# SECURITY INFORMATION AND EVENT MANAGEMENT (SIEM) ENHANCEMENT USING MACHINE LEARNING FOR DETECTING CYBER ATTACKS

**THESIS**

*Submitted as fulfilment of the requirements for the completion of*

*Master of Computer Science Program*

**DIMAS ERLANGGA**

20210130034

**MASTER OF COMPUTER SCIENCE PROGRAM**

**SCHOOL OF COMPUTER SCIENCE**

**2024**

## STATEMENT OF AUTHENTICITY

The undersigned below:

| | | |
|---|---|---|
| Name | : | Dimas Erlangga |
| ID of student | : | 20210130034 |
| Faculty | : | Computer Science |
| The Title of Thesis | : | Security Information and Event Management (SIEM) Enhancement Using Machine Learning for Detecting Cyber Attacks |

Stating truthfully that this thesis has nothing in common with other thesis. Thus this statement is made without coercion from any party. If this statement is not true, it will be sanctioned by the faculty leader.

Sukabumi, January 2024

Dimas Erlangga
NIM. 20210130034

**APPROVAL OF THESIS**

| | | |
|---|---|---|
| Title | : | Security Information and Event Management (SIEM) Enhancement Using Machine Learning for Detecting Cyber Attacks |
| Name | : | Dimas Erlangga |
| ID of student | : | 20210130034 |

The thesis has been reviewed and approved

Sukabumi, January 2024

| Head of Study Program, | Supervisor |
|---|---|
| Prof. Ir. Teddy Mantoro, M.Sc., PhD. | Prof. Ir. Media Anugerah Ayu, M.Sc., PhD |
| NIDN. 0323096491 | NIDN. 0315046903 |

## THESIS APPROVAL

| | | |
|---|---|---|
| Title | : | Security Information and Event Management (SIEM) Enhancement Using Machine Learning for Detecting Cyber Attacks |
| Name | : | Dimas Erlangga |
| ID of student | : | 20210130034 |

This Thesis has been tested and defended in front of the Board of Examiners in Thesis session on 27 January 2024. In our review, this Thesis adequate in terms of quality for the purpose of awarding the Master of Computer Degree.

Sukabumi, January 2024

| | |
|---|---|
| Supervisor 1, | Examiner 1 |
| Prof. Ir. Media Anugerah Ayu, M.Sc., PhD | Umar Aditiawarman, S.T.,M.Sc., PhD. |
| NIDN. 0315046903 | NIDN. 0424068107 |
| Supervisor 2, | Examiner 2, |
| Dini Oktarina Dwi H., S.T., M.Sc., PhD. | Haris Al Qodri Maarif, M.Sc., PhD, |
| NIDN. 0415108006 | NIDN. 0418068505 |

## PUBLICATION APPROVAL

As a member of the academic community of Nusa Putra University, i undersigned:

Name : Dimas Erlangga
ID of Student 20210130034
Study Program : Computer Science
Type of Work : Thesis

For the sake of scientific development, agree to grant to the University of Nusa Putra the Non-Exclusive Royality-Free Right for my scientific work entitled: **Security Information and Event Management (SIEM) Enhancement Using Machine Learning for Detecting Cyber Attacks.**

Along with existing devices (if needed). With this non-exclusive royalty-free right, Nusa Putra University has the right to store, transfer media/formats, process in the form of a database, maintain and publish my thesis as long as I keep my name as the author/creator and as the copyright owner.

This statement I made in truth.

