A Cognitive Cyber Threat Intelligence: Harnessing the Power of Big Data Analytics for Advanced Cyber Security

A. Kanthimathinathan
Research Scholar
Dept. of Computer Science and Engineering
Annamalai University
Annamalainagar – 608002
Chidambaram, Tamilnadu
Email: kanthi_88@yahoo.co.in
Dr.S.Saravanan
Assistant Professor
Dept. of Computer Science and Engineering
Annamalai University
Annamalainagar – 608002
Chidambaram, Tamilnadu
Email: ssaravau@gmail.com
Dr. G. Ramachandran
Associate Professor
Dept. of Computer Science and Engineering
Annamalai University
Annamalainagar – 608002
Tamilnadu, India
Email: gmrama1975@gmail.com

ABSTRACT

Cyber Threat Intelligence (CTI) is the process of collecting, analyzing, and utilizing information about potential cyber threats to an organization. The goal of CTI is to provide organizations with the knowledge and understanding, it needs to prevent, detect, and respond to cyber attacks. CTI involves collecting and analyzing data from a variety of sources, including Open-Source Intelligence (OSINT), social media, and specialized intelligence feeds. The data is then used to create a comprehensive view of the current threat landscape, including information on the Tactics, Techniques, and Procedures (TTPs) used by attackers, as well as the types of attacks and vulnerabilities that are most exploited. This research paper proposes a novel approach to CTI by leveraging the power of big data analytics. The proposed approach, called CogCyber, a Cognitive Cyber Threat Intelligence (CCTI), integrates machine learning and natural language processing techniques to collect, analyze, and visualize massive amounts of structured and unstructured data from various sources, including social media, forums, blogs, news feeds, and dark web. By analyzing and correlating this data, CCTI can provide more accurate and timely threat intelligence, identify emerging threats, and support proactive defense strategies. The paper presents a detailed architecture of the CCTI framework, including data collection, pre-processing, feature extraction, modelling, and visualization. The effectiveness of the proposed approach is evaluated through a series of experiments on real-world datasets, demonstrating significant improvements in threat detection and response. This research contributes to the development of advanced Cyber security solutions that can cope with the growing complexity and sophistication of modern cyber threats. Organizations can use this information to improve their overall security posture, prioritize security investments, and respond more effectively to threats. Hence, this work proposes an integrated CTI architecture which can also be used to develop proactive defense strategies and enhance incident response capabilities, helping organizations to better manage the risks posed by cyber threats.

Keywords - Information gathering, Cyber Threat Intelligence (CTI), Open-Source Intelligence (OSINT), Incident Response, Cyber-attacks, and Security.

1. INTRODUCTION

In recent years, cyber attacks have become more frequent, sophisticated, and destructive. As the number of devices and systems connected to the internet continues to grow, so does the number of potential attack vectors. The need for effective cyber security solutions has become increasingly important to protect organizations from the growing threat of cyber attacks. In response to this need, CTI has emerged as a vital tool for organizations to proactively identify, analyze, and respond to cyber threats.

CTI is a process of gathering, analyzing, and disseminating information about potential or actual cyber threats to an organization. It provides valuable insights into the tactics, techniques, and procedures used by threat actors, as well as the Indicators of Compromises (IoCs) that can be used to detect and mitigate cyber attacks. However, the sheer volume and velocity of data generated by modern information systems and networks make it difficult for organizations to effectively collect, process, and analyze CTI. This is where big data comes in. big data refers to the large and complex sets of data that are generated by modern information systems and networks. The volume, velocity, and variety of data generated by these systems make it difficult for traditional data processing techniques to effectively analyze and extract insights from the data. However, with the emergence of big data technologies, it is now possible to store, process, and analyze vast amounts of data in real-time.

In this research paper, it was proposed a novel approach that combines CTI and big data to enable organizations to effectively detect, analyze, and respond to cyber threats. This will explore how big data technologies such as Hadoop, Spark, and NoSQL databases can be used to store, process, and analyze large volumes of CTI data. This will also discuss how machine learning algorithms and data mining techniques can be applied to this data to extract valuable insights and improve the accuracy and speed of threat detection.

1.1 Cyber Threat Intelligence (CTI): A Preliminaries

Cyber Threat Intelligence (CTI) is the process of collecting, analyzing, and sharing information about potential cyber threats and risks to an organization's IT infrastructure, assets, and systems. According to market research, the global cyber security industry's revenue exceeded $173.5 billion in 2022 and is projected to increase at a Compound Annual Growth Rate (CAGR) of 8.9% to exceed $266.2 billion by the end of 2027. As organisations continue to prioritise cyber security measures to reduce the risks caused on by cyber attacks, this may contribute to a considerable rise in the industry [1]. Recent years have seen a significant rise in the complexity and sophistication of cyber threats [2]. Attackers are continuously looking at new attack methods, from gaining victims to run infected files to making use of zero-day vulnerabilities to their benefit. Attackers are also creating and exploiting different methods to hide identities within trustworthy and efficient processes [3][23].

CTI, in which security researchers gather and analyse from a variety of sources to handle these developing threats, it is a massive body of knowledge. Security experts can learn about the danger landscape and find probable threat detection hints by combining data from several sources, which may not reveal them separately [4]. Through the early detection and mitigation of potential risks, CTI is a crucial component of a comprehensive cyber security strategy that aids organisations in staying one step ahead of cybercriminals. Data for CTI is gathered from a variety of sources, including internal network logs, dark web sources, open-source intelligence, and other proprietary data sources.

Then, using this data, patterns, trends, and indicators of possible dangers are found and used to design preventive measures and reaction strategies. The main objective of CTI is to give organisations the information and resources they need to more thoroughly comprehend the risks they face, spot possible dangers, and take preventative action against attacks. This could be putting in place improved security measures, teaching staff about cyber security best practises, or creating incident response strategies to swiftly identify and eliminate threats [5]. In order to stay current with the

most recent threats and vulnerabilities, CTI is a continual process that needs constant monitoring and analysis. It is an essential part of any successful cyber security strategy and can assist organisations in staying one step ahead of cybercriminals in a threat environment that is always changing.

