Investigating the Application of Artificial Intelligence and Machine Learning in Threat Intelligence Platforms for Small and Medium Enterprises (SMEs) and Developing Economies

Abstract

Due to budget limitations and limited access to sophisticated threat intelligence solutions, Small and Medium Enterprises (SMEs) and developing economies are especially vulnerable in the quickly changing world of cybersecurity threats. To strengthen these companies' cybersecurity posture, this study investigates the strategic integration of Artificial Intelligence (AI) and Machine Learning (ML) into Cyber Threat Intelligence (CTI) platforms. We carry out an extensive analysis of a collection of research on CTI, finding shortcomings in its application to SMEs and developing nations. We also investigate the possibility of open-source threat feeds as a useful, affordable source of intelligence.

The AI and ML techniques may automate the processing of these data feeds through systematic examination and comparison, offering improved predictive capabilities, anomaly detection, and actionable insights customized to the unique requirements and restrictions of SMEs and developing countries. According to the research, utilizing AI and ML in CTI platforms helps to promote a more proactive cybersecurity approach in addition to increasing threat visibility. The article concludes with suggestions for putting AI/ML-enhanced CTI solutions into practice that address issues including algorithmic bias, data quality, and the ever-changing nature of cyber threats.

Keywords: Cyber Threat Intelligence, Artificial Intelligence, Machine Learning, SMEs, developing economies, developing nations, small and medium enterprises,

Table of Contents

Abstract 1
Table of Contents 2
List of Abbreviations and Acronyms 2
Introduction 3
Related Work 4
Methodology 7
Discussion 7
Recommendation 9
Conclusion 11
References 12

List of Abbreviations and Acronyms

SME: Small and Medium Enterprises

ML: Machine Learning

AI: Artificial Intelligence

CTI: Cyber Threat Intelligence

DS: Data science

CVE: Common Vulnerabilities and Exposures

CWE: common weakness enumeration

DDoS: Distributed Denial of Service

TIP: Threat Intelligence Platform

IoC: Indication of Compromise

OSINT: Open-source Intelligence

Introduction

Most organizations in the today use digital tools and procedures from the Internet. Traditional enterprise security perimeters have been undermined by changes in IT architecture and usage norms, such as virtualization, cloud computing, and mobility. This has left a large attack surface open to hackers and other threat actors[1]. Coupled with the escalating complexity and frequency of cyber threats, particularly highlighted by recent prominent supply chain attacks, underscores the imperative for robust Cyber Threat Intelligence (CTI) strategies[2]. While there are many data sources comprising intelligences communities, security organization, social media, dark web and security blogs, the data exist scattered in various places and format including text, web database, email subscription, excel sheets usually and require huge resources, time, and expertise to convert it to intelligence information[1, 3]. Existing Threat Intelligence Platforms (TIPs) offer several benefits that help organizations swiftly establish internal processes for gathering, evaluating, enriching, normalizing, and disseminating threat-related data. But there are several issues with the current TIPs that keep them from being widely used [1] and specifically being adopted by SMEs and developing economies. As the cyber threats are growing exponentially, the SMEs and developing economies are battled with no resources and expertise to safeguard their digital assets and their security hugely jeopardized due to lack of threat visibility[2, 4].

The advent of Artificial Intelligence (AI) and Machine Learning (ML) technologies has significantly transformed the cybersecurity landscape, introducing advanced tools and methods to address and mitigate the continuously evolving security threats [5].By combining machine learning and artificial intelligence with cyber threat information it would ease the process of gathering useful threat intelligence from a variety of sources [3]. This method has significantly enhanced the future of cybersecurity information sharing, enabling deeper insights from open source intelligence[4, 6]. Despite these improvements, there is a notable gap in research on leveraging these threat data to support SMEs and developing economies, a sector that is struggling to keep up with threats.

Thus, the focus of this paper is to find out how we can leverage open-source intelligence with artificial intelligence and machine learning to aid SMEs and developing economies.

