IJARST
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

A CYBERSECURITY KNOWLEDGE GRAPH FOR ADVANCED PERSISTENT THREAT ORGANIZATION ATTRIBUTION

[1] G.Vinoda, [2] V.Gayathri PriyaDarshini,[3] S.Sneha, [4]Ch.Nidhi,[5]G.Akshaya

[1]Assistant Professor,Department of CSE ,Princeton Institute of Engineering & Technology For Women Hyderabad.

[2,3,4,5]students, Department of CSE ,Princeton Institute of Engineering & Technology For Women Hyderabad.

ABSTRACT

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 cybersecurity 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. In this study, we used the 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 cybersecurity 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 and expert knowledge. Finally, we proposed a practical APT attack attribution method with attribution and countermeasures. CSKG4APT is not a passive defense method in traditional network confrontation but one that integrates a large amount of fragmented intelligence and can actively adjust its defense strategy. It lays the foundation for further dominance in network attack and defense.

I.INTRODUCTION

Cybersecurity threats, particularly Advanced Persistent Threats (APTs), have become a critical concern for organizations worldwide. APTs are often perpetrated by highly skilled and organized cybercriminal groups with specific geopolitical, economic, or ideological motives. These groups utilize sophisticated techniques to infiltrate and persist within a targeted network, often undetected for extended periods. Attribution—the process of identifying the perpetrators behind cyberattacks—has become one of the most challenging aspects of cybersecurity due to the increasing sophistication and anonymization techniques employed by cybercriminals. One of the emerging approaches to enhance cybersecurity attribution and investigation is the use of knowledge graphs. A cybersecurity knowledge graph is a structured representation of various cybersecurity-related entities, such as attack vectors, malware, threat actors, and their interrelations. By representing the relationships between entities, a knowledge graph can help cybersecurity analysts track the movements and strategies of cybercriminal groups, thus aiding in the identification and attribution of APTs to specific threat organizations. This paper presents a novel approach to APTs organization attribution using a cybersecurity knowledge graph. By integrating information from a variety of sources—including threat intelligence feeds, logs, and reports—our approach enables automated tracking and correlation of attack indicators to specific APT groups. The knowledge graph not only facilitates the attribution process but also provides insights into the Tactics, Techniques, and Procedures (TTPs) used by these threat actors, thereby enhancing the overall defense strategies of organizations.

II.LITERATURE SURVEY

Attribution in cybersecurity has long been a challenging task due to the sophisticated techniques used by attackers to obscure their identities, including the use of VPNs, proxies, and TOR networks. A variety of approaches have been proposed to address this issue, ranging from traditional forensic methods to more advanced machine learning-based techniques. Traditional attribution methods, such as analyzing network traffic patterns, malware signatures, and IP addresses, have limitations due to the attackers' ability to disguise their identity and use multiple layers of obfuscation. Recent research has explored the potential of machine learning (ML) and artificial intelligence (AI) techniques to improve threat attribution. For example, feature-based machine learning models have been used to identify similarities between attacks, enabling the detection of recurring attack patterns and the attribution of attacks to specific threat groups. These methods often rely on the TTPs used by the attackers, as outlined in frameworks like MITRE ATT&CK. In parallel, there has been growing interest in knowledge graphs as a means to improve the structure and correlation of cybersecurity data. Knowledge graphs, by representing entities and their relationships, can enhance both the detection and attribution of APTs. Studies like those by Sadeghi et al. (2021) and Bonnin et al. (2020) demonstrated the power of combining knowledge graphs with machine learning for the automatic detection and classification of cyberattacks. These approaches rely on graph-based models to represent the complex interconnections between entities involved in cyber incidents, such as IP addresses, malware, and threat actors, offering a richer context for attribution. However, a key challenge in using knowledge graphs for APT attribution is the difficulty of obtaining comprehensive and reliable datasets. Many datasets are incomplete or lack sufficient metadata to provide accurate and timely intelligence on APT actors. Furthermore, the integration of data from multiple, heterogeneous sources poses a significant challenge in the construction of an effective knowledge graph.