Cyber attacks take advantage of weaknesses in software, computing hardware, or user behaviour. For it to be understandable and useful, CTI must offer context and a course of action [22]. CTI services manage the pertinent data to battle cyber risks while tracking, analysing, interpreting, and mitigating, continuously increases cyber security threats and attacks. Investigation, detection, and repair efforts are further complicated by shifting threat patterns. Attacks are becoming more dynamic, enabling complex adversarial attacks, despite the development of improved security measures [6]. This demonstrates the necessity for increased security measures and the sharing of relevant information for prevention through a better understanding of potential risks. CTI expedites security operations, prioritises the implementation of security controls, and assists organisational detection and knowledge of current attacks.

Adopting an intelligence driven method possesses its primary goals to avoid the occurrence of successful attacks, the identification of appropriate responses to incidents, and the gathering of technical data about attacks for better understanding and attribution of the Tactics, Techniques, And Procedures (TTPs), attack principal, and attack motives [7] [24]. This makes it possible to have proactive, rapid and intelligence driven cyber resilience. To be effective, CTI must disseminate information about current or potential dangers that is pertinent, timely, accurate, and practical. To counteract cyber risks, risk-based mitigation strategies are implemented, then detection strategies are developed to lessen the effects of attacks as soon as possible [8][9][25]. Several phases make up the CTI life cycle, which enables organisations to recognise, evaluate, and respond to cyber threats in an efficient manner. The typical phases of the CTI life cycle are depicted in Figure 1. Each stage of the CTI life cycle informs the one after it, and feedback from the assessment stage is used to improve the programme and increase its efficacy over time.

Figure 1 CTI Life Cycle

1.2 Problem Statement

The ever changing environment of cyber security threats and attacks necessitates a comprehensive method to gather, analyse, and distribute knowledge in a timely and useful manner, according to the problem statement of the cyber threat knowledge framework. In order to prevent, detect, and respond to cyber incidents, organisations must be proactive in detecting possible risks, comprehending their strategies and objectives, and putting them into practice. Lack of an organised approach to CTI can lead to knowledge gaps and inefficient resource use, leaving organisations open to threats and attacks via the internet. In order to improve an organization's cyber security posture, there is a need for a framework that offers a systematic approach for the collection, analysis, dissemination, and use of CTI.

1.3 Motivation of the Work

The CTI framework was created with the intention of providing organisations a structured approach for collecting, analysing, and disseminating data regarding potential cyber threats. Organisations require a proactive strategy to cyber security that goes beyond conventional security measures like firewalls and antivirus software in light of the rising number of cyber attacks and the constantly evolving nature of cyber threats. Organisations can recognise possible risks, comprehend the plans and objectives of attackers, and create efficient ways to stop, and respond to cyber attacks with the use of CTI. Organisations may improve the security of their digital assets and reduce the risk of data breaches and other cyber incidents by using a CTI framework.The main contribution of this work is as follows:

- Improving Threat Awareness: The CTI framework aids organizations in staying informed about the newest and most dangerous cyber threats. Organizations can comprehend the strategies, methods, and processes used by threat actors, their objectives, and the possible impact of these threats by gathering and analyzing threat intelligence. They can proactively guard against possible cyber attacks.
- Improved threat detection: By leveraging big data technologies such as machine learning and data mining, the framework is able to detect patterns and anomalies in vast amounts of data that would otherwise be impossible for human analysts to process. This enhances the accuracy and speed of threat detection, reducing the risk of cyber attacks and minimizing the damage caused by successful attacks.
- Real-time threat intelligence: The framework is designed to provide real-time threat intelligence, enabling organizations to quickly respond to potential threats and minimize the time window during which attackers can cause damage. This can be particularly valuable in fast-moving industries such as finance or healthcare, where any delay in responding to a cyber attack can have severe consequences.
- Enabling Proactive Defence: By foreseeing risks before they materialize, the CTI framework enables organizations to adopt a proactive stance towards security. Organizations can discover possible risks and vulnerabilities in their networks and systems and take action to resolve them before an attack by analyzing threat intelligence data.
- Supporting Incident Response: By giving organizations the knowledge they need to react appropriately to a cyber attack, the CTI framework helps incident response. Organizations can swiftly determine the origin of an attack, the scope of the damage, and the necessary actions to contain and repair the incident by having a thorough understanding of the threat landscape.
- Facilitating Risk Management: By providing intelligence data to assist in decision-making, the CTI framework enables organizations to efficiently manage cyber risks. Organizations may prioritize their security investments and make educated decisions to successfully manage cyber risk by evaluating the possible hazards and consequences of cyber threats.
- Improved incident response: The framework provides security teams with a centralized platform for incident response, allowing them to collaborate and share information more easily. This can improve the speed and effectiveness of incident response, reducing the time required to investigate and remediate security incidents.
- Scalability: The framework is designed to be scalable, allowing it to handle large amounts of data and adapt to changing threat landscapes. This means that it can be applied to a wide range of industries and organizations, from small businesses to large enterprises.

2. RELATED WORKS

To increase the precision and effectiveness of CTI and response, the author [10] suggests a system that blends deep learning and data fusion approaches. The three key components of the suggested framework are data gathering, data fusion, and classification using deep learning. Data from multiple sources, including social media, the dark web, and honeypots, are gathered and preprocessed in the data collection step to remove noise and irrelevant data. The preprocessed data are joined using a data fusion algorithm in the data fusion step to produce a unified view of the cyber threat landscape. The data fusion method prioritizes the data for analysis by taking into an account, a number of factors, including source dependability, timeliness, and relevancy.

