***Exploring Supervised and Unsupervised Learning Approaches for Cyber Threat Intelligence***

Sai Sravanthi Racharla

University of Houston-Victoria

College of Natural & Applied Science

Victoria, TX United States

racharlas@uhv.edu

Asfandyar Kazi

University of Houston-Victoria

College of Natural & Applied Science

Victoria TX, United States

Kazia@uhv.edu

Sri Vishnu Gopinathan

University of Houston-Victoria

College of Natural & Applied Science

Victoria TX, United States

gopinathansv@uhv.edu

Malavkumar R Patel

University of Houston-Victoria

College of Natural & Applied Science

Victoria TX, United States

PatelMR1@uhv.edu

*Abstract*—Cyber Threat Intelligence is accumulating over time as cyber-crime incidents are escalating on a large scale. Cybercrime poses a severe threat to organizations all over the world. There is a great deal of information about cyber-attacks available, which, if analyzed and utilized properly can foster organizations to shield themselves from potential cyber threats and take preventive measures. However, the vast amount of data present makes it difficult for the organizations to convert the available information into actionable intelligence. Failure to protect themselves from such potential threats might compromise the organization’s security, reputation, data, and assets. Organizations are rapidly adopting cloud services and cyber threats on cloud technologies are inevitable. To overcome this problem, we perform comparative analysis of clustering and classification techniques of CTI using Machine Learning. This study provides efficient methods in classifying emerging threats into their respective categories and clustering of similar threats, based on several factors including tactics, techniques, and procedures (TTP’s), domain specific threats, threat actor profiles, etc. This approach helps organizations to identify common patterns and characteristics of threats in the cloud and on-premises. Cyber Threat Intelligence (CTI) processing can play an essential role in enhancing cloud security by providing actionable insights to organizations on potential threats to their cloud infrastructure.

Keywords—cyber threat intelligence, cyber defense, clustering, cloud security.

# Introduction

The sophistication of cyberattack methods has increased recently. These days organized attacks are on the rise for the purpose of committing financial fraud. Additionally, targeted attacks on the cloud have become more common for the purpose of data exploitation. The rapidly expanding number of new malware kinds cannot be detected by conventional malware detection techniques.

In order to handle this issue, it is necessary to anticipate cyberattacks and take the required precautions in advance, and using intelligence is crucial to making this feasible. The threat information-sharing system in the cloud has been widely adopted, so cloud threats can be detected in advance and appropriate measures can be taken to secure the cloud environment. We anticipate that much intelligence, including these instances of illegal content, exists in cyberspace because many black hat hackers frequently share information and tools that can be used for assaults on the dark web or in particular forums. It is anticipated that employing threat intelligence will enable early attack detection and active protection in cloud. However, such intelligence is currently retrieved manually, and various research efforts are being carried out to improve this process.

The goal of this project is to develop and compare techniques for automatically extracting and processing significant posts in order to increase the effectiveness of cloud security using cyber threat intelligence. The movement to actively defend against cyberattacks by using intelligence is currently most prevalent in the industrial world, and some of these efforts have produced impressive outcomes.

# BACKGROUND

## Related Works

[1] The data was collected from a leaked hacker forum (Nulled.IO). Data conversion and cleansing such as removing HTML tags, replacing invalid characters, lower case conversion etc., were carried out. Later, the data was analyzed by Latent Dirichlet Allocation for the purpose of topic modelling.

[2] The published study collects posts from forums on the dark web that might be connected to cyberattacks, such as personal data and hacking methods. On the data gathered, they use Bag-of-Words, a tool that counts the words that exist in a document and uses the frequency of words as a feature. They classify the subjects that effectively characterize the provided content using SVM and logistic regression. They suggested a technique that combines supervised learning and semi-supervised learning to extract forum postings about cyberattacks in order to decrease the labeling time when creating training data.