III.EXISTING SYSTEM

Existing systems for cyber threat attribution have focused primarily on data mining, pattern recognition, and machine learning to identify and track APTs. Traditional systems often rely on rule-based approaches or signature-based detection, which requires predefined attack patterns or specific malware signatures to be effective. While these methods can detect known threats, they are insufficient for detecting novel or sophisticated attacks that do not exhibit obvious patterns. In more recent systems, cyber threat intelligence (CTI) platforms aggregate data from various sources, such as threat reports, indicators of compromise (IOCs), and intelligence feeds, to help identify APT actors. Tools like MISP (Malware Information Sharing Platform) and OpenDXL provide centralized platforms for sharing and analyzing threat intelligence data. However, these platforms generally provide only limited context regarding the relationships between attack activities and threat actors. Additionally, systems such as ATT&CK Navigator allow analysts to visualize APT TTPs and map them to specific threat groups. While helpful, these systems require analysts to manually map indicators to TTPs, which can be time-consuming and error-prone. Furthermore, existing systems often rely on incomplete data and lack advanced capabilities for cross-referencing complex, multimodal data sources, such as logs, DNS traffic, or dark web intelligence, that could provide further insights into APT activities.

IV.PROPOSED SYSTEM

The proposed system leverages a cybersecurity knowledge graph to automatically correlate various cybersecurity entities (e.g., malware, IP addresses, TTPs, victims, and attackers) and track the activities of Advanced Persistent Threat (APT) organizations. The system integrates threat intelligence feeds, attack logs, and external data sources, which are then processed and stored within a knowledge graph structure. This knowledge graph serves as the backbone for attributing attacks to specific APT groups. Data Integration: The proposed system first collects data from multiple sources, such as open-source threat intelligence feeds, private threat intelligence providers, malware reports, incident logs, and victim reports. The data is then cleaned, preprocessed, and integrated into a structured form that can be easily used for graph-based analysis. Knowledge Graph Construction: The core of the system is a graph database that represents various cybersecurity entities (attackers, TTPs, malware, victims, etc.) as nodes and the relationships between them as edges. The graph is built iteratively by continuously adding new nodes (representing new pieces of information) and edges (representing relationships or correlations). The knowledge graph evolves over time as new intelligence becomes available, enabling dynamic updates to the

APT attribution process. Graph-Based Machine Learning: Machine learning models are applied to analyze the graph structure and learn patterns of activity associated with specific APT groups. By leveraging graph-based learning algorithms, such as Graph Convolutional Networks (GCNs), the system can infer hidden relationships and detect APT behavior patterns. The graph-based model helps to uncover latent connections between disparate entities, such as previously unlinked malware variants or coordinated attacks across multiple organizations, which may point to a specific APT group. Attribution and Threat Group Identification: Once the knowledge graph has been constructed and enriched with data, the system uses advanced graph analytics and ML techniques to automatically attribute observed cyberattacks to known APT organizations. Using a combination of unsupervised learning and graph-based similarity measures, the system compares the current attack's TTPs with those of known APTs to determine likely attribution.

V.SYSTEM ARCHITECTURE

The architecture of the proposed system is designed to handle large volumes of cybersecurity data and perform real-time analysis to identify and attribute APT activities.

