CSKG4APT: A Cyber security Knowledge Graph for Advanced Persistent Threat Organization Attribution

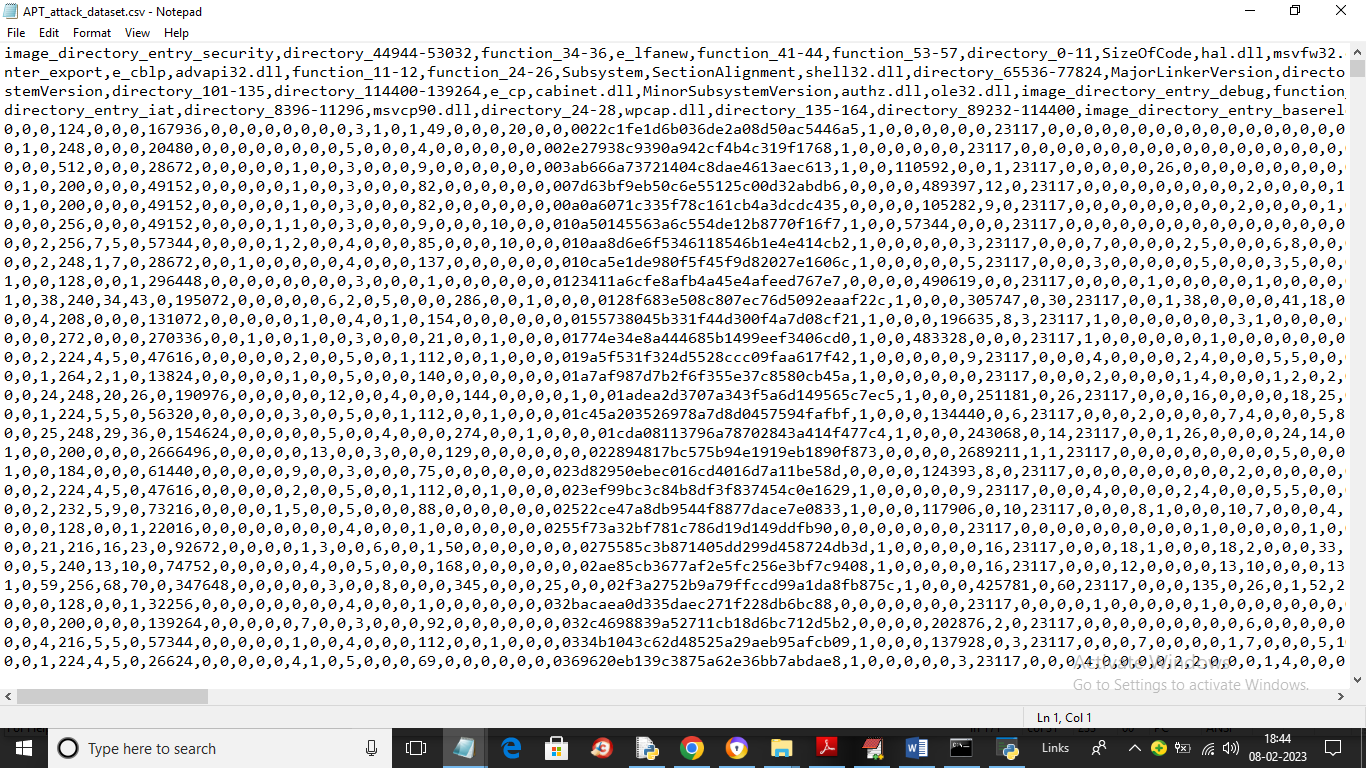
Networks always faces security issues with different types of attack in which some are permanent and some are non-permanent. APT (advance Persistent Attack) remain in network permanently. Existing algorithms on cyber threat intelligence (CTI) focus on automating the extraction of threat entities from public sources that describe attack events but this technique is not feasible so in propose paper author employing Knowledge Graph on APT attack dataset to discover APT attacks.

Building ontology based knowledge graph from APT dataset to extract network features and then employing deep learning BI-LSTM with GRU layers algorithm to train a model on APT graph features and this model can be applied on any network test data to identify whether test data is normal or contains any APT attacks.

To implement this project author has used APT Text base network dataset and then apply BERT (bidirectional encoder representations from transformers) algorithm on text data to convert into numeric vector and this vector contains average frequency of each words from the dataset. This BERT vector will be input to BI-LSTM with GRU algorithm to train a model and this model will be applied on test data to calculate prediction accuracy, precision, recall and FSCORE.

Open-source cyber threat intelligence (OSCTI) is becoming more influential in obtaining current network security information. Most studies on cyber threat intelligence (CTI) focus on automating the extraction of threat entities from public sources that describe attack events. The cyber security knowledge graph aims to change the expression of threat knowledge so that security researchers can accurately and efficiently obtain various types of threat information for preliminary intelligent decisions. The attribution technology can not only assist security analysts in detecting advanced persistent threats, but can also identify the same threat from different attack events. Therefore, it is important to trace the attack threat actor. Proposed paper apply knowledge graph technology, considered the latest research on cyber threat attack attribution, and thoroughly examined key related technologies and theories in the process of constructing and applying the advanced persistent threat (APT) knowledge graph from OSCTI. We designed a cyber-security platform named CSKG4APT based on a knowledge graph. Inspired by the theory of ontology, we constructed CSKG4APT as an APT knowledge graph model based on real APT attack scenarios. We then designed an APT threat knowledge. extraction algorithm for completing and updating the knowledge graph using deep learning (BI-LSTM with GRU layers) and expert knowledge.