[3] This paper proposes a method for enhancing Open-Source Intelligence (OSINT) processing from threat intelligence platforms to produce high-quality enriched Indicators of Compromises (IoC). This enhanced intelligence is produced by correlating and merging IoC’s from various OSINT feeds that provide details on the same threat, grouping them into clusters, and then portraying the threat data contained in those clusters in a single enriched IoC.

## Algorithms

### Unsupervised Approach:

### K-means:

Data mining and machine learning use the cluster analysis technique to organize related objects into clusters. The goal of K-means clustering, a popular technique for cluster analysis, is to divide a set of objects into K clusters. The algorithm clusters in a way such that the least possible sum of the squared distances between the data points and the assigned cluster mean is minimized.

The centroid of each cluster in K-means is chosen at random after the number of clusters, k, is defined. The next step is to allocate each data point to the cluster with the closest centroid (mean). Then, the average of the data in each cluster is taken as the new centroid, and each data point is reassigned to the cluster with the nearest centroid.  Until the centroid of each cluster does not change, these procedures are repeated.

### Supervised Approach:

### 1) Logistic Regression:

Classification issues are resolved using logistic regression as it achieves this by forecasting categorical outcomes. Data and the relationship between one dependent variable and one or more independent variables are described using logistic regression. There are two types of classification: binomial(two outcomes) and multinomial(more than two outcomes).

### Random Forest Classifier:

Random forest is a supervised machine learning model which is frequently employed in classification and regression issues.

It makes use of ensemble learning, a method for solving complicated issues by combining a number of classifiers. It constructs decision trees on various samples, and takes the majority of votes from the trees to determine the result. The Random Forest Algorithm's ability to handle data sets with continuous variables, as in regression, and categorical variables, as in classification, is one of its most crucial qualities.

# proposed methodology

The first step in this process is data collection. We plan to collect data from OSI (Open-Source Intelligence) which is a publicly accessible resource available on the internet. However, after collection of data from the forum, the data must be cleansed, formatted and analyzed to design an efficient system supporting cyber defense. This raw data will be subjected to various data pre-processing steps such as identifying and replacing null values, removing stop words, punctuation, special characters. The next step is feature extraction where important aspects of cyberattack for each record such as type of attack (phishing, malware, or DDoS), Indicators of Compromises (IOCs), Tactics Techniques and Procedures (TTPs), targeted sector, severity and threat actors are analyzed and extracted from the data. Once feature extraction is completed, the process of classification and clustering comes into picture. For unsupervised approach, we employed K-means clustering as it is one of the simplest and popular unsupervised machine learning algorithms. For supervised methods, we tested Random Forest classifier as it is best known for its performance in multi-class classification problems and Logistic Regression for its elegancy and simplest modelling features. All these methods are experimented for finding an optimal algorithm for CTI threat based categorization. With the help of CTI processing, real-time alerts and relevant threat intelligence can be provided to incident response teams, enabling them to respond quickly to cloud security incidents. Organizations can proactively mitigate the risk of potential attacks on their cloud infrastructure by identifying vulnerabilities through the analysis of clusters of similar threats and taking appropriate measures.

# experimental evaluation

### Data collection and pre-processing:

For this experiment, we collected the dataset from [OSI](https://www.kaggle.com/datasets/alextamboli/unsw-nb15) which is available over the internet. The data set had around 1,73,500 records with several attributes such as Start time, Last time, Attack category, Attack subcategory, Protocol, Source IP, Source port, Destination IP, Destination Port, Attack name and Attack Reference.

