Received: 15 May 2023 | Revised: 14 July 2023 | Accepted: 24 July 2023 | Published online: 25 July 2023

RESEARCH ARTICLE

Archives of Advanced Engineering Science

yyyy, Vol. XX(XX) 1–10

DOI: 10.47852/bonviewAAES32021068

Optimization of Security Information and Event Management (SIEM) Infrastructures, and Events Correlation/Regression Analysis for Optimal Cyber Security Posture

Akhigbe-mudu Thursday Ehis[1]*

1 Computer Science/Information Technology Department, African Institute of Science Administration and Commercial Studies Lome, Togo, akhigbe-mudut@iaec-university.tg.

***Corresponding author:** Akhigbe-mudu Thursday Ehis, Computer Science/Information Technology Department, African Institute of Science Administration & Commercial Studies Lome, Togo. Email: akhigbe-mudut@iaec-university.tg

Abstract: This work integrates logical and physical security processes, and simplifies the manageability of the security infrastructure. The process increases visibility to resources, which makes it easier to prevent security incidents, and provides a platform to manage the response and recovery after an incident occur. Log collection is the heart and soul of a SIEM. Log correlation is employed to identify particular sequences of log events from devices. The comparison between network level and host level events automatically perform initial validation that would not normally be performed. It considers movement of data between systems where it would not normally accounts logging on at unusual times or from unusual places, these may not generate specific security alerts, but can be much more easily spotted and flagged by a log correlation solution that sees everything in the environment. It shows some enhancements to event log normalization and significantly improves correlation rule execution. The event monitoring algorithm and SIEM correlation rules result in false positives or false negatives. Security managers, therefore, may waste time and resources that could be used to respond to real threats and assaults if there are too many false positives. This study hereby, strikes a compromise between lowering false positive alerts and not ignoring any potential abnormalities that could indicate a cyberattack when establishing SIEM correlation rules. In order to decide which data is pertinent and which data is irrelevant in an event pipeline, this research employs the use of filters. Through this examination, it can be inferred that the conditions are advantageous for promoting investment in the growth and enhancement of this technology as an essential component of industrial control systems with security operation centers, as well as offering cyber security management for small and medium-sized enterprises (SMEs) with restricted security expertise and capabilities.

Keywords: attackers, correlation rules, event log, false positive, false negatives

1. Introduction

Due to growing activity by nation-states and cybercriminals, cyber security hazards affecting industrial control systems (ICT) have significantly escalated over the past few years. Attackers are now more hazardous and technically advanced, making it difficult to identify them in time. The media, the security community, and the IT sector have in recent times focused on security due to extremely significant cyberattacks. The Solar Winds breach in 2019 affected the network management systems of numerous organizations (Luigi et al., 2022). This resulted in significant data leaks and caused a great deal of damage. As a result, these sectors have seen an increase in the number of complicated problems requiring optimization solutions. Companies must adapt rapidly to properly detect, respond to, and safeguard their surroundings as cyber-attacks steadily

increase in sophistication and frequency (Oyinloye et al., 2021).

1.1. Physical and logical security

Physical security refers to preventive measures put in place to prevent trespassers from physically entering the area. Physical security is effective in preventing unauthorized visitors and undesirable trespassers. It refers to physical or electronic equipment that safeguards the site's visible components. Physical security includes things like security vaults, closed-circuit televisions (CCTVs), and alarms. Unauthorized visitors can be attempting to steal equipment like computers and monitors, damage staff members, or extract company data (Demetrio 2017).

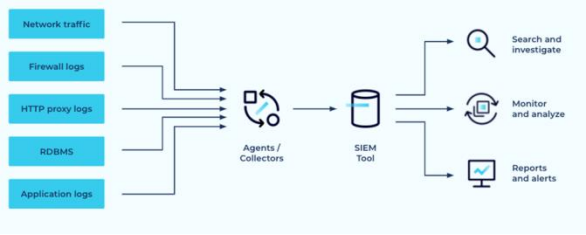
1.2. How SIEM works

Information security and network operations make up the two components of logical security. Information security focuses on ensuring the security of data, including data at rest and in transit. Network security is a concern for network operations. One can more effectively protect your assets by integrating the logical and physical security domains. When logical and physical security processes and infrastructures are combined, resource visibility is improved, making it simpler to detect and stop security events and giving you a platform to manage the response and recovery after they happen. Security Information and Event Management is referred to as SIEM. The best elements of Security Event Management (SEM) and Security Information Management (SIM) are combined in services that offer SIEM to enable real-time monitoring, notifications, event correlation and analysis accessible to its users (Gonzalez-Granadillo et al., 2017). Concisely, SIEM is a security tool that aids companies in spotting potential security flaws and threats before they have a chance to affect daily operations. It reveals unusual user behavior and has established itself as a mainstay in modern security operation centers (SOCs) for security and compliance management use cases. Artificial intelligence is used to automate many manual processes related to threat detection and incident response.