In propose paper we have process text data to get below network information as BERT vector



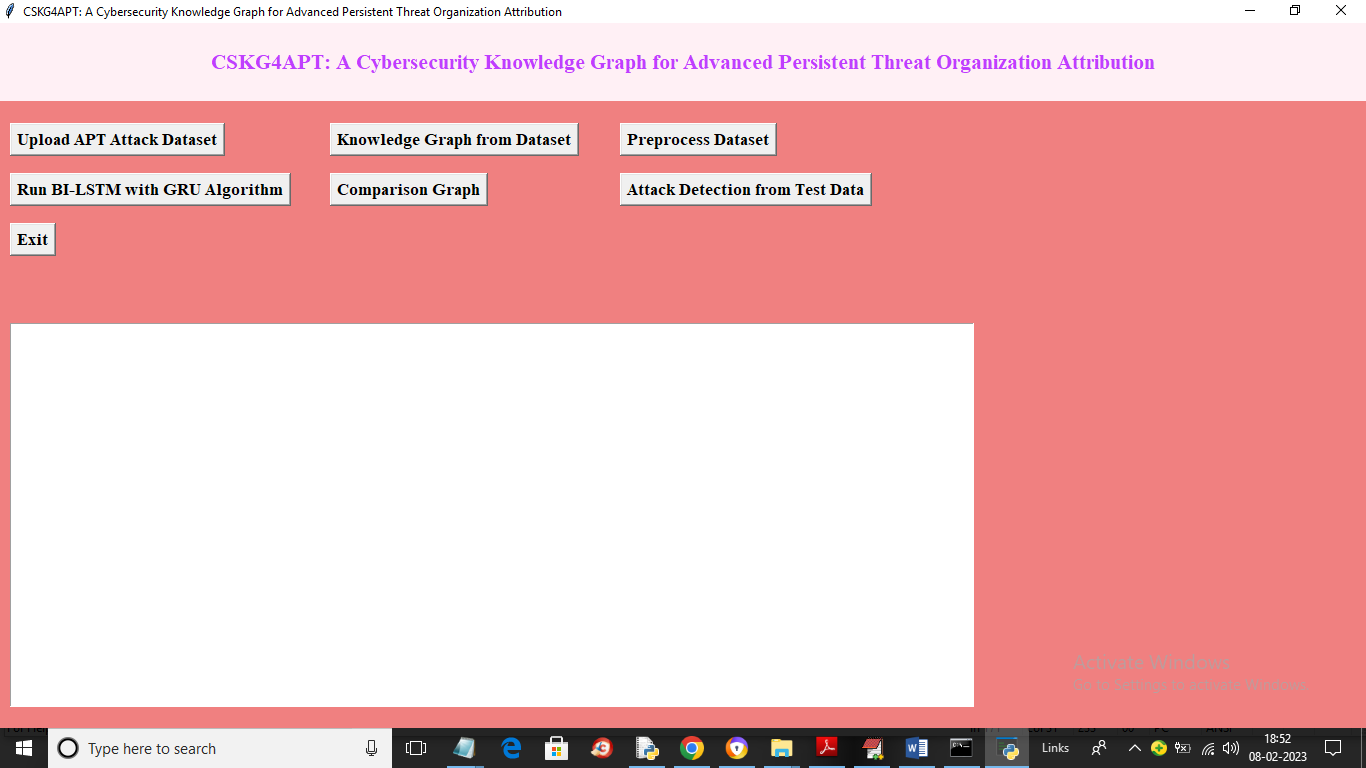
In above dataset screen first 4 rows contains words from the dataset and remaining words contains average frequency of each words under that column word name. This dataset will be input to deep learning algorithm to train a APT attack detection model.