TABLE I

RAW DATA

|  |  |
| --- | --- |
| COLUMNS | VALUES |
| Start time | 1421927418 |
| Last time | 1421927419 |
| Attack category | Exploits |
| Attack subcategory | Cisco IOS |
| Protocol | tcp |
| Source IP | 175.45.176.2 |
| Source Port | 26939 |
| Destination IP | 149.171.126.10 |
| Destination Port | 80 |
| Attack Name | Cisco IOS HTTP Authentication Bypass Level 64 |
| Attack Reference | CVE 2001-0537 [Reference Link](http://cve.mitre.org/cgi-bin/cvename.cgi?name=2001%2d0537)BID%202936%20(http://www.securityfocus.com/bid/2936)OSVDB%20578%20(http://www.osvdb.org/578)CVSS-High%20(https://strikecenter.bpointsys.com/bps/reference/CVSS/9.3%20%28AV%3aN%2fAC%3aM%2fAu%3aN%2fC%3aC%2fI%3aC%2fA%3aC%29) |

From examining the data, we determined that the Attack Reference column has no significance on categorizing the threats. The Attack subcategory column had considerable null values which were replaced with the most frequently occurring value. The protocol attribute gives information regarding the type of protocol used to carry out the cloud threat. Attack name has organization and infrastructure specific threat information.

TABLE II

ATTACK CATEGORIES

|  |  |
| --- | --- |
| CATEGORY | COUNT |
| Exploits | 68217 |
| Fuzzers | 33638 |
| DOS | 24582 |
| Reconnaissance | 20136 |
| Generic | 19860 |
| Backdoor | 4097 |
| Analysis | 1881 |
| Shellcode | 1511 |
| Backdoors | 256 |
| Worms | 169 |

The first seven cyber-attack categories from TABLE II were considered for this study because of their significant count. The attack sub-category and protocol columns were subjected to enumeration for numerical conversion. Attack name was converted using Term frequency Inverse Document Frequency method. The data was split into training for model learning and testing sets for validation in the ratio of 8:2. The dependent variable is Attack Category which is determined from the remaining columns (independent variables).

### Supervised Approach:

In our experiment, three iterations were carried out for training and evaluating the models. In the first iteration, first five attack categories from TABLE II were considered. The total number of categories are then increased to six and seven for the second and third iterations respectively. The accuracies are determined for each iteration and are compared. This process helps in testing the stability of the model at each level.

### Logistic Regression:

The logistic regression model is tested with “max\_iter” as 500, having other parameters as default. The model accuracies are 57.43% , 58.59% and 57.04% for the successive iterations.

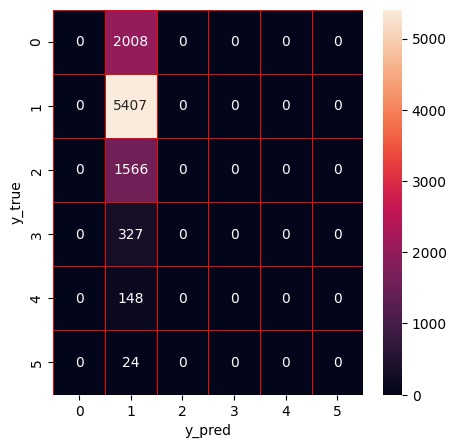


Fig. 1. Confusion matrix for Logistic Regression with the highest accuracy (58.59%).

### Random Forest Classifier:

The random forest model is evaluated with “n\_estimators” as 500, having other parameters as default. The model accuracies are 75.52%, 77.31% and 75.19% for the successive iterations.

The accuracies obtained from Random Forest classifier are better compared to the Logistic Regression model. Fig 3 demonstrates a graphical comparison of the two classifier models. The graph obtained shows the performance analysis based on the number of categories taken into consideration.

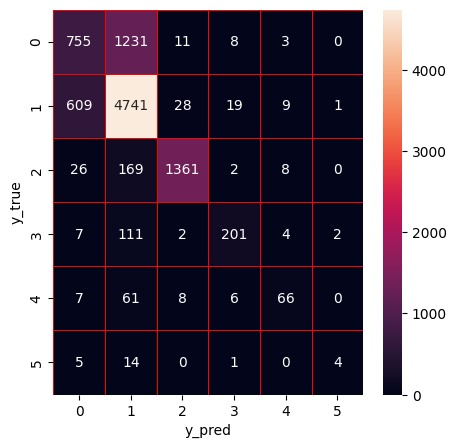


Fig. 2. Confusion matrix for Random Forest with the highest accuracy (77.31%).

