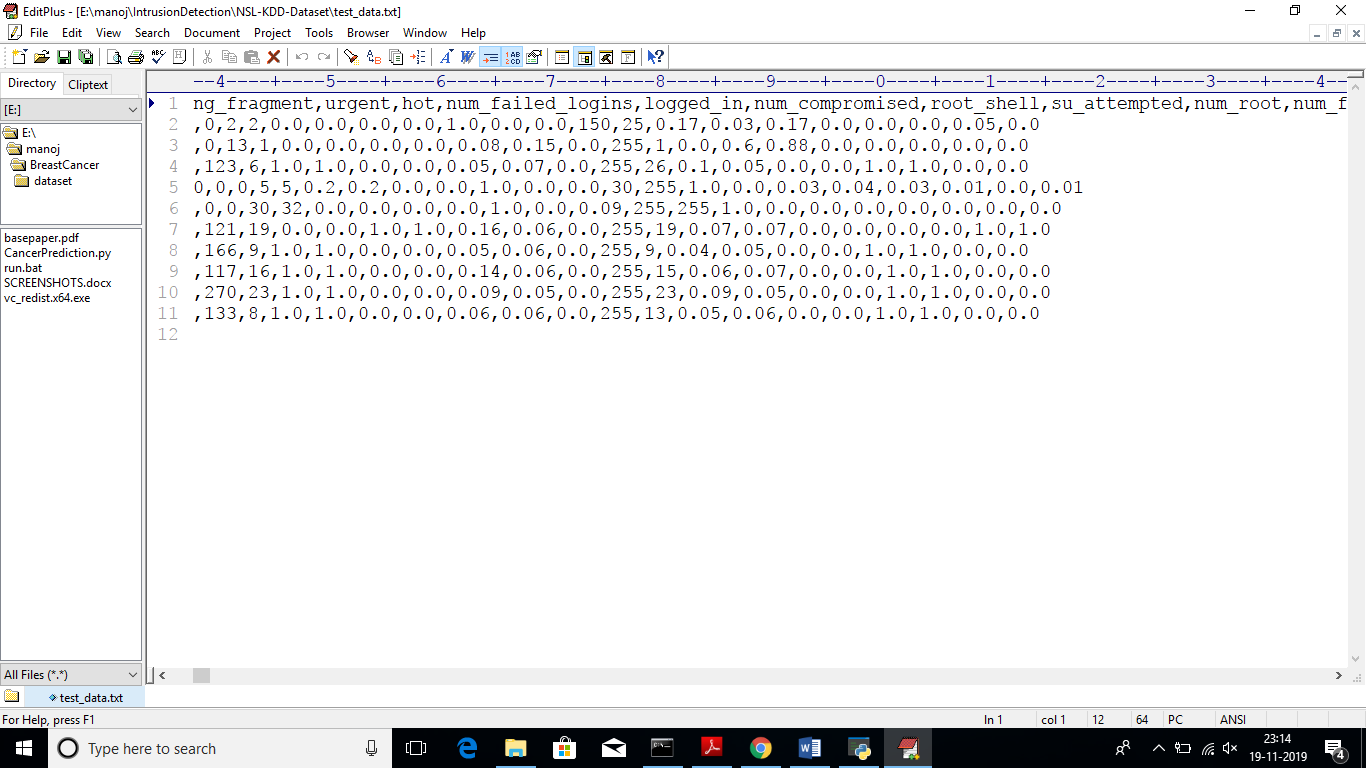
In above screen we can see dataset contains total 1244 records and 995 used for training and 249 used for testing. Now click on ‘Run SVM Algorithm’ to generate SVM model and calculate its model accuracy



