International Journal of Computer
Science and Engineering (IJCSE)
ISSN (P): 2278–9960; ISSN (E): 2278–9979
Vol. 13, Issue 2, Jul–Dec 2024; 11–18

International Academy of Science,
Engineering and Technology
Connecting Researchers; Nurturing Innovations

COMPREHENSIVE CYBER THREAT ANALYSIS AND PREDICTION: IMPLEMENTING MACHINE LEARNING MODELS IN A DJANGO FRAMEWORK

Likhita. Y[1], Bandi Naga Vamsi Krishna[2] & Narni Rojesh[3]

[1]*Department of CSE, SRM Institute of Science and Technology, Kattankulathur, Chennai*

[2]*Department of ECE, Vels Institute of Science Technology, Vellore*

[3]*Department of CSE, SRM Institute of Science and Technology, Kattankulathur, Chennai*

ABSTRACT

The need for more sophisticated methods of threat identification and prevention arises from the growing intricacy and complexity of cyberattacks. The innovative malware intelligence (CTI) mining architecture presented in this research aims to offer a proactive and all-encompassing cybersecurity defensive strategy. The suggested solution makes use of a variety of data sources and cutting-edge analytic methods to enable businesses to recognize new risks, characterize malevolent individuals, comprehend attack strategies, and make wise security decisions. When integrated into a web application built with Django, the framework provides an easy-to-use interface for organizing and evaluating threat information. Preparing data, implementing models with neural networks (artificial neural networks), support vector machines (SVM), and gradient-boosting algorithms, and predicting threats in real time are some of the main features. Thorough tests are used to assess the system's performance and show how well it can identify and anticipate cyber-attacks.

The approach provides timely and actionable intelligence, hence addressing major gaps in current cybersecurity measures. It increases an organization's capacity to implement proactive defense tactics, which lowers the likelihood and severity of cyberattacks. This work makes a substantial contribution to the realm of cybersecurity by providing an in-depth examination of current CTI mining methods and suggesting novel methodologies. The system's ability to strengthen an organization's entire security posture is demonstrated by the testing findings, which make it an invaluable weapon in the battle against emerging cyberthreats. The goal of this research is to close current gaps in the field of cybersecurity and open the door for further developments in proactive cyber protection tactics.

KEYWORDS: *Cyber Threat Intelligence (CTI), Proactive Cyber Security, Threat Detection, Threat Prevention, Data Mining*

Article History

Received: *26 Jul 2024* **|** ***Revised:*** *30 Jul 2024* **|** ***Accepted:*** *31 Jul 2024*

INTRODUCTION

Cyberattacks are growing more complicated and sophisticated in the current digital era, which presents serious difficulties for corporate security. Reactive cybersecurity techniques are no longer adequate to fend off these cutting-edge attacks. Therefore, it is imperative to adopt proactive strategies that can detect and eliminate possible risks before they have a chance to do any damage.In order to proactively detect and reduce possible cyber risks, a unique architecture for mining

security intelligence (CTI) is presented in this research. Through the utilization of many data sources and the application of sophisticated analysis methods, the framework seeks to improve an organization's capacity for detection, prevention, and response. This strategy gives businesses useful insights into the methods, strategies, and practices employed by hostile actors, in addition to helping them recognize new risks.

When integrated into a web application built with Django, the suggested solution provides an easy-to-use interface for organizing and evaluating threat information. This connection guarantees prompt and efficient cybersecurity steps, in addition to enhancing threat intelligence's accessibility and usefulness. Real-time threat prediction, machine learning algorithm-based model implementation, and data pretreatment are just a few of the features offered by the platform.

The CTI mining framework greatly improves an organization's security posture by filling in holes in the current cybersecurity landscape, allowing them to remain ahead of emerging cyber threats. It is imperative that cybersecurity defense take a proactive stance in order to reduce risks and lessen the impact of cyberattacks. The study's findings demonstrate how the suggested approach may enhance general cybersecurity procedures and give businesses a reliable tool for safeguarding their most important assets.