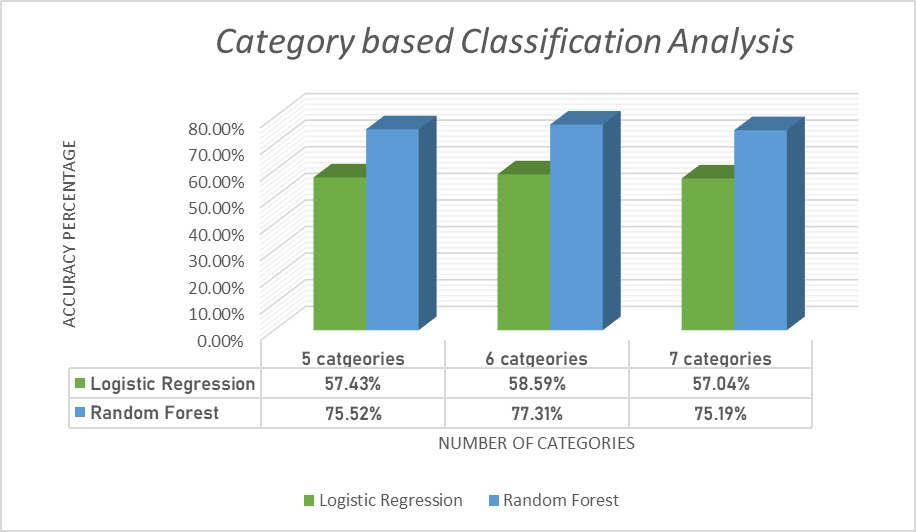


Fig. 3. Comparison between Logistic Regression and Random Forest Classifiers.

From the graph, in Fig.3, we observe that the performance of the model is not hampered when classifying CTI with different number of categories. This demonstrates the flexibility of the model, to have optimal classification results on any future data with reduced or additional cyber threat categories.

### Unsupervised Approach:

The K-Means model was developed with 7 clusters, with each cluster corresponding to a threat type. This clustering serves the purpose of eradicating a new threat by referring to similar threats in its cluster and finding an appropriate solution.

From the results provided in Fig.4 and Fig.5, we understand that the results from clustering were not appropriate. Our research aimed to cluster the CTI into separate categories but the data was not suitable for clustering.