To implement this project we have designed following modules

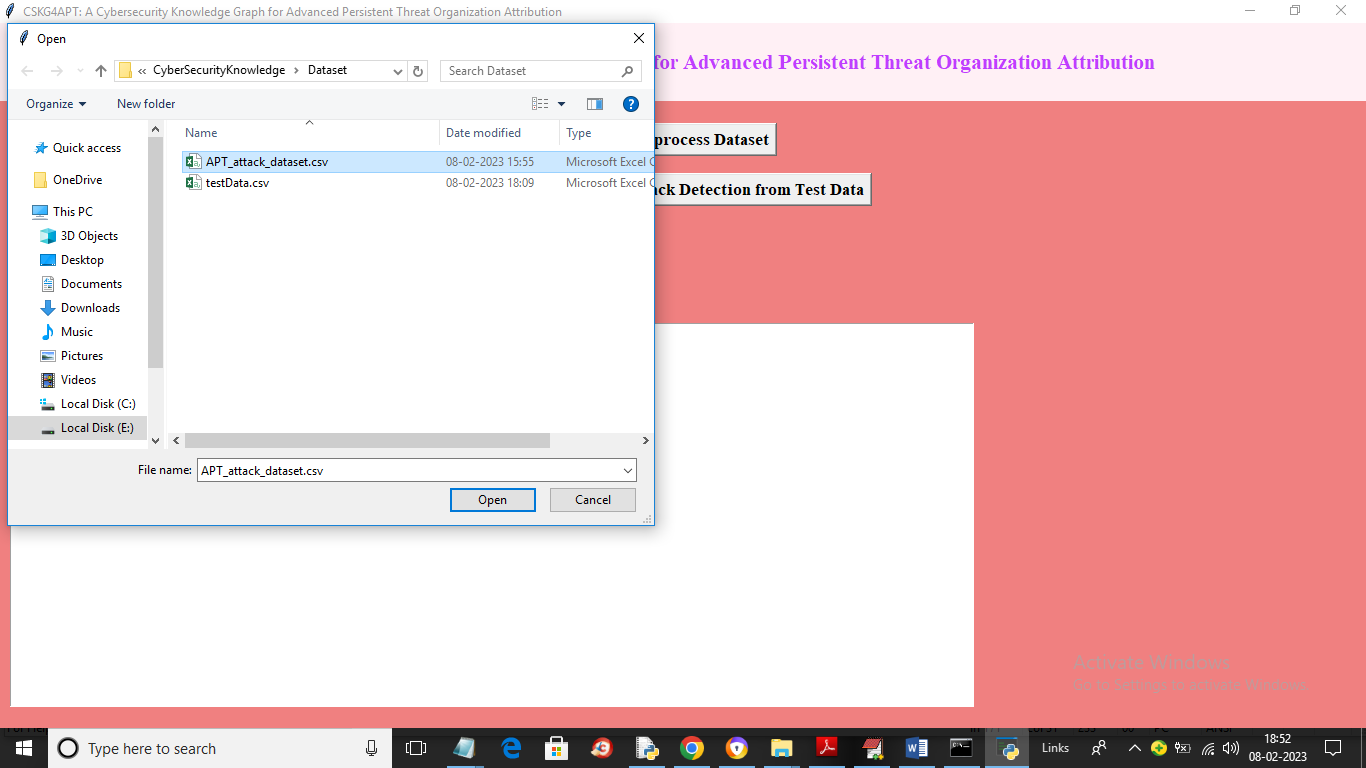
1. Upload APT Attack Dataset: using this module we will upload APT dataset to application and then find various cyber security attacks found in dataset and then plot a graph with all those attack names and their appearance frequency
2. Knowledge Graph from Dataset: using this module we will input entire dataset to graph algorithm to build a knowledge graph and this graph will display how attacks using network features
3. Preprocess Dataset: using this module we will remove missing values and then shuffle, normalize and split dataset into train and test where deep learning algorithm will take 80% dataset for training and 20% for testing
4. Run BI-LSTM with GRU Algorithm: 80% dataset will be input to BI-LSTM algorithm to train a model and this model will be applied on test data to calculate prediction accuracy
5. Comparison Graph: using this module we will plot propose algorithm accuracy and other metric comparison graph
6. Attack Detection from Test Data: using this module we will upload test data and then propose algorithm will analyse test data to predict APT attacks

SCREEN SHOTS

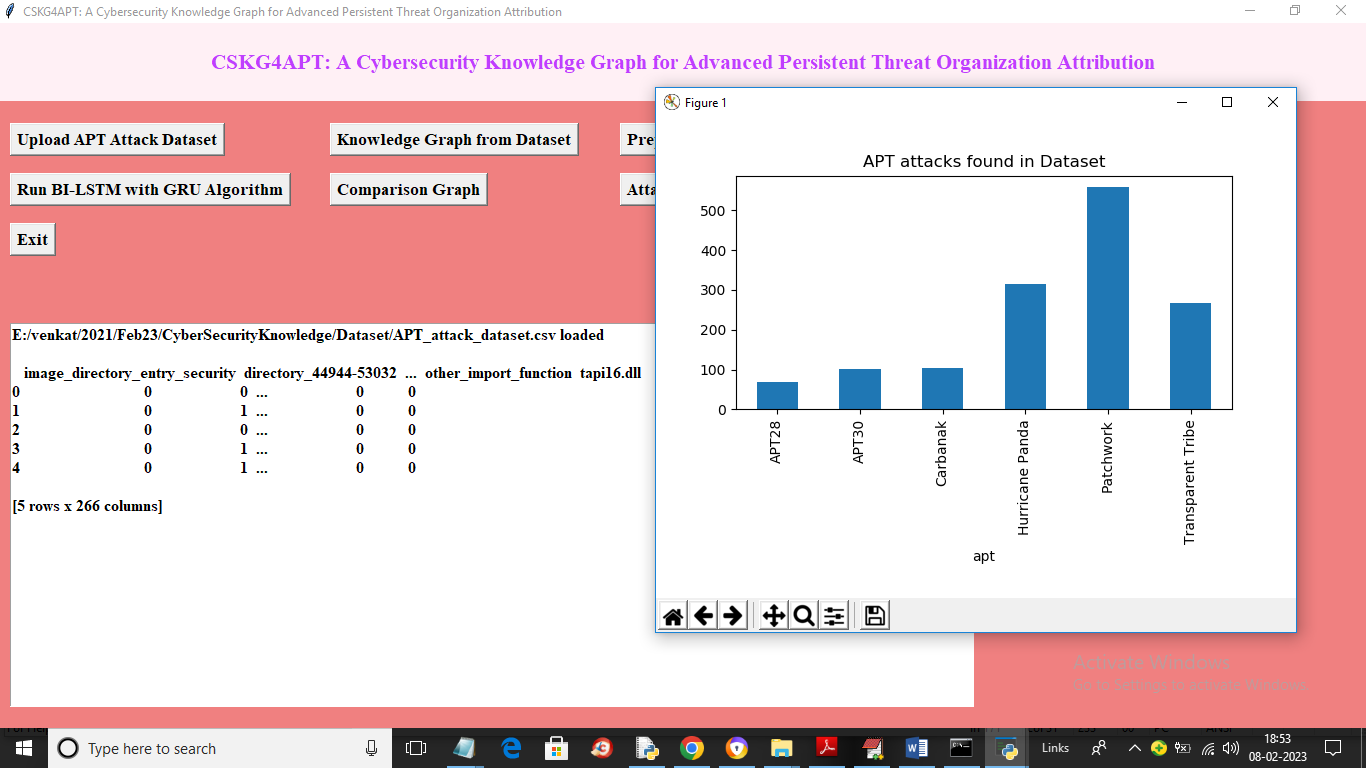
To run project double click on ‘run.bat’ file to get below screen