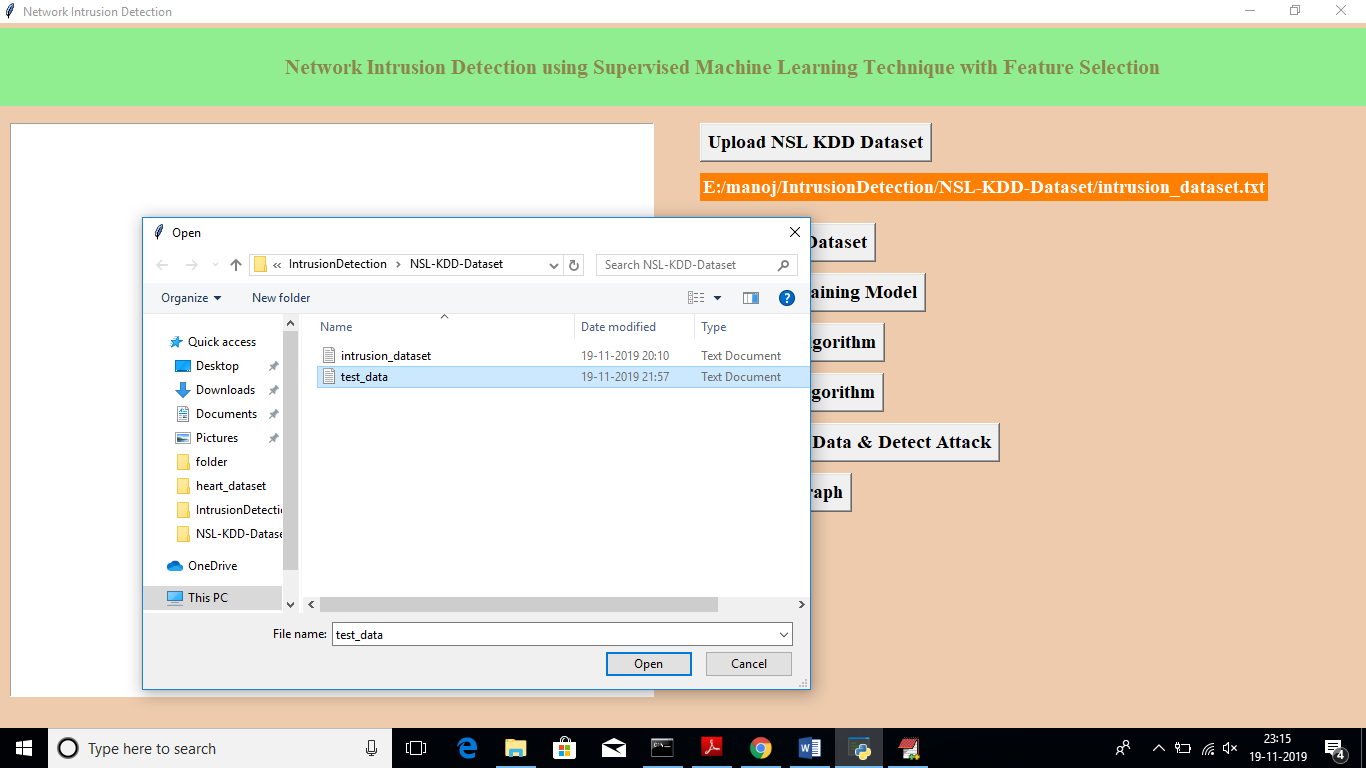
In above screen we can see with SVM we got 84.73% accuracy, now click on ‘Run ANN Algorithm’ to calculate ANN accuracy



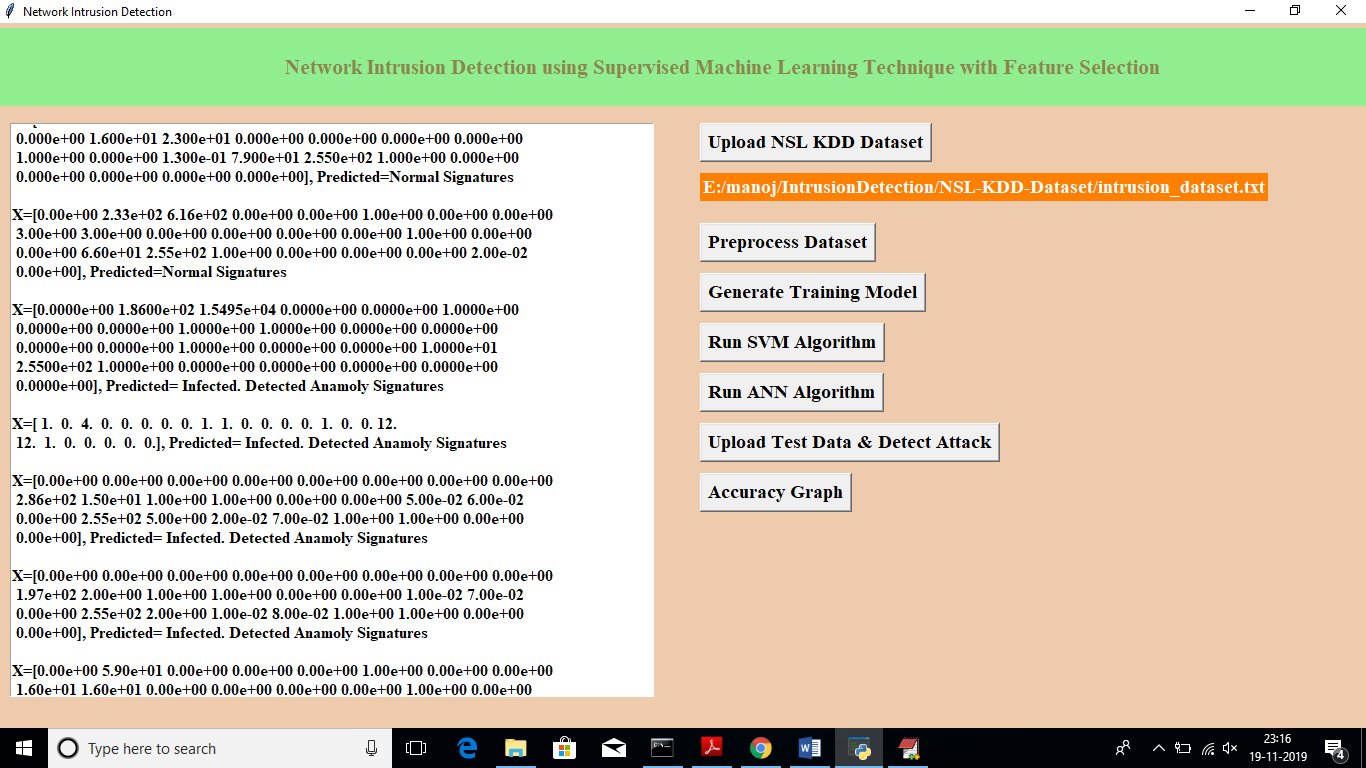
In above screen we got 96.88% accuracy, now we will click on ‘Upload Test Data & Detect Attack’ button to upload test data and to predict whether test data is normal or contains attack. All test data has no class either 0 or 1 and application will predict and give us result. See below some records from test data



