A Cyber Security Knowledge Graph for Advanced Persistent Threat Organization Attribution

Mr.P.Vinay Kumar[1] and Mr.M.Sandhya Vani [2]
[1] [1] Assistant Professor, Department of Information Technology, MREC (A), Hyderabad-500100

Abstract

Facing the dynamic complex cyber environments, internal and external cyber threat intelligence, and the increasing risk of cyber-attack, knowledge graphs show great application potential in the cyber security area because of their capabilities in knowledge aggregation, representation, management, and reasoning. However, while most research has focused on how to develop a complete knowledge graph, it remains unclear how to apply the knowledge graph to solve industrial real challenges in cyber-attack and defense scenarios. In this review, we provide a brief overview of the basic concepts, schema, and construction approaches for the cyber security knowledge graph. To facilitate future research on cyber security knowledge graphs, we also present a curated collection of datasets and open-source libraries on the knowledge construction and information extraction task. In the major part of this article, we conduct a comparative review of the different works that elaborate on the recent progress in the application scenarios of cyber security knowledge graph. Furthermore, a novel comprehensive classification framework is created to describe the connected works from nine primary categories and eighteen subcategories. Finally, we have a thorough outlook on several promising research directions based on the discussion of existing research flaw.
Keywords: Cyber security, high efficiency data transmission.

I. INTRODUCTION

With the development of new information technologies and applications, the scale of the cyber space is gradually expanding from the traditional internet to a variety of areas such as manufacturing, healthcare, agriculture, aviation, business, etc. As a result, cyber space can comprise interactions between industrial physical systems, human social systems, and network information systems and has become an increasingly complex infrastructure for social development.

The opportunities left for attackers are increasing. Due to their combination of cyber as well as many physical assets, the consequence of cyber-attacks become more and more serious. The cyberattack experienced by Colonial Pipeline is an example of how a cyberattack can impact the physical world. The cyberattack shut down a pipeline that supplies 45% of the East Coast's fuel, leading to a $5 million economic loss , fuel delivery disruption, and panic buying across the United States . Given the increasing

number and intensity of attacks and malware, the lack of qualified cyber security personnel is a cause for concern. Since neither the number of available people nor the required skills can be increased overnight, companies must increase the development of technologies for modeling experts' knowledge and experience. The integration of automation, intelligent technology, and attack defense technology has become one of the inevitable trends in the development of cyber security technology. Cyberattacks and defenses against them are conducted in dynamic complex environments, with numerous factors contributing to attack success and mission impacts. The network environments are continually changing, with the applications installed, machines added and removed, etc., which is one of the main obstacles . On the other hand, the information asymmetry between the offensive and defensive sides in cyberspace is becoming more and more obvious as advanced persistent threat (APT) attacks , make the limitations of traditional defense technologies based on expert rules, machine learning, and deep learning have become increasingly apparent. The relatively simple tasks, such as feature extraction , anomaly detection, and data classification , can no longer restore the full picture of attack behavior. Expert knowledge hidden in cybersecurity data is still a very important breakthrough to solve the above problems.

II. LITERATURE SURVEY

Deep learning for anomaly detec by Kumar K, Pande

Anomalies, often referred to as outliers, abnormalities, rare events, or deviants, are data points or patterns in data that do not conform to a notion of normal behavior. Anomaly detection, then, is the task of finding those patterns in data that do not adhere to expected norms, given previous observations. The capability to recognize or detect anomalous behavior can provide highly useful insights across industries. Flagging unusual cases or enacting a planned response when they occur can save businesses time, costs, and customers. Hence, anomaly detection has found diverse applications in a variety of domains, including IT analytics, network intrusion analytics, medical diagnostics, financial fraud protection, manufacturing quality control, marketing and social media analytics, and more.