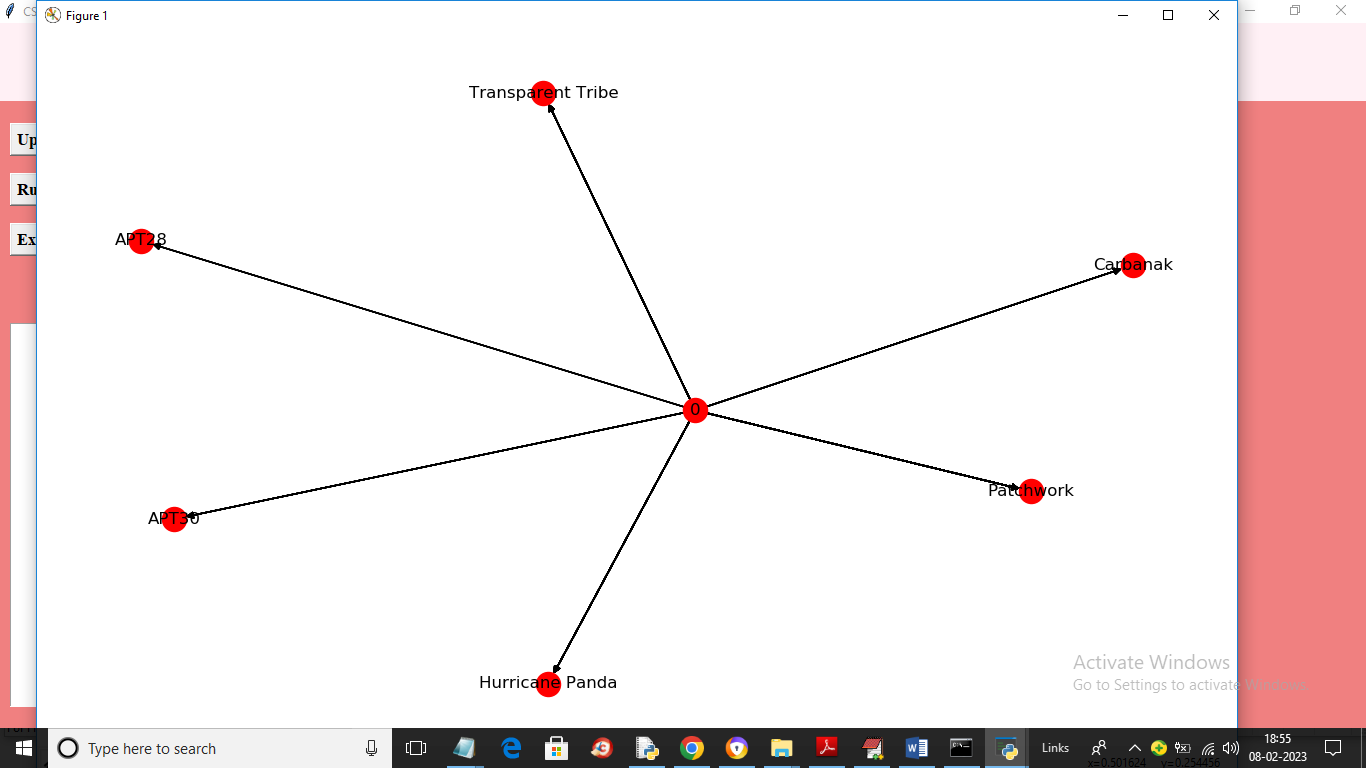
In above screen click on ‘Upload APT Attack Dataset’ button to upload APT dataset and get below output