Made in : Sukabumi
At the Date of : January 2024

That States

(Dimas Erlangga)

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## FOREWORD

Praise and gratitude to Allah SWT Almighty, because only with His blessings and grace can the writer complete this thesis. Writing this thesis is one of the requirements to achieve a Master's degree in Computers at Nusa Putra University. I realize that, without the help and guidance of various parties, from the lecture period to the preparation of this thesis, it is very difficult for the writer to complete this thesis. Therefore, I would like to thank:

1. Dr. Kurniawan, ST., M.Sc., MM. as Chancellor of Nusa Putra University;
2. Anggy Pradiftha Junfitrana, MT. as Vice Chancellor 1 for Academic Affairs;
3. Prof. Ir. Teddy Mantoro, M.Sc., PhD as Head of School Computer Science Nusa Putra University and Examiner 1;
4. Prof. Ir. Media Anugerah Ayu, M.Sc., PhD as Supervisor 1;
5. Dini Oktarina D.H., S.T., M.Sc., PhD. as Supervisor 2;
6. All Masters of Computer Science Lecturers who have provided very useful knowledge during lectures;
7. Mr. Lutfil Khakim, S.Kom., M.Si. for the support and encouragement to writer for completing study in Nusa Putra University;
8. Fellow comrades in Master of Computer Science batch 2021 who always give encouragement and support for the writer to complete this thesis;
9. All parties who have helped the writer in writing this thesis; For further improvement, suggestions and constructive criticism will be gladly accepted.

Sukabumi, January 2024

Writer

## ABSTRACT

Network security is a crucial component of Information Technology, yet organizations continue to grapple with meeting established security benchmarks. Given the rise in cyber attacks and the continuous emergence of new attack types, it's practically infeasible to persistently update attack patterns or signatures within security parameters. Key tools such as Intrusion Detection Systems (IDS) and Security Information and Event Management (SIEM) are instrumental in monitoring network traffic and identifying potential threats. However, these tools face limitations, such as the high volume of alerts produced by IDS and the use of rule-based method, also the inability of SIEM tools to analyze logs comprehensively to identify inappropriate activities. This research will conduct anomaly detection using machine learning process to classify cyber attacks network flow collected from IDS that installed inside network infrastructure. This process will also be integrated with SIEM. The algorithm used in this research is Random Forest Classifier using CSE-CID-IDS2018 dataset analyzed with Principal Component Analysis (PCA). Results of research shows that the application of PCA on balanced and imbalanced datasets demonstrates its effectiveness in dimensionality reduction, achieving high accuracy across training/testing splits, while balanced datasets, despite a slight decrease in accuracy, ensure fair class representation and efficient data management, particularly vital in resource-limited settings.



**Keywords—Network Security, IDS, SIEM, Machine Learning, Principal Component Analysis.**

# CHAPTER I
# INTRODUCTION

## 1.1 Research Background

Nowadays, Network security is a crucial component of Information Technology, yet organizations continue to grapple with meeting established security benchmarks. Predominant threats to public safety, such as identity attacks, intrusions, and hacking incidents, underscore the pressing need for effective information security solutions (Khan et al., 2017). Through targeting both internal and external network threats, network security can prevent the intrusion and proliferation of these threats within the network. A secured network entails an intricate array of hardware devices, including but not limited to firewalls, routers, and anti-malware utilities.

Within a network infrastructure, the task of logging data from network and server equipment, as well as monitoring and informing users about the system status, falls on the admin. As such, it is crucial to implement a comprehensive, centralized log management strategy in a infrastructure network. This strategy enables the analysis of events stemming from thousands of nodes, converging onto a few dedicated servers for centralized scrutiny. Real-time analytics can help identify potential security events from future occurrences through event correlation and other sophisticated surveillance methods. Additionally, it can serve as a retrospective forensic tool, allowing past incidents to be investigated to better comprehend any security breaches that have transpired (Isa et al., 2021). The core tenets of any network and security reporting system analysis involve collating data from various sources, pinpointing specific threats, and executing suitable responses. For example, the system can produce additional log information, trigger alerts, and guarantee the proactive monitoring and mitigation of security controls when such issues are identified. The term 'log management infrastructure' encompasses the network components, software, hardware, and media employed to create, disseminate, store, evaluate, and delete log data. Most organizations possess one or multiple log management infrastructures.