1.3. Understanding SIEM architecture

SIEM technologies enable the search and analysis of security incidents as well as the application of particular criteria for attack detection by combining data from various log sources. A feature of SIEM solutions called security analytics consists primarily of real-time dashboards that logically display security data as graphs and charts. The security team can quickly identify malicious activity and address security concerns thanks to these dashboards' automatic updates. Agents stationed at the infrastructure under surveillance gather the data. These agents are the components responsible for collecting data sent by various nodes or devices in the infrastructure being monitored and normalizing it into security events.

Figure 1
Understanding SIEM Architecture



SIEM architecture is focused on the development of SIEM systems and its essential elements. In a nutshell, SIEM architecture includes the following elements:

Keeping track of logs: This relates to data management, data collection, and the preservation of historical data. As shown in Figure 1 above, the SIEM gathers contextual data in addition to event data. Organizing systems like installed devices, network protocols, storage protocols (Syslog), and streaming protocols are used by SIEM architecture to gather event data.

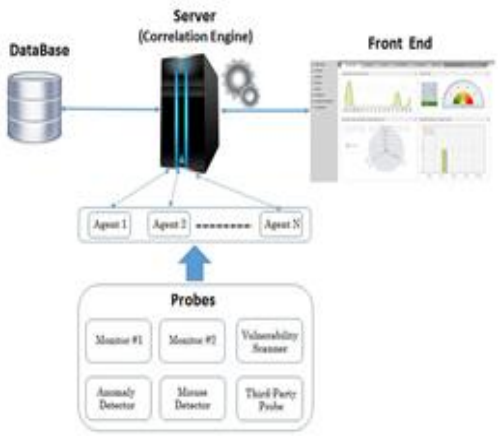
Log normalization: Figure 1 clearly shows that the event and contextual data are input into SIEM. However, it is necessary to normalize such. This relates to the process by which event data is converted into pertinent security insights.

Sources of Logs: Networking applications, security systems, and cloud systems all produce logs. This procedure is concerned with how organizations feed logs into the SIEM.

1.4. SIEM as a service

The fact is that cyber-attacks pose a constant threat to businesses and organizations. A real-time threat-monitoring program, like a Security Information and Event Management (SIEM) system, that offers visibility into the network is now a crucial layer of defense (Amantie, 2018). SIEM as a service is a collection of integrated security services, log management, and monitoring tools that deliver immediate incident response and threat detection. SIEM as a service helps businesses identify attacks and data breaches in their network more quickly. The objective of Managed SIEM is to reduce the risk of data breaches by enabling organizations to efficiently gather and analyze log data from all their digital assets. A top-notch SIEM solution brings together processes and technologies, ensuring that every interaction is secure and visible— and ensuring that the correct information is readily available to respond to potential threats, faster. With the assistance of a cloud-based SIEM, all network devices, servers, applications, users, and infrastructure components can be thoroughly and effectively monitored (figure 2).

Figure 2
SIEM as a Service



A company may have moved some of its workloads and workflows to the cloud in these circumstances. If this is the case, one thing needs to be understood: the risk surface a company faces will change. The danger surface has changed, necessitating an adjustment in the threat detection and response strategy. When all workloads and processes, including all physical hardware and data storage, are implemented on premises, it has the sole responsibility of managing the security infrastructure. Due to the divided duty model created by the deployment of things in the cloud, this role has undergone significant change (Lao et al., 2018). The business has a responsibility to safeguard and maintain the data on those systems. If one does not properly manage area of responsibility, the organization's attack surface will have a significant visibility gap. One won't be able to see the numerous security gaps. Clouds are naturally very active in nature. A workload in the cloud can be deployed or removed with a few simple clicks.

1.5. Motivation

No organization, no matter how big or little, is safe from knowledgeable and persistent cybercriminals. The idea that too many businesses simply invest in industry-standard cyber security measures like firewalls, antivirus software, and virtual private networks is even more unsettling. For instance, the global Marriott hotel chain experienced a series of significant data breaches that exposed the personal information of over 300 million visitors. But the fact that hackers hid in plain sight for more than four years astounded industry experts in managed IT and cyber security. The hotel company would have likely identified hackers before any serious monetary or reputational harm was done if it had properly implemented a Security Information Event Management (SIEM) system.

1.6. Statement of the problem

Organizations becoming inundated with security alerts is one of the most persistent SIEM challenges. Security event correlation generates alerts in traditional SIEM solutions when it detects potential incidents. However, these security alerts often misclassify regular actions and behaviors as linked attacks. These warnings, referred to as false positives, drain the investigative efforts of IT security teams, depleting their time, resources, and motivation. It allows real threats to persist for longer periods and contribute to fatigue.

These SIEM problems could, naturally, still emerge even with a threat intelligence source. The issue in this instance is a dearth of pertinent threat intelligence. The truth is that not all threat intelligence is created equal. For instance, ransomware that targets Internet of Things (IoT) in manufacturing may not have an impact on retail business. It requires a cybersecurity solution that offers several threat intelligence feeds relevant to the company in order to overcome these SIEM difficulties. Additionally, the threat intelligence feeds need to grow with the IT infrastructure and change along with the threat environment.

Update mechanisms are present in almost all operating systems. This notification system needs to be activated. Although the issue of whether updates ought to be installed automatically is up for discussion. At the very least, updates should be communicated to system administrators. Given that patches and updates have a history of causing more issues than they resolve, they might not want to have them loaded automatically. Administrators should install updates as soon as possible; waiting too long can expose systems to attack.

1.7. Event correlations

What is event correlation? Any size organization can have a lot of strange activity going on in its network, and keeping an eye on this activity can help protect network from threats. Security administrators may mark a user’s account as suspicious if there are 100 unsuccessful login attempts prior to a successful login, for instance. It can be challenging to determine the precise level needed to identify suspicious conduct. If you put up a rule of notification after 100 unsuccessful login attempts followed by a successful login, it will go undetected in the scenario above if the hacker cracks the password on the 90th attempt (Muhammad et al., 2022). Its requires more effective and dependable method of identifying potential risks to fix this.

Before offering logical solutions, event correlation analyzes a large number of events, adds business context to those events, and sequentially connects them. A set of rules are used to compare activity sequences in correlation. With the help of these guidelines, SIEM can determine which

suspicious activity ought to be taken seriously as a security concern.

For instance, one can specify a correlation rule to check for events X and Y that take place in a particular sequence, where X is the number of unsuccessful login attempts from a user account using a specific IP address and Y is the successful login using the same IP address to any network device. It will receive notifications if this rule is in effect each time a series of these occurrences takes place in the network. It can distinguish between possible threats and everyday occurrences using the preset variables in these events (Li et al., 2022). It requires the rules established by the SIEM solution or write its own criteria based on the requirements of the organization. The key to accurately detecting incidents is configuring the correlation engine of the security solution based on the nature of business, securing past and present experiences (Motorga et al., 2022). There are two types of correlation: **static and dynamic**.

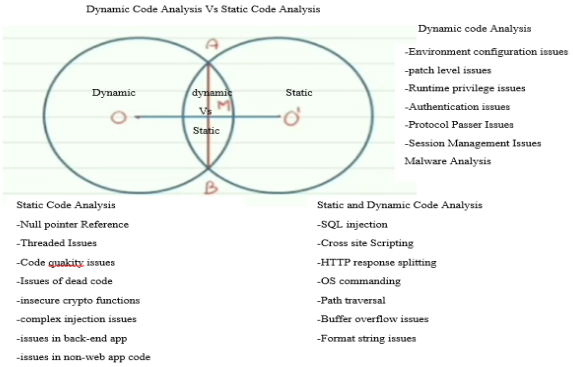
1.8. Static correlation

Enterprises cannot consistently rely on a preventive security approach. Unavoidably, breaches happen, and when they do, it's critical to investigate how and why in order to avoid recurrence and lessen the harm. Investigating old logs to examine breach activities after an occurrence is known as static correlation. Static correlation allows for the analysis of log data and the discovery of intricate historical patterns. This can inform you of an ongoing attack or assist you in identifying dangers that may have jeopardized the security of your network.

1.9. Dynamic analysis versus static analysis

Static analysis seeks to predict program behavior at the time of compilation, before the program is ever run. In contrast, the aim of dynamic analysis is to make inferences about run-time behavior of programs. Dynamic and static analyses can complement one another by supplying details that would not otherwise be available. It is more effective to perform both analyses simultaneously rather than one after the other (and possibly iterating) (figure 3). Since events are subject to correlation rules as they happen, a SIEM solution can immediately examine incoming log data and look for attack patterns. Static and dynamic correlation can be used to ensure that organization's network has a timely defense against security attacks (Motorga et al., 2022).

Figure 3
Dynamic and Static Analysis



Early bug detection and repair benefits the project in many ways. It can speed up development, lower costs, and guard against data breaches and other security flaws. Testing using both dynamic and static analysis are useful in this situation. Each one has a different function and offers a special return on investment (ROI) that is almost immediate for any development team. A two-pronged strategy using static and dynamic analysis enhances the development process' dependability, bug detection, efficiency, and security. Static analysis offers data that can be used to anticipate potential outcomes of the integration and execution of code. Based on what the tool defines as a defect, it finds flaws.

2. Related Work

Numerous proposals for the examination and assessment of security systems have been made in the literature. Others concentrate on the technological aspects that could be enhanced in present SIEM solutions, while some study focuses on the business aspects. For instance, well-known organizations like Gartner (Cruz-Duarte et al., 2022) present a commercial analysis of SIEM systems based on the market and main vendors, for which a report is produced on an annual basis to position SIEM companies as market leaders, challengers, niche players, or visionaries. To the best of our knowledge, there is no systematic evaluation of these systems, their capabilities, and the open gaps despite the fact that organizations like Gartner periodically review the competence of SIEMs (Xinjian et al., 2020).