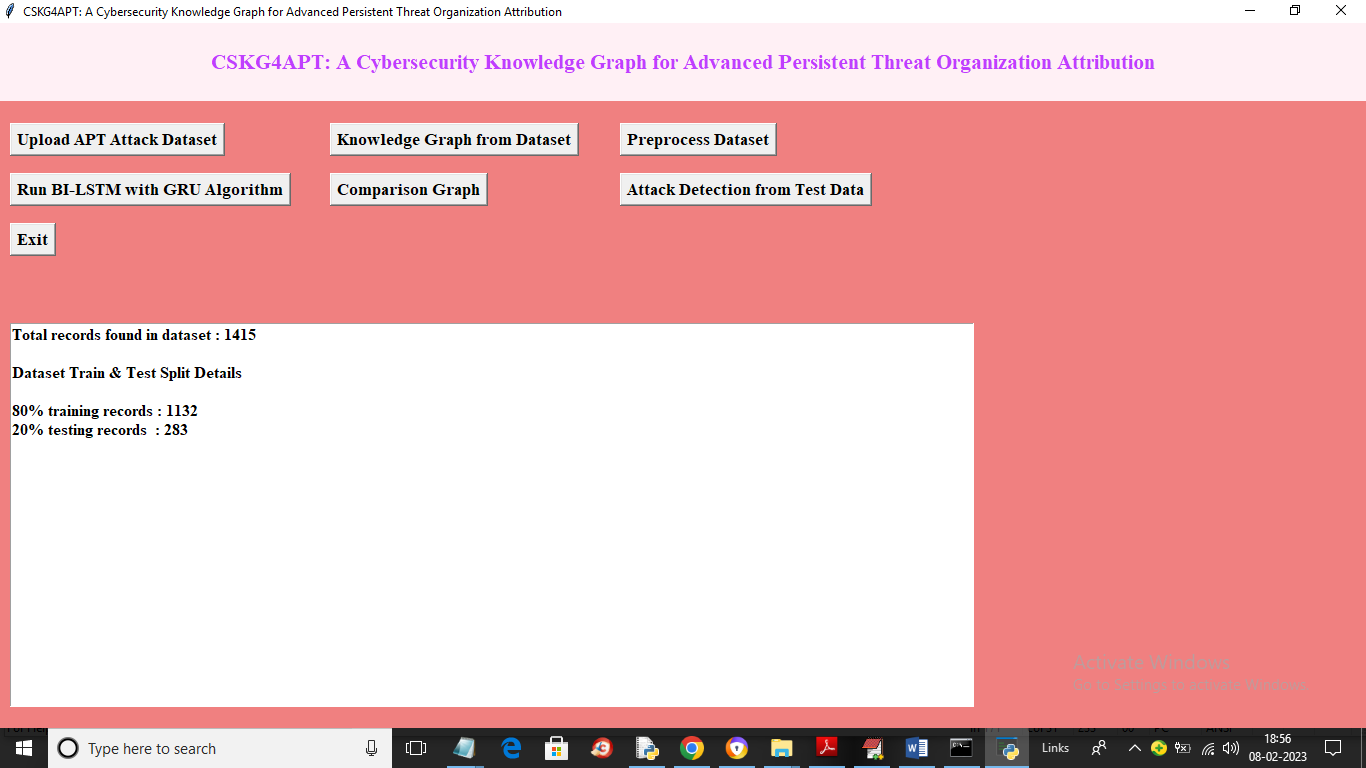
In above screen selecting and uploading APT dataset and then click on ‘Open’ button to load dataset and get below output



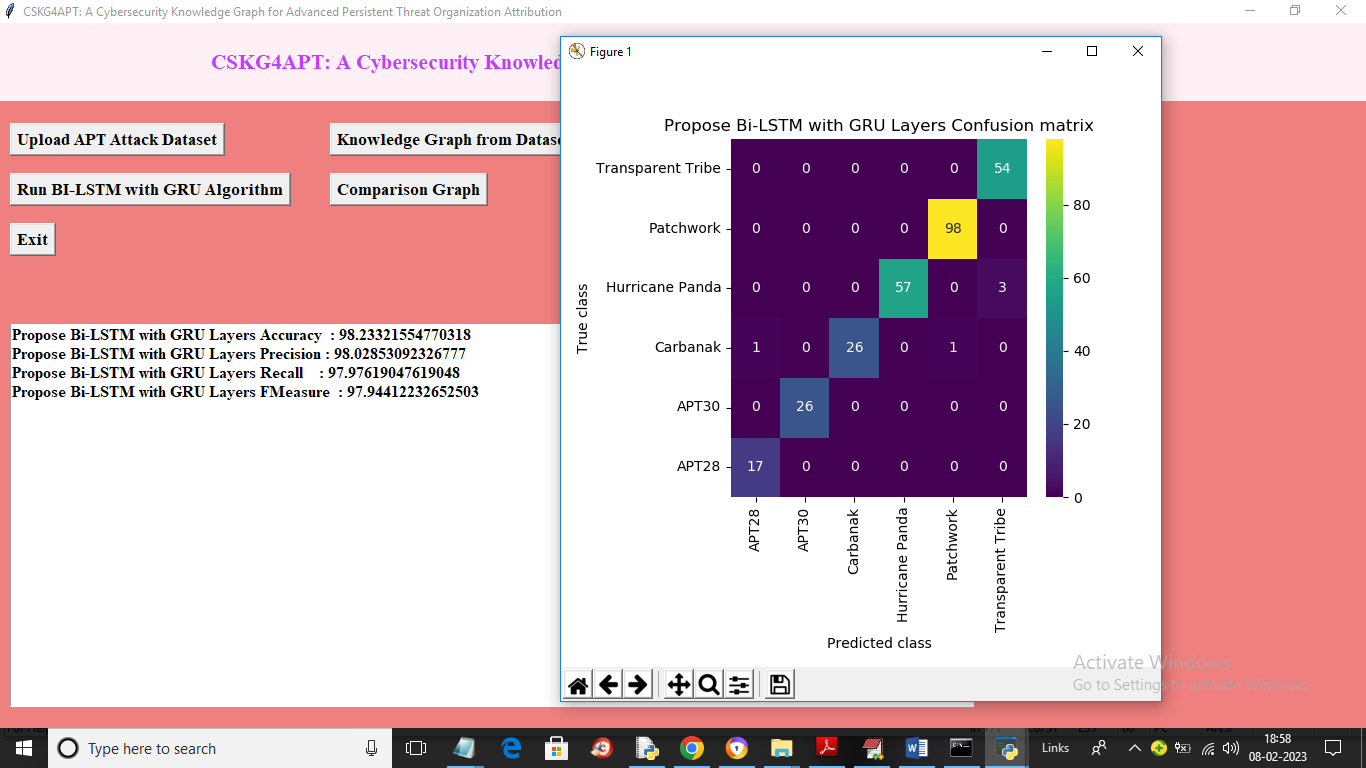
In above screen in text area we can see dataset loaded and in graph we can see x-axis contains APT names and y-axis contains attack count and now close above graph and then click on ‘Knowledge Graph from Dataset’ button to build graph and get below output



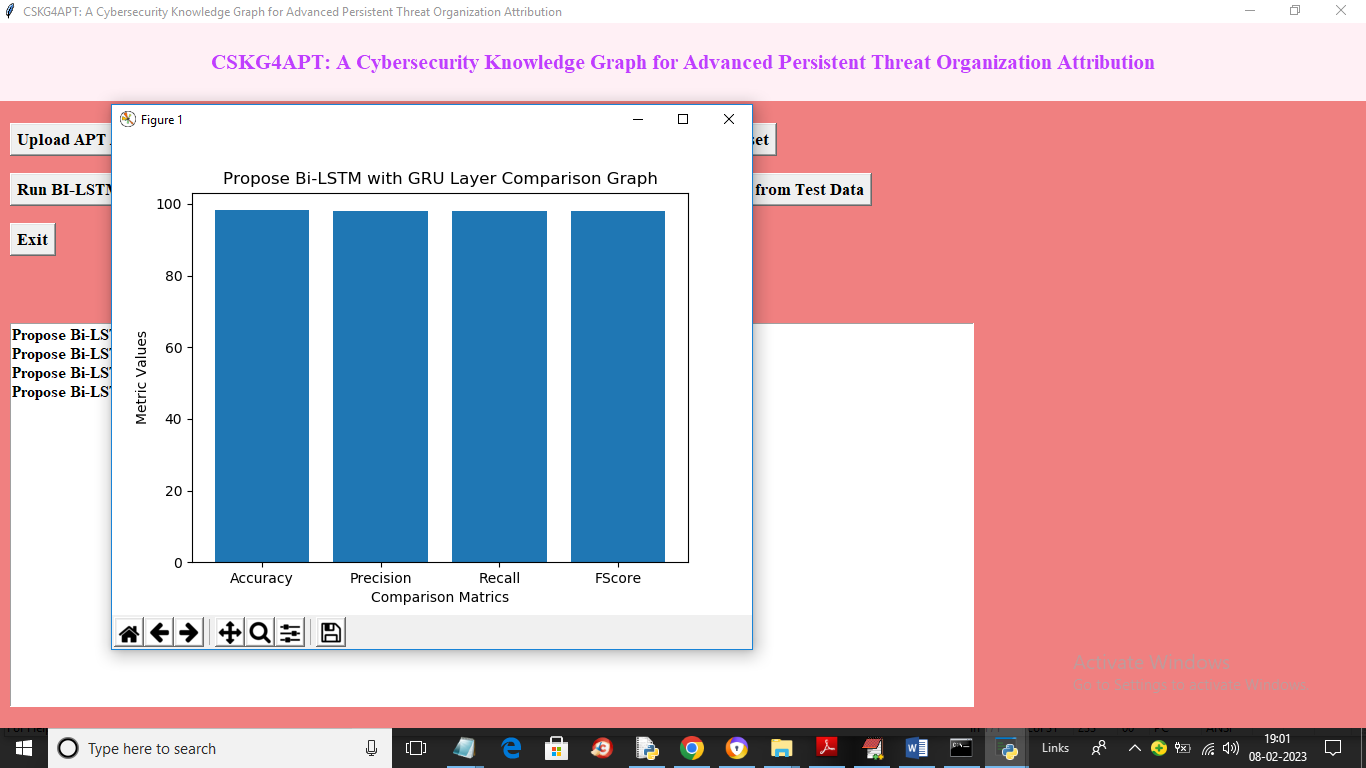
In above screen from dataset we got knowledge graph with various attacks and now close above graph and then click on ‘Preprocess Dataset’ button to process dataset and get below output



