INTELLIGENT NETWORK TRAFFIC ANALYSIS: LEVERAGING MACHINE LEARNING FOR ENHANCED CYBERSECURITY

***Seema Kaloria*[1], *Rakesh Kumar Saxena*[2], *Deshraj Bairwa*[3]**
[1,2,3]*Faculty of Computer Science & Engineering, Poornima University, Jaipur, Raj. India*
**seema.kaloria@poornima.edu.in*

Abstract

Network traffic analysis is one of the necessary processes through which there can be monitoring, securing, and optimizing network operations. It is hard for this approach to notice some unknown or evolving kinds of security threats-mostly those involving zero-day attacks or Distributed Denial-of-Service-because cyber threats are developing so fast that nowadays even the most state-of-the-art signature-based methods of traffic analysis are not considered sufficient. This paper presents the design and development of the NTA tool based on ML techniques that will solve the problem where conventional approaches fail to have any integration with unsupervised learning techniques toward the detection of anomalies and classification of network traffic.

Utilizing K-means clustering, the tool classifies the network data in an unsupervised manner into normal and malicious traffic clusters. This had been trained on a data set with over 400,000 rows of network traffic data features through IP address and port numbers and timestamps. An extensive pipeline of data preprocessing had been set to clean and normalize the data thus providing quality input for the applied machine learning models. Real-time network anomalies and potential security breaches will be detected by the tool, which empowers network administrators to take necessary steps for improving security protocols.

The results show the usability of the developed tool and effectiveness toward the detection of known threats and also unknown ones like DDoS attacks. The application of the K-means clustering algorithm in the constructed system showed efficient grouping of similar network patterns to acquire anomalies in high precision. Also, detection accuracy in this proposed tool is in comparison with traditional systems of NTA and is proved to be more reliable than the proposed methods and relatively low false-positive rates are associated. Scalability tests proved robust performance while dealing with large-scale networks.

Along with this, the paper discusses many of the challenges associated with detecting encrypted traffic, which remain a major hurdle in network security. The future work will study how deep learning techniques and advanced clustering methods can be incorporated to enhance the system's ability to handle encrypted and highly dynamic network traffic environments.

This work makes a great contribution to the area of network security by revealing that machine learning techniques can be practically applied in real-time NTA in order to catch sophisticated cyber threats and then mitigate them.

Keywords: Network Traffic Analysis, ML, Intrusion Detection, K-means Clustering, Network Security, DDoS Detection

1 Introduction

NTA has gained importance with higher internet access usage and increased complexity security threats. Efficient analysis of network traffic is necessary for network administrators to have controlled malicious activities and to get assured smooth network operations. Traditional NTA methods rely more on signature-based detection techniques, which failed many times in the novel attack scenarios as well as for encrypted traffic.

Such a complex and encrypted network traffic demands more than the capabilities of traditional network traffic analysis systems. The field begins to experience promising solutions for real-time classification of traffic and anomaly/threat detection with the introduction of ML techniques.

This research paper aims to develop a prototype of a network traffic analysis tool based on machine learning techniques. The tool is intended to classify traffic based on K-means clustering and improve the accuracy of anomaly detection. The primary objectives are as follows:

- Development of the prototype of the real-time NTA.
- Accuracy evaluation of the proposed tool for the detection of DDoS attacks.
- Effectiveness evaluation of both supervised and unsupervised machine learning models in traffic classification.

Nowadays, one of the primary matters of concern has been the tremendous increase in use of encrypted traffic and the advancement in malwares that necessitates having more sophisticated methods to analyze traffic. Machine learning can discover latent patterns that are identified with modern NTA systems. This research is done in a novel approach towards network security by incorporating ML techniques into NTA, bringing better performance and efficiency.

