HCIS | Human-centric Computing and Information Sciences

RESEARCH (Research Manuscript) **Open Access**

Human-centric Computing and Information Sciences (2023) 13:17
DOI: https://doi.org/10.22967/HCIS.2023.13.017
Received: April 26, 2022; Accepted: November 1, 2022; Published: April 30, 2023

Effective Security Monitoring Using Efficient SIEM Architecture

Muhammad Sheeraz[1], Muhammad Arsalan Paracha[1], Mansoor Ul Haque[1], Muhammad Hanif Durad[1], Syed Muhammad Mohsin[2,3], Shahab S. Band[4,*], and Amir Mosavi[5,6,*]

Abstract

The unprecedented advances and myriad benefits of the internet have made it indispensable for almost every organization. With its growing popularity and widespread use, the problem of security threats has emerged to the forefront, while attacks are constantly on the rise. Therefore, an organization must continuously monitor its security status to take immediate remedial measures. Security information and event management (SIEM) systems in tandem with security orchestration, automation, and response (SOAR) systems are an integral part of a security operation center (SOC) because this not only further helps organizations gain a holistic view of their security status but also protects their IT infrastructure. In this research paper, we will provide discussions on the latest and most advanced and widely used SIEM systems. These include both open-source and proprietary solutions. However, as documented in literature, no comprehensive SIEM system architecture is available. The main contribution of this research work is that we have proposed a comprehensive, well-defined and modular architecture of the SIEM system. Each module has been discussed in detail with reference to its input parameters, processing, and output details. This modular approach will help developers extend the functionality of the SIEM system without compromising the overall performance and integration issues, while also helping end users make better decisions to select a SIEM system.

Keywords

Security Information and Event Management, Security Operation Center, Data Aggregation, Log Formats, Data Normalization, Event Correlation, Correlation Engine, Big Data

1. Introduction

Every firm and business organization in the modern world requires a high-performance network to function properly. Any network failure can possibly incur a loss of time, money, and other valuable resources. A network monitoring software solution is highly needed if the purpose is to better monitor

***Corresponding Authors:** Shahab S. Band (shamshirbands@yuntech.edu.tw), Amir Mosavi (amir.mosavi@uni-obuda.hu)
[1]Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences (PIEAS), Islamabad, Pakistan
[2]Department of Computer Science, COMSATS University Islamabad, Islamabad, Pakistan
[3]College of Intellectual Novitiates (COIN), Virtual University of Pakistan, Lahore 55150, Pakistan
[4]Future Technology Research Center, College of Future, National Yunlin University of Science and Technology, Douliu, Taiwan
[5]Institute of Information Society, University of Public Service, Budapest, Hungary
[6]Faculty of Informatics, Obuda University, Budapest, Hungary

the current state of the network processes. It can alert as to any irregularity regarding the network functioning, and thus provides many benefits like more time for essentials, added security, better control, and increased potential and financial savings [1]. To that end, the security personnel have to analyze a flow of security alerts, since none of these can be ignored. Thus, this has caused a substantial rise in the volume of security alerts. Nowadays, the volume of security alerts has turned from manageable forms to unmanageable ones unless some sort of prioritization of alerts is performed. Given the fact that no current and previous (e.g., for a specific period of at least 6 months) security alert can be ignored for security analysis, this is another momentous task we should not neglect.

Various devices, such as routers, switches, firewalls, antivirus, intrusion detection system (IDS), intrusion prevention system (IPS), servers, personal computers (PCs), etc., have been widely used in every organization. These devices generate a huge number of logs and alerts. To view, categorize, and analyze this massive amount of data, a corresponding amount of time and human effort is highly needed [2]. This could still produce errors and miss important logs or alerts. Critical security information must be provided in time for correct and immediate remedial measures to be taken accordingly. A delay of report in a security event tends to reduce its usefulness to a great extent. From an operator's perspective, it is neither feasible to analyze such a massive quantity of real time data nor match that data against a set of rules and policies [3].

Innumerable types of cyberattacks have emerged during the last two decades. Some examples of common attacks include denial-of-service (DoS), distributed denial-of-service (DDoS), SQL injection (SQLi), cross-site scripting (XSS), cyber vandalism, espionage, web session hijacking attacks, etc. [4–6]. Ransomware is another example that has targeted the world for the worse [7]. According to [8], ransomware attackers extorted US$412 million in payouts in 2020. Various security systems, such as firewalls, IDS, IPS, etc., have been deployed worldwide for protection against such threats. Moreover, to improve the performance of these security systems, the use of machine learning-based techniques has been also highly considered [9]. Despite this, most of the time, such attacks remain undetected by a single security device, so there is a need for a multilayered solution to be made.

The security information and event management (SIEM) system is a perfect solution for this problem because this solution can analyze a massive amount of data rapidly and then produce the required outcome in a minimum amount of time. It keeps security personnel aware of the overