Enhancing Cyber Threat Intelligence and Security Automation: A Comprehensive Approach for Effective Protection

Amit Kumar Bairwa[1] Rohan Khanna,[1] Sandeep Joshi[1], and Pljonkin Anton Pavlovich[2]

[1] Manipal University Jaipur, India,
amitbairwa@gmail.com,
[2] Southern Federal University, Russia

Abstract. Network management gets more difficult as technology develops, and internet information is more readily available. As a result, cyberattacks have increased as well, making it harder for organisations to identify and stop threats. It has become essential to create a cyber threat intelligence team. In order to improve threat identification and response using cyber threat intelligence, this project will make use of artificial intelligence and machine learning. In order to increase the precision and speed of threat detection, it investigates powerful techniques including supervised and unsupervised learning. The report recommends investing in cutting-edge technology, knowledgeable personnel, and encouraging advancement to address issues of combining CTI and security automation. It evaluates the effects of security automation and CTI on incident response, effectiveness, cost savings, and compliance. The study uses machine learning approaches to recognise threat actors and classify them according to the nature of the assault and their organisational affiliation. The accuracy of the KNN model was 93%, whereas the accuracy of the logistic regression was 97.5%. The paper uses the same technique as the research article "Threat Actor Type Inference and Characterization within Cyber Threat Intelligence," which it builds upon.

Keywords: Cyber, Threat, Intelligence, Automation, Malware, Threat actor, Machine Learning, Logistic Regression, KNN.

1 Introduction

Cyber Threat Intelligence (CTI) and Security Automation are two critical components of modern information security. CTI involves the collection, analysis, and dissemination of threat intelligence to identify potential threats and respond to them proactively. Security automation, on the other hand, leverages advanced technologies such as Artificial Intelligence (AI), Machine Learning (ML), and big data analytics to automate routine security tasks, improve incident response times, and reduce the risk of human error. Together, CTI and security automation enable organizations to stay ahead of evolving threats and protect their sensitive information in real-time.

Cyber threat intelligence refers to information that is based on knowledge, skill, and experience concerning the occurrence and evaluation of both cyber and physical threats and threat actors, with the aim of aiding in the mitigation of potential attacks and harmful cyberspace occurrences.[1].

The risk of organisational computer networks getting attacked by threat actors has increased exponentially due to various factors. This is where the role of threat intelligence comes. CTI is a very crucial part of any cyber security ecosystem [2]. A well defined CTI program can help an organisation in:

1. **Prevention of Data Loss**: CTI can help organisations spot cyber threats and prevent data breaches which might lead to release of sensitive information.
2. **Implementation of Proper Security measures**: CTI helps the organisation in understanding the pattern used by the hackers. Based on these patterns it helps in implementing the appropriate security measures[3].
3. **Compromise Assessment**: CTI is can also be used to analyse a organisation's environment for possible compromised assets. A compromise assessment can be helpful in the timely recovery of compromised assets.

Threat intelligence is knowledge about current or future risks to assets that is supported by facts. It contains substantial instruction, tools, points, and pertinent background. Unauthorised access, unauthorised asset usage, the release of sensitive data, unauthorised asset alterations, and access denial are all examples of these dangers. Cyber threat intelligence (CTI) specialises in researching cutting-edge adversary tactics, methods, and practises. CTI has several uses, including the detection of breaches or anomalous behaviour and the ability to follow and stop threats even before they manifest [4]. Cyber defence may become more successful by reducing false positives (and false negatives) by utilising cyber threat intelligence. Before adding CTI into security procedures, though, it's important to fully comprehend it [5] [6].

Security automation is the automation of security duties including administrative work and incident detection and response. Security automation makes it possible for security teams to expand to handle growing workloads, which has several benefits for the business.

In response to the number and complexity of assaults that are on the rise, the zero trust security architecture was created to manage business cyber risk. Before giving or maintaining access to applications and data, it requires authentication, authorisation, continual security configuration, and posture checks for all users, regardless of their location. Zero trust security bases access choices on role-based constraints, in contrast to conventional methods that implicitly trust internal persons and systems.

Zero trust architecture comes with increased overhead even if it provides granular safety. Security automation becomes crucial for ensuring a zero trust approach that is safe, scalable, and long-lasting. Automating security procedures makes zero trust deployment more efficient and enables organisations to manage their cyber risk more successfully in the long run.[7] [8].

Due to emerging malware with numerous layers and self-updating algorithms to escape detection, threat intelligence and analysis are complicated [9]. Due to the exponential growth in device numbers, manual compromise evaluation is difficult. In cybersecurity, machine learning has made great progress. Automated security software now uses ML to identify sophisticated malware and improve scanning processes [10]. This opinion is backed up by a number of studies that emphasise how ML methods are used in cyber threat intelligence. Effective ML algorithms for threat assessments present organisations with potential solutions.