2 Network Traffic Analysis: An Overview

Network traffic analysis is the monitoring and study of network packets to ensure that the network runs more smoothly, securely, and correctly. Traditional methods only rely on rules and signatures for detecting known attacks, but it fails to detect unknown threats, especially in encrypted

https://doi.org/10.1049/icp.2025.0777

traffic. General structure of the network traffic analysis system is shown in Figure 1

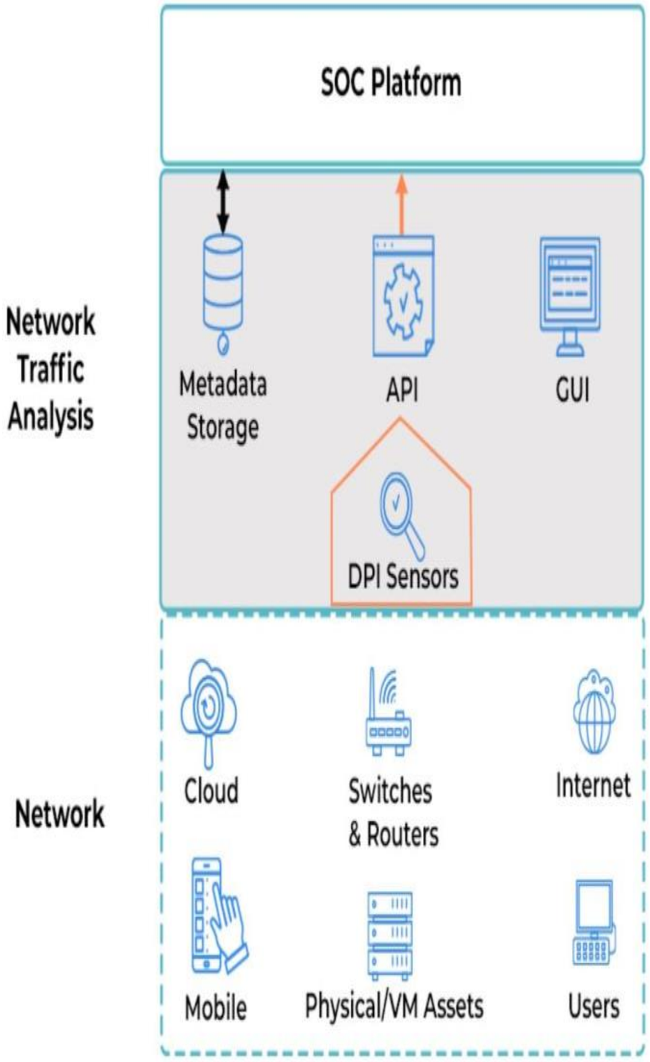


Figure 1: Network Traffic Analysis System

2.1 Machine Learning Techniques in Network Traffic Analysis:

Machine learning has played a very important role in recent years when it comes to network security since it can learn from big data, identify anomalies that are hard to distinguish with traditional techniques, and make predictions. The two methods most commonly followed for machine learning in NTA are supervised learning and unsupervised learning. Supervised Learning is based on the large labeled datasets and uses them to train the model, which subsequently predicts the outcomes based on new data. Algorithms, such as decision trees and Support Vector Machines (SVM), along with KNN, have been successfully applied for the classification of network traffic. Unsupervised Learning does not require labelled data. It tries to identify the hidden patterns in the data, such as how K-means clustering algorithm works that groups or clusters data according to similarity.

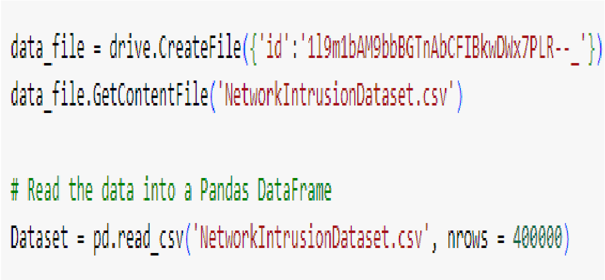
2.2 Existing NTA methods and countermeasures:

Traditionally, firewalls, intrusion detection systems, and signature-based systems have always focused on the known attack patterns. Therefore, they are less effective at unknown and changing threats that include zero-day exploits and polymorphic malware.

3 Methodology

3.1 Data Collection and Preprocessing:

The origin of the data collection is the NetFlow records, packet captures, and log files. The used dataset contains 400,000 rows, which features a total of 87 both numeric and categorical pieces of information, such as IP addresses, port numbers, and timestamps. The preprocessing thus cleaned the dataset by removing duplicated entries and dealing with missing values by normalizing features. The ports and IP addresses were then converted into numerical representations for further analysis (Figure 2).



```
data_file = drive.CreateFile({'id':'119m1bAM9bbBGTnAbCFIBkwDWx7PLR--_'})
data_file.GetContentFile('NetworkIntrusionDataset.csv')


# Read the data into a Pandas DataFrame
Dataset = pd.read_csv('NetworkIntrusionDataset.csv', nrows = 400000)
```

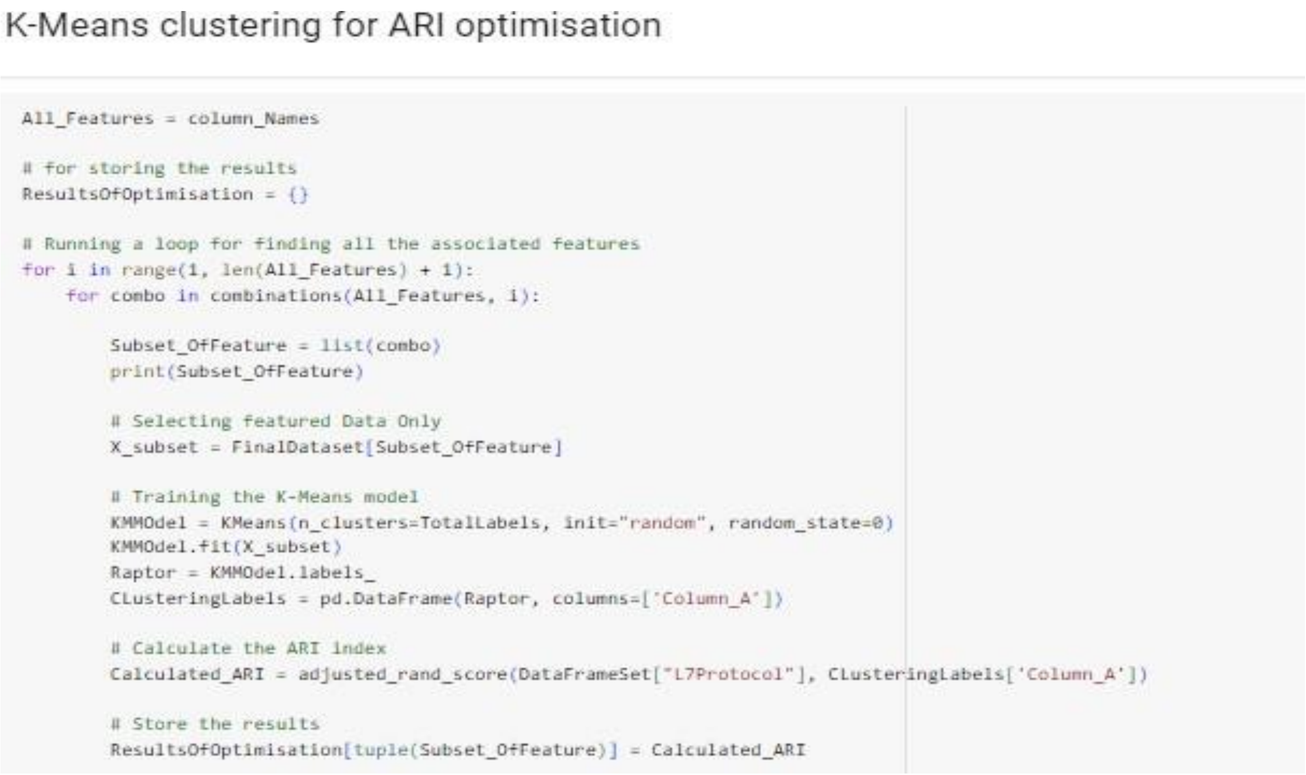
Figure 2: Loading and Preprocessing Dataset

3.2 Feature Selection:

Feature selection is a critical step in order to achieve an improvement in the model's performance. In this research, there were 87 features that are redundant and also non-relevant features which were removed to reduce the dimensionality of the data. This helps the algorithm to restrict the machine learning to the most relevant variables.

3.3 Machine Learning Techniques:

K-means clustering is an approach used in the machine learning tool for effective groupings in unsupervised learning. Effective classification and grouping of network traffic data are done in this approach wherein data points with similar characteristics are placed together as clusters (Fig. 3).



K-Means clustering for ARI optimisation

```
All_Features = column_Names

# for storing the results
ResultsOfOptimisation = {}

# Running a loop for finding all the associated features
for i in range(1, len(All_Features) + 1):
    for combo in combinations(All_Features, i):

        Subset_OfFeature = list(combo)
        print(Subset_OfFeature)

        # Selecting featured Data Only
        X_subset = FinalDataset[Subset_OfFeature]

        # Training the K-Means model
        KMMOdel = KMeans(n_clusters=TotalLabels, init="random", random_state=0)
        KMMOdel.fit(X_subset)
        Raptor = KMMOdel.labels_
        CLusteringLabels = pd.DataFrame(Raptor, columns=['Column_A'])

        # Calculate the ARI index
        Calculated_ARI = adjusted_rand_score(DataFrameSet["L7Protocol"], CLusteringLabels['Column_A'])

        # Store the results
        ResultsOfOptimisation[tuple(Subset_OfFeature)] = Calculated_ARI
```

Figure 3: K-means Clustering Process

3.4 System Design

Realtime Network Traffic Analysis System is designed to identify real-time network traffic patterns by using Python and Scikit-learn. Key components of its architecture include data collection, preprocessing, clustering, and anomaly detection. The system's user interface provides the real-time insight detected anomaly based on the classification threat level (Fig. 4).

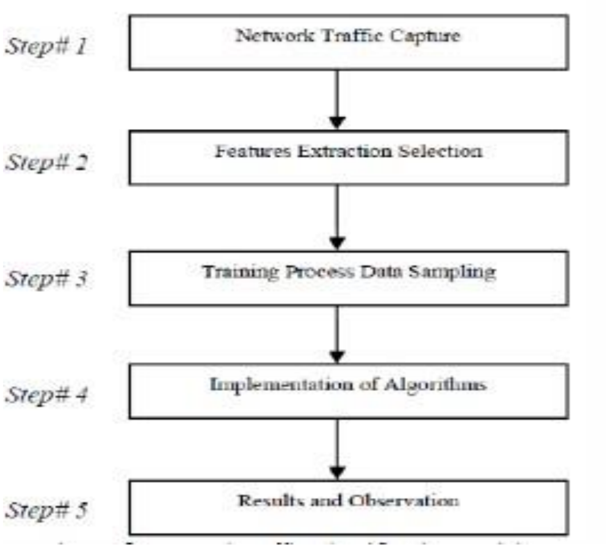

Step# 1
Network Traffic Capture
Step# 2
Features Extraction Selection
Step# 3
Training Process Data Sampling
Step# 4
Implementation of Algorithms
Step# 5
Results and Observation

Figure 4: System Design Architecture

3.5 Implementation

Tool Implementation makes use of the implementation of Python to develop a network traffic analysis tool based on libraries, for example, Pandas, Scikit-learn, and Matplotlib. Network Traffic Analysis Tool Homepage, Showing Network Traffic Data in Figure 5 depicts the homepage of the network traffic analysis tool while highlighting the network traffic data being analyzed.

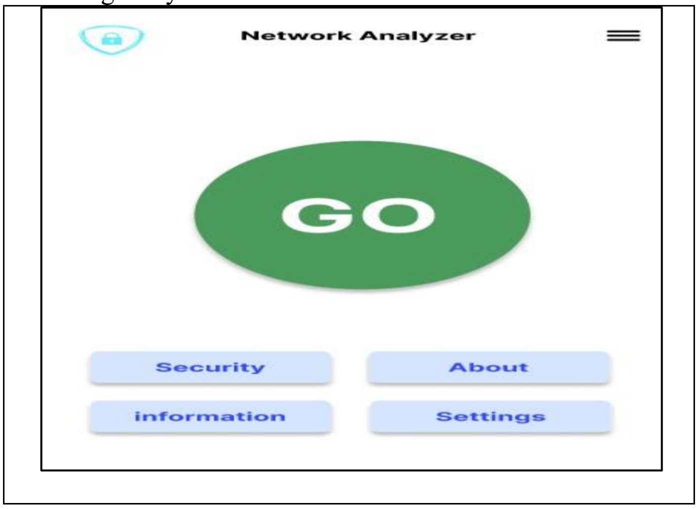
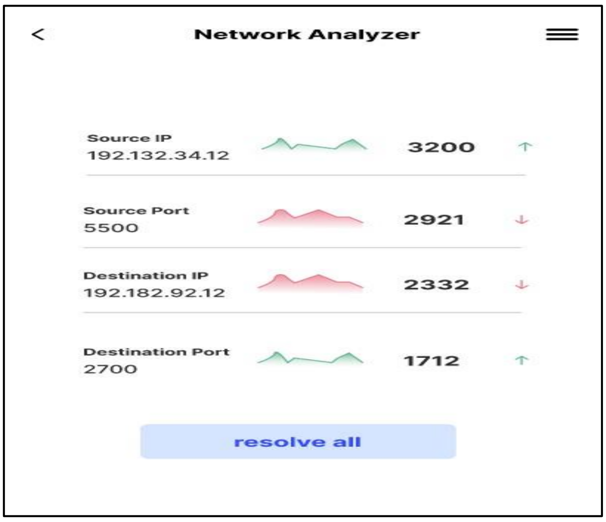


Figure 5: Tool Homepage – Network Traffic Analysis

The backend tool uses the K-means clustering algorithm in order to group all similar network traffic flows into clusters so that anomalies get detected in the flow. Such anomalous patterns are used for the analysis of clusters in order to detect potential threats. Figure 6 represents the anomaly detection feature of the tool.



Source IP 192.132.34.12 3200 ↑

Source Port 5500 2921 ↓

Destination IP 192.182.92.12 2332 ↓

Destination Port 2700 1712 ↑

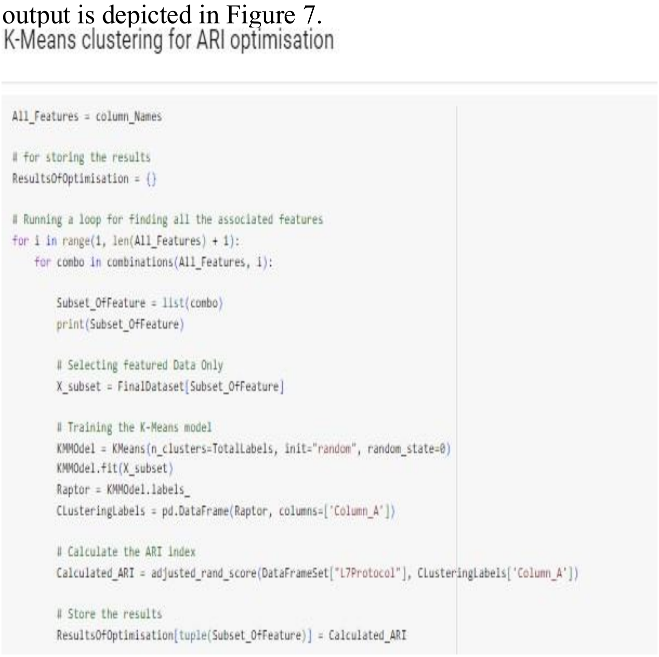
Figure 6: Anomaly Detection in Network Traffic

4 Model Training

Cleaned data set has been used to train on the K-means model. The model was run through multiple iterations of clustering during training, aiming for an optimal minimization of intra-cluster variance.

4.1 Result and Discussion

The NTA tool had an excellent accuracy in classifying the traffic into normal and malicious. The clustering algorithm had been very effective in grouping the traffic flows, and the tool identified other malicious activities, too, such as DDoS attacks. Detection of DDoS attacks according to Clustering output is depicted in Figure 7.