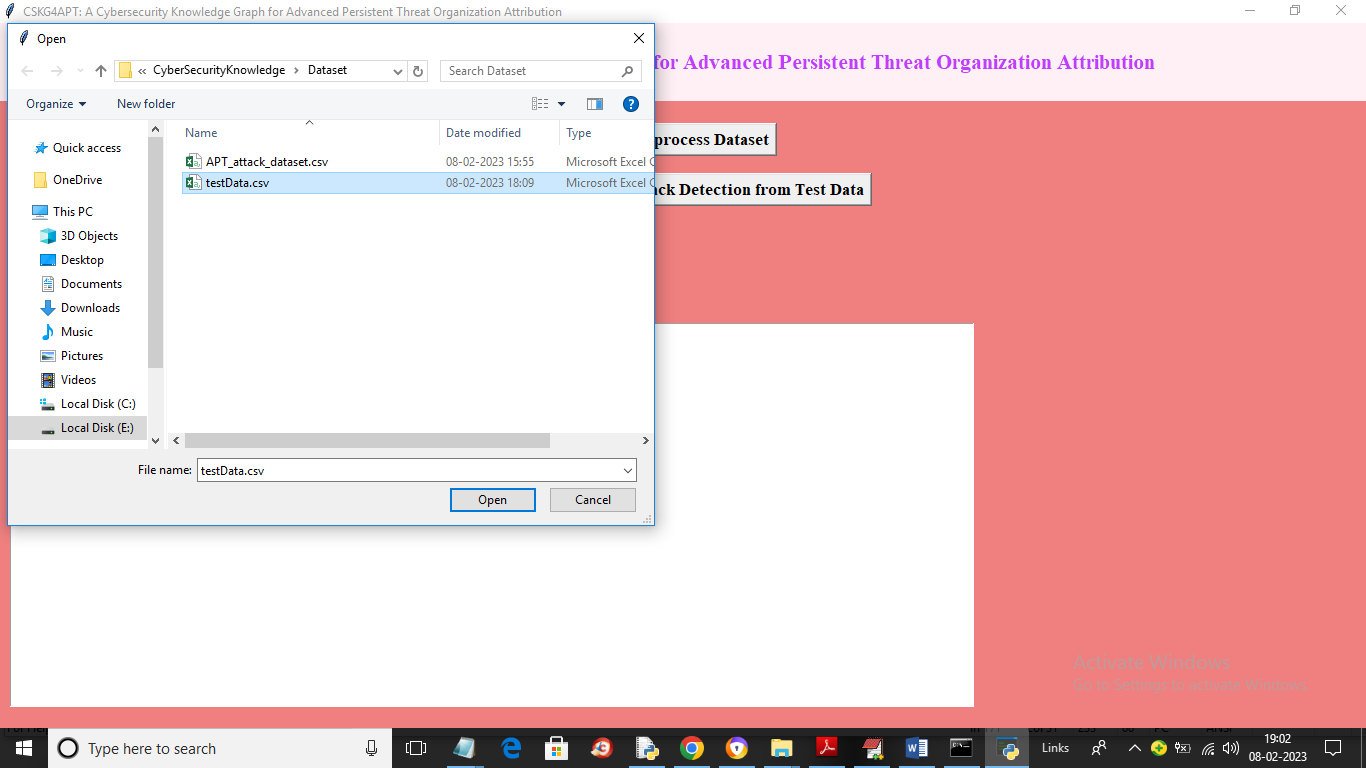
In above screen dataset processing completed and we can see dataset contains 1415 records and then application using 80% (1132 records) dataset for training and 283 (20% records) dataset values for testing and now click on ‘Run BI-LSTM with GRU Algorithm’ button to train deep learning algorithm and get below output



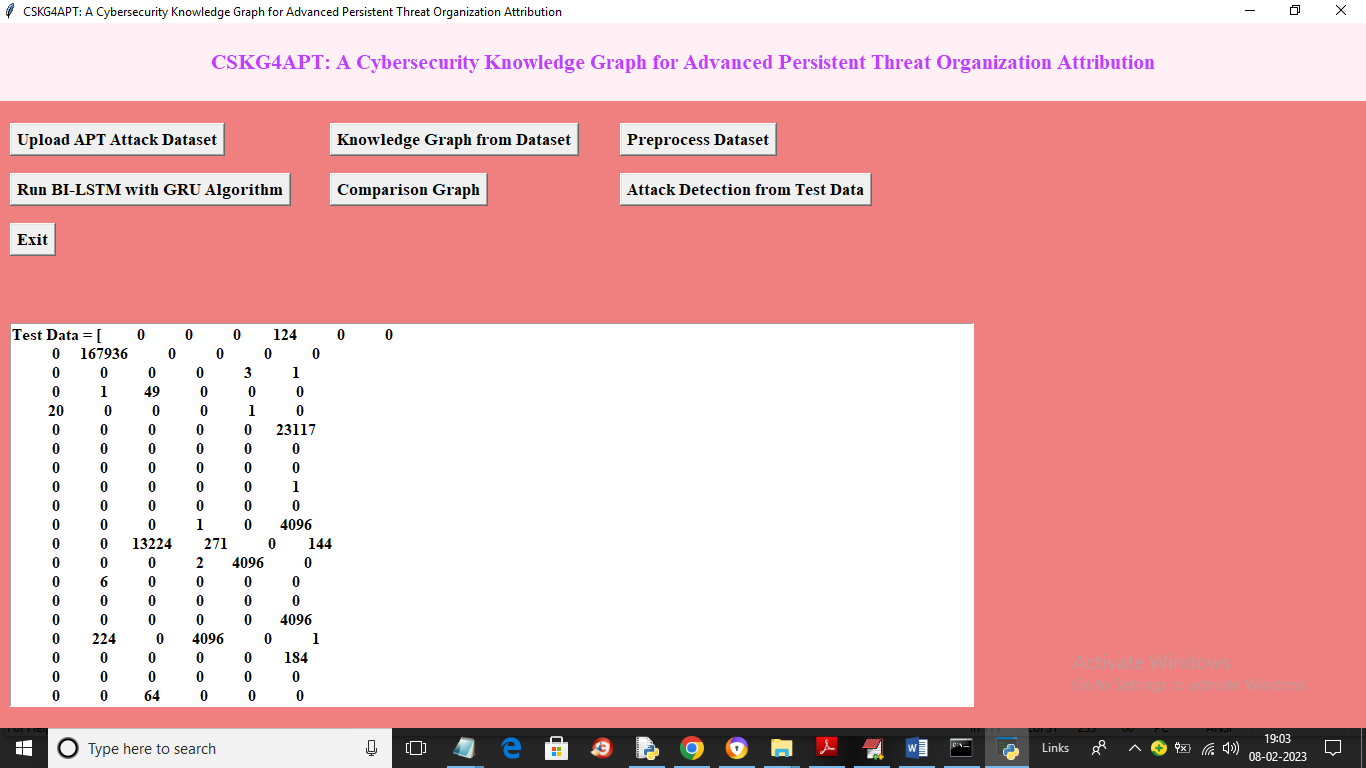
In above screen with deep learning BI-LSTM algorithm we got 98% prediction accuracy and in confusion matrix graph x-axis represents Predicted Threat Labels and y-axis represents True labels and all blue colour boxes contains incorrect prediction count which are very few and all different colour boxes in diagnol represents correct prediction count. So deep learning algorithm can predict APT threat with an accuracy of 98%. Now close above graph and then click on ‘Comparison Graph’ button to get below graph