In addition, numerous articles on the capabilities of SIEM solutions and on how SIEM providers can be compared and evaluated have been published by various security organizations (e.g., Techtarget (Li et al., 2022) and Info-Tech Research Group (Gonzalez-Granadillo et

al.,2021). On the one hand, Techtarget regularly publishes electronic guidelines on SIEM system security as well as how to define SIEM strategy, administration, and success in the business. On the other hand, Info-Tech offers technical reports on the SIEM vendor landscape (Adrian Olaru 2023) with an emphasis on the advantages and disadvantages of significant commercial SIEMs. Both firms use the Gartner Magic Quadrant as their starting point for study (Balayla Jacques 2020), putting the more complex factors to the side for consideration in upcoming SIEMs. Similar to this, businesses like Solutions Review by (Luigi et al.,2022) provides periodic studies to help SIEM purchasers choose the best SIEM solution for their companies. The authors conduct a vendor comparison map focused on compliance, log management, and threat detection, three essential SIEM characteristics. Although the study enables linking prospective customers with providers, it neither provides technical information about the tools nor examines potential enhancements to present SIEMs' capabilities or outside factors that might have an impact on their performance in the future (Muhammad et al., 2022).

Chicco et al., (2023) proposed that modeling attacks and evaluating security components will enable more precise and rapid assessments of network security elements. Apart from a few technical factors, no additional features are considered for the advancement of modern SIEM systems (Oyinloye et al., 2021). Based on the aforementioned limitations, we propose in this article a technical and commercial examination of current SIEM systems that could lead to enhancements in the design, development, and utilization of the following generation of SIEMs. The examination focuses on the deficiencies of current SIEMs and the external factors that may ultimately impact them. It provides an analysis and comparison of various commercial SIEMs from the past ten years (Perera et al., 2020).

A recent study by Gustavo et al., (2021) explores different perspectives in the 2020 SIEM vendor map based on three primary capabilities: (i) threat intelligence detection, (ii) compliance, and (iii) log management. In addition to threat intelligence, compliance, and log management, SIEM developers are considering security capabilities and intelligent dashboards as innovative additions to their solutions. Consequently, new SIEM systems will assist security administrators with pre-built dashboards, reports, incident response workflows, advanced analytics, correlation searches, and security indicators (Papastergiou et al., 2021). Furthermore, a comprehensive analysis of SIEMs by Subach et al., (2019) revealed that current SIEM solutions need to enhance features such as behavioral analysis, risk analysis and deployment, visualization, data storage, and response capabilities, in order to keep pace with the market. Therefore, there is an urgent need to devise new systems for efficient handling and providing a comprehensive and shared understanding of cyber-attack situations in a timely manner. In conclusion, the main limitations of the existing approaches can be summarized as follows: (i) the traditional linear incident response models are too slow, ineffective, and do not support the highly efficient capability required to handle and manage today's incidents. In contrast, our work presents a novel integrated approach that detects malicious activities and provides a thorough analysis of the identified anomalies in a more efficient, flexible, and scalable manner. To achieve this, it combines proactive approaches, which detect and analyze abnormal activities and attacks in real-time, with reactive approaches, which provide a comprehensive analysis of the underlying infrastructure to evaluate the reported incident.

3. Methodology

3.1. Scalability method

Before a system's security can be ensured, it must be scalable. In order to examine the scaling strategy from the perspective of security, let us look at a web server. The theoretical calculation of the web server demand is shown in Figure 4a. You must take into account a crucial question in order to fully comprehend the workload on a web server: if the average time between incoming requests is 100 ms (milliseconds), how many requests are received on average of one second? Let us label the unknown value in order to mathematically explain this. Think of T as representing, for instance, the interval between server requests (Murtaza & Wooguil 2021).

Figure 4a
Communications between two parties surfing the web

Up until now, the existing communication channel, illustrated in figure 4a, has exerted significant effort in safeguarding all its systems from cyber-attacks, malware variations, and various other malicious methods. In this regard, the objective of combating cyber-attacks and safeguarding critical national infrastructures has led to the deployment of a series of security sensors at multiple points within the port's technological ecosystem, which is comprised of Information Technology (IT). These sensors are devices or software programs that monitor and gather network data, as well as activity data from systems, and offer

mechanisms for early identification of attacks and system vulnerabilities. This enables prompt implementation of countermeasures to mitigate these threats (Kalyan et al., 2022).

3.2. Mathematical model of correlation coefficient

The aforementioned example shows that node perception data have some degree of similarity within a specific range. It determines the average amount of information delivered by a packet by computing the entropy of each node (Cruz-Duarte et al., 2022). To express node data correlation, the joint entropy can be derived by comparing the correlation entropy value with a predetermined threshold. Expressions of information entropy are described as follows:

$$H(x) = -\sum_{i=1}^{q} p(x_i) x \log p(x_i) \quad 1$$

$$H(y) = -\sum_{i=1}^{q} p(y_i) x \log p(y_i) \quad 2$$

H(x), H(y) separately represent the node's entropy in the two formulas above. $p(x_i)$ is the likelihood of events detected by the node i. $p(y_j)$ is the likelihood that events will be detected by node j, where q is the total number of events being gathered. The following is how the joint entropy between nodes i and j is expressed:

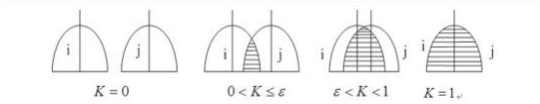
$$H(x,y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log p(x_i, y_j) \quad 3$$

In formula 3, $p(x_i, y_j)$ is the joint probability consisting of two-dimensional random variables (x_i, y_j). In general, $H(x,y) \le H(x) + H(y)$ if $H(x,y) = H(x) + H(y)$, which means information collected by two nodes are independent. H(x, y) \ (H(x) + H(y) represents the correlation degree of data collection for i^{th} node and the j^{th} nodes, and K is the correlation coefficient:

$$K = 1 - \frac{H(x,y)}{H(x) + H(y)} \quad 4$$

K's data range is obviously [0:1]. K = 0 denotes the independence of the data acquired by nodes i and j. The higher the K-value, the greater the correlation between the two nodes. K = 1 indicates that the information gathered by nodes i and j is the same. On the basis of conforming to the requirements of the network application, nodes are separated according to the correlation coefficient K achieved. The threshold value is set to the parameter = 0.8. Node i is capable of sensing information flow and the area it is associated with. Fig. 4b displays a node perceptual relevance schematic (Arul et al., 2023).

Figure 4b
Network conformity requirements



The similarity of nodes is low when $0 < k < \varepsilon$. The similarity of the nodes increases when $\varepsilon < k < 1$. The correlation coefficient between nodes can be determined by calculating their combined entropy and individual information entropies. The nodes that satisfy $0 < k < 1$ are separated into groups. All of the nodes in the similarity coefficient, which make up the relevant area (correlated area, CA), are automatically divided into numerous smaller sections. According to Vasillii et al. (2022), the cycle formulas 4 and 5 are used to calculate the merits of priority for each node within the pertinent area. A representative node (RN) with a cycle length of T is selected for the relevant area. To prevent the interruption of original data transmission brought on by the switch of RN, the value of T must be greater than and at least twice as large as the value of the EA-MAC protocol listener sleep cycle. Only RN will communicate the collected data to the sink node.

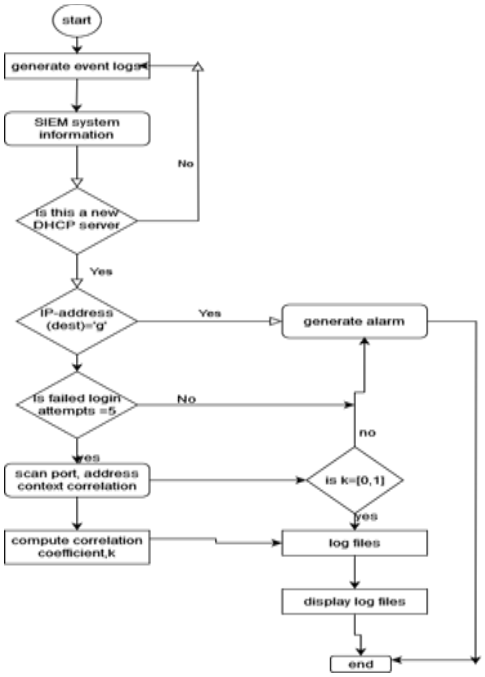
3.3. Correlation algorithm

INPUT: Sink Nodes, SIEM systems
OUTPUT: correlation Coefficient,
(i) Generate event logs,
(ii) Nodes receive (messages, sync)
(iii) Check for new DHCP servers, destination IP address and check if it is within the list (“Z”)
(iv) If five failed attempts login are tried with different usernames from the same IP address within fifteen minutes (“x”), send an alarm to the administrator,
(v) Scan port and add context to correlation if the IP address of the scan port and the login attempts are the same.
(vi) Computing entropy according to formula 1, 2, 3
(vii) Compute correlation coefficient, k, using formula 4
(viii)Is k = [0,1]
(ix) Analyze a whole lot of event logs with mundane activities,

(x) To make sense of them, read the event log format. The goal of the event log format is to standardize the event log formats used by various vendors and network components across the entire network.

(xi) The goal of the event log format is to standardize the event log formats used by various vendors and network components across the entire network.

Figure 5
Flowchart computing correlation Coefficient



3.4 Result and discussion

Platforms for (SIEM) offer real-time analysis of security events generated by network hardware and software. The core elements of a typical SIEM solution are depicted in Figure 5 and the pseudocode. The SIEM is comprised of several functional components, each capable of operating independently but requiring collaboration for optimal efficiency. These components encompass the source device, log aggregation, parsing standardization, rule engine, log storage, and event monitoring. Network security devices such as Intrusion Detection Systems (IDS), Intrusion Prevention Systems (IPS), and firewalls are a few examples of systems that generate extensive logs. The architecture effectively integrates two crucial components/features (data processing APIs), one for continuous processing and the other for batch processing (Tharwat 2018). In order to swiftly detect abnormal incidents and/or patterns, these two complementary attributes must be present. Significantly, the batch processing capabilities provide stream processing features with the necessary contextual/domain knowledge to identify deviations in observed events/values from established norms (Kalyan et al., 2022).

Like any event-monitoring algorithm, SIEM correlation rules might provide false positives. Security managers may waste time and resources that could be used to respond to real threats and assaults if there are too many false positives. The SIEM correlation rules' setting aims to achieve a balance between lowering false-positive alert rates and not ignoring any potential anomalies that might point to a cyberattack. A security administrator won't be as likely to make mistakes or overlook events thanks to the better event log normalization that is provided here.

3.5. Evaluation

To overcome classification problems, we use the Area under Curve - Receiver's Operating Characteristics (AUC-ROC) curve. It is one of the most important evaluation criteria for determining whether a classification model is effective (Marcelo et al., 2022). On the other hand, the false positive rate (FPR) measures the proportion of available negative samples that contain false positive results. The x and y- axes of a ROC space represent the relative trade-offs between true positives and false positives as described by FPR and TPR.

As a result, the metrics listed below (Eqs. (5) to (10)) are defined. Equation (5)'s definition of accuracy (ACC) shows that it is the proportion of correctly classified samples to all samples of data. When the training dataset contains an equal number of data samples for all classes, ACC can be used as an objective evaluation metric:

$$Accuracy(ACC) = \frac{TP + FN}{TP + TN + FP + FN} \quad 5$$

The proportion of typical behaviors that are mistakenly identified as malicious or anomalous is known as the false positive rate (FPR; equation (6)). By dividing FP with the total of FP and TN, FPR is calculated.

$$FPR = \frac{FP}{FP + TN} \quad 6$$

The True Positive Rate (TPR) (Equation (7) determines the percentage of actual malicious or anomalous activities identified as malicious or anomalous. TPR is determined by dividing TP by the total of FN and TP with an emphasis on FN.

$$TPR = \frac{TP}{TP + FN} \quad 7$$

The F1 score (equation (8)), which incorporates FN and FP, expresses the ideal ratio between TPR and Precision. Another evaluation metric, precision, calculates the

percentage of data samples that are labeled as malicious or anomalous.

$$FI = \frac{2 \times \Pr ecision \times TPR}{\Pr ecision + TPR} \quad 8$$

where $$\Pr ecision = \frac{TP}{TP + FP}$$

In other words, it depends on how well the risk of mistaking a real attack for a false one and the risk of mistaking a fake attack for a real one are balanced. Different criteria are applied when making decisions. The effectiveness of each risk balancing technique must therefore be carefully considered. The study makes use of phrases like "true positive" and "false positive." As displayed in table 1 below: "True negative" and "false negative," respectively.

Table 1
shows True Positive, False Positive, False Negative and True negative in Tabular Format

	Positive attack	False attack
Positive Attack	True positive	False positive
False Negative	False negative	True negative

The rates of type i errors and type ii errors are respectively, the power and false positive rate are defined as follows:

$$power = \frac{TP}{TP + FN} = 1 - \beta \quad 9$$

$$FalsePositiveRate = \frac{TP}{FP + TN} = \alpha \quad 10$$

The ROC curve is the curve with the power on the vertical axis and the false positive rate on the horizontal axis. The higher the curve rises to the upper left corner of the plot, or the larger the area under the ROC curve (AUC, maximum of 1), the better the test performance, as shown in figure 6 (Li et al., 2022).

Table 2

An explanation of area under the curve

Area Under Curve Interpretation	
AUC Value	Interpretation
⩾0.9	Excellent Model
0.8-0.9	Good Model
0.7-0.8	Fair Model
0.6-0.7	Poor Model
＜0.6	Very Poor Model

The AUC needs to be higher than 0.5 for a diagnostic test to be considered useful. is typically regarded as acceptable (Kalyan et al., 2022).

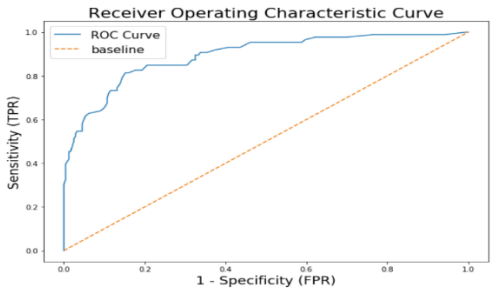
3.6. Pseudocode for AUC-ROC curve implementation

Having defined most of the metrics that involve the evaluation of the models, the probability threshold that gives the best performance for the situation is hereby stated. This is where the ROC or Receiver Operating Characteristic Curve comes into play. It is a graphical representation of how two of these metrics (the Sensitivity or Recall and the Specificity) vary as we change this probability threshold. Intuitively, it is a summarization of all the confusion matrices that we would obtain as this threshold varies from 0 to 1, in a single, concise source of information.

Figure 6
Pseudocode for the ROC Implementations

```
N_0 = 1000; N_1000
Mu_1 = mu_ 0 =1#TPR: 1; FPR: 0
Var_1; var_0 =1
X = np. Random. Normal (mu_1; var_0; N_0
Y = np. Random (mu_1; var_1, N_1) theta.seq = np. Exp {(np. Arrange (-10,100.1)}
U = [J; V = [ ]
For I in range {Len (theta. Seq)};
U = np. Sum (stats. Norm pdf (x, mu_1, var_1)/stats. Norm pdf (x, mu_1, var_0)}>theta. Seq [i]}/N_0
# (TPR) treated as (FPR)
V = np. Sum (stats. Norm pdf (y. mu_1, var_1) / stats pdf (y, mu_0, var_0)} > theta. Seq [i] /N_1
# treated real attack (TPR) as positive attack
u. append (u); append (v);
AUC = 0 # estimate the area;
For I in range (Len (theta. Seq) = 1);
AUC = AUC + np. Abs (u (u (i+1) - v[i] + v[i];
Plt.plot (u, y)
Plot label {"false positive Rate}
Plot x label {"True _ Positive}
Plot title {"ROC CURVE"}
Plot text (0.3, 0.5, AUC = { }' format {AUC font size = 1.5}
Text (0.3, 0.5, 'AUC = 0.93'}
```

Figure 7
ROC curve shows all the performance of the test for acceptable false positive (kitchen et al., 2023).



An ROC curve is a graph that shows the performance of the receiver operating characteristic curve (ROC) at all the classification thresholds. The ROC curve plots the True Positive Rate (PPR) and False Positive Rate (FPR). The first step in plotting the ROC curve is to calculate the Recall and FPR for the different thresholds. Then, plot the two values against each other. Fig. 7 shows the ROC curve of a random model: The line that goes from 0,0 to 1,1 is the ROC curve. This is the ROC for a random model that predicts a half-time value of 0 and a 1-half-time value of 1, regardless of its inputs.

4. Conclusion

This study further supports the notion that SIEM aggregates logs through data correlation, enabling security analysts to sift through billions of logs generated by network devices (González-Granadillo et al., 2021). Security concerns can then be sorted by risk factors, providing immediate action to minimize the attack surface and improve network health. This research is armed with the features and capabilities to fight suspicious network activity while producing report logs to comply with industry policies and practices. Network security equipment such as IDS, IPS and firewall devices generate many logs. This course shows how SIEM alerts security analysts about events and trends they need to be aware of. One of the most important components of a functional SIEM is good and logical correlation rules. Improving event log normalization significantly improves SIEM's and its correlation rules' functionality. If SIEM can normalize event logs, it is less likely to make errors or ignore events a security analyst needs to be concerned about (Oyinloye, 2021). To differentiate between valuable and insignificant data in the event pipeline, this research employs the utilization of rules. It becomes more captivating when employing log-based activity information and correlation inspired by security events to address business issues. This function as a tool for managing security has provided organizations and their networks with a strong security base.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

References

Adrian Olaru (2023): "Dynamic Modeling and Simulation for Control Systems." Mathematics, ISBN978-3-0365-7104-1; ISBN 978-3-0365-71058, March 2023 Pages: 242 https://doi.org/10.3390/books978-3-03657105-8

Balayla, Jacques (2020). "Prevalence threshold (ϕe) and the geometry of screening curves". PLoS One. 15 (10). https://doi.org/10.1371/journal.pone.0240215

Cruz-Duarte, J.M.; Toledo-Hernández, P. (2022): "Fractional Calculus in Mexico: The 5th Mexican Workshop on Fractional Calculus (MWFC)". Comput. Sci. Math. Forum 2022, 4, 4007. https://doi.org/10.3390/cmsf2022004007

Chicco D.; Jurman G. (2023). "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification". BioData Mining. 16 (1). https://doi.org/10.1186/s13040-023-00322-4.

Demetrio's Psaltopoulos; Andrew J. Wade; Dimitris Skuras; Martin Kernan; Emmanouli Tyllianakis; Martin Erlandsson (2017):" Science of The Total Environment". Science of The Total Environment, Volume 575, 1 January 2017, Pages 1087-1099. https://doi.org/10.1016/j.scitotenv.2016.09.181

González-Granadillo, G.; González-Zarzosa, S.; Diaz, R. (2021): Security Information and Event Management (SIEM): Analysis, Trends, and Usage in Critical Infrastructures. Sensors 2021, 21, 4759. https://doi.org/10.3390/s21144759

Hélio Amante Miot (2018):" Correlation analysis in clinical and experimental studies". J Vasc Bras. 2018 Oct-Dec; 17(4): 275–279. https://10.1590/1677-5449.174118

Kalyan, C.N.S.; Goud, B.S.; Bajaj, M.; Kumar, M.K.; Ahmed, E.M.; Kamel, S. (2022):"Water-Cycle-Algorithm- Tuned Intelligent Fuzzy Controller for Stability of Multi-Area Multi-Fuel Power System with Time Delays. Mathematics 2022, 10, 508. https://doi.org/10.3390/math10030508

Kitchenham B.A., Madeyski L., Budgen D. (2023):"SEGRESS: Software engineering guidelines for reporting secondary studies. "IEEE Trans. Softw. Eng. (2023), https://doi.org/10.1109/TSE.2022.3174092

Li, Y.; Liang, H. Robust (2022):" Finite-Time Control Algorithm Based on Dynamic Sliding Mode for Satellite Attitude Maneuver." Mathematics 2022, 10,111. https://doi.org/10.3390/math10010111

Lao, X., Liu, X., Deng, H., Chan, T., Ho, K., Wang, F., Yeoh, E. (2018). Sleep Quality, Sleep Duration, and the Risk of Coronary Heart Disease: A Prospective Cohort Study with 60,586 Adults. Journal of Clinical Sleep Medicine, 14(1), 109-117. https://doi.org/10.5664/jcsm.6894