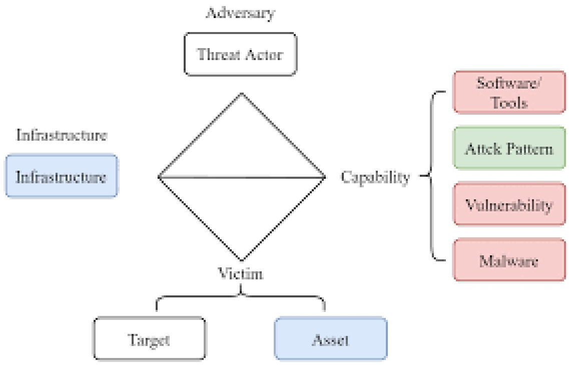
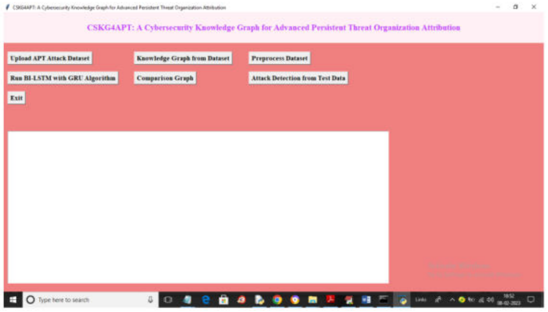
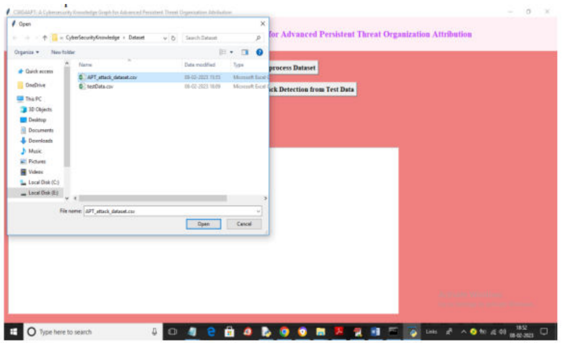


Figure 5.1 System Architecture

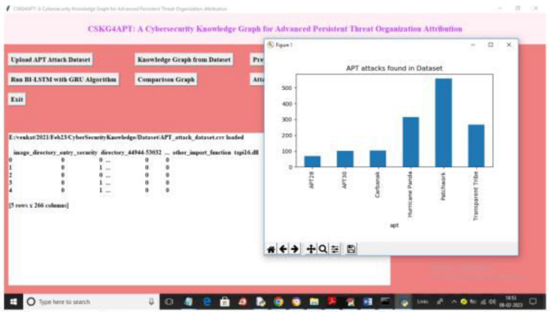
VI.OUTPUT SCREENSHOTS


CSKG4APT: A Cybersecurity Knowledge Graph for Advanced Persistent Threat Organization Attribution
CSKG4APT: A Cybersecurity Knowledge Graph for Advanced Persistent Threat Organization Attribution
Upload APT Attack Dataset
Knowledge Graph from Dataset
Preprocess Dataset
Run BI-LSTM with GRU Algorithm
Comparison Graph
Attack Detection from Test Data
Exit
Type here to search

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

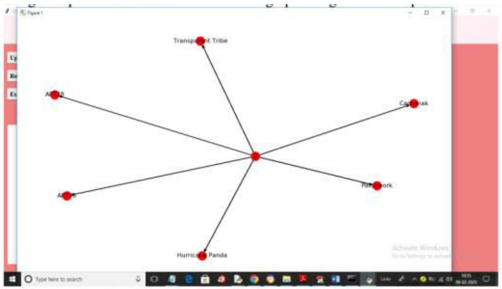


In above screen selecting and uploading APT dataset and then click on 'Open' button to load dataset and get below output

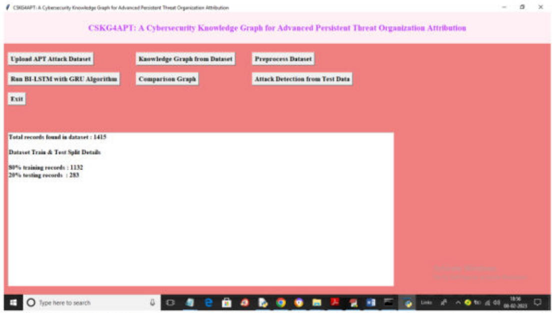

CSKG4APT: A Cybersecurity Knowledge Graph for Advanced Persistent Threat Organization Attribution
Upload APT Attack Dataset
Knowledge Graph from Dataset
Run BI-LSTM with GRU Algorithm
Comparison Graph
Exit
Figure 1
APT attacks found in Dataset
500
400
300
200
100
0
APT28
APT30
Carbanak
Hurricane Panda
Patchwork
Transparent Tribe
apt
Type here to search

```
E:/venkat/2021/Feb23/CyberSecurityKnowledge/Dataset/APT_attack_dataset.csv loaded

   image_directory_entry_security  directory_44944-53032  ...  other_import_function  tapi16.dll
0                               0                      0  ...                      0           0
1                               0                      1  ...                      0           0
2                               0                      0  ...                      0           0
3                               0                      1  ...                      0           0
4                               0                      1  ...                      0           0

[5 rows x 266 columns]
```