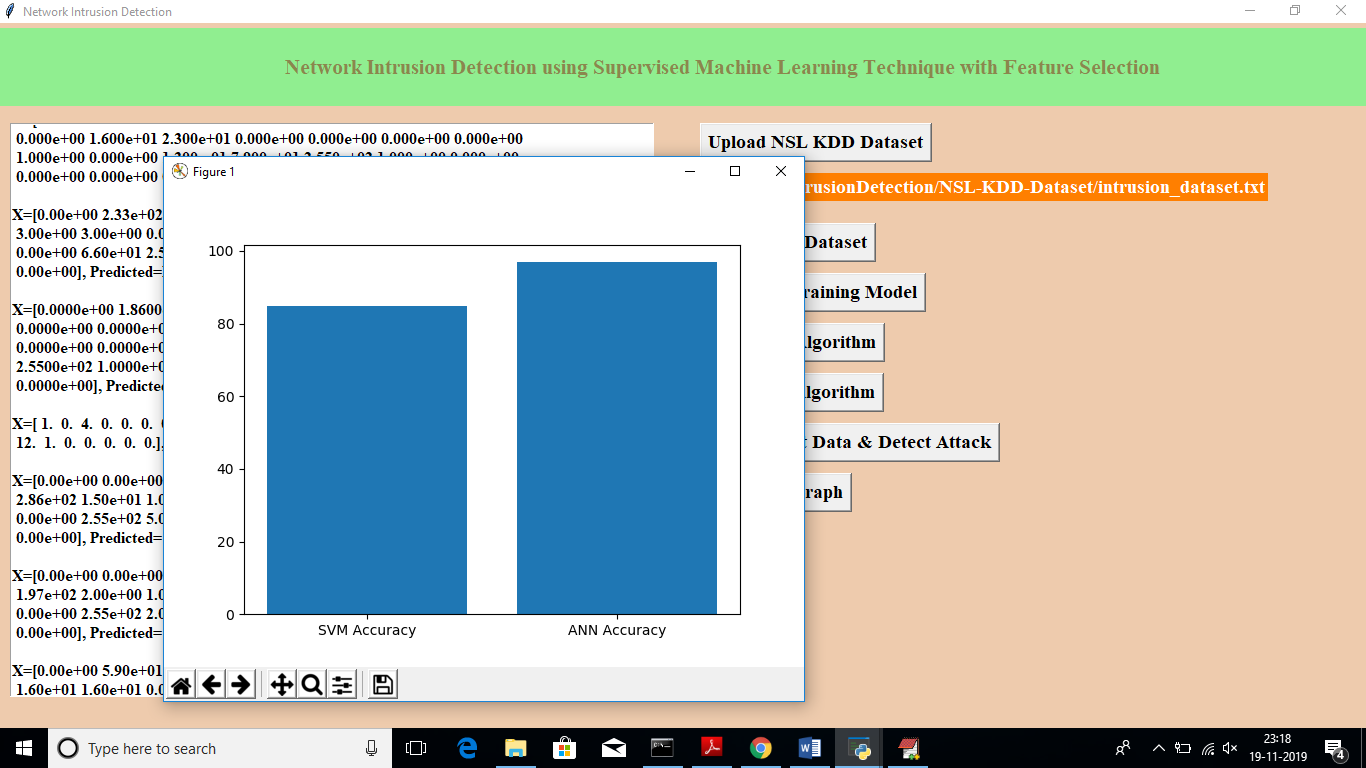
In above test data we don’t have either ‘0’ or ‘1’ and application will detect and give us result



In above screen I am uploading ‘test\_data’ file which contains test record, after prediction will get below results



In above screen for each test data we got predicted results as ‘Normal Signatures’ or ‘infected’ record for each test record. Now click on ‘Accuracy Graph’ button to see SVM and ANN accuracy comparison in graph format



From above graph we can see ANN got better accuracy compare to SVM, in above graph x-axis contains algorithm name and y-axis represents accuracy of that algorithms