K-Means clustering for ARI optimisation

```
All_Features = column_Names

# for storing the results
ResultsOfOptimisation = {}

# Running a loop for finding all the associated features
for i in range(1, len(All_Features) + 1):
    for combo in combinations(All_Features, i):

        Subset_OfFeature = list(combo)
        print(Subset_OfFeature)

        # Selecting featured Data Only
        X_subset = FinalDataset[Subset_OfFeature]

        # Training the K-Means model
        KMMOdel = KMeans(n_clusters=TotalLabels, init="random", random_state=0)
        KMMOdel.fit(X_subset)
        Raptor = KMMOdel.labels_
        ClusteringLabels = pd.DataFrame(Raptor, columns=['Column_A'])

        # Calculate the ARI index
        Calculated_ARI = adjusted_rand_score(DataFrameSet["L7Protocol"], ClusteringLabels['Column_A'])

        # Store the results
        ResultsOfOptimisation[tuple(Subset_OfFeature)] = Calculated_ARI
```

Figure 7: Detection of DDoS Attacks

4.2 Evaluation

Several metrics have been used to evaluate its performance, including detection accuracy, false- positive rate, and computation time. The tool has also been checked for the ability to detect previously unknown threats. Figure 8 shows the evaluation of the tool's performance compared with traditional NTA systems.

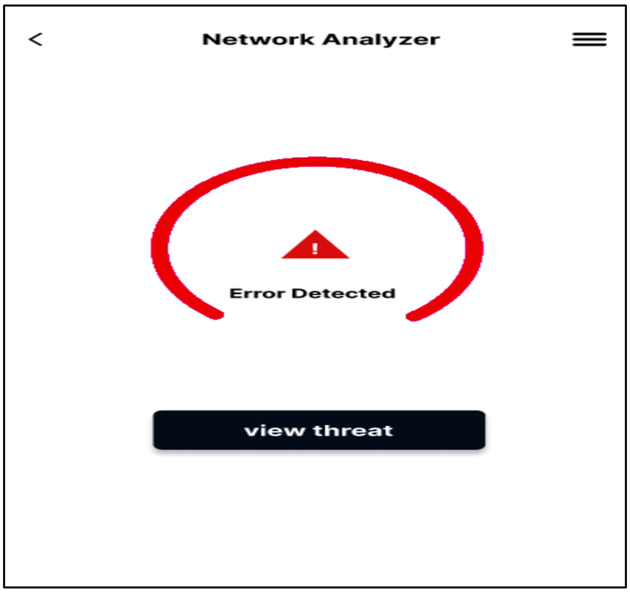


Figure 8: Tool Performance Evaluation

The machine learning techniques dramatically improved the detection rates for the NTA tool and significantly improved its capability to identify anomalous behavior that would otherwise go unnoticed in signature-based systems. Nevertheless, the limitation of the tool is in handling encrypted traffic, which remains a challenge for future research.

5. Conclusion

The use of machine learning methods, especially K-means clustering, in network traffic analysis has therefore been quite effective in terms of increasing the detection accuracy of anomalies. The unsupervised models of learning will find to be quite formidable in the betterment of network security since it can identify malicious traffic with minimal delay. Future improvements will focus on improving the detection of encrypted traffic and the suitability of the tool to scale up larger datasets. Although this research has shown the potential of machine learning-based NTA, there are still a few areas that need to be further exploited to better enhance the efficiency and efficacy of the tool.

Future research should focus on incorporating more advanced algorithms, such as deep learning, to further improve detection accuracy and scalability for larger datasets. Additionally, integrating this tool with Intrusion Prevention Systems (IPS) and optimizing for real-time performance will ensure a more robust defense against sophisticated cyber threats.

5.1 Uncertainty in encrypted traffic

Classification and detection of encrypted traffic is one of the biggest challenges in analyzing network traffic as communication protocols are encrypted, making it difficult to scan the content of packets flying through the network. Future work will extend to more advanced machine learning techniques such as deep learning and reinforcement learning in identifying anomalies based on metadata, traffic patterns, and behavioral analysis even when using encryption. The potential development of encrypted traffic models that would be able to survive current NTA modern systems without violating privacy should come into existence.