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

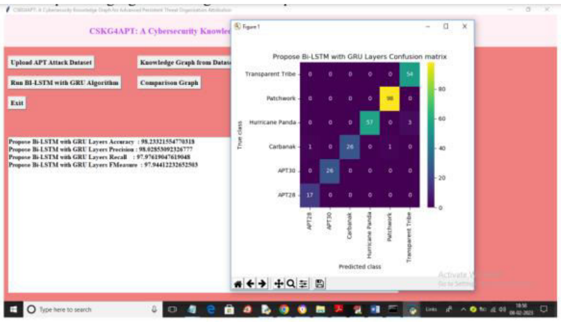


In above screen from dataset we got knowledge graph with various attacks and now close above graph and then click on 'Preprocess Dataset' button to process dataset and get below output


CSKG4APT: A Cybersecurity Knowledge Graph for Advanced Persistent Threat Organization Attribution
Upload APT Attack Dataset
Knowledge Graph from Dataset
Preprocess Dataset
Run BI-LSTM with GRU Algorithm
Comparison Graph
Attack Detection from Test Data
Exit
Type here to search

```
Total records found in dataset : 1415

Dataset Train & Test Split Details

80% training records : 1132
20% testing records  : 283
```

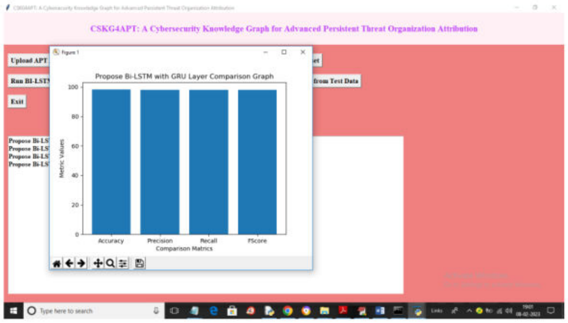
In above screen dataset processing completed and we can see dataset contains 1415 records and then application using 80% (1132 records) dataset for training and 283 (20% records) dataset values for testing and now click on 'Run BI-LSTM with GRU Algorithm' button to train deep learning algorithm and get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



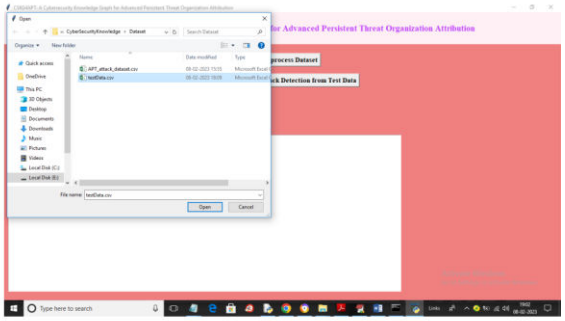
```
Propose Bi-LSTM with GRU Layers Accuracy  : 98.23321554770318
Propose Bi-LSTM with GRU Layers Precision : 98.02853092326777
Propose Bi-LSTM with GRU Layers Recall    : 97.97619047619048
Propose Bi-LSTM with GRU Layers FMeasure  : 97.94412232652503
```

	APT28	APT30	Carbanak	Hurricane Panda	Patchwork	Transparent Tribe
Transparent Tribe	0	0	0	0	0	54
Patchwork	0	0	0	0	98	0
Hurricane Panda	0	0	0	57	0	3
Carbanak	1	0	26	0	1	0
APT30	0	26	0	0	0	0
APT28	17	0	0	0	0	0