A method to increase the explainability of machine learning models used in CTI is suggested by the author [11]. According to the authors, increasing the explainability of machine learning models can aid analysts in understanding how a model arrived at a specific conclusion, which can enhance the precision and utility of CTI. In order to better understand the problems and potential for enhancing the

explainability of machine learning models in this field, the authors first survey the body of research on explainable machine learning in CTI. The authors then put out a paradigm for incorporating explainability into CTI that entails the crucial processes of feature selection, model training, assessment, and model improvement.

To improve data mining methods for CTI, according to the authors [12], information that can be utilized to recognize possible threats can be extracted from the data, gathered from different sources using data mining techniques. This strategy makes use of machine learning methods for data preprocessing, feature selection, and classification. The modified Naive Bayes algorithm is the new approach that the authors suggest to classify the data. The method was evaluated using a dataset of cyber threat indicators, and the findings demonstrated that it performed better in terms of accuracy and precision than previous classification algorithms.

The research paper [13] compares the features of different CTI sharing platforms. This paper highlights the necessity of CTI sharing and the difficulties of the businesses experience while putting such platforms in a place. Based on elements including their features, data sources, data sharing strategies, and integration process with other security solutions, the authors assess several CTI sharing platforms. [14] Talks about the benefits, difficulties, and several sharing methods and frameworks that are now in use for CTI sharing. According to this study, CTI sharing can have a big impact on enhancing cyber security, but there are still a lot of obstacles to overcome, including those related to law and regulation, trust and privacy issues, and technical compatibility.

The K-CTIAA approach offers a knowledge graph based automatic analysis methodology for CTI. The system is made to gather, archive, and analyze CTI from a variety of sources, including news articles, social media, and forums on the dark web. The CTI data is then incorporated into a knowledge graph to capture dependencies and relationships among various CTI domain components. In an interactive dashboard, the authors' suggested rules and algorithms for analyzing the knowledge graph and spotting potential dangers are displayed to analysts [15].

Using unstructured text sources including news, stories, blogs, social media posts, and other textual data, the Vulcan system is an automated platform for gathering and analyzing CTI. The system employs machine learning algorithms and NLP techniques to discover and extract pertinent threat intelligence data from huge amounts of text data [16]. The Vulcan system is made to carry out the following actions automatically: Such as entity extraction, event detection, threat actor identification, and threat intelligence analysis are some of the data collecting and preprocessing techniques used. The CTI frameworks, APT threats [18] methodologies, and tools are important for gathering and analyzing a large amount of threat-related data from multiple sources in order to spot new trends and patterns. Key conclusions from CTI research include the following:

- CTI is a crucial part of any cyber security strategy, offering insightful information about the current threat landscape and assisting organizations in identifying potential risks and vulnerabilities.
- CTI frameworks and tools are constantly evolving to keep up with the constantly shifting threat landscape. The accuracy and speed of threat analysis are being improved by researchers employing cutting-edge techniques like machine learning and data mining.
- CTI is a collective endeavor, and the sharing of threat intelligence among organizations is crucial to enhancing overall security posture. Legal and regulatory obstacles, data privacy issues, and a lack of standardisation make it difficult to share CTI.
- Explainable AI (XAI) is becoming more and more crucial in CTI research to shed light on the justifications for AI-driven decisions and to guarantee the accountability and transparency of CTI tools.

Table 1 Techniques, Parameters Measured and the Limitations of Related Work

Techniques Used	Parameters Measured	Limitations
Deep Learning, Data Fusion [10]	CTI Accuracy, CTI Latency, False Positives, False Negatives	Limited to specific types of cyber threats, such as malware, phishing, and DDoS attacks
Explainable Machine Learning [11]	Explainability, CTI Accuracy	Specific machine learning algorithms and use cases
Data Mining [12]	CTI Accuracy, CTI Latency, False Positives, False Negatives	Limited to structured data sources and may not capture all types of cyber threats
Review of Cyber Threat Intelligence Sharing Platforms [13]	CTI Sharing Mechanisms, Security and Privacy Controls	Limited to existing platforms and may not consider all available options
Systematic Literature Review [14]	CTI Sharing Challenges, CTI Sharing Benefits, CTI Sharing Best Practices	Limited to the published literature and may not capture all current CTI sharing practices
K-CTIAA Tool [15]	Precision and recall, Mean Reciprocal Rank (MRR) and Hit@k, Execution time	May not capture all current CTI sharing practices
Vulcan automated tool [16]	Extraction Time, Data Sources, Performance, Scalability	It is limited to the threats on the software

Table 1 says the list some of the related work with techniques, parameters measured and limitations. A tool for CTI called cyber active [17] makes it possible to include CTI into intricate models. It is based on the STIX format, which is used to represent and communicate information about cyber threats. In order to extract pertinent data from unstructured text sources like social media, news articles, and other internet sources, the application employs NLP algorithms. A knowledge graph depicts the connections between different entities in the cyber threat landscape. It is created using the retrieved data, which is subsequently transformed into STIX format. The overview of the most popular CTI protocols may be found in Table 2. Organisations frequently utilise these protocols to exchange and analyse cyber threat intelligence, improving threat detection and response.