To consolidate log management, many organizations typically resort to Security Information and Event Management (SIEM) tools. These tools are specifically engineered to streamline corporate compliance reporting through a centralized logging solution. Each operational host is required to maintain a security log record for the report, and this log data can be transmitted to the SIEM server. A single SIEM server has the capability to gather log data from an unlimited number of devices, generate comprehensive reports, and manage all security events from each log it receives. In the current landscape, each system necessitates regular manual data retrieval from every device to ensure a centralized configuration can be generated for report production. The SIEM system server acts as a tool for identifying obscure events. Most of the equipment in use does not adhere to safety regulations and lacks depth in tracking events or logs. Even though such tools can recognize, monitor events and produce audit log entries, they fall short in analyzing logins to detect inappropriate activities. However, devices such as personal computers and laptops can alert users when an event transpires. SIEM devices can conduct advanced detection by correlating events or logs from the deployed equipment. By aggregating the events or logs from connected equipment, the SIEM system can detect multifaceted attacks that manifest differently across various devices, thus enabling it to record events or logs to determine the nature of the attack and its effectiveness (Isa et al., 2021).

SIEM is a system designed to handle logs of cyberattacks, which are generated from various data sources, often receiving substantial cyberattack data from IDS. Beyond this, SIEM can function as a live monitor, tracking real-time network traffic flow from multiple IDS. Serving as a sensor, IDS captures the network traffic flow and identifies network anomalies. Essentially, IDS detects cyberattacks within the network using a rule-based method and alerts the IT or cybersecurity team within the organization (Muhammad et al., 2023). An IDS serves as a security barrier, overseeing and analyzing computer and network activities to uncover any abnormal behavior or attacks (Tasneem et al., 2018). However, standalone IDS can pose challenges due to the sheer volume of alerts it generates. Often, IDS produces an overwhelming number of alerts,

making them unmanageable for the analyst or operator and consequently, preventing the IDS from being utilized to its maximum capacity. Furthermore, alarm quality may decline as regular network behavior constantly evolves and new threats continually emerge. Consequently, the IDS may miss actual attacks' alarms and report an excessive number of false alarms amid regular behavior. (Panda et al., 2011).

Given the rise in cyber attacks and the continuous emergence of new attack types, it's practically infeasible to persistently update attack patterns or signatures within security parameters. This is due to the constant repetition of this process, which is not efficient. (Hubballi & Suryanarayanan, 2014). As such, there is a need for an Intrusion Detection System (IDS) incorporating an anomaly detection technique. Anomaly detection is a method that employs machine learning techniques to discern if a network packet exhibits suspicious behavior akin to a cyber-attack. Utilizing anomaly detection enhances the network's efficiency in thwarting attacks by learning user patterns, allowing the system to identify any packets that deviate from the norm. The system then categorizes these unusual packets as anomalous. (García-Teodoro et al., 2009).

As the SIEM technology is constantly improve, there is still much differrence between traditional SIEM and next-generation SIEM's. Traditional SIEM (Security Information and Event Management) solutions are primarily focused on the collection and indexing of log data from various applications and devices. This log data is pivotal for conducting specific searches, such as retrieving all log entries for a particular device on a given day. These searches can yield extensive results, ranging from tens to hundreds of pages, and in cases where there is an issue with the device, the output can be even more substantial, sometimes reaching a thousand pages or more. To manage this vast amount of data, SIEM systems offer advanced filtering options. These filters allow users to refine their searches with greater precision, such as narrowing down to logs from a specific device at a particular time or focusing on certain types of log events. However, effectively utilizing these filters typically requires a high level of expertise from the user, as setting the correct parameters is crucial for obtaining relevant results. Another significant capability of SIEM systems is

their ability to correlate logs from multiple sources. For instance, when investigating a particular device identified by its IP address, a SIEM can aggregate logs from various sources related to that device. This feature is especially valuable for detailed forensic analyses and for auditing compliance event reporting, where understanding the interplay of different log entries is crucial. Some SIEM solutions also incorporate network data, but they often struggle to use this information effectively. Network data, such as flow data for a device, can create an overwhelming amount of additional information, adding thousands of line items to the search results along with the log data. Due to this inundation of data, network data integration is not commonly utilized in many SIEM systems. This limitation is a significant concern, as network data is crucial for detecting the most active threats. The network component provides critical context and visibility that complements log data, offering a more comprehensive view of the security landscape and enhancing the ability to detect and respond to threats effectively (Nichols, 2020).