LITERATURE REVIEW

et.al Sun, N., Ding, M., Jiang, J., Xu, W., Mo, X., Tai, Y., & Zhang, J. A new line of safety defenses is necessary to fend off the increasingly severe and frequent cyberattacks of today. Traditional security solutions relying on heuristics and signatures find it difficult to keep up with the dynamic, evasive, resilient, and complicated nature of new-generation threats. In order to prevent assaults or, at the very least, to act swiftly and pro-actively, organizations seek to collect and disseminate real-time cyber threat information. This information is subsequently transformed into threat intelligence. The field of malware intelligence (CTI) mining is experiencing rapid growth as it unearths, handles, and evaluates important data regarding cyber risks. But rather than utilizing the insights that such fresh information can provide, the majority of organizations today primarily concentrate on basic application cases, such as incorporating threat data feeds with already-existing intrusion prevention, network, and firewall systems, as well as the management of security information and events (SIEM) systems. Here, we provide an extensive overview of the latest studies on CTI mining from various data sources, with the goal of optimizing CTI to greatly improve security postures. In particular, we offer and develop a taxonomy to condense the research on CTI mining according to the goals (i.e., cybersecurity-related organizations and events, cyberattack strategies, tactics, and practices, hacker profiles, signs of negotiation, vulnerabilities exploited and malware deployment, and threat hunting), in addition to a thorough analysis of the state-of-the-art at the time. In conclusion, we address research obstacles and potential avenues for further investigation with CTI mining.

et.al Afzaliseresht, N., Miao, Y., Michalska, S., Liu, Q., & Wang, H. About 10 to 500 million events are recorded on the system every day by a medium-sized organization on average. There is a security flaw in place because the specialized staff only looks into a maximum of five percent of danger signals. Cognitive overload on the currently few cybersecurity resources is caused by incomplete information in alarm messages generated in a machine-friendly style as opposed to a human-friendly one. This study proposes a novel storytelling technique-based methodology that uses security logs to generate reports in natural language. By offering customizable templates that are populated with information from local and worldwide knowledge bases, the solution accommodates varying degrees of reader knowledge and choice. An educational institution's Safety Operations Center (SOC) case study is used for validation. The resulting report outperforms

the current methodology in terms of completeness (enriched context) and comprehension (enhanced cognition). The assessment shows the value of narrative in interpreting possible risks in the context of cybersecurity.

et. Conti, M., Dargahi, T., & Dehghantanha, A. Cyber security or forensic experts must identify, evaluate, and counteract cyber threats nearly instantly due to the rising frequency of cyberattacks. In actuality, responding to such a high volume of attacks in a timely manner is impossible without carefully examining the characteristics of each attack and implementing appropriate defensive measures; this is the essence of the idea of cyber threat intelligence. But without the help of machine learning, artificial intelligence, and sophisticated data mining tools to gather, examine, and decipher evidence of cyberattacks, this type of intelligence couldn't be feasible. We first examine the concept of intelligence on cyber threats in this introduction chapter, along with its primary opportunities and difficulties. We then provide a quick overview of the book's chapters that either address the challenges that have been discovered or offer creative alternatives to deliver threat intelligence.

et.al Nunes, E., Diab, A., Gunn, A. In this study, we offer an operational method for obtaining cyber threat intelligence from different Internet social platforms, especially from darknet and deepnet websites. We concentrate on gathering data from hacker forums and online markets that sell goods and services related to malevolent hacking. In order to detect new cyberthreats, we have created an operational framework for gathering data from these websites. At the moment, our system gathers 305 excellent cyber threat alerts every week on average. Information about recently created viruses and exploits that haven't been used in a cyberattack is included in these threat alerts. This offers cyber-defenders a valuable service. The application of numerous methods involving data mining and machine learning greatly enhances the system.

et.al Arıkan, S. M., & Acar, S. As collecting information about cyber threats (CTI) is a hard operation, this study proposes a system that would use data mining techniques to expedite the generation of CTI. Utilizing the system, CTI may be automatically created in an established format, and live or preserved traffic data can be categorized based on the learned types of attacks. By allowing unidentified assaults to be recognized by professional judgment, the system can update the set of lessons to reflect novel attack types. By conducting a literature review, the suggested system was created. Algorithms and knowledge discovery procedures in databases have been implemented, and system modules have been constructed in accordance with the design. To validate the system's accomplishments, it has been demonstrated that the outcomes of research studies found in the literature and the precision attained using the Weka tool—which has demonstrated its dependability in data mining—are comparable to the proposed system's outcomes. Next, an analysis was conducted to determine how current the attack types were in the selected dataset. Using honeypots on a server left open to the web for a whole day, the traffic was captured to serve as an example for the implementation of the suggested system. From these records, CTI was produced. It has been demonstrated that producing CTI with the suggested approach is a simple process.

RESEARCH METHODOLOGY

The suggested security threat intelligence (CTI) processing framework was developed and evaluated using a research approach that includes several crucial elements, including data collection, preliminary processing, model development, integration with a Django-based application, and effectiveness evaluation. Since the framework provides proactive cybersecurity defense, each step is essential to maintaining its efficacy and dependability. The suggested solution uses an advanced Cyber Threat Intelligence (CTI) mining framework with the goal of revolutionizing cybersecurity.