Table 2 Cyber Threat Intelligence (CTI) Protocols and its Descriptions

Protocol	Description
STIX [18]	The XML-based language known as Structured Threat Information eXpression (STIX) is used to communicate threat intelligence. A standard for defining cyber threat intelligence, including malware, campaigns, and incidents is provided by STIX.
TAXII	A mechanism for exchanging cyber threat intelligence is called Trusted Automated eXchange of Indicator Information (TAXII). Organisations can communicate cyber threat intelligence in a secure, automated fashion by using the services defined by TAXII.
OpenIOC [19]	An XML-based standard called Open Indicators of Compromise (OpenIOC) is used to share threat intelligence. A common language for describing indicators of compromise, such as file hashes, IP addresses, and domain names, is provided by OpenIOC.
CybOX	Cyber observables eXpression CybOX that can be seen in network traffic, system logs, and other sources of security data. Cyber Observable eXpression (CybOX) is an XML-based language for defining these objects. It is a standardised approach to describe cyber observables, such as file objects, network connections, and email messages.
MISP	A platform for exchanging threat intelligence is called the Malware Information Sharing

	Platform (MISP). MISP offers a suite of tools for gathering, archiving, and disseminating cyber threat intelligence, such as malware samples, indicators of compromise, and threat actor data.
IODEF	Information concerning security incidents can be sent using the Incident Object Description Exchange Format (IODEF), an XML-based format. IODEF offers a common terminology for describing security incidents, together with information on their timeframe, impact, and affected assets.
CIF [20]	An open-source framework for gathering and disseminating information on cyber threats is called Collective Intelligence Framework (CIF). CIF compiles threat intelligence from various sources and offers a selection of tools for data analysis and visualisation.
CRITs	An open-source platform called Collaborative Research Into Threats (CRITs) is used to gather, evaluate, and share information about cyber threats. CRITs offer a set of technologies for gathering, storing, and correlating threat intelligence, such as malware samples, indicators of compromise, and threat actor data.

Depending on the organization's size, cyber security requirements, and resources available, different types of CTI and Cyber Intelligence (CI) are used. Different strategies, including conventional, centralised, and dispersed strategies, have been put out and employed.

Table 3 Cyber Threat Intelligence and Cyber Intelligence Implementation Approaches

Cyber Intelligence & Implementation Approaches	**Description**
Risk-Based	Determining the risk profile of an organisation and putting intelligence measures in place based on the dangers found. Critical assets must be identified, and the possibility and impact of cyber attacks on those assets must be evaluated.
Defense-in-Depth	Setting multiple security layers to defend against cyber attacks. By using tools like firewalls, intrusion detection systems, and incident response plans, it combines preventive, detective, and reactionary measures.
Proactive	This strategy involves proactively addressing cyber security by spotting and reducing possible threats before they can be taken an advantage. To find and fix vulnerabilities, it entails putting continuous monitoring, vulnerability scanning, and penetration testing into practise.
Intelligence-Driven	Intelligence to inform cyber security decisions and strategies. It involves collecting, analyzing, and disseminating intelligence to relevant stakeholders, and using that intelligence to inform security measures and responses.
Indicator-Based	Threat indicator identification is necessary to recognise and prevent cyber attacks. To find possible threats, it entails gathering, analysing, and exchanging threat intelligence indications including IP addresses, domain names, and hashes.
Behavioural Analytics	This method examines system and user behaviour to find any irregularities that might point to a cyber-attack. To find patterns of behaviour that differ from typical behaviour, machine learning techniques and statistical analysis are used.
Threat Hunting	Actively looking for threats that conventional security procedures may have missed. In order to look for signs of compromise and potential vulnerabilities, it makes use of both human skills and cutting edge equipment.
Fusion Center	Integration of threat intelligence data from many sources to provide a centralised hub for analysis and sharing. It entails gathering, analysing, and

	sharing intelligence with the appropriate stakeholders using automation and advanced analytics.

Table 3 presents some of the manual methods used in the traditional approach, which can be time consuming and frequently leads to assaults being responded. The centralised method makes it simpler to manage and distribute information within the organisation by using a centralised platform to collect, analyse, and disseminate CTI [20]. The distributed strategy uses a variety of CTI sources that are combined and analysed to give a thorough picture of the threat landscape [21].

3. COGCYBER: THE PROPOSED ARCHITECTURE

The architecture depicted in Figure 2 represents an automated Cyber Threat Intelligence (CTI) system, showcasing the collection, processing, analysis, and dissemination of threat intelligence in a streamlined manner. This automated CTI system utilizes various applications and technologies to gather threat intelligence data from multiple sources, such as social media, Open-Source Intelligence (OSINT) feeds, network traffic logs, and system event logs. These sources provide valuable information that helps in understanding potential threats and risks. Once the data is collected, it undergoes a series of processing steps to extract relevant and actionable information. This involves employing sophisticated techniques like data mining, Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL) to analyze and make sense of the collected data. These approaches help in identifying patterns, detecting anomalies, and recognizing potential dangers within the vast amount of information gathered. The automated CTI system depicted in Figure 2 also incorporates data ingestion from various security tools and technologies

Security Information and Event Management (SIEM) systems collect and aggregate security event data from various sources within an organization's network infrastructure. Integrating SIEM data into the CTI system allows for comprehensive threat intelligence analysis by incorporating valuable information from security events, alerts, and incident logs.

Intrusion Detection Systems (IDS) and Intrusion Prevention Systems (IPS) systems monitor network traffic for suspicious activities and potential intrusions. Ingesting data from IDS and IPS logs provides valuable insights into detected threats, attack patterns, and attempted intrusions, enabling proactive threat intelligence analysis.

Firewalls are critical security components that monitor and control network traffic based on predefined security rules. Ingesting firewall logs into the CTI system enables the identification of potential malicious activities, unauthorized access attempts, or suspicious traffic patterns that may indicate an ongoing attack.

Malware Detection systems such as antivirus and anti-malware systems help identify and prevent the execution of malicious software within an organization's infrastructure. Integrating data from these units into the CTI system allows for the analysis of malware trends, detection rates, and patterns of malicious behaviour.

Forensics tools are used to investigate security incidents, analyze compromised systems, and gather evidence. Ingesting data from forensics tools, such as memory dumps, disk images, or network captures, provides valuable contextual information for threat intelligence analysis, allowing for the identification of attack vectors and understanding the tactics employed by threat actors.