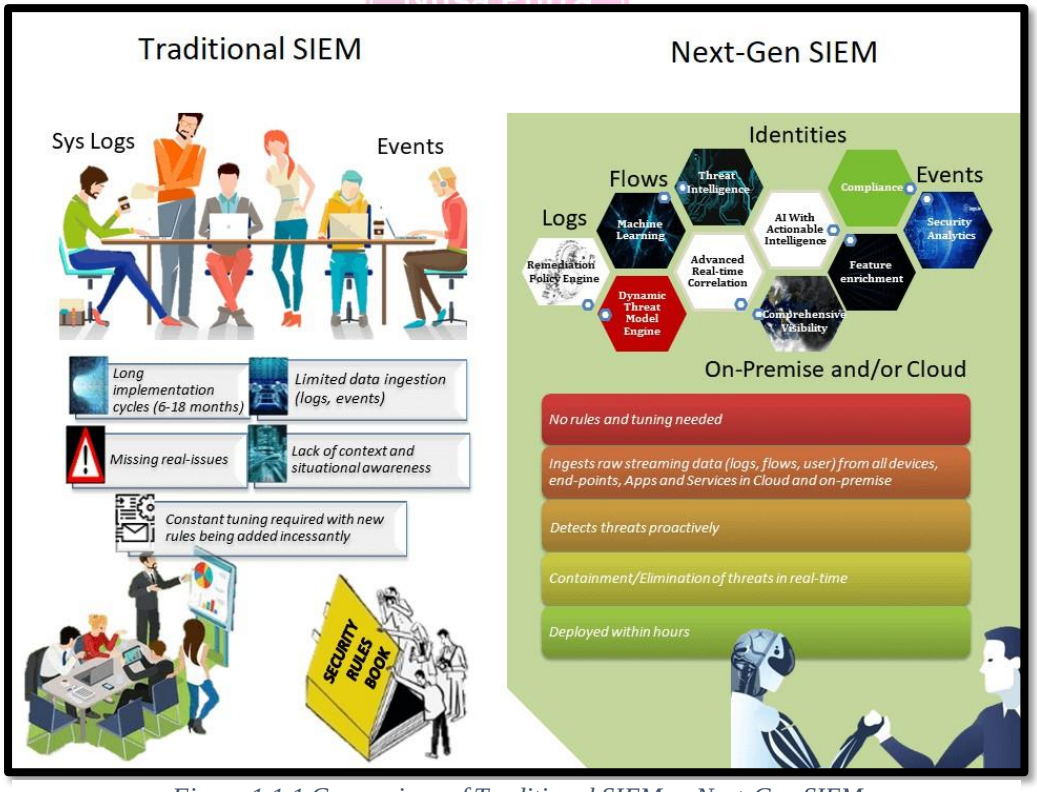


*Figure 1.1.1 Comparison of Traditional SIEM vs Next-Gen SIEM*

Traditional SIEM (Security Information and Event Management) systems are adept at gathering and providing a wealth of information, including some level of analysis that can hint at potential security incidents. For example, they might flag events like a user's credentials being changed, or a single user logging in from multiple devices simultaneously. However, a common challenge with these systems is the sheer volume of data they present. In trying to investigate a specific user or device, a user might have to sift through hundreds or even thousands of lines of information to understand the full context and identify what exactly is happening. In contrast, Next-Generation SIEM systems represent a significant evolution in this technology. These advanced systems not only ingest both log and flow data, but they also employ sophisticated threat models to automate the detection of threats, reducing the reliance on manual analysis by human operators.mThese models are complex and designed to identify and differentiate various types of threats, such as DDoS attacks, brute force attacks, malware infections, Advanced Persistent Threats (APTs), credential loss, or insider attacks. Next-Gen SIEMs use Machine Learning algorithms to recognize behaviors that are abnormal for a particular device, application, or user. These behaviors are then correlated with other rule triggers, forming a comprehensive threat model. When a match to a known threat model is found, the system generates an alert.

One of the key advancements in Next-Gen SIEMs is the way they present this information. Instead of inundating the user with hundreds or thousands of lines of data, they consolidate relevant threat behaviors into a single line alert on the user interface. This streamlined alert not only identifies the type of threat and the devices or users involved but also often includes recommendations for remedial actions. This approach significantly enhances the efficiency and effectiveness of the threat detection and response process, enabling quicker and more accurate decision-making. It represents a substantial improvement over traditional SIEMs, both in terms of data management and the actionable intelligence provided.

This research will conduct anomaly detection using machine learning process to classify cyber attacks network flow collected from IDS that installed

insiden network infrastructure. The analysis of IDS using machine learning, integrated with SIEM as described previously, constitutes a composite system encompassing numerous processes and services. The right selection of components, coupled with a well-configured system deployment, is vital to craft a robust and reliable analysis technique for IDS utilizing machine learning and integrated with SIEM. Additionally, the availability of open-source systems that are simple to deploy for industrial applications is also crucial.

## 1.2 Problem Statement

Key tools such as Intrusion Detection Systems (IDS) and Security Information and Event Management (SIEM) are instrumental in monitoring network traffic and identifying potential threats. However, these tools face limitations, such as the high volume of alerts produced by IDS and the use of rule-based method, also the inability of SIEM tools to analyze logs comprehensively to identify inappropriate activities. Given these challenges and the impracticality of continually updating attack signatures, there is a need to enhance the functionality of these systems.

In response to these issues, this research aims to investigate the application of machine learning techniques to improve the performance of IDS and SIEM systems in detecting and classifying network anomalies indicative of cyber attacks. By leveraging machine learning for anomaly detection, we can enhance the efficiency of these systems in thwarting attacks and managing the massive data they generate. The study will provide insights into creating a robust and reliable technique for IDS analysis, utilizing machine learning and integrated with SIEM, considering the vital role of component selection and system configuration.

## 1.3 Research Objectives

1. To Design and Develop Machine Learning Models for Detecting Cyber Attacks;
2. To Evaluate the machine learning performance based on ratio on training and testing datasets;

3. Perform detection using the developed model to enchance SIEM detection of cyber attacks.

## 1.4 Significance of Research

1. This research is significant as it utilizes machine learning algorithms to identify malicious traffic that can be identified as cyber attacks in network infrastructure.
2. The research's importance lies in enchancing capabilities of SIEM and IDS to detect cyber attacks using machine learning algorithm in order to increase capability of IT security team to adapt and mitigate various cyber attack in network infrastructure.

## 1.5 Scope of Limitations

The scope of this research focuses on implementing machine learning algorithm which is Random Forest Classifier to identify cyber attacks mainly in network infrastructure. The obtained model will be trained and evaluated for accuracy using CSE-CIC-IDS2018 dataset and utilized to analyze cyber attacks on network infrastructure with various simulated attack scenarios.



## 1.6 Thesis Structure

The rest of thesis is organized as follows:

1. Chapter I describes the background of problem that will be discussed in the thesis.
2. Chapter II describes the literature review of thesis.
3. Chapter III describes the methodology of thesis.
4. Chapter IV presents the expereiment result and discussion.
5. Chapter V presents the conclusion the thesis and future work.