In this work, the available dataset will be used to train different machine learning models. Now, let's look at the machine learning concepts which will be applied to the dataset used in the research[11].

1. **Logistic Regression**: One of the most popular machine learning techniques under supervised machine learning is logistic regression. It analyses categorical dependent variables using a collection of independent factors. In place of discrete numbers, it provides probabilistic values between 0 and 1, as shown in Figure 1. When attempting to solve classification difficulties, logistic regression is employed. Two maximum values between 0 and 1 are predicted by fitting a "S"-shaped logistic function. The equation for Logistic Regression can be represented as shown in 1:

$$\log[\frac{y}{1-y}] = b_0 + b_1 \times x_1 + b_2 \times x_2 + \cdots + b_n \times x_n \tag{1}$$

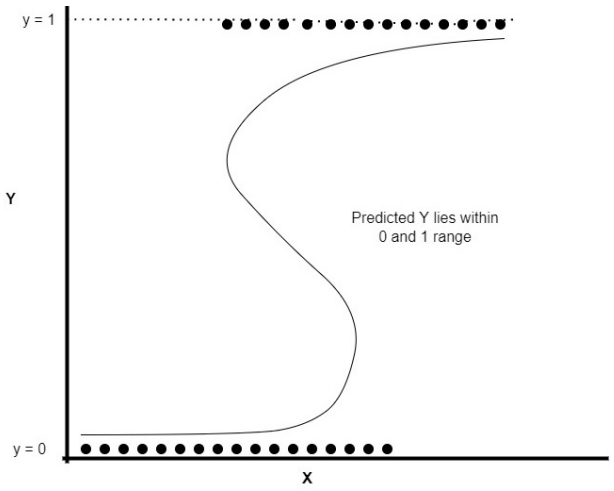

Fig. 1. Logistic Regression

2. **K-Nearest Neighbour's Algorithm**:A supervised machine learning technique, utilised for classification and regression problems. By selecting the

K closest neighbours from the training dataset, it may predict the label or value for a new data point. KNN assigns the label that appears the most frequently among the K closest neighbours while performing classification tasks. It determines the mean or median of the data from the K closest neighbours while doing regression tasks.

Algorithm 1 K- Nearest Neighbour's Pseudocode

```
Load the training and test data
Choose the value of K
For each point in test data:
    find the Euclidean distance to all training data points.
    store the Euclidean distances in a list and sort it
    choose the first k points
    assign a class to the test point based on the majority of classes present in the
chosen points
End
```

1.1 Problem Statement

"Researching on the importance of Cyber Threat Intelligence and Security Automation in today's world and classification of threat actors using various machine learning algorithms . Comparing these models based on several criteria, including as accuracy, efficiency, and other factors. After careful evaluation, choosing the best model."

1.2 Contribution of the Study

The primary objective of this research is to gather significant insights from various research papers published in numerous conferences and journals. These papers will be directly linked to the problem statement defined by our study. This will enable us to understand various techniques that can be utilized to address the given problem. By categorizing threats and threat actors based on motivation, target, and intent, organizations can apply effective procedures to mitigate these threats. The study also aims to build efficient Machine Learning models for the classification of threat actors.

2 Motivation

In today's digital age, cyber threats are becoming increasingly sophisticated, frequent, and targeted. These threats can come from various sources such as hackers, cybercriminals, and state-sponsored actors, and can target organizations of all sizes and across all industries. These threats can lead to data breaches, financial losses, reputation damage, and even business disruption or shutdown.

CTI provides a proactive approach to cybersecurity by collecting and analyzing data from various sources such as open-source intelligence, dark web monitoring, and internal network data to identify potential threats and vulnerabilities. By leveraging CTI, organizations can gain a better understanding of the threat landscape and the tactics, techniques, and procedures (TTPs) used by threat actors. This information can then be used to develop and implement more effective security measures to protect against potential threats [12].

Moreover, CTI can help organizations to identify emerging threats and vulnerabilities, enabling them to take proactive steps to mitigate potential risks. This can include implementing security controls, conducting security awareness training for employees, and developing incident response plans to minimize the impact of security incidents. Thus we can conclude that the motivation behind using CTI is to stay ahead of evolving cyber threats, proactively identify and mitigate potential risks, and protect an organization's systems, data, and reputation. By leveraging CTI, organizations can improve their overall security posture and minimize the impact of potential security incidents.