EXISTING SYSTEM

Cyber security ontology is used to describe cyber security concepts and relationships between concepts in a cyber security field or even a wider range. These concepts and relationships have a common, unambiguous, and unique definition that everyone agrees on in the shared range, which makes humans and machines can communicate with each other . Unified ontologies, such as STUCCO , Unified Cybersecurity Ontology (UCO) , were created in the field of cyber security to incorporate and integrate heterogeneous data and knowledge schemas from various cybersecurity systems, as well as the most commonly used cybersecurity standards for information sharing and exchange. For different specific application scenarios, researchers have developed different ontologies, such as intrusion detection , malware categorization and behavior modeling , cyber threat intelligence (CTI) analysis

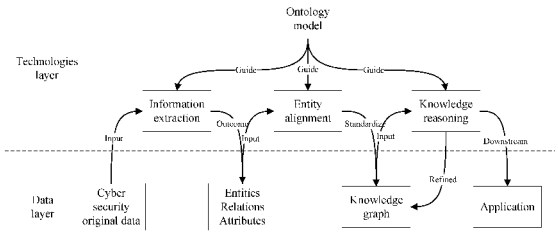
III. PROPOSED SYSTEM

Networks always face 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 thatdescribe 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 layersalgorithm 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.

SYSTEM ARCHITECTURE

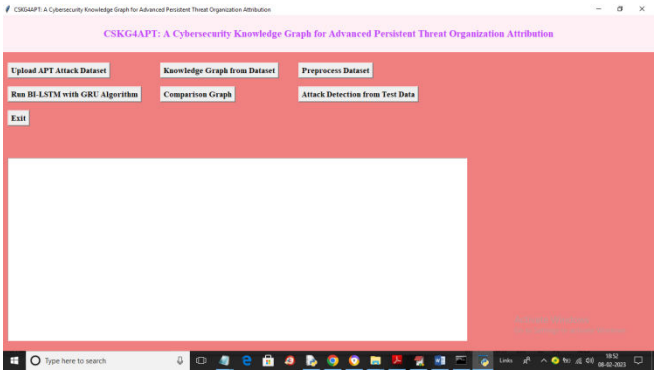


IV. RESULTS DISCUSSION

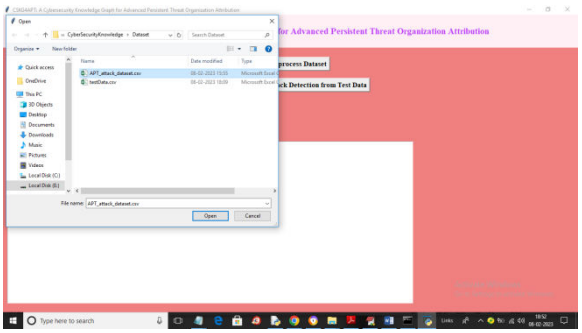
To implement this project we have designed following modules

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

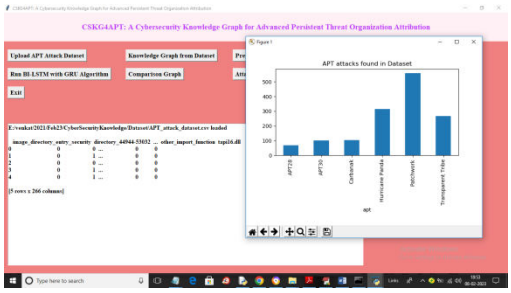
To run project double click on 'run.bat' file to get below screen



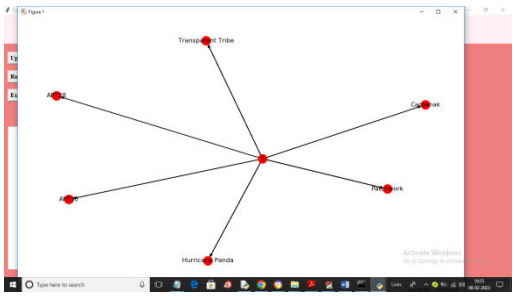
In above screen click on 'Upload APT Attack Dataset' button to upload APT dataset and get below output



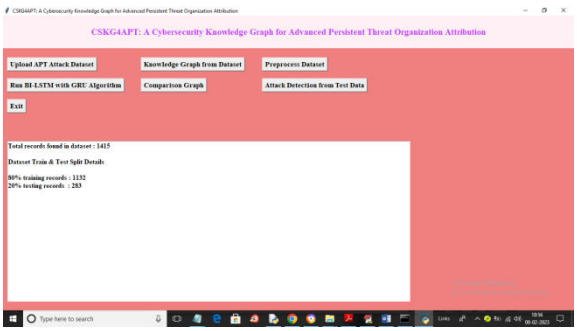
In above screen selecting and uploading APT dataset and then click on 'Open' button to load dataset and get below output



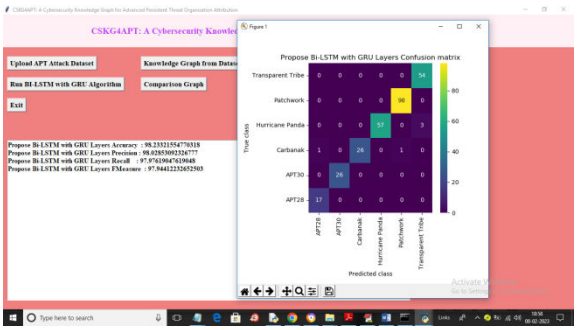
In above screen in text area we can see dataset loaded and in graph we can see x-axis contains APT names and y-axis contains attack count and now close above graph and then click on 'Knowledge Graph from Dataset' button to build graph and get below output



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

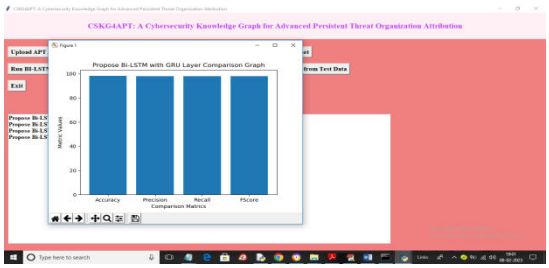


In above screen dataset processing completed and we can see dataset contains 1415 records and then application using 80% (1132 records) dataset for training and 283 (20% records) dataset values for testing and now click on 'Run BI-LSTM with GRU Algorithm' button to train deep learning algorithm and get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

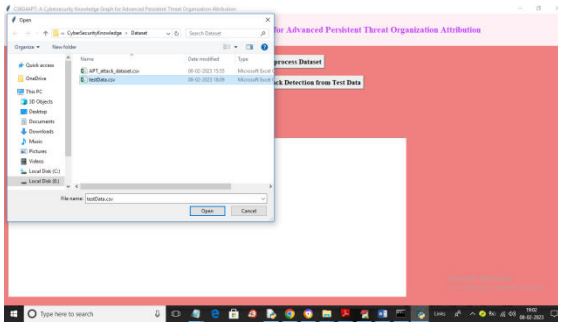


In above screen with deep learning BI-LSTM algorithm we got 98% prediction accuracy and in confusion matrix graph x-axis represents Predicted Threat Labels and y-axis represents True labels and all blue colour boxes contains incorrect prediction count which are very few and all different colour boxes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

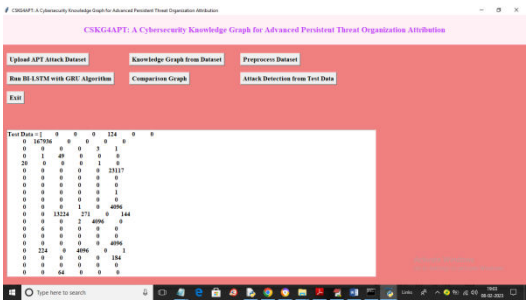
98%. Now close above graph and then click on 'Comparison Graph' button to get below graph



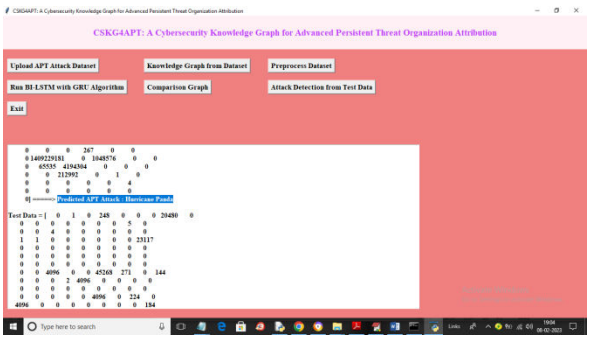
In above graph x-axis represents deep learning BI-LSTM metric names like accuracy and other and y-axis represents values and in above graph we can see all metrics of algorithm is closer to 1. So we can say this algorithm is best in performance and now close above graph and then click on 'Attack Detection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



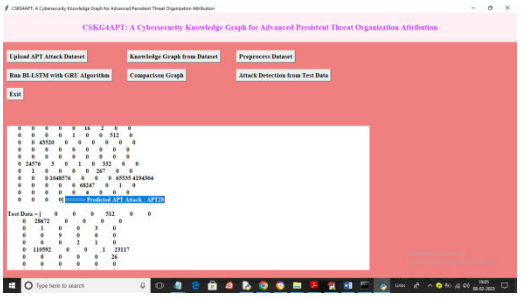
In above screen we are selecting and uploading 'testData.csv' file and then click on 'Open' button to get below output

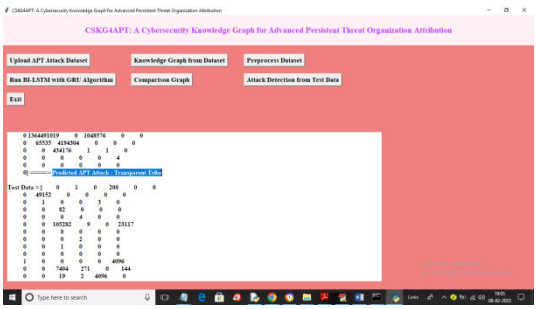


In above screen in square bracket we can see test data and after arrow symbol =➔ we can see predicted Threat which is showing in below screen



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





V. CONCLUSION

In this review, we have provided a critical overview of the various works on the application scenarios of the cyber security knowledge graph. To begin, this paper introduces a brief overview of the background, concepts, and construction technologies of the cyber security knowledge graph. Then, several open-source datasets that are available for building cyber security knowledge graph and the information extraction task, and their drawbacks are illustrated. In the fourth part of this paper, we carried out a comparative study of the different works that elaborate on the recent progress in the application scenarios of CSKG. A novel comprehensive classification framework was developed for describing the related works from nine main aspects and eighteen subclasses. Finally, based on the discussion of shortcomings of existing research, future research directions have been prospected. Security managers can use KG to intuitively understand security intelligence, network situation, entity relationships, and then discover the attributes of security entities, which could serve as a foundation for understanding cyber security knowledge, analyzing cyber security data, and discovering attack patterns and abnormal characteristics related to cyber-attacks. It is hoped that this research will contribute to a deeper understanding of how to apply cyber security knowledge graph

REFERENCE

[1] Osborne, Charlie. Colonial Pipeline paid close to $5 million in ransomware blackmail payment, Zero Day, 13 May 2021, https://www.calvin.edu/library/knightcite/index.php.

[2] Auer, Markus. Lack of experts in cyber security, ThreatQuotient, Inc., 14 July 2020, https://www.threatq.com/lack-of-experts-in-cyber-security/.

[3] Kumar K, Pande B P. Applications of Machine Learning Techniques in the Realm of Cybersecurity[J]. Cyber Security and Digital Forensics, 2022: 295-315.

[4] Liebetrau T. Cyber conflict short of war: a European strategic vacuum[J]. European Security, 2022: 1-20.

[5] Cole E. Advanced persistent threat: understanding the danger and how to protect your organization[M]. Newnes, 2012.

[6] Sriavstava R, Singh P, Chhabra H. Review on Cyber Security Intrusion Detection: Using Methods of Machine Learning and Data Mining[M]//Internet of Things and Big Data Applications. Springer, Cham, 2020: 121-132.

[7] Pang G, Shen C, Cao L, et al. Deep learning for anomaly detection: A review[J]. ACM Computing Surveys (CSUR), 2021, 54(2): 1-38.

[8] Perdisci R, D Ariu, Fogla P, et al. McPAD: A multiple classifier system for accurate payload- based anomaly detection[J]. Computer Networks the International Journal of Computer & Telecommunications Networking, 2009, 53(6):864-881.

[9] Llorens, Audrey. 5 Best Practices to Get More from Threat Intelligence, ThreatQuotient, Inc., 26 Oct. 2021, https://www.threatq.com/5-best-practices-more-threat-intelligence/.

[10] Xue R, Tang P, Fang S. Prediction of Computer Network Security Situation Based on Association Rules Mining[J]. Wireless Communications and Mobile Computing, 2022.