Related Work

1. **Cyber threat Intelligence**

   A Cyber Threat Intelligence (CTI) platform is an advanced tool intended to gather, examine, and share data on known or unknown cyberthreats and vulnerabilities[2, 7, 8]. To give a thorough picture of the cyber threat landscape, these platforms compile information from a variety of sources, such as social media, forums on the dark web, open-source intelligence (OSINT), technical data streams, and more[1, 9]. A CTI platform's main goal is to help organizations comprehend the risk they face and make informed decisions about proactive defense measures.

2. **Opensource Threat Data**

   open-source intelligence (OSINT) is far less expensive than traditional information-gathering technologies, which is one of its main advantages[2, 9]. OSINT , apart from its cost-effectiveness, offers numerous advantages in terms of information sharing and access, since it can be freely and lawfully shared with any party and is constantly updated[9, 10]. On the other hand, there are certain limitations to OSINT, such as the large volume of data that must be analyzed to produce accurate intelligence, necessitating a significant amount of effort to separate the valuable information from the noise. To separate genuine, validated information from fraudulent, deceptive, or erroneous data, security experts must perform a significant amount of analytical work[1].

3. **The challenge**

   The sheer volume and complexity of data being generated and processed each day with rapidly changing digital landscape has led to threat sophistication and heightened risk to organization and individuals despite their size and geographic location[1].

   Cyber Threat Intelligence (CTI), aimed at the proactive identification and analysis of cyber threats, has become a key strategy for businesses tackling the growing number and complexity of security events, yet subscribing to multiple sources of threat feeds can result in an overload of information especially for entities with limited resources [1, 11]. As SMEs and developing economies embark into their digital transformation journey with

limited resources, the risk is even higher with limited protection capabilities. Data science, along with AI and ML, holds the power of enhancing decision-making for SMEs by leveraging informed, data-driven strategies [12]. The use of AI and ML in cybersecurity spans multiple areas, such as detecting anomalies, conducting predictive analyses, analyzing behavioral patterns, and gathering threat intelligence [5].

4. Existing research on AI and ML in CTI.

Leveraging their ability to swiftly and accurately process immense data volumes, AI and ML technologies enable real-time identification of anomalies and threats, while their adaptive algorithms ensure security systems continuously evolve, keeping pace with changing threats and reducing vulnerabilities [5]. CTI involves the collection, analysis, and dissemination of data to recognize, track, and forecast cyber threats [11]. AI and ML technologies in threat intelligence can be adapted and applied in cost-effective, scalable, and efficient ways to meet the cybersecurity needs of SMEs and developing economies [4].

However, challenges such as data privacy concerns, the need for skilled personnel to manage and interpret AI/ML outputs, and the initial setup costs must be carefully addressed to fully leverage these technologies in resource-constrained environments [2].

- Comparative analysis of Models used in earlier research.

Paper Title	AI/ML Model Used	Strengths	Limitations	Use Case
An Automatic Generation Approach of the Cyber Threat Intelligence Records Based on Multi-Source Information Fusion[10]	Decision Trees, Random Forest, Support Vector Machines (SVM), MLP, XGBoost	- Wide range of models tested for threat-article classification.	- Each model has specific conditions where it performs best, limiting general applicability	- Classify threat articles vs nonthreat articles from various sources.

Cyber Threat Intelligence Discovery using Machine Learning from the Dark Web[13]	Weka machine learning tools used to test performance of Support Vector Machine (SVM), Artificial Neural Networks (ANN), Decision Trees (including Random Forest and J48 in Weka) and Naïve Bayes	- High accuracy for specific exploit type detection using SVM and ANN. - Effective in extracting actionable intelligence from semi-structured data.	- Performance varies significantly with the data's nature (e.g., system vs. web exploit types). - Some models showed low MCC (Matthews correlation coefficient), indicating moderate prediction quality.	- Discovering cyber threat intelligence from forums and dark web sources.
An Overview of Cyber Threat Intelligence Platform and Role of Artificial Intelligence and Machine Learning [3]	Naïve Bayes Classifier	- High accuracy for classifying text data related to cyber threats. - Efficient in processing and analysing textual threat intelligence.	- Simplistic model assumptions may not capture complex relationships in data. - Primarily effective for text classification, limiting its utility for other data types.	- Extracting high-level threat intelligence from structured data and unstructured data sources.