3 Brief literature review

The main aim of the research by Roumen Trifonov et al. [13] is to assess risks and examine strategies, techniques, and processes related to Advanced Persistent Threats (APTs), this research uses artificial intelligence in Operational Cyber Threat Intelligence. By defining the operational environment, evaluating adversaries, and identifying probable antagonistic courses of action, the objective is to reduce risks to an organisation and its assets. The CTI model makes use of a multi-agent system and machine learning and signature matching methods. The behavioural model entails choosing pertinent traits and refining classifiers to categorise behaviours as hostile or non-hostile. The model features an offline calibration phase for training and an online phase for behaviour recognition and command conversion. Preprocessing makes use of Sequential Feature Selection and the Echo State Network technique (a kind of RNN known as Reservoir Computing) ensures feasbility.

The research by Florian K. Kaiser et al. [14] looks at the benefits and cons of automated incident response utilising cyber threat intelligence. It starts by discussing the present condition of international cyber security, highlighting the rise in cyberattacks and the shortcomings of current defensive tactics. The paper's main objective is to suggest an automated incident responder that makes use of straightforward heuristics to find efficient defences. The idea of a computer network's biological immune system serves as the inspiration for the paper's further investigation of the pairing of static base defence and adaptive incident response.The main elements of an automated incident response system—data collecting, data analysis, decision-making, and action implementation—are described in the study. It emphasises how crucial human monitoring is at every stage of the procedure. Examples from the real world that demonstrate how automated incident responses with cyber threat intelligence have been effective in

controlling and reducing the effects of malware outbreaks are given to support the research. The research also uses honeypots and other deception tactics to entice threat actors into a controlled setting.

The research by Vasileios Mavroeidis et al. [15] uses a clear methodology to systematically detect and classify threat actors. Data processing, data analysis, and data collecting make up the methodology's three steps. Collecting data entails obtaining information from a variety of places, including social media, open-source intelligence, and dark web forums. The data is then cleaned, organised, and processed in order to make it ready for analysis. Techniques like clustering and classification are used during the data analysis step to find patterns and characteristics in threat actors. The study article offers actual examples to show how these stages are applied in real-world situations. Threat actors are divided into six groups in the paper: competitors, nation-states, cybercriminals, hacktivists, insiders, and terrorists. The authors categorise the threat actors using traits including motive, capacity, and targeting.

The research by Amira M. Aljuhami et al.[16] looks into how risk management is impacted by cyber threat intelligence. The importance of CTI in assisting a company in identifying, assessing, and lowering possible cyber threats is emphasised in the article. Benefits including the ability to assess risks, identify new hazards, and develop effective risk management strategies were also discussed.The issues with CTI have, however, been the difficulty in keeping up with changing threats and the requirement for specialist knowledge and abilities.Later in the report, the author offers suggestions for integrating CTI into risk management procedures. According to the author, it is suggested to first create a CTI team. The objectives and requirements of the CTI are then explained. The CTI is connected with a task management system following the creation of a strategy for CTI collecting and analysis.

3.1 Research Gap

This paper also aims at overcoming some of the limitations that are being faced by the previous research works. These are as follows:

1. Contextuality issues and poor data quality make it difficult to train the model. The employment of AI-based systems in CTI raises ethical questions about privacy, prejudice, discrimination, and other issues.
2. The analysis's reach is constrained, and CTI's efficacy is reliant on the calibre of the data it uses. CTI implementation in practical situations might be technically challenging.
3. The technique mainly depends on the study of statistical data, ignoring qualitative elements that could affect the classification of threat actors. This constrained method might not offer a thorough knowledge of danger actors and their behaviour.
4. It is difficult to gather and analyse accurate and relevant CTI since it necessitates access to a variety of sources, from closed, proprietary data to

open-source intelligence. Although CTI offers insightful data, it is not a perfect solution. To effectively mitigate risks in organisations, effective security controls and processes are still required.

4 Methodology

4.1 System Architecture

1. RAM: 8 to 16 GB minimum
2. CPU: Intel Core i7 or later is preferred as the CPU.
3. Storage: 128 GB or more minimum.
4. OS: Mac, Windows, and Linux
5. Internet: High-speed internet connectivity required.

4.2 Technology used

1. Python
2. Sckit-Learn
3. Pandas
4. Matplotlib
5. Jupyter Notebook/ Google Colab
6. Basic Knowledge Of Machine Learning And Data Science

4.3 Algorithms/ Techniques

Machine learning models may be used for categorization in a variety of ways. The two categories of these models are supervised and unsupervised, respectively. Logistic Regression, Decision Trees, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbor's Algorithm are a few examples of supervised machine-learning models. On the other hand, K-Means Clustering and Principal Component Analysis are two examples of unsupervised machine-learning methods. For the research we will be using Logistic Regression and K-Nearest Neighbor's Algorithm.

Logistic Regression A form of supervised machine learning algorithm is logistic regression. This statistical model simulates a binary dependent variable using a logistic function. When the target or the dependent variable is categorical, this is one of the most widely utilised strategies.