Luigi Coppolino1, Luigi Sgaglione1, Salvatore D'Antonio1, Mario Magliulo, Luigi Romano, Roberto Pacelli (2022):" Risk Assessment Driven Use of Advanced

SIEM Technology for Cyber Protection of Critical e-Health Processes". SN Computer Science (2022) 3:16 https://doi.org/10.1007/s42979-021-00858-4

Muhammad Bilal Khan, Gustavo Santos-García, Hatim Ghazi Zaini, Savin Treanță and Mohamed S. Solim (2022):" Some New Concepts Related to Integral Operators and Inequalities on Coordinates in Fuzzy Fractional Calculus". Mathematics 2022, 10(4), 534; https://doi.org/10.3390/math10040534

Murtaza Ahmed Siddiqi and Wooguil Pak (2021):" An Agile Approach to Identify Single and Hybrid Normalization for Enhancing Machine Learning-Based Network Intrusion Detection." Journal: IEEE Access, 2021, Volume 9, Page 137494. https://doi.org/10.1109/ACCESS.2021.3118361

Motorga, R.; Mures, an, V.; Ungures, an, M.-L.; Abrudean, M.; V˘alean, H.; Clitan, I. (2022): "Artificial Intelligence in Fractional-Order Systems Approximation with High Performances: Application in Modelling of an Isotopic Separation Process. Mathematics 2022, 10, 1459. https://doi.org/10.3390/math10091459

Mahmudul Hoque Mahmud, Md. Tanzirul Haque Nayan, Dewan Md. Nur Anjum Ashir and Md Alamgir Kabir (2022):" Software Risk Prediction: Systematic Singh, P., Sreenivasan, S., Szymanski, B. et al. Threshold-limited spreading in social networks with multiple initiators. Sci Rep 3, 2330 (2013). https://doi.org/10.1038/srep02330Literature Review on Machine Learning Techniques. Journal: Applied Sciences, 2022, Volume 12, Number 22, Page 11694. https://doi.org/10.3390/app122211694

Marcelo Cerqueiraa*, Paulo Silvab, Sergio Fernandes (2022);" Systematic Literature Review on the Machine Learning Approach in Software Engineering." American Academic Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS) (2022) Volume 85, No1, pp 370- 396

Oyinloye, D.P.; Teh, J.S..; Alawida, M.; Jamil, N. (2021):" Block chain Consensus: An Overview of Alternative Protocols. Symmetry 2021, 13, 1363. https://doi.org/10.3390/sym13081363

Perera A., Aleti A., Böhme M., Turhan B. (2020):" Defect prediction guided search-based software testing." Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering, ACM (2020), pp. 448-460, https://doi.org/10.1145/3324884.3416612

S. Arul Jothi, R. Venkatesan and V. Santhi (2023):" Rule-Based Outlier Detection with a Modified Variation Auto Encoder for Enhancing Data Accuracy in Wireless Sensor Networks." Journal: International Journal of Fuzzy Systems, 2023. https://doi.org/10.1007/s40815-023-01496-z

Szymon Stradowski, Lech Madeyski (2023): "Industrial applications of software defect prediction using machine learning: A business-driven systematic literature review. Information and Software Technology." Volume 159, July 2023, 107192. https://doi.org/10.1016/j.infsof.2023.107192

Spyridon Papastergiou, Haris Mouratidis, Eleni Maria Kalogeraki (2021):"Handling of advanced persistent threats and complex incidents in healthcare, transportation and energy ICT infrastructures". March 2021 Evolving Systems 12(5), https://doi.org/10.1007/s12530-020-09335-4

Subach, I., Mykytiuk, A., & Kubrak, V. (2019). Architecture and functional model of a perspective proactive intellectual SIEM for cyber protection of objects of critical infrastructure. Collection "Information Technology and Security", 7(2), 208–215. https://doi.org/10.20535/2411-1031.2019.7.2.190570

Tharwat A. (August 2018). "Classification assessment methods". Applied Computing and Informatics. https://doi.org/10.1016/j.aci.2018.08.003

Vasilii Mosin, Miroslaw Staron, Yury Tarakanov and Darko Durisic (2022):" Comparing autoencoder-based approaches for anomaly detection in highway driving scenario images." Journal: SN Applied Sciences, 2022, Volume 4, Number 12. https://doi.org/50.10.1007/s42452-022-05160-3

Xinjian Yu, Siqi Lai, Hongjun Chen and Ming Chen (2020): "Protein–protein interaction network with machine learning models and multiomics data reveal potential neurodegenerative disease-related proteins". Journal: Human Molecular Genetics, 2020, Volume 29, Number 8, Page 1378, https://doi.org/10.1093/hmg/ddaa065

How to Cite: Thursday Ehis, A.- mudu. (2023). Optimization of Security Information and Event Management (SIEM) Infrastructures, and Events Correlation / Regression Analysis for Optimal Cyber Security Posture. Archives of Advanced Engineering Science, 1–10. https://doi.org/10.47852/bonviewAAES32021068