Nusa Putra
UNIVERSITY
2002
AMOR DEUS, PARENTIUM, CONCERVIS

# CHAPTER V
# CONCLUSIONS AND RECOMMENDATIONS

## 5.1 Conclusions

| No | Research Objective | Research Findings |
|---|---|---|
| 1 | To Design and Develop Machine Learning Models for Detecting Cyber Attacks | Applied Principal Component Analysis (PCA) and Balancing the dataset to develop faster Machine Learning Model |
| 2 | To Evaluate the machine learning performance based on ratio on training and testing datasets. | Datasets Train with PCA and Balanced Datasets are prove to be less resourceful and produce high accuracy results. |
| 3 | Perform detection using the developed model to enchance SIEM detection of cyber attacks. | Model deployed can detect DDos Attacks that simulated in Virtual Environment and can be visualized in SIEM. |

*Table 5.1.1 Research Findings based on Objective*

From the analysis of the model training results utilizing PCA on balanced and imbalanced datasets, several key conclusions emerge. Firstly, PCA proves to be an effective technique for dimensionality reduction, allowing for the retention of critical information even in significantly reduced datasets. This effectiveness is evident in the high accuracy scores achieved across various training/testing splits in both the initial and main research phases, irrespective of data balance.

Secondly, while balanced datasets exhibit a marginal decrease in accuracy compared to imbalanced ones, this trade-off is crucial for ensuring a more equitable representation of classes within the model. This aspect is particularly important in applications where the accurate recognition of minority classes is as important as overall accuracy. Lastly, the reduction in training time with smaller datasets underscores the importance of efficient data

management. In resource-constrained environments, the ability to maintain high accuracy with reduced computational load and time is invaluable.

During the model deployment phase, the trained algorithms utilize the previously captured traffic data to make informed predictions. This stage is a critical part of a layered and robust network security strategy. It weaves together continuous traffic surveillance, the predictive power of machine learning, and the advanced pattern recognition capabilities of a SIEM system. Such an integrated system is adept at detecting and neutralizing cyber threats. Enhanced by the capabilities of traffic visualization, the system's efficiency in identifying imminent threats is significantly improved. Together, these elements create a formidable barrier against cyber adversaries, safeguarding the server. In today's digital landscape, marked by increasingly complex and frequent cyber threats, such a proactive and adaptive security approach is vital to preserve the integrity of the network and the protection of sensitive data.

## 5.2 Recommendations

PCA's effectiveness in reducing dataset size while maintaining accuracy is clear, making it an essential tool in scenarios with large datasets and limited computational resources. The high accuracy levels achieved across various training/testing splits, irrespective of dataset balance, attest to PCA's robustness. However, a slight accuracy trade-off observed in balanced datasets highlights the importance of considering dataset composition, especially in applications where class representation is crucial for fairness and accuracy, such as in medical diagnostics or fraud detection.

Continuation the use of PCA for efficient data management and the prioritization of balanced datasets to ensure equitable class representation. In resource-constrained environments, smaller, PCA-processed datasets are recommended for their efficiency in computational load and training time. Further research into PCA's optimal application in complex datasets is suggested, alongside continuous monitoring and updating of models to maintain their relevance in dynamic environments. Also, ethical considerations are crucial when balancing dataset accuracy and representation, ensuring that

machine learning model deployments are not only efficient and effective but also ethically sound and contextually appropriate.

Meanwhile, research is designed to explore and enhance network security strategies by integrating machine learning algorithms with Security Information and Event Management (SIEM) systems, supplemented by traffic visualization techniques. The objective is to investigate the effectiveness of this integrated approach in detecting and neutralizing cyber threats in an increasingly complex digital landscape. The methodology involves a combination of exploratory and applied research, utilizing both qualitative and quantitative methods. Based on the findings, the machine learning model trained with less resource data could provide cyber attack detection against attacks to network infrastructure. But, there's still a need to research more in order to create Machine Learning model that could detect cyber attacks against applications or websites based from its traffic behavior to improve attack detection comprehensively in enterprise infrastructure that also include applications.