4.4 K-Nearest Neighbour's Algorithms

KNN (K-Nearest Neighbors) is a supervised machine learning algorithm that is used for classification and regression tasks. In order to forecast the label or value for the new data point, the KNN method locates the K data points (neighbours) that are in the training dataset that are the closest to the new input data point. The KNN method assigns the label that appears the most frequently among the K nearest neighbours in classification tasks. It determines the mean or median of the values of the K nearest neighbours while doing regression tasks[17].

Table 1
Literature Review: Summary Table

Author and Year	Contribution	Scope	Limitations
Roumen Trifonov et al., 2018 [13]	Use of AI in CTI(Automated Threat Detection, threat hunting, predictive analysis, NLP, Malware Detection and Incident Response)	The information obtained can be used to build a behaviour model	The lack of context and data quality is limited for training the model. AI based systems in CTI also raise ethical concerns related to privacy, bias, discrimination, etc.
Florian K. Kaiser et al., 2022 [14]	Exploring the potential of Cyber Threat Intelligence (CTI) to enable automated attack incident response and improve its efficiency and effectiveness.	A framework used to integerate CTI with incident response for rapid and effective response to cyber attacks.	Use of limited scope and not covering the full range. Effectiveness depends on the quality of data used for analysis. Implementation in real world scenarios might be techinally very complex.
Vasileios Mavroeidis et al., 2021 [15]	Defining a methodology using various machine learning algorithms for CTI and threat actor classification	Using a framework and to characterise threat actors based on their motivation, capability, and intent.	Methodology emphasises statistical data analysis and ignores qualitative factors that may affect threat actor characterization. Threat actors' behaviour may be incomplete with this approach.
Amira M. Aljuhami et al., 2021 [16]	Use of AI to implement CTI in an organisation. Establishing a CTI team, defining CTI objectives and requirements, developing a CTI collection and analysis plan, and integrating CTI into risk management process.	CTI collects, analyses, and shares cyberthreat and attacker data. CTI helps organisations identify, assess, and rank threats. CTI can also help organisations develop effective risk management plans by revealing attackers' strategies, tactics, and procedures (TTPs).	CTI requires access to open-source intelligence and closed, proprietary data, making it difficult to gather and analyse.CTI is useful but not perfect. Security controls and procedures are still needed to reduce risks.

Table 2
Pros & Cons: Logistic Regression

PROS	CONS
One of the most straightforward methods to implement.	Poor performance in the case of Unstructured Data (such as images, audio, videos)
No need for hyperparameter tuning.	Does not work well with features that are highly correlated with each other.
Usually effective in the case of Structured Data (such as tables and lists).	Not the best algorithm in context to power.
Can work with both Scaled and Unscaled Data. Feature Scaling is not necessary.	Assumes linearity between dependent and independent variables.

Table 3
Pros & Cons: K-Nearest Neighbour's Algorithm

PROS	CONS
Simple to understand and implement. To classify new data point KNN reads through whole dataset.	Does not scale well, takes up more memory and data storage.
A memory-based approach. Immediately adapts as we collect new training data. Allows the algorithm to respond quickly to changes in the input during real-time use.	Curse of dimensionality,does not perform well with high-dimensional data inputs.
Few hyperparameters(only requires k value and distance metric).	Prone to overfitting.

5 Implementation and Result

5.1 Data Preprocessing

Preparing the Dataset For this study, a wide variety of datasets from various regions of the Internet that are relevant to threat actors, malicious files, and benign files have been investigated. The final dataset was created by combining these datasets from several sources. Only the most pertinent or significant file-related attributes are included in this final dataset[18]. The dataset was then divided into two parts based on internal actors and external actors.

The final dataset has 9156 rows(data points) and 2443 columns(features).

Exploratory Data Analysis This phase entailed familiarising ourself with the dataset entirely . The dataset's many facets have been investigated. All of the significant characteristics or qualities in the dataset were found through this step. All of the unnecessary characteristics were removed from the dataset at the same time. To find trends or patterns, more analysis of the chosen characteristics

