Research Exhibition in Mathematics and Computer Sciences (REMACS 6.0)/
ISBN: 978-629-97440-5-4

ANALYSIS OF MACHINE LEARNING (ML) ALGORITHM ON SYSTEM INFORMATION AND EVENT MANAGEMENT (SIEM) LOGS

Afif Haziq Haris, Mohd Faris Mohd Fuzi and Hafizah Hajimia
College of Computing Informatics and Mathematics,
Universiti Teknologi MARA, Perlis Branch
afifhaziq3078@gmail.com, farisfuzi@uitm.edu.my and hafizah.hajimia@uitm.edu.my

ABSTRACT - Security Information and Event Management (SIEM) is one of the essential security measures for enhancing the network's cybersecurity. The SIEM system which is used by Security Operation Centre (SOC) analysts as the central location where all security notifications from various security technologies, such as firewalls, IPS/IDS, and Anti-Virus logs, are gathered and visualized. However, the increasing frequency of cybercrime incidents and a shortage of cybersecurity specialists highlight the need for more effective detection methods. The objective is to conduct a comparative analysis of multiple ML algorithms based on accuracy, F1 scores, recall, precision, computer resource utilization, and feature importance to determine the most effective algorithms for SIEM log analysis. Three algorithms, namely Random Forest, XGBoost, and Isolation Forest are utilized in the research. According to the results, Random Forest has the highest accuracy, precision, recall, and processing speed. XGBoost also performs admirably, with perfect accuracy, excellent precision, and recall, but at a slower rate. Isolation Forest is inferior in terms of precision, accuracy, and F1 score, as well as processing time. This research is hoped to contribute to the field of cybersecurity and can guide future research and the selection of ML algorithms for SIEM log analysis.

Keywords: Machine Learning, System Information and Event Management, Random Forest, XGBoost, Isolation Forest.

1. INTRODUCTION

Providing strong cybersecurity is essential in the quickly changing digital landscape of today. However, organizations face significant challenges as a result of the complexity and amount of security incidents that keep growing. To monitor and identify potential security issues, SIEM systems are necessary. However, the SIEM log live analysis can be tedious and require a lot of resources. SIEM alerts SOC analysts and offers contextual information to aid in the investigation of a security event or incident by using correlation and statistical models to identify occurrences that could be security incidents (Skendzic et al., 2022). AI-based anomaly detection can help overcome these challenges by automating the process of identifying anomalies in log data (Kumar et al., 2022). With ML algorithms applied to SIEM logs analysis, hidden patterns, anomalies and useful insights for threat detection and response can be identified.

2. METHODOLOGY

There are several essential phases in the process for the analysis of machine learning algorithms on SIEM logs. First, raw data is gathered, which includes the SIEM system's logs. In order to address missing values, outliers, and inconsistencies and maintain the quality and integrity of the information, the data is then put through a data cleaning procedure. Then, the dataset is divided into training and testing sets. The parameters of the machine learning algorithms, in particular Random Forest, XGBoost, and Isolation Forest, are adjusted throughout the training phase to maximize performance. The separate testing dataset is used to test the models after they have been trained to determine their accuracy and performance.

3. RESULTS AND DISCUSSION

Based on the result, Random Forest and XGBoost perform better than Isolation Forest while analyzing SIEM logs. Random Forest was able to create accurate positive predictions with few false positives, as evidenced by its remarkable accuracy of 99.99% and a high precision score of 0.9996. XGBoost demonstrated 100% accuracy and received a precision score of 1. These findings demonstrate the accuracy with which Random Forest and XGBoost can detect anomalies and security threats in SIEM logs. Contrarily, Isolation Forest displayed a lower accuracy of 64.79% as well as precision and recall scores of 0.6479. Although Isolation Forest may have specific uses, its limits in reliably detecting anomalies in SIEM logs are indicated by its lesser accuracy and precision. Additionally, compared to the other algorithms, Isolation Forest required much more CPU time and elapsed time, suggesting possible scaling issues.

Therefore, Random Forest or XGBoost should be considered as the preferred algorithms by organizations looking for effective SIEM log analysis.

4. NOVELTY OF RESEARCH

This research presents contributions in the analysis of machine learning algorithms on SIEM logs. It offers insights into the effectiveness of Random Forest, XGBoost, and Isolation Forest for anomaly detection and security analysis by studying their implementation in the analysis process. The study includes detailed performance indicators including accuracy, precision, recall, and F1 score as well as measures of resource utilization, like CPU and RAM consumption. The study makes use of machine learning to automate the procedure and strengthen cybersecurity defenses in order to overcome the difficulties of manual analysis. Overall, this research advances the field by presenting fresh viewpoints on SIEM log analysis, analyzing certain ML algorithms, utilizing extensive performance indicators, and highlighting the demand for effective and automated methods to deal with cybersecurity issues.

5. CONCLUSION

In summary, ML integration with SIEM systems has a tremendous potential to improve cybersecurity threat detection and response. However, issues like interpretability, model selection and data bias and quality are still a limitation for ML development. For the ML integration in SIEM systems to be as effective as possible, it is crucial to carefully weigh these advantages and difficulties, as well as to implement the right setup and conduct ongoing monitoring.

REFERENCES

Kumar, G. R., Karthik, J., Rao, B. S., & Prasad, C. (2022). Anomaly Identification Performed Independently in Explanatory Machine using Log-based Method. *Proceedings of the International Conference on Electronics and Renewable Systems, ICEARS 2022*, 1160–1166. https://doi.org/10.1109/ICEARS53579.2022.9751876

Perera, A., Rathnayaka, S., Perera, N. D., Madushanka, W. W., & Senarathne, A. N. (2021, April 2). The Next Gen Security Operation Center. *2021 6th International Conference for Convergence in Technology, I2CT 2021*. https://doi.org/10.1109/I2CT51068.2021.9418136

Skendzic, A., Kovacic, B., & Balon, B. (2022). Management and Monitoring Security Events in a Business Organization - SIEM system. *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology, MIPRO 2022 - Proceedings*, 1203–1208. https://doi.org/10.23919/MIPRO55190.2022.9803428