## REFERENCES


Ahmed, N., Ngadi, A. B., Sharif, J. M., Hussain, S., Uddin, M., Rathore, M. S., Iqbal, J., Abdelhaq, M., Alsaqour, R., Ullah, S. S., & Zuhra, F. T. (2022). Network Threat Detection Using Machine/Deep Learning in SDN-Based Platforms: A Comprehensive Analysis of State-of-the-Art Solutions, Discussion, Challenges, and Future Research Direction. *Sensors*, *22*(20), 7896. https://doi.org/10.3390/s22207896

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Fitni, Q. R. S., & Ramli, K. (2020). Implementation of Ensemble Learning and Feature Selection for Performance Improvements in Anomaly-Based Intrusion Detection Systems. *2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, 118–124. https://doi.org/10.1109/IAICT50021.2020.9172014

García-Teodoro, P., Díaz-Verdejo, J., Maciá-Fernández, G., & Vázquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. *Computers & Security*, *28*(1–2), 18–28. https://doi.org/10.1016/j.cose.2008.08.003

Hubballi, N., & Suryanarayanan, V. (2014). False Alarm Minimization Techniques in Signature-Based Intrusion Detection Systems: A Survey. *Computer Communications*, *49*. https://doi.org/10.1016/j.comcom.2014.04.012

*IDS 2018 | Datasets | Research | Canadian Institute for Cybersecurity | UNB*. (n.d.). Retrieved July 28, 2023, from https://www.unb.ca/cic/datasets/ids-2018.html

Isa, M. R. M., Khairuddin, M. A., Sulaiman, M. A. B. M., Ismail, M. N., Shukran, M. A. M., & Sajak, A. A. B. (2021). SIEM Network Behaviour Monitoring Framework using Deep Learning Approach for Campus Network Infrastructure. *International Journal of Electrical and Computer Engineering Systems*, 9–21. https://doi.org/10.32985/ijeces.12.si.2

Khan, A., Khan, R., & Nisar, F. (2017). Novice threat model using SIEM system for threat assessment. *2017 International Conference on Communication Technologies (ComTech)*, 72–77. https://doi.org/10.1109/COMTECH.2017.8065753

Masduki, B. W., Ramli, K., & Murfi, H. (2018). Implementation and Analysis of Combined Machine Learning Method for Intrusion Detection System. *International Journal of Communication Networks and Information Security (IJCNIS)*, *10*(2), Article 2. https://doi.org/10.17762/ijcnis.v10i2.3375

Mohammed, S. B., Khalid, A., Osman, S. E. F., & Helali, R. G. M. (n.d.). Usage of Principal Component Analysis (PCA) in AI Applications. *International Journal of Engineering Research*, *5*(12).

Muhammad, A. R., Sukarno, P., & Wardana, A. A. (2023). Integrated Security Information and Event Management (SIEM) with Intrusion Detection System (IDS) for Live Analysis based on Machine Learning. *Procedia Computer Science*, *217*, 1406–1415. https://doi.org/10.1016/j.procs.2022.12.339

Nichols, C. (2020, June 9). Traditional SIEM vs. Next-Generation SIEM. *Cybriant*. https://cybriant.com/traditional-siem-vs-next-generation-siem/

Panda, M., Abraham, A., Das, S., & Patra, M. (2011). Network intrusion detection system: A machine learning approach. *Intelligent Decision Technologies (IDT) Journal, IOS Press, 5*, 347–356. https://doi.org/10.3233/IDT-2011-0117

Prihantono, Y., & Ramli, K. (2022). *Model-Based Feature Selection for Developing Network Attack Detection and Alerting System*. *6*(2). https://doi.org/10.29207/resti.v6i2.3989

Tasneem, A., Kumar, A., & Sharma, S. (2018). Intrusion Detection Prevention System using SNORT. *International Journal of Computer Applications*, *181*(32), 21–24. https://doi.org/10.5120/ijca2018918280

Yang, C.-T., Chan, Y.-W., Liu, J.-C., Kristiani, E., & Lai, C.-H. (2021, September 3). *Cyberattacks Detection and Analysis in a Network Log System Using XGBoost with ELK Stack*. https://doi.org/10.21203/rs.3.rs-838650/v1