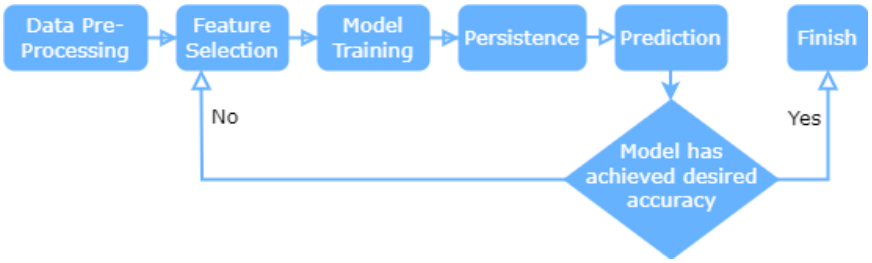

Fig. 2. Methodology Overview

Table 4
Internal Actors

by	enum	x	n	freq
actor.Internal	End-User	56	109.0	0.51376
actor.Internal	System Admin	14	109.0	0.12844
actor.Internal	Cashier	11	109.0	0.10092
actor.Internal	Other	10	109.0	0.09174
actor.Internal	Executive	6	109.0	0.055055
actor.Internal	Manager	6	109.0	0.055055
actor.Internal	Finance	3	109.0	0.02752
actor.Internal	Developer	2	109.0	0.01835
actor.Internal	Call Center	1	109.0	0.00917
actor.Internal	Doctor or Nurse	1	109.0	0.00917

Table 5
External Actors

by	enum	x	n	freq
actor.External	Organized Crime	743	2229.0	0.33333
actor.External	Unaffiliated	526	2229.0	0.23598
actor.External	Activist	486	2229.0	0.21803
actor.External	state-Affliated	221	2229.0	0.09925
actor.External	Former Employee	77	2229.0	0.03454
actor.External	Other	53	2229.0	0.02378
actor.External	Nation-State	51	2229.0	0.02288
actor.External	Customer	24	2229.0	0.01077
actor.External	Force Majeure	22	2229.0	0.00987
actor.External	Competitor	18	2229.0	0.00808

was done. Understanding the provided dataset and gaining as many insights as you can are crucial. Prior to delivering the data to the model for analysis, this is a very important step [19].

On further studying the data, we got the following results.

We further classified the actors based on external and internal and got the following results.

Table 6
Statistical representation of the dataset

	x	**n**	**freq**
count	120.000000	112.000000	84.000000
mean	84.083333	546.401786	0.071783
std	361.367933	886.988116	0.118319
min	0.000000	0.000000	0.0000000
25%	0.000000	0.000000	0.0000000
50%	1.000000	57.000000	0.021855
75%	17.250000	564.250000	0.099593
max	2846.000000	2229.000000	0.666670

Table 7
Actor Type Count

Actor	**Type**	**Count**
0	External	4722
1	Internal	4116
2	Partner	376
3	Unknown	212

Table 8
External Actor Type Count

External Actor	**Type**	**Count**
0	Organized Crime	743
1	Unaffiliated	526
2	Activist	486
3	State-Affiliated	221
4	Former Employee	77

Table 9
Internal Actor Type Count

Internal Actor	**Type**	**Count**
0	End-User	792
1	System admin	319
2	Other	235
3	Developer	125
4	Executive	108

The graphical representation of the dataset based on the number of attacks:

To be more accurate we decided to study the dataset based on the attacks caused by different internal and external actors.

After gathering all the results we performed the Train-test Split on the dataset

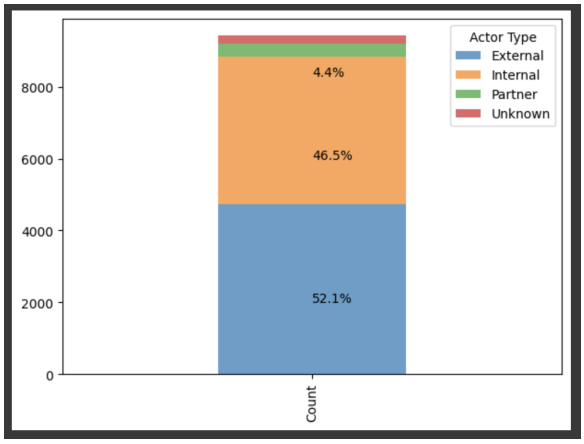

Fig. 3. Incidents of Actors

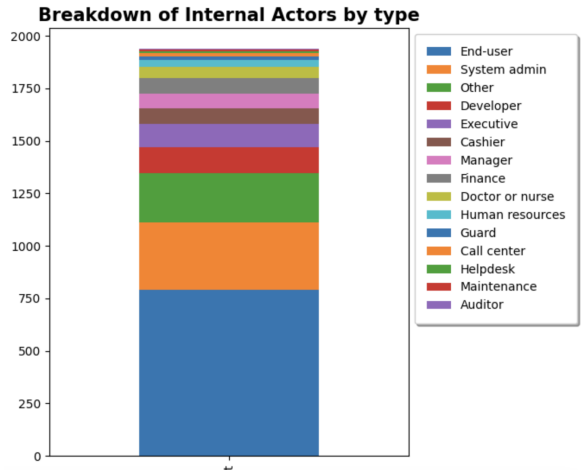

Fig. 4. Incidents of Internal Actors

5.2 Result

Imbalanced Data During data analysis and visualisation we observed that the data in the dataset was imbalanced. There is a very high likelihood that this points to some sort of abnormality, such as model overfitting. Imbalanced data issues can be solved in a variety of ways.