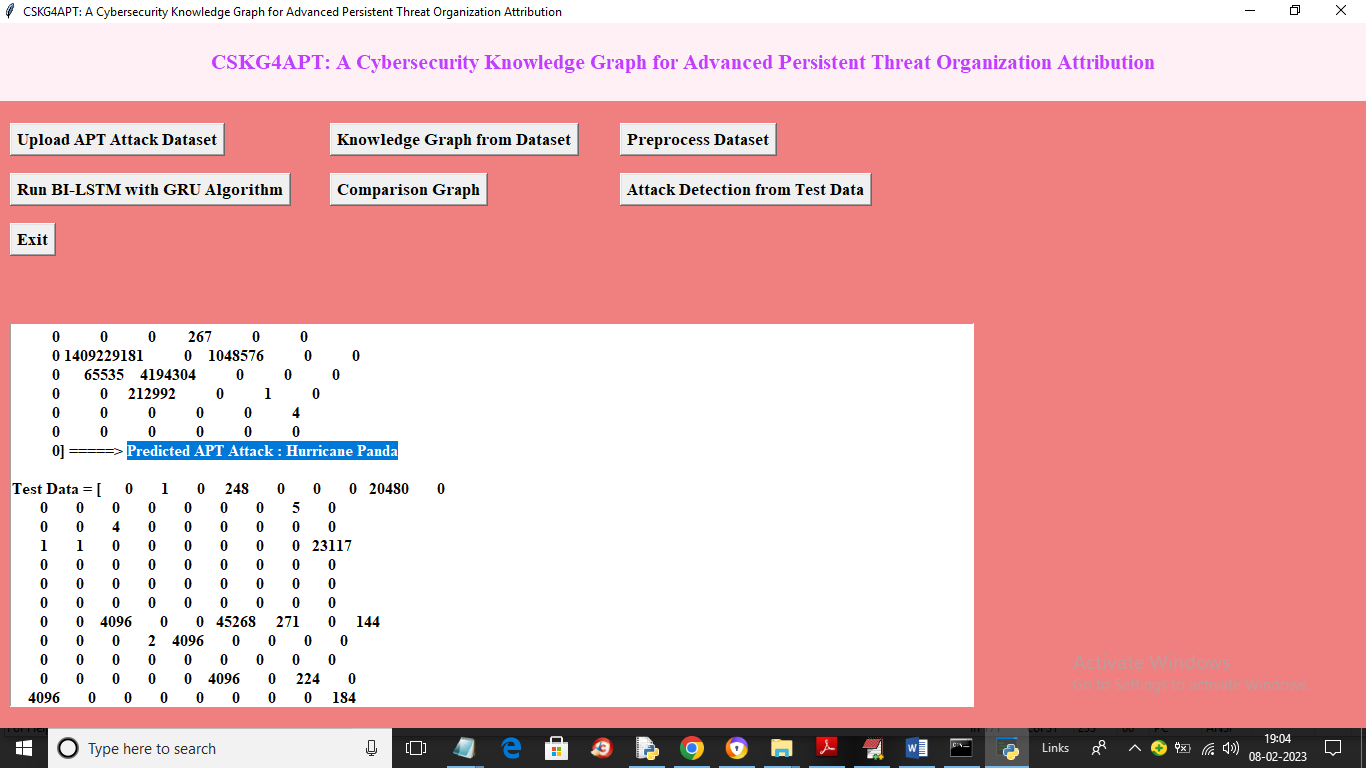
In above graph x-axis represents deep learning BI-LSTM metric names like accuracy and other and y-axis represents values and in above graph we can see all metrics of algorithm is closer to 1. So we can say this algorithm is best in performance and now close above graph and then click on ‘Attack Detection from Test Data’ button to upload test data and get Threat prediction output



In above screen we are selecting and uploading ‘testData.csv’ file and then click on ‘Open’ button to get below output



In above screen in square bracket we can see test data and after arrow symbol =🡺 we can see predicted Threat which is showing in below screen



In above screen in blue colour text we can see predicted APT as ‘Hurricane’ and similarly scroll down above screen to view all threats