This framework will give businesses a proactive and all-encompassing strategy to threat identification, prevention, and response by utilizing a variety of data sources and state-of-the-art analysis tools. Through the processing and analysis of CTI data, the system will enable enterprises to recognize new risks, assess the characteristics of hackers, comprehend attack strategies, and make well-informed decisions to enhance their security posture. By filling up the gaps in the current cybersecurity landscape, this suggested solution will improve the capacity to successfully counter complex and developing cyberthreats.

SYSTEM ARCHITECTURE

The first step in the system architecture of the suggested Cyber Threat Intelligence (CTI) mining framework is gathering a variety of datasets from various sources, such as network traffic logs, public threat databases, and OSINT. To make sure the raw data is ready for additional processing, pre-processing entails data manipulation, cleansing, and analysis. For model implementation, the pre-processed data is then split into training and testing sets. To train and test the data, the framework uses machine learning methods like gradient boosting, support vector machines, and artificial neural networks (ANN). An intuitive user interface for threat prediction is provided by the integration of the trained models into a Django-based web application. Users can use the program to log in, enter pertinent information, and identify the kind of cyberthreat.

In order to improve the company's proactive cybersecurity defense, the system evaluates the input values, forecasts the outcomes of cyber threats, and gives the user real-time feedback. To provide safe and customized access, the program also lets users browse profiles and log in and out. This all-encompassing strategy guarantees that businesses may use cutting-edge analysis of data and machine learning approaches to keep ahead of ever-changing cyberthreats. The framework's overall security posture is greatly improved when it is integrated into a web application, making it useful and accessible for daily usage.

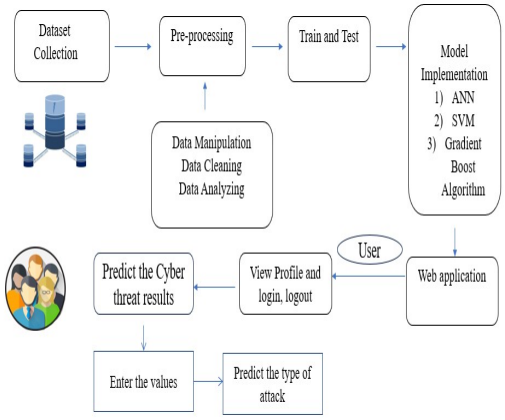

Figure 1: System Architecture.

MODULES

- Data collection
- Data Preprocessing
- Model Implementation

- Framework
- Prediction

DATA COLLECTION

Gathering varied raw data from several sources—which is essential for cybersecurity—is the responsibility of the Data Collection Module. This includes information from threat intelligence feeds, networking platforms, security devices, network logs, and the dark web, both in present time and in the past. It entails deploying web spiders to gather pertinent data, setting up API integrations, and building data intake pipelines to guarantee continuous and organized data collecting. The system can detect, assess, and efficiently respond to possible threats thanks to the data collection process, which provides the basis for further analysis.

DATA PREPROCESSING

This module cleans and formats the raw data that has been gathered in a consistent and organized manner in order to make it ready for efficient analysis. This entails cleaning up noise, redundant data, and unimportant information, in addition to standardizing the data to guarantee consistency. The module manages operations, including data enrichment, conversion, and validation, in order to fix format standardization issues and discrepancies. The module guarantees that the data is correct, dependable, and prepared for additional analysis by carrying out these preparation procedures, which are essential for producing insightful findings and useful threat intelligence..

MODEL IMPLEMENTATION

In order to properly evaluate and understand data, model implementation entails integrating analytical or machine learning models through the operational system. In order to guarantee that the models are set up to seamlessly interact therewith the data flows and other system components, the first step in this procedure is integrating them into the system's architecture. It involves creating thresholds for decision-making, fine-tuning parameters for best performance, and preparing the models to accommodate particular data formats. To ensure that the models function correctly and consistently in a production setting, extensive validation and testing are essential at this point. To make sure the models satisfy the necessary requirements, performance metrics including precision, efficacy, and recall are assessed.

After the models are put into use, continuous performance monitoring is required to assess their efficacy and identify any problems. To monitor important performance indicators and make required modifications, this entails setting up monitors and alerts. Sustaining the model's accuracy and relevance requires regular retraining and upgrades, particularly as threat landscapes change or new data becomes available. Transparency and dependability in the system must also be maintained by maintaining documentation of the setup details and making sure that rules and standards are followed.