1. **Oversampling :** Using replacement, we enhance the proportion of occurrences pertaining to the Minority Class in this strategy. By doing this, the proportion of instances of the majority and minority classes is at its most equal.

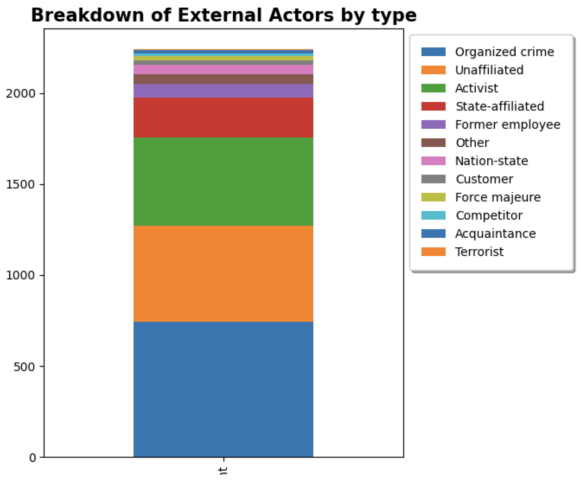

Fig. 5. Incidents of External Actors

2. **Undersampling :** Using this method, a number of rows pertaining to the Majority Class are randomly deleted. The outcome will be a dataset with about equal numbers of instances associated with the two output classes.
3. **Synthetic Minority Oversampling Technique (SMOTE) :** The Minority Class is likewise oversampled in this method. This technique uses pre-existing data to create new records or instances. The K Nearest Neighbour Method is used for this synthesis.
4. **Balanced Bagging Classifier :** This classifier enables extra balancing in addition to being a normal classifier. This method is an additional step in the process of balancing the training set while fitting the data.

We used the Oversampling technique to overcome the problem of imbalanced data.

5.3 Logistic Regression

On the training dataset, accuracy of 98.75% has been attained. On the other side, the Test Dataset has a 97.5% accuracy. The Test Dataset's F-1 Score is 0.9775.

5.4 K-Nearest Neighbours Algorithms

The Training Dataset yielded an accuracy of 98.75%. On the Test Dataset, however, an accuracy of 93% was attained. The F-1 Score of the Test Dataset is 0.93 .

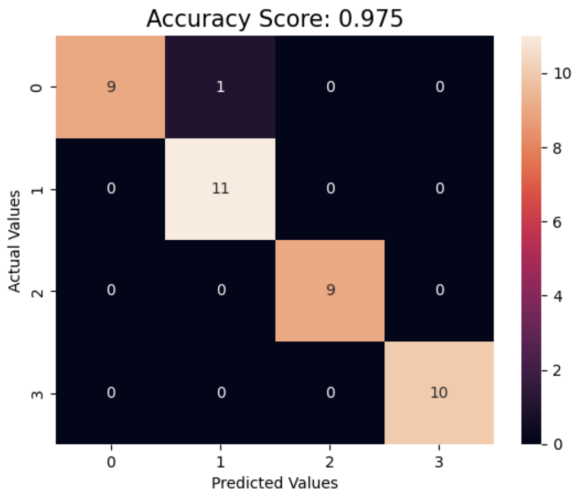

Fig. 6. Confusion Matrix: Logistic Regression

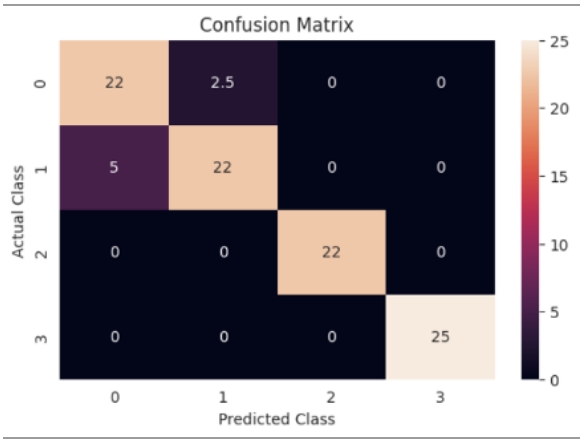

Fig. 7. Confusion Matrix: KNN

Comparison With Previous Work [15] Our work extends the work done in the research paper "Threat Actor Type Inference and Characterization within Cyber Threat" [15]. Work in this research paper provides a methodology for identifying and categorizing threat actors. The methodology is divided into three stages: data collection, data processing, and data analysis. The data collection stage involves gathering data from various sources, including open-source intelligence, social media, and dark web forums. The data processing stage involves filtering, cleaning, and structuring the data to enable analysis. The authors pro-

vide guidance on how to pre process the data, including techniques for data normalization, entity extraction, and relationship mapping.