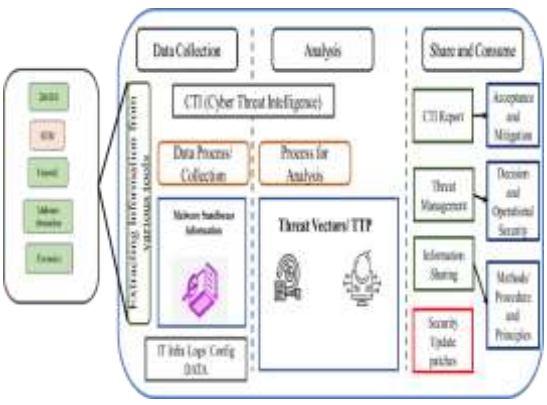


Figure 2 Proposed Framework of CTI

CTI tools and techniques are used to identify, analyze, and respond to cyber threats. Table gives some examples of CTI tools and techniques, as well as descriptions of some common cyber-attacks. Table 4 list the automated tools and techniques used to detect the vulnerabilities on the organization.

Table 4 Some of the Tools, Techniques, and Attack Detection of CTI

Tools	Techniques	Attack Examples
Maltego	Open-Source Intelligence (OSINT)	Phishing
Shodan	Dark web monitoring	Ransomware
VirusTotal	Signature-based detection	Distributed Denial of Service (DDoS) attacks
Snort	Network-based intrusion detection	Man-In-The-Middle (MITM) attacks
Bro	Network security monitoring	Advanced Persistent Threats (APTs)
Yara	Rule-based detection	Malware infections
FireEye	Threat intelligence platform	Spear-phishing
IBM X-Force	Security research and insights	Zero-day exploits
CrowdStrike	Endpoint detection and response	Fileless attacks

4. EXPERIMENTAL SETUP AND REAL TIME RESULTS

The experimental setup involves the integration of CTI and big data technologies to build an effective Security Operations Center (SOC) that can detect and respond to cyber threats in real-time. The SOC is built using open source big data tools such as Apache Hadoop, Apache Spark, and Apache Kafka, and CTI tools such as MISP, OpenCTI, and TheHive. The SOAR component of the SOC is built using open source tools such as Cortex, Elasticsearch, and Kibana. The experimental setup consists of the following components:

1. Data Ingestion: In this component, the data from various sources such as network traffic logs, system logs, and security logs is collected and ingested into the big data platform using Apache Kafka. The data is then stored in Apache Hadoop Distributed File System (HDFS) for further processing. Table 5 gives the illustration of data ingestion modules and its descriptions.

2. Data Processing: In this component, the ingested data is processed using Apache Spark to perform data cleaning, normalization, and feature engineering. The processed data is then stored in HDFS for further analysis.

3. Threat Intelligence Integration: In this component, the processed data is correlated with external threat intelligence data using MISP and OpenCTI. The threat intelligence data is then used to enrich the processed data with contextual information about known threats, vulnerabilities, and Indicators of Compromise (IOCs).

4. Machine Learning (ML) Models: In this component, ML models such as Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Support Vector Machine (SVM) are trained on the enriched data to detect and classify cyber threats.

5. Real-time Threat Detection: In this component, the trained ML models are used to detect and classify cyber threats in real-time. The results of the ML models are then fed into the SOAR component of the SOC for automated response.

6. Automated Response: In this component, the SOAR component of the SOC, built using Cortex, Elasticsearch, and Kibana, responds to detected threats by taking automated actions such as blocking IP addresses, quarantining systems, and sending alerts to security personnel.

Table 5 Data Sources and Technologies

Data Sources	Description
Social Media	Gather threat intelligence data from various social media platforms.
Open-Source Intelligence (OSINT) Feeds	Collect publicly available information and intelligence feeds.
Network Traffic Logs	Capture and analyze network traffic data to identify potential threats.
System Event Logs	Monitor system events and logs to detect security-related activities.
SIEM	Ingest security event data from Security Information and Event Management systems.
IDS	Collect intrusion detection system logs for analyzing detected threats.
IPS	Capture intrusion prevention system data to identify and prevent network intrusions.
Firewalls	Extract logs and data from firewalls to analyze network traffic and security events.
Malware Detection Units	Integrate antivirus and anti-malware systems to analyze malware detection data.
Forensics Tools	Utilize forensics tools to investigate security incidents and gather evidence.

4.1 Testing and Validating the Experimental Setup

The experimental setup is evaluated using real-world datasets containing network traffic logs, system logs, and security logs. The performance of the ML models is evaluated using metrics such as accuracy, precision, recall, and F1-score. The effectiveness of the SOAR component of the SOC is evaluated based on its ability to automate the detection and response process, reduce the response time, and improve the overall security posture of the organization.

To test the effectiveness of the experimental setup, a series of simulated attacks were conducted. Threat intelligence data was ingested into MISP and OpenCTI, and rules were created to trigger automated responses in Splunk Phantom & TheHive through Cortex. The logs generated by the different tools were collected and analyzed using Elasticsearch, Apache Hadoop, and Apache Spark.

To test the CTI and SOAR setup, it was created a use case for detecting and responding to a phishing email. The following process is carried out:

- Ingested the email logs into the big data platform using Apache Kafka.
- Then it was processed the ingested email logs using Apache Spark to perform data cleaning, normalization, and feature engineering.
- The processed data was then stored in Apache Hadoop Distributed File System (HDFS) for further analysis.

After the above process, the processed email logs with external threat intelligence data using MISP and OpenCTI was correlated. The threat intelligence data was then used to enrich the processed data with contextual information about known threats, vulnerabilities, and Indicators of Compromise (IoCs). Then the ML models were trained such as Random Forest and Navie Bayes (NB) on the enriched data to detect and classify phishing emails. In addition, we used the trained ML models to detect and classify phishing emails in real-time. The results of the ML models were then fed into the SOAR component of the SOC for automated response. Finally, a robust setup is deployed with an automated response in Splunk Phantom & TheHive through Cortex to respond, to detected phishing emails by taking automated actions such as blocking email addresses, quarantining emails, and sending alerts to security personnel.

The results of the simulated attacks and the use case demonstrated the effectiveness of the CTI and SOAR setup in detecting and responding to cyber threats in real-time. The combination of CTI and Big Data technologies enabled us to

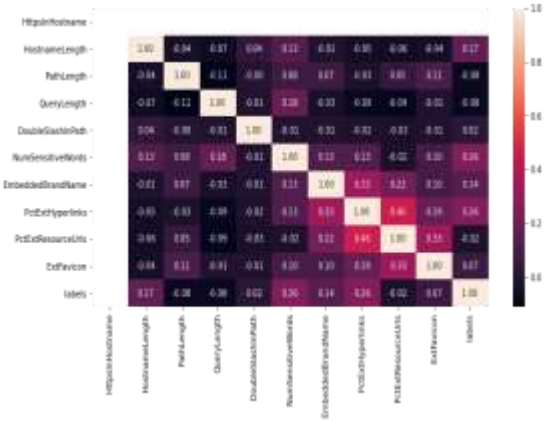
collect, process, and analyze large amounts of data from various sources, and uses the information to identify and respond to cyber threats in a timely and efficient manner.

4.2 Metrics to Measure the CTI

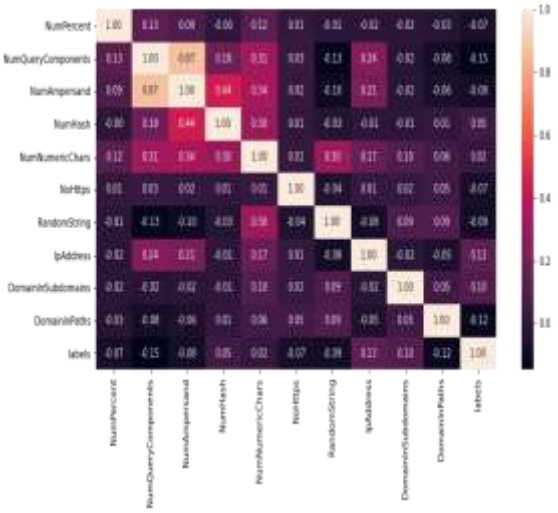
The following measures can be used to evaluate the rate at which cyber threat intelligence is working:

1. Time to detect: The amount of time it takes to identify a threat or an attack can be a good indicator of how effective a program. This metric track how long it takes for a threat to be discovered after it first manifests.
2. Rate of false positives: The system's accuracy can be determined by the number of false positives the CTI programme produces. A high false positive rate indicates an excessive number of alerts that end up being false alarms, wasting resources and adding to the workload of analysts.
3. The number of incidents that were successfully resolved: This indicator assesses how well the CTI programme handled incidents. A programme is more effective if there are more incidents that are successfully resolved.
4. Threat coverage: The range of threats the CTI programme covers might be a good indicator of its efficacy. It is more beneficial to have a programme that addresses a variety of risks and vulnerabilities than one that concentrates on just a few.
5. Actionable intelligence: The quantity of alerts that result in actionable intelligence can be a good indicator of how well the CTI programme is working. Information that can be used to stop an attack or lessen its effects is known as actionable intelligence.
6. Response Time: The amount of time it takes to respond to an incident might be a good indicator of how well the CTI programme is working. A programme is more efficient if it responds more quickly.
7. Cost-effectiveness: The CTI program's cost-effectiveness can be determined by weighing the advantages it provides against the costs associated with creating and sustaining the programme.

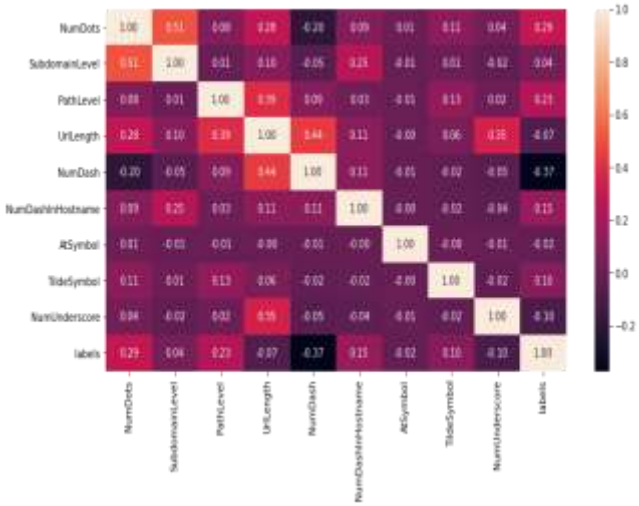
These indicators can be used to assess a program's effectiveness and pinpoint areas for development in CTI.



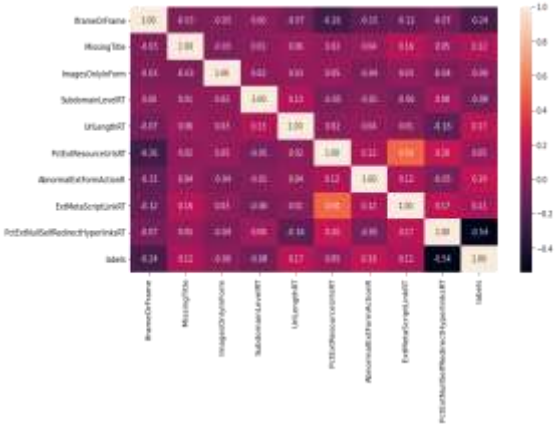
a) Confusion matrix for the classified host features



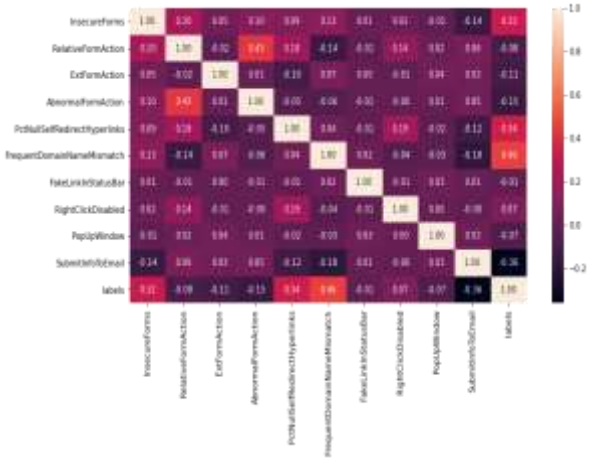
b) Confusion matrix for the classified URL data



c) Confusion matrix for the classified URL data (subdomain details)



d) Confusion matrix for the URL features



e) Confusion matrix for the URL Form data

Figure 3 Confusion matrixes for the overall data

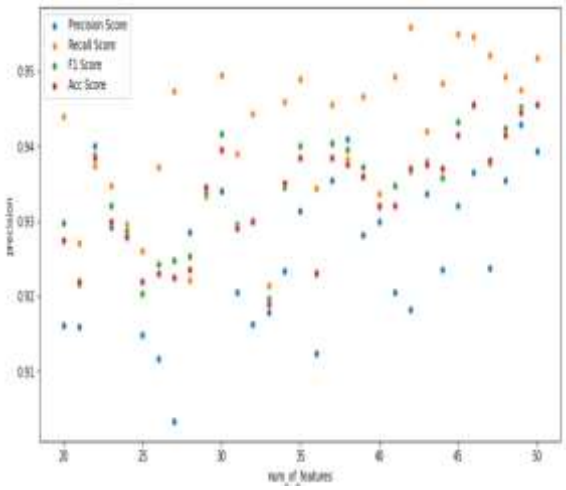


Figure 4 Performance Metric for the (NB) ML Model

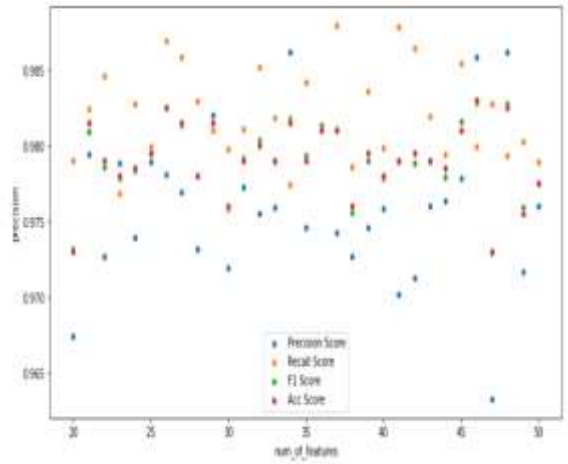


Figure 5 Performance Metric for the (RF) ML Model

Figure 3 represents the confusion matrix for the overall data. The performance metric of a ML model with respect to features considered is an important aspect that must be evaluated before the model can be deemed useful. The experimental setup involving the integration of CTI and big data technologies to build an effective Security Operations Center (SOC) that can detect and respond to cyber threats in realtime. The performance of the Naive Bayes (NB) model is evaluated based on different features. The NB model is one of the most commonly used machine learning algorithms in the field of cyber security for classification tasks which is based on Bayes' theorem, which

assumes that the features are independent of each other, and it works well with high-dimensional datasets.

In the experimental setup, the NB and RF models are trained on data which is collected in real time containing network traffic logs, system logs, and security logs. The model is evaluated based on different features such as the number of features considered, the type of features (e.g., binary, categorical, numerical), and the level of feature engineering performed.

The performance metrics used to evaluate the NB and RF models include accuracy, precision, recall, and F1-score (Refer Figure 4 & 5). These metrics are essential for measuring the effectiveness of the model in detecting and classifying cyber threats accurately.

Accuracy measures the proportion of correctly classified instances over all instances. Precision measures the proportion of correctly classified instances of a specific class over all instances classified at that class. Recall measures the proportion of correctly classified instances of a specific class out of all instances that truly belong to that class. F1-score is a weighted average of precision and recall, where F1-score ranges from 0 to 1, with 1 being the best possible score.

To evaluate the performance of the NB and RF models with respect to the number of features considered, different experiments are conducted by varying the number of features used to train the model. The results of the experiments are then recorded and analyzed using various performance metrics.

The performance metrics are plotted against the number of features considered, and a scatter plot is generated. The scatter plot as shown in the Figure 4 and 5 shows how the performance of the NB and RF models changes as the number of features considered increases. The results obtained can be used to determine the optimal number of features to use in training the model. It is to be noted that the results for training and testing are processed using real time data.

5. CONCLUSION AND FUTURE WORK

From the experimental results it is inferred that the performance metrics of the NB and RF models are crucial in evaluating the effectiveness of the model in detecting and classifying cyber threats accurately. The experimental setup involving the integration of CTI and big data technologies provides a robust framework for evaluating the performance of the NB and RF model. By varying the number of features considered, the type of features, and the level of feature engineering performed, the optimal configuration for training the NB and RF model can be determined. This information is essential for building an effective SOC that can detect and respond to cyber threats in real-time.

Also, in the current cyber security perspective, CTI has grown to be an essential element. It is essential in giving businesses the knowledge they need to guard against cyber attacks. Over the years, CTI has changed, and several research studies have suggested various frameworks and strategies to increase CTI's effectiveness. Machine learning, data mining, knowledge graphs, and natural language processing are a few of these methods. In order to standardise the gathering and sharing of CTI among various organisations, CTI protocols have also been developed. The need for more automated CTI analysis tools and issues with data quality, lack of standardisation, and other issues must yet be resolved. The goal of the upcoming effort is to raise the standard of CTI, included by creating better tools for automated CTI analysis. To ensure that the CTI sharing is of good quality and can be easily shared across different organisations, there is a need for increased standardisation in CTI collection and sharing protocols.