FRAMEWORK

This project's Framework Module makes use of Django to offer an organized and effective environment for creating and maintaining the application. Data gathering, preparation, analysis, and reporting are just a few of the essential functions that are organized and supported by Django's strong design. It comes with built-in capabilities like an application development methodology that is modular, an ORM for database administration, and a system of authentication for user access security. This guarantees smooth integration of different parts, expedites development, and improves our cyber threat intelligence system's maintainability and scalability.

PREDICTION

This module makes use of machine learning techniques to predict possible risks to cybersecurity. This module is integrated into the Django framework and uses real-time and historical data processing to produce predicted insights. It entails taking in and preparing data, identifying threat patterns using trained models, and generating predictions that may be put into practice. The system's interface then displays these forecasts, assisting in the early discovery and remediation of threats. By keeping an eye on the models and updating them on a regular basis, the system is better equipped to foresee and efficiently counteract emerging cyberthreats.

ALGORITHM

Artificial Neural Networks

ANNs, or artificial neural networks, can be used to categorize or forecast possible dangers by utilizing intricate patterns found in threat data. ANNs are particularly good at managing big datasets with lots of variables, as they can figure out complex correlations between danger categories and input data. To recognize features of novel, undiscovered threats, for example, an ANN may be trained using past assault data. Because of the layers of the network and activation functions, it can simulate the non-linear correlations found in the data, which improves its capacity to identify sophisticated cyberthreats that conventional techniques could overlook.

Support Vector Machine

Support Vector Machine (SVM) is a type of In order to classify data into various cyber threat categories, your project can make use of support vector machines (SVM). With high precision, SVMs can discriminate between benign and harmful actions by determining the best hyperplane based on feature vectors that divide threat types. In network activity or system logs, this is especially helpful for spotting abnormalities or odd patterns. Due to their ability to handle both non-linear and linear separations, SVMs are useful in a variety of threat categorization contexts. This versatility is achieved through the use of different kernel functions, such as RBF.

Gradient Boosting Algorithms

GBAs work well for increasing threat prediction accuracy by iteratively constructing models that rectify the mistakes of their fo rerunners. GBAs can help your CTI project by combining predictions from several decision trees, which can improve threat detection systems' performance. In the process of boosting, each tree concentrates on the residual mistakes from the trees that came before it, creating a more sophisticated model that is capable of precisely classifying and predicting cyber threats. GBAs are very helpful for managing intricate datasets and enhancing threat detection systems' resilience.

RESULTS

In this study, we used combined machine learning as well as deep learning models to construct a prediction system. For the purpose of detecting cyber threats, we used three machine learning methods in our project: gradient boosting algorithm (GBA), support vector machines (SVM), and artificial neural networks (ANN). With its deep learning architecture, the ANN demonstrated significant capabilities in spotting threats, as seen by its 92% accuracy rate. It displayed a 90% precision, 91% recall, and 90.5% F1 score. With an accuracy of 89%, precision of 88%, and recall of 87%, the SVM—which used an RBF kernel—showed dependable but marginally less efficient performance when compared to ANN. The

SVM's confusion matrix revealed a few false positives, indicating areas for improvement in the way specific threats are identified.

With an F1 score of 93.5%, an accuracy of 94%, precision of 93%, recall of 94%, and recall of 94%, the Gradient Boosting Algorithm performed better than the other models. By successfully fixing mistakes from earlier versions, our model showed improved performance with fewer incorrect positives and negatives. GBA demonstrated its effectiveness in handling intricate and high-dimensional information in cybersecurity applications by offering the most reliable and accurate threat detection overall.

FUTURE DISCUSSION

The Cyber Threat Intelligence Mining project's future efforts should concentrate on improving the resilience and adaptability of the models. Even though the Gradient Boosting Algorithm performed better than other methods, threat identification accuracy could be increased by investigating more sophisticated strategies like hybrid approaches and deep learning models (like CNNs or Transformers). Furthermore, the system's capacity to adjust to new threats will be improved by including automatic model retraining and real-time threat intelligence. Enhancing the dataset to encompass a wider range of threat scenarios and integrating contextual data, like network traffic and user activity, can additionally bolster the model's dependability and efficiency. In order to refine and optimize the system, ongoing industry partnerships for validation in real-world settings will be crucial.