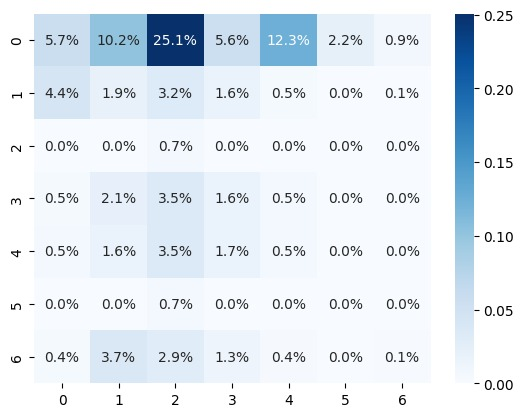


Fig. 4. Confusion matrix for K-Means

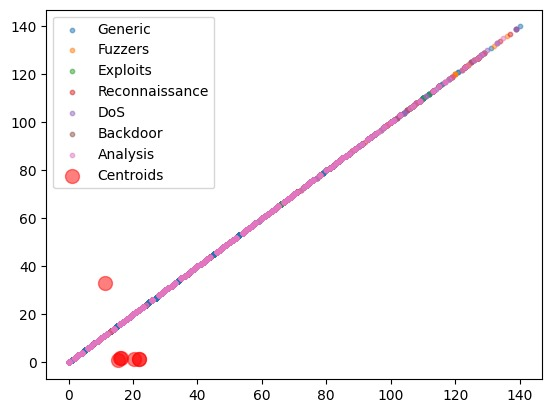


Fig. 5. Clustering map by K-Means

# conclusion

In this study, we discussed efficient methods for processing cloud-based CTI into different categories. For the OSI data we collected, supervised approaches show better outcomes compared to unsupervised approaches. Random Forest Classifier has the highest accuracy in supervised approaches. Though clustering the CTI did not provide optimal results, we still studied the practical implications in clustering for the dataset. The large volume of hacker forum posts requires automation, as manual analysis is a time-consuming, expensive, and error-prone procedure that is not viable for obtaining CTI efficiently. With the help of classification and clustering techniques in machine learning, we can identify, and group similar threats based on their category and domain. This helps the organizations in reducing the incident response time in the case of a new attack in cloud and on-prem, thereby enhancing security. The future scope of this experiment would be extracting and processing CTI threats related to specific industry such as medical, finance, education etc., along with clustering of domain specific threats. This will support organizations with similar goals to improve their security in cloud environment and have effective response strategies to cloud attacks.

##### References

1. E. Nunes et al., "Darknet and deepnet mining for proactive cybersecurity threat intelligence," 2016 IEEE Conference on Intelligence and Security Informatics (ISI), Tucson, AZ, USA, 2016, pp. 7-12, doi: 10.1109/ISI.2016.7745435.
2. M. Kadoguchi, H. Kobayashi, S. Hayashi, A. Otsuka and M. Hashimoto, "Deep Self-Supervised Clustering of the Dark Web for Cyber Threat Intelligence," 2020 IEEE International Conference on Intelligence and Security Informatics (ISI), Arlington, VA, USA, 2020, pp. 1-6, doi: 10.1109/ISI49825.2020.9280485.
3. R. Azevedo, I. Medeiros and A. Bessani, "PURE: Generating Quality Threat Intelligence by Clustering and Correlating OSINT," 2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), Rotorua, New Zealand, 2019, pp. 483-490, doi: 10.1109/TrustCom/BigDataSE.2019.00071.
4. W. Fan et al., "Enabling Privacy-Preserving Sharing of Cyber Threat Information in the Cloud," 2019 6th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/ 2019 5th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom), Paris, France, 2019, pp. 74-80, doi: 10.1109/CSCloud/EdgeCom.2019.00-15.
5. Thomas D. Wagner, Khaled Mahbub, Esther Palomar, Ali E. Abdallah,Cyber threat intelligence sharing: Survey and research directions,Computers & Security,Volume 87,2019,101589,ISSN 0167-4048
6. E. Nunes et al., "Darknet and deepnet mining for proactive cybersecurity threat intelligence," 2016 IEEE Conference on Intelligence and Security Informatics (ISI), Tucson, AZ, USA, 2016, pp. 7-12, doi: 10.1109/ISI.2016.7745435.
7. H. Suryotrisongko, H. Ginardi, H. T. Ciptaningtyas, S. Dehqan and Y. Musashi, "Topic Modeling for Cyber Threat Intelligence (CTI)," 2022 Seventh International Conference on Informatics and Computing (ICIC), Denpasar, Bali, Indonesia, 2022, pp. 1-7, doi: 10.1109/ICIC56845.2022.10006988.
8. Abu, Md Sahrom, et al. "Cyber threat intelligence–issue and challenges." Indonesian Journal of Electrical Engineering and Computer Science 10.1 (2018): 371-379.
9. I. Deliu, C. Leichter and K. Franke, "Collecting Cyber Threat Intelligence from Hacker Forums via a Two-Stage, Hybrid Process using Support Vector Machines and Latent Dirichlet Allocation," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 5008-5013, doi: 10.1109/BigData.2018.8622469.
10. Hossen, Md Imran, et al. "Generating Cyber Threat Intelligence to Discover Potential Security Threats Using Classification and Topic Modeling." arXiv preprint arXiv:2108.06862 (2021).