Additionally, utilising cutting-edge technology like blockchain and AI may improve the efficacy of CTI. Hence, there is a need for increased standardization in CTI collection, sharing protocols and the adoption of cutting edge technologies like blockchain and artificial intelligence to enhance the efficacy of CTI. However, further research is needed to explore advanced feature engineering techniques, data

mining techniques, and the mentioned tech stack to improve the accuracy of the models.

REFERENCES

1. https://www.marketsandmarkets.com/Market-Reports/cyber-security-market-05.html.
2. EMBROKER. 2021 must-know cyber-attack statistics and trends. https://www.embroker.com/blog/cyber-attack-statistics/.
3. Wang, Q., Hassan, W.U., Li, D., Jee, K., Yu, X., Zou, K., Rhee, J., Chen, Z., Cheng, W.,Gunter, C.A., et al., 2020. You are what you do: hunting stealthy malware viadata provenance analysis. NDSS.
4. Jo, Hyeonseong, Yongjae Lee, and Seungwon Shin. Vulcan: Automatic extraction and analysis of cyber threat intelligence from unstructured text. Computers & Security 120 (2022): 102763.
5. Kanakogi, K., Washizaki, H., Fukazawa, Y., Ogata, S., Okubo, T., Kato, T., Kanuka, H., Hazeyama, A., & Yoshioka, N. (2021). Tracing CAPEC Attack Patterns from CVE Vulnerability Information using Natural Language Processing Technique. Academic Press.
6. Li J. Liu Y. Chen T. Xiao Z. Li Z. Wang J. (2020). Adversarial Attacks and Defenses on Cyber–Physical Systems: A Survey. IEEE Internet of Things Journal, 7(6), 5103–5115.
7. Ramsdale A. Shiaeles S. Kolokotronis N. (2020). A Comparative Analysis of Cyber Threat Intelligence Sources, Formats and Languages. Electronics (Basel), 9(5), 824.
8. Veerasamy, N., Mashiane, T. T., & Pillay, K. J. (2019). Towards Cyber Incident Response Strategic Planning. In THREAT 2019 Cyber security Summit, Sandton, South Africa.
9. Mavroeidis, V., &Bromander, S. (2021). Cyber Threat Intelligence Model: An Evaluation of Taxonomies, Sharing Standards, and Ontologies within Cyber Threat Intelligence. arXiv:2103.03530v4 [cs.CR].
10. Aryal, R., Jang, Y. M., Lee, S., & Kim, J.A Novel Cyber Threat Intelligence Framework Based on Deep Learning and Data Fusion, Journal of Information Processing Systems, 2020, Volume: 1 Issue: 6, Pages: 1495-1507.
11. Cho, K., Shin, J., & Kim, Y. (2021). Enhancing cyber threat intelligence with explainable machine learning. IEEE Transactions on Information Forensics and Security, 16, 2047-2061.
12. Sinha, A., Yadav, V., & Yadav, N. (2020). An approach to improve cyber threat intelligence using data mining techniques. International Journal of Network Security, 22(2), 216-225.
13. Arul, R., Karthikeyan, K., & Chandrasekaran, K.A Comprehensive Analysis of Cyber Threat Intelligence Sharing Platforms, Journal of Information Security and Applications, 2021, Volume: 62, Pages: 102887. DOI: https://doi.org/10.1016/j.jisa.2021.102887.
14. Alsaleh, M., Alturki, R., & Alhomoud, F. (2021). Cyber Threat Intelligence Sharing: A Systematic Review and Future Directions. IEEE Access, 9, 45212-45227.
15. Li, Zong-Xun, et al. "K-CTIAA: Automatic Analysis of Cyber Threat Intelligence Based on a Knowledge Graph." Symmetry 15.2 (2023): 337.
16. Jo, Hyeonseong, Yongjae Lee, and Seungwon Shin. Vulcan: Automatic extraction and analysis of cyber threat intelligence from unstructured text. Computers & Security 120 (2022): 102763.
17. Czekster, Ricardo M., Roberto Metere, and Charles Morisset. cyberaCTIve: a STIX-based Tool for Cyber Threat Intelligence in Complex Models. arXiv preprint arXiv:2204.03676 (2022).
18. Zhou, Yinghai, et al. CTI view: APT threat intelligence analysis system. Security and Communication Networks 2022 (2022): 1-15.
19. Sakellariou, Georgios, Panagiotis Fouliras, and Ioannis Mavridis. SECDFAN: A Cyber Threat Intelligence System for Discussion Forums Utilization. Eng 4.1 (2023): 615-634.
20. Lin, P. C., Hsu, W. H., Lin, Y. D., Hwang, R. H., Wu, H. K., Lai, Y. C., & Chen, C. K. (2023). Correlation of cyber threat intelligence with sightings for intelligence assessment

and augmentation. Computer Networks, 228, 109736.

21. Sibi Chakkaravarthy Sethuraman, Aditya Mitra, Kuan-Ching Li, Anisha Ghosh, M Gopinath, Nitin Sukhija, Loki: A Physical Security Key Compatible IoT Based Lock for Protecting Physical Assets, Vol. 10, Pages. 112721-112730, IEEE Access, 2023.
22. Czekster, R. M. (2023). Leveraging Cyber Threat Intelligence in Smart Devices. In Information Security and Privacy in Smart Devices: Tools, Methods, and Applications (pp. 71-95). IGI Global.
23. Devi Priya V S, Sibi Chakkaravarthy Sethuraman, Containerized cloud-based honeypot deception for tracking attackers, Scientific Reports, Nature, 2023.
24. Dedipyaman Das, S.Sibi Chakkaravarthy, Suresh Chandra Satapathy, A Decentralized Open Web Cryptographic Standard, Computers and Electrical Engineering, Elsevier, Vol. 99, 107751, April, 2022.
25. Gopinath M, SC Sethuraman, A comprehensive survey on deep learning based malware detection techniques, Computer Science Review, Vol. 47, February 2023, Elsevier.