The data analysis stage involves applying various techniques to identify patterns and characteristics of threat actors. Clustering and classification techniques to group similar threat actors together based on their characteristics, such as motivation, capability, and targeting were used.The results for the models were obtained by implementing this methodology.

Our work uses machine learning algorithms primarily Logistic Regression and KNN to implement the analysis of data to not just characterise the threat actors but also predict what type threat actor was behind the attack. Our dataset has been picked up from the VERIS Community Database[18]. It is a compilation of cyber threat instances from all around the world. The database consists of 9156 rows(data points) and 2443 columns(features).

The threat actors in "Threat Actor Type Inference and Characterization within Cyber Threat Intelligence" were categorised in six categories:- action-state, cyber criminal, hacktivist, insider, terrorist, and competitor. Our work has divided into mainly four types:- internal, external, partner and Unknown with external and internal threat actors having subcategories:

1. **External Threat Actors:** Organized Crime, Unaffiliated, Activist, State-Affiliated, Former Employee, Other, Nation-State, Customer, Force majeure, Competitor, Acquaintance, Terrorist
2. **Internal Threat Actors:** End-user, System Admin, Other, Developer, Executive, Cashier, Manager, Finance, Doctor Nurse, Human Resource, Guard, Call Centre, Helpdesk, Manintenance, Auditor.

5.5 Discussion

On the initially acquired dataset, two distinct machine learning models—Logistic Regression and K-Nearest Neighbor's Algorithm —were trained. Accuracy rates of 99.99% and 98.28% respectively. However, it was discovered that the initial dataset was quite unbalanced. A balanced dataset has been produced using a variety of methods. These include balanced bagging classifier, synthetic minority oversampling technique (SMOTE), undersampling, and oversampling. After balancing the dataset, the KNN algorithm and logistic regression were used.

The results of these two models on the balanced dataset were 97.5% and 93% respectively. The models implement the methodolgy stated in the previous research paper.

5.6 Challenges Faced

The following are some of the major challenges faced during the implementation:

1. The resulting dataset contained several problems, including rows with null values for various characteristics and an absence of standard data formatting throughout a particular column. Several of the data preprocessing techniques listed in section5.1 were used to solve these.

2. The trained models' first predictions were not particularly accurate. Multiple dataset modifications and model hyperparameter tweaks were used to address this issue. All of this allowed us to significantly increase the accuracy of the models.
3. The obtained dataset had a large number of attributes (columns) that were present in the categorical form. The items under these characteristics were transformed into numerical values using effective data encoding techniques so that they could be utilised to train the models.
4. In the course of the research, an imbalance in the dataset was discovered. As a result, the trained models were producing predictions that were more favourable to one class than the other. Utilising the oversampling approach, this was solved. The results of the method are discussed in section 5.2.

6 Tradeoffs

utilising machine learning models for threat actor categorization has a number of disadvantages compared to utilising traditional methods:

1. Utilising machine learning to categorise any type of threat actor takes much longer than utilising traditional techniques. This is because it takes time for the machine-learning algorithms to pick up new information from the extremely complex threat actor dataset.
2. For machine learning models to function effectively, a massive quantity of data comprising details about all potential threat actors in the globe is needed. These models could be unable to recognise new threat actors if they are presented with data that was not included in the training dataset.
3. Since it comprises data on the numerous criteria used to classify threat actors, threat actor datasets frequently contain a lot of information. This necessitates the employment of several resources.
4. When it comes to some categorization tasks, it's feasible that traditional approaches will outperform machine learning strategies in terms of accuracy.

7 Conclusion

The collected dataset was utilised in this study to train several machine learning models. Logistic Regression and K-Nearest Neighbor's Algorithm are two of the primary supervised machine learning models that were applied.

The Logistic Regression Model was shown to be the most accurate of all these models. For the test dataset, it was 97.5% accurate. This was followed by K-Nearest Neigbour's Algorithm (93%). Finally, a comparison between the findings of this investigation and earlier studies was done.

The research findings suggest that Cyber Threat Intelligence (CTI) has emerged as a critical element for organizations to safeguard their networks from potential threats and threat actors. CTI facilitates early detection through incident response and risk management, thus reducing the maintenance costs of networks.

The integration of machine learning in cyber security can prove highly advantageous for organizations as it minimizes the chances of human errors, improves efficiency, and provides real-time visibility into security incidents. By adopting the latest automation technologies, organizations can enhance the protection of their critical assets and respond promptly and efficiently to security threats.

8 Future Work

Many suggestions in the case of my project can be put into practise both immediately and in the long run. The main objective in the near future will be to continually test several machine learning models to find the model that best suits the demands. The prediction accuracy of each model used also has to be improved at the same time. Other factors that need to be taken into account include preventing multi-collinearity and model overfitting. The objective of these multiple processes will always be the utmost accuracy, and problems like multicollinearity and overfitting will be avoided at all costs.

A number of things can also be put into practise over the long run. For instance, the capability to recognise the threat actor type, the type of assault, and the intensity of the attack may all be integrated. Similar to that, it is also possible to incorporate the notion of figuring out what sort of malware the threat actor has employed into the project.

References

1. Santosh Jhansi Kattamuri, Ravi Kiran Varma Penmatsa, Sujata Chakravarty, and Venkata Sai Pavan Madabathula. Swarm optimization and machine learning applied to pe malware detection towards cyber threat intelligence. *Electronics*, 12(2):342, 2023.
2. Kelsie Nabben. Governance by algorithms, governance of algorithms: Human-machine politics in decentralised autonomous organisations (daos). *puntOorg International Journal*, 8(1):36–54, 2023.
3. Vamsikrishna Bandari. Enterprise data security measures: A comparative review of effectiveness and risks across different industries and organization types. *International Journal of Business Intelligence and Big Data Analytics*, 6(1):1–11, 2023.
4. J Jasmine Hephzipah, Ranadheer Reddy Vallem, M Sahaya Sheela, and G Dhanalakshmi. An efficient cyber security system based on flow-based anomaly detection using artificial neural network. *Mesopotamian Journal of Cybersecurity*, 2023:48–56, 2023.
5. Alok Bansal, Amit Kumar Bairwa, and Saroj Hiranwal. Security issues in cloud computing: A review. In *Proceedings of International Conference on Communication and Computational Technologies: ICCCT-2019*, pages 515–521. Springer, 2021.
6. Avishek Bose. *Learning representations for information mining from text corpora with applications to cyber threat intelligence*. PhD thesis, 2023.

7. Amit Kumar Bairwa and Sandeep Joshi. Mutual authentication of nodes using session token with fingerprint and mac address validation. *Egyptian Informatics Journal*, 22(4):479–491, 2021.
8. Kevin Morio, Ilkan Esiyok, Dennis Jackson, and Robert Künnemann. Automated security analysis of exposure notification systems. In *32st USENIX Security Symposium (USENIX Security 23)*, pages 1–18. USENIX Association, 2023.
9. Akshit Kamboj, Priyanshu Kumar, Amit Kumar Bairwa, and Sandeep Joshi. Detection of malware in downloaded files using various machine learning models. *Egyptian Informatics Journal*, 24(1):81–94, 2023.
10. Amit Kumar Bairwa and Sandeep Joshi. Mla-rpm: A machine learning approach to enhance trust for secure routing protocol in mobile ad hoc networks. *Int J Adv Sci Technol*, 29(04):11265–11274, 2020.
11. Amit Kumar Bairwa and Sandeep Joshi. Mutual authentication of nodes using session token with fingerprint and mac address validation. *Egyptian Informatics Journal*, 22(4):479–491, 2021.
12. Jeril Kuriakose, Sandeep Joshi, and Amit Kumar Bairwa. Embn-manet: A method to eliminating malicious beacon nodes in ultra-wideband (uwb) based mobile adhoc network. *Ad Hoc Networks*, 140:103063, 2023.
13. Roumen Trifonov, Ognyan Nakov, and Valeri Mladenov. Artificial intelligence in cyber threats intelligence. In *2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC)*, pages 1–4, 2018.
14. Florian K. Kaiser, Leon J. Andris, Tim F. Tennig, Jonas M. Iser, Marcus Wiens, and Frank Schultmann. Cyber threat intelligence enabled automated attack incident response. In *2022 3rd International Conference on Next Generation Computing Applications (NextComp)*, pages 1–6, 2022.
15. Vasileios Mavroeidis, Ryan Hohimer, Tim Casey, and Audun Jesang. Threat actor type inference and characterization within cyber threat intelligence. In *2021 13th International Conference on Cyber Conflict (CyCon)*, pages 327–352, 2021.
16. Amira M. Aljuhami and Doaa M. Bamasoud. Cyber threat intelligence in risk management. *International Journal of Advanced Computer Science and Applications*, 12(10), 2021.
17. Basyir Al Musthoqfirin Majid, Abdul Mubarak, and Salkin Lutfi. Classification of device addiction to students using sas-sv with k-nearest neighbor algorithm method. *Journal of Computer Engineering, Electronics and Information Technology*, 1(2):43–50.
18. VERIS. Veris community database. Accessed: 2023-04-01.
19. Guangjin Wang, Bing Zhao, Bisheng Wu, Chao Zhang, and Wenlian Liu. Intelligent prediction of slope stability based on visual exploratory data analysis of 77 in situ cases. *International Journal of Mining Science and Technology*, 33(1):47–59, 2023.