- This comparative table highlights that while there's a diverse application of AI/ML models in CTI, each comes with its own set of strengths and limitations. The choice of model significantly depends on the specific CTI task at hand, whether it's real-time threat detection, extracting intelligence from text data, or identifying specific exploit types[3, 13]. The limitations mainly revolve around the models' applicability to different data types, the requirement for extensive data pre-processing, and the models' inherent assumptions that may not always hold true in the complex domain of cyber threats.

5. Gap in the literature, particularly in the context of SMEs and developing countries.

SMEs and developing economies frequently encounter cyberattacks but often lack the resources and knowledge to effectively counter these threats. This fundamental challenge underlines the need for solutions that are not only automated but also adaptable to the specific contexts of the sector[2]. During the review process there were limited research addressing the specific needs of this sector. The existing research in cybersecurity intelligence sharing suggests that current tools such as MISP for utilizing shared incident data may not adequately serve the needs of these sectors[2]. However, using automation capabilities of modern techniques with AI and ML, the existing tools offer foundational elements that could be adapted to develop platforms beneficial for SMEs.

Methodology

As part of its methodological framework, the study used a systematic desktop research strategy that involved analysing many publicly accessible information sources to ensure a comprehensive analysis of relevant literature and data. Through a review of scholarly works, including research papers and journals, the study analysed the uses of machine learning (ML) and artificial intelligence (AI) in cyber threat intelligence with the aim of fully understanding their functions and applicability to Deakin Threat Mirror. The primary goal of this study was to compile findings, conclusions, and tactical suggestions that follow a vendor-neutral path. This position emphasizes the dedication to ensuring an open-source ecosystem by giving priority to initiatives that are both financially feasible and free from proprietary restrictions.

Most importantly, the study was purposefully designed to the requirements of developing economies and small and medium-sized businesses (SMEs). Understanding the difficulties and limitations these organizations encounter, especially around cyber threat visibility and intelligence, served as the basis for this review. Also, the process was modified to obtain insights and suggestions that are easily adjustable in addition to being economical and resource efficient. The objective was to enable the creation of adaptable and scalable threat intelligence systems with the use of AI/ML that are especially made to reflect and offset the changing threats that small and medium-sized businesses and developing nations encounter.

By using this methodology, the study aims to significantly advance the field by offering guidance to the Deakin Threat Mirror development team on how to include AI/ML to make the tool more effective for the targeted Sector. To strengthen these sectors' cybersecurity posture and encourage sustainability and resilience during global cyberthreats, the objective is to provide them with the visibility, information, and resources they require to secure their digital space despite the resource constraints.

Discussion

6. **Enhanced threat intelligence from opensource threat feeds using AI/ML:**

   Open-source security feeds are a valuable resource for cybersecurity threat intelligence especially for SMEs and developing economies that do not have internal data sources[1,

2]. These feeds provide timely information on threats, vulnerabilities, and incidents, which can be crucial for the sector with limited resources to invest in proprietary threat intelligence services and building security operation centre[2, 11]. AI and ML can transform the use of these feeds by:

1. Automating Data Analysis:

   AI and ML can automate the analysis of vast volumes of data from open-source security feeds, identifying relevant threats more efficiently than manual processes. This automation is crucial, where cybersecurity staffing may be limited[6].

2. Enhancing Predictive Capabilities

   By applying ML algorithms to historical data from open-source feeds, SMEs can predict potential future attacks and prepare defences in advance. This predictive capability is especially beneficial for developing economies where reactive measures may be less feasible due to resource constraints[6, 7, 12].

7. Addressing Challenges and Ethical Considerations

While the application of AI and ML to open-source security feeds offer substantial benefits, it also presents several challenges:

1. Data Quality and Relevance

   Open-source feeds vary widely in quality and relevance. AI and ML models require high-quality, relevant data to be effective. Ensuring the quality and relevance of data from these feeds is crucial for developing accurate and reliable cybersecurity measures[1, 7].

2. Algorithmic Bias and Data Privacy

   The risk of algorithmic bias and data privacy concerns are particularly pertinent when processing open-source security feeds. The platform must implement strategies to mitigate bias and protect any sensitive information that may be included in these feeds[11-13].

3. Need for Continuous Adaptation

   The cybersecurity threat landscape is dynamic, necessitating continuous updates to AI and ML models. The platform must establish processes for regularly updating the models based on the latest data from open-source feeds to maintain the effectiveness[1, 5].

Recommendation

Following use case could be explored further in future to help in analyzing, categorizing, and predicting threats from vast amounts of data collected using different sources efficiently.

1. **Clustering and outlier Detection Model:**
   - *Use Case:*
     To detect unusual patterns that may indicate new or evolving cyber threats.
   - *Model:*
     - K-Means Clustering for grouping similar types of threats together and identifying outlier data points[6, 9]. The models can run through collected OSINT data, grouping similar threats, and identifying outliers and unusual patterns indicative of new or targeted cybersecurity threats.

2. **Classification Models**
   - *Use Case:*
     To classify and tag data into predefined categories such as "spam", "phishing", "malware", etc. Currently the data is supplied static classification where missing, which could be inefficient when data sources grow.

   - *Models:*
     - Random Forest for their robustness and ability to handle imbalanced datasets, which is common in threat intelligence data [6, 9-11, 13]
     - Support Vector Machines (SVM) for high-dimensional data classification tasks, useful for feature-rich OSINT data [6, 7, 9-11, 13].
     - Naïve Bias for multiclass classification and textual threat data extraction[3]

3. **Natural Language Processing (NLP) Models**
   - *Use Case:*
     To complement the data collected using intelmq, NLP can be used to analyse and extract meaningful information from unstructured text data sources.

   - *Models:*
     - BERT (Bidirectional Encoder Representations from Transformers) for deep understanding of contextual relationships in text data[7, 11]. The model can be

used to analyse the content of blocklists, forums, and reports to identify emerging threats.

4. **Time Series Analysis and Prediction Models**
    - *Use Case:* To forecast trends in cyber threats and prepare defences accordingly.
    - *Modes:*
        - LSTM (Long Short-Term Memory) networks, a type of recurrent neural network, suitable for making predictions on time-series data and capturing long-term dependencies to analyze time-based patterns in threat occurrences, helping predict future threat landscapes and allowing SMEs to prepare or strengthen their defenses proactively[6, 7].

Thus, in continuation to this report, it is advised to investigate and test the aforementioned machine learning models for automated reporting. These models can shorten the period between data collection and actionable intelligence by automatically generating threat intelligence reports from analysed data.
Creating interactive data visualisation tools that make use of machine learning insights is additionally recommended to provide product users the ability to study and comprehend cyber threat data in an intuitive manner.
While choosing the models the most important parameter to consider is not just the accuracy but also time and resource required to run the models keeping in mind the target users of the platform.

Conclusion

Therefore, integrating AI and ML with open-source security feeds presents a promising avenue for enhancing cybersecurity in SMEs and developing economies. By automating the analysis of these feeds, Deakin Threat Mirror aims to uplift the threat intelligence and predictive capabilities of SMEs and developing economies, enabling more effective and proactive cybersecurity measures. However, to fully realize these benefits, the platform will also centralize the navigational challenges related to data quality, algorithmic bias, and the dynamic nature of cyber threats. Through careful implementation and ongoing adaptation, the platform will empower SMEs and developing economies to leverage AI and ML to build more resilient cybersecurity defences with open-source threat intelligence data.

Thus, combining open-source security feeds with AI and ML offers an opportunity to improve cybersecurity in SMEs and nations with limited resources. Deakin Threat Mirror aspires to improve the threat intelligence and prediction capacities of SMEs and developing economies by automating the analysis of data feeds, hence enabling more efficient and proactive cybersecurity responses. To maximise the benefits, the platform will also centralise the navigational difficulties pertaining to algorithmic bias, data quality, and the ever-changing nature of cyber threats.

With careful implementation and ongoing adaptation, the platform will empower SMEs and developing economies to leverage AI and ML to build more resilient cybersecurity visibility and response with open-source threat intelligence data.

References

[1] C. Martins and I. Medeiros, "Generating quality threat intelligence leveraging osint and a cyber threat unified taxonomy," *ACM Transactions on Privacy and Security,* vol. 25, no. 3, pp. 1-39, 2022.

[2] M. Van Haastrecht *et al.*, "A shared cyber threat intelligence solution for SMEs," *Electronics,* vol. 10, no. 23, p. 2913, 2021.

[3] A. Dutta and S. Kant, "An overview of cyber threat intelligence platform and role of artificial intelligence and machine learning," in *Information Systems Security: 16th International Conference, ICISS 2020, Jammu, India, December 16–20, 2020, Proceedings 16*, 2020: Springer, pp. 81-86.

[4] M. Van Haastrecht, B. Yigit Ozkan, M. Brinkhuis, and M. Spruit, "Respite for SMEs: A systematic review of socio-technical cybersecurity metrics," *Applied sciences,* vol. 11, no. 15, p. 6909, 2021.

[5] R. Evren and S. Milson, "The Cyber Threat Landscape: Understanding and Mitigating Risks," EasyChair, 2516-2314, 2024.

[6] I. H. Sarker, A. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, "Cybersecurity data science: an overview from machine learning perspective," *Journal of Big data,* vol. 7, pp. 1-29, 2020.

[7] P. Ranade, A. Piplai, S. Mittal, A. Joshi, and T. Finin, "Generating fake cyber threat intelligence using transformer-based models," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021: IEEE, pp. 1-9.

[8] A. Dehghantanha, G. Dietrich, and K.-K. R. Choo, "Introduction to the Minitrack on Machine Learning and Cyber Threat Intelligence and Analytics," 2021.

[9] A. Yadav, A. Kumar, and V. Singh, "Open-source intelligence: a comprehensive review of the current state, applications and future perspectives in cyber security," *Artificial Intelligence Review,* vol. 56, no. 11, pp. 12407-12438, 2023.

[10] T. Sun, P. Yang, M. Li, and S. Liao, "An automatic generation approach of the cyber threat intelligence records based on multi-source information fusion," *Future Internet,* vol. 13, no. 2, p. 40, 2021.

[11] S. Saeed, S. A. Suayyid, M. S. Al-Ghamdi, H. Al-Muhaisen, and A. M. Almuhaideb, "A systematic literature review on cyber threat intelligence for organizational cybersecurity resilience," *Sensors,* vol. 23, no. 16, p. 7273, 2023.

[12] S. Rautenbach, I. H. de Kock, and J. Grobler, "DATA SCIENCE FOR SMALL AND MEDIUM-SIZED ENTERPRISES: A STRUCTURED LITERATURE REVIEW," (in eng), *South African Journal of Industrial Engineering,* vol. 33, no. 3, pp. 83-95, 2022, doi: 10.7166/33-3-2797.

[13] A. Zenebe, "Cyber Threat Intelligence Discovery using Machine Learning from the Dark Web," *Communications of the IIMA,* vol. 20, no. 2, 2022.