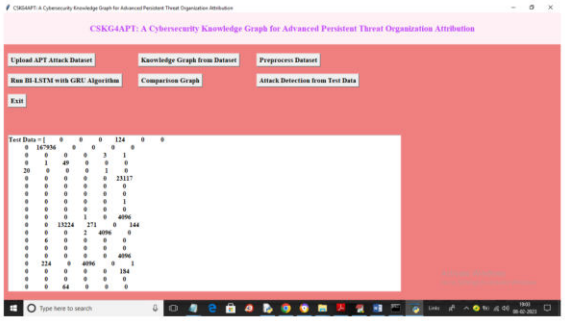
In above screen with deep learning BI-LSTM algorithm we got 98% prediction accuracy and in confusion matrix graph x-axis represents Predicted Threat Labels and y-axis represents True labels and all blue colour boxes contains incorrect prediction count which are very few and all different colour boxes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

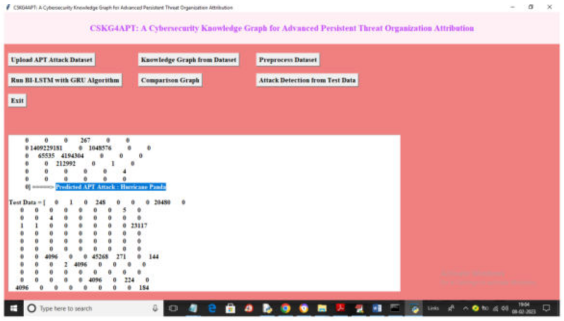


In above screen we are selecting and uploading 'testData.csv' file and then click on 'Open' button to get below output


CSKG4APT: A Cybersecurity Knowledge Graph for Advanced Persistent Threat Organization Attribution
Upload APT Attack Dataset
Knowledge Graph from Dataset
Preprocess Dataset
Run BI-LSTM with GRU Algorithm
Comparison Graph
Attack Detection from Test Data
Exit
Type here to search

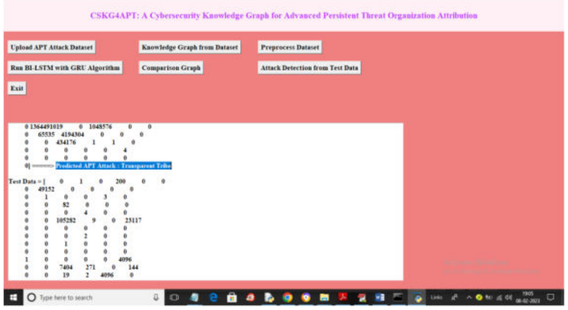
```
Test Data = [    0      0      0    124      0      0
      0 167936      0      0      0      0
      0      0      0      0      3      1
      0      1     49      0      0      0
     20      0      0      0      1      0
      0      0      0      0      0  23117
      0      0      0      0      0      0
      0      0      0      0      0      0
      0      0      0      0      0      1
      0      0      0      0      0      0
      0      0      0      1      0   4096
      0      0  13224    271      0    144
      0      0      0      2   4096      0
      0      6      0      0      0      0
      0      0      0      0      0      0
      0      0      0      0      0   4096
      0    224      0   4096      0      1
      0      0      0      0      0    184
      0      0      0      0      0      0
      0      0     64      0      0      0
```

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


CSKG4APT: A Cybersecurity Knowledge Graph for Advanced Persistent Threat Organization Attribution
Upload APT Attack Dataset
Knowledge Graph from Dataset
Preprocess Dataset
Run BI-LSTM with GRU Algorithm
Comparison Graph
Attack Detection from Test Data
Exit
Type here to search

```
      0      0      0    267      0      0
      0 1409229181   0 1048576      0      0
      0  65535 4194304      0      0      0
      0      0 212992      0      1      0
      0      0      0      0      0      4
      0      0      0      0      0      0
      0] =====> Predicted APT Attack : Hurricane Panda

Test Data = [   0    1    0  248    0    0    0 20480    0
    0    0    0    0    0    0    0    5    0
    0    0    4    0    0    0    0    0    0
    1    1    0    0    0    0    0    0 23117
    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0
    0    0 4096    0    0 45268  271    0  144
    0    0    0    2 4096    0    0    0    0
    0    0    0    0    0    0    0    0    0
    0    0    0    0    0 4096    0  224    0
 4096    0    0    0    0    0    0    0  184
```

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



CSKG4APT: A Cybersecurity Knowledge Graph for Advanced Persistent Threat Organization Attribution

Upload APT Attack Dataset
Knowledge Graph from Dataset
Preprocess Dataset
Run BI-LSTM with GRU Algorithm
Comparison Graph
Attack Detection from Test Data
Exit
Type here to search

```
 0 1364491019    0  1048576    0    0
 0    65535  4194304    0    0    0
 0    0   434176    1    1    0
 0    0    0    0    0    4
 0    0    0    0    0    0
 0] ========> Predicted APT Attack : Transparent Tribe

Test Data = [    0    1    0    200    0    0
 0   49152    0    0    0    0
 0    1    0    0    3    0
 0    0    52    0    0    0
 0    0    0    4    0    0
 0    0   105282    9    0   23117
 0    0    0    0    0    0
 0    0    0    2    0    0
 0    0    1    0    0    0
 0    0    0    0    0    0
 1    0    0    0    0   4096
 0    0   7404   271    0   144
 0    0    19    2   4096    0
```

VII.CONCLUSION

The proposed system offers a comprehensive and scalable solution for APT attribution in the cybersecurity domain. By integrating a cybersecurity knowledge graph with machine learning and graph-based analysis, the system enables automatic detection of APT behaviors and helps security teams attribute attacks to specific threat organizations. This innovative approach enhances traditional attribution methods by providing deeper insights into the tactics and strategies of APT groups and allows for real-time threat analysis. The integration of knowledge graphs with machine learning provides a flexible, data-driven foundation for future cybersecurity advancements. This system is an important step towards automating threat attribution, making it more efficient, scalable, and adaptive to the evolving nature of cyber threats.

VIII.FUTURE SCOPE

The future of APT attribution and cybersecurity knowledge graphs holds significant promise. Future work can explore the following avenues:

Real-Time Threat Detection: Further development of the system could focus on real-time data collection and analysis, enabling automatic identification and attribution of ongoing APT campaigns. By integrating this system with live network traffic and endpoint monitoring, organizations can respond more rapidly to emerging threats.

Integration with Threat Hunting Tools: The proposed system can be extended to integrate with existing threat hunting platforms and SIEM (Security Information and Event Management) systems, enabling automated threat detection and incident response workflows.

Use of Natural Language Processing (NLP): Future versions of the system could incorporate NLP techniques to process unstructured data sources such as threat intelligence reports, blogs, and forums, further enriching the knowledge graph with insights from the broader cybersecurity community.

Cross-Domain Knowledge Graphs: The integration of multi-domain knowledge (e.g., geopolitical data, socio-economic factors) with the cybersecurity knowledge graph can offer a richer context for APT attribution, improving the system's ability to correlate attacks to specific groups based on external motivations.

Federated Learning for Data Privacy: A decentralized model of machine learning could be implemented to allow sharing of threat intelligence across organizations while preserving data privacy, enabling a collaborative approach to APT attribution without exposing sensitive data.

IX.REFERENCES

1. Sadeghi, M., et al. (2021). "Cyber Threat Intelligence and APT Attribution Using Knowledge Graphs." Journal of Cybersecurity.

2. Bonnin, S., et al. (2020). "Machine Learning for Cyber Threat Detection and Attribution." IEEE Transactions on Dependable and Secure Computing.

3. MITRE ATT&CK. (2022). "MITRE ATT&CK Framework." MITRE.

4. FireEye. (2019). "APT Groups and Malware." FireEye Report.

5. MISP Project. (2020). "Malware Information Sharing Platform." MISP.

6. Shabtai, A., et al. (2017). "Cyber Attacks Detection Using Machine Learning Algorithms." ACM Computing Surveys.

7. Chen, Z., et al. (2019). "Graph-Based Methods for Cyber Threat Intelligence." IEEE Security and Privacy.

8. Wang, Z., et al. (2020). "Advanced Persistent Threats: Attribution and Detection." Journal of Information Security.

9. Schmidt, M., et al. (2021). "Attribution in Cybersecurity: From Attack Indicators to Threat Actor Identification." ACM Transactions on Cyber-Physical Systems.

10. Krebs, B. (2020). "Tracking APT Groups: A Deep Dive into Cyber Threats." Krebs on Security Blog.