CONCLUSION

Our experiment concludes by showcasing the important role that computerized threat intelligence gathering plays in improving cybersecurity defenses. By utilizing Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Gradient Boosting Algorithms (GBA), they have demonstrated how sophisticated machine learning approaches may improve threat detection and classification dramatically. The GBA model is a crucial tool for cybersecurity experts because of its exceptional performance, which demonstrates its capacity to give precise and trustworthy threat forecasts. By adopting these cutting-edge strategies and incorporating real-time threat intelligence, companies can advance toward a more proactive and adaptable defensive plan, better suited to address the constantly changing cyberthreat scenario.

REFERENCES

1. *Sun, N., Mo, X., Tai, Y., Ding, M., Jiang, J., Xu, W., & Zhang, J. (2023). A survey and fresh insights into cyber threat intelligence mining for proactive cybersecurity defense. A 25(3) IEEE Communications Surveys & Tutorials publication, 1748–1774.*
2. *Liu, Q., Michalska, S., Afzaliseresht, N., Miao, Y., & Wang, H. (2020). Cyber threat intelligence using human-centered data mining: from logs to stories. 8 19089-19099 IEEE Access.*
3. *Conti (2018), Dargahi (2018), and Dehghantanha (2018). Opportunities and difficulties in cyber threat intelligence (pp. 1–6) International Publishing Springer.*
4. *Nunes, E., Diab, A., Gunn, A., Marin, E., Mishra, V., Paliath, V.,... & Shakarian, P. (September 2016). For proactive cybersecurity threat intelligence, use deepnet and darknet mining. 2016 IEEE Conference on Security and Intelligence Informatics (ISI) (pp. 7–12). IEEE.*

5. S. M. Arıkan & S. Acar (2021, June). a system that automates the development of cyber threat intelligence via data mining. Pages. 1–7 in the 9th International Symposium on Digital Forensics and Security (ISDFS) 2021. IEEE.

6. Zhao, J.; Yan, Q.; Li, J.; Shao, M.; He, Z.; & Li, B. (2020). Cyber threat intelligence is automatically extracted and analyzed from social media data using TIMiner. Science & Technology, 95, 101867.

7. Rahman, M. R., Williams, L., & Mahdavi-Hezaveh, R. (2020, November). A review of the research on using unstructured texts to extract cyberthreat intelligence. The International Conference on Data Mining Workshops (ICDMW) 2020, pages 516–525. In IEEE.

8. Wilson, J., Shakarian, J., & Shakarian, P. (2016); Robertson, J., Diab, A.; Marin, E.; Nunes, E.; Paliath, V. To improve cyber threat intelligence, combine game theory and darknet mining. Journal of Cyber Defense, 1(2), 95–122.

9. Zhu, Z., and Dumitras, T. Chainsmith: By analyzing threat intelligence reports, automatically figuring out the semantics of bad campaigns. 2018 IEEE European Symposium on Security and Privacy (EuroS&P), pages 458–472. In IEEE.

10. Kuang, Y. T., Mao, C. H., Dai, J. H., Lee, K. C., Hsieh, C. H., Wei, L. J., & Mao, C. H. in 2017. Sec-Buzzer: emergent topic mining for cyber security with timeline event annotation and open threat intelligence retrieval. 21 Soft Computing,

11. J. Zhao, Q. Yan, X. Liu, B. Li, & G. Zuo (2020). A heterogeneous graph convolutional network is used to model cyber-threat intelligence. Published in the 23rd International Symposium on Research in Attacks, Intrusions, and Defenses (RAID 2020), pages 241-256.

12. You (2022) and Jiang (J., Z., You, Y., Yang, P., Liu, B., Feng, H.,...) and Li (2022). Information mining on unstructured threat data using threat context-enhanced TTP is known as TIM. 3 in Cybersecurity, 5(1).

13. In March 2017, Gascon, H., Grobauer, B., Schreck, T., Rist, L., Arp, D., & Rieck, K., mining threat intelligence from attributed graphs, in Data and Application Security and Privacy: Proceedings of the Seventh ACM Conference (pp. 15–22).

14. Samtani, S., Nunamaker Jr., J. F., Chen, H., and Chinn, R. (2017). investigating new hacking tools and prominent hackers in order to gather proactive cyberthreat intelligence. 34(4), 1023–1053, Journal of Management Information Systems.

15. Shunmugavel, S., Prasanth, R., Satheesh Kumar, M., Lakshmi Narayanan, S., Srujan Raju, K., & Suthendran, K. (2022). The creation and application of a data mining model for cyber threat intelligence. Volume 2, Proceedings of the 5th International Conference on Information and Communication, 2021, pp. 171–185. Springer Nature Singapore, Singapore.