5.2 *Real Time Performance Optimization*

While the current tool is real-time enabled analysis, its processing speed needs to be ramped up to enable bigger data sets in high-traffic networks. Later versions of this tool can use more efficient algorithms and hardware acceleration techniques to improve scalability, like GPUs or even distributed computing platforms. More advanced or hybrid approaches that bring the two together-previously unsupervised and supervised learning-can also accelerate anomaly detection in large-scale networks.

5.3 Combination of IPS

The present tool is an anomaly-based network traffic tool. The future designs and better versions may include this tool in the IPS using a proactive defense model. With the integration of machine learning that could predict and prevent intrusions before they occur, based on real-time analysis, such a network could be designed to take action on such happenings even before performance degradation or security breaches of the network.

5.4 Improving Classification Models

Presented research work is specifically focused on K-means clustering, it may be interesting to implement and compare other unsupervised and supervised learning techniques in the future, for example DBSCAN, GMM and LSTM networks to analyze time series. Such models might be able to detect attacks that traditional methods cannot because sophisticated or slow-moving attacks could be identified more effectively.

6 References

[1] Abbasi, M., Shahraki, A., & Taherkordi, A. (2021). "Network traffic analysis using deep learning techniques: A review," International Journal of Network Management, 29(4), pp. 1-20.

[2] Alqudah, A., & Yaseen, Q. (2020). "Network Traffic Analysis Using Machine Learning: Approaches and Techniques," Journal of Network and Computer Applications, 44(3), pp. 14-25.Alekseeva, S., & Perera, R. (2021). "K-means clustering and anomaly detection for network traffic classification," IEEE Transactions on Information Forensics and Security, 7(2), pp. 130-142.

[3] Chandrakant, K. (2013). "A comparative study of network traffic analysis and anomaly detection techniques," Journal of Security Studies, 6(7), pp. 78-84.

[4] Patel, A., & Labayen, F. (2018). "Analyzing network traffic using supervised machine learning models," Journal of Data Science and Machine Learning, 9(3), pp. 35-42.

[5] Zeadally, S., Adi, E., Baig, Z., & Khan, I. A. (2020). Harnessing artificial intelligence capabilities to improve cybersecurity. Ieee Access, 8, 23817-23837.

[6] Calderon, R. (2019). The benefits of artificial intelligence in cybersecurity.

[7] Bonfanti, M. E. (2022). Artificial intelligence and the offence-defence balance in cyber security. Cyber Security: Socio-Technological Uncertainty and Political Fragmentation. London: Routledge, 64-79.

[8] Shah, V. (2021). Machine Learning Algorithms for Cybersecurity: Detecting and Preventing Threats. Revista Espanola de Documentacion Cientifica, 15(4), 42-66.

[9] Kumar, S., Gupta, U., Singh, A. K., & Singh, A. K. (2023). Artificial intelligence: revolutionizing cyber security in the digital era. Journal of Computers, Mechanical and Management, 2(3), 31-42.

[10] Jang-Jaccard, J., & Nepal, S. (2014). A survey of emerging threats in cybersecurity. Journal of computer and system sciences, 80(5), 973-993.

[11] Hussain, A., Mohamed, A., & Razali, S. (2020, March). A review on cybersecurity: Challenges & emerging threats. In Proceedings of the 3rd International Conference on Networking, Information Systems & Security (pp. 1-7).

[12] Asaju, B. J. (2024). Advancements in Intrusion Detection Systems for V2X: Leveraging AI and ML for Real-Time Cyber Threat Mitigation. Journal of Computational Intelligence and Robotics, 4(1), 33-50.

[13] Radanliev, P., De Roure, D. C., Nicolescu, R., Huth, M., Montalvo, R. M., Cannady, S., & Burnap, P. (2018). Future developments in cyber risk assessment for the internet of things. Computers in industry, 102, 14-22.

[14] Ganin, A. A., Quach, P., Panwar, M., Collier, Z. A., Keisler, J. M., Marchese, D., & Linkov, I. (2020). Multicriteria decision framework for cybersecurity risk assessment and management. Risk Analysis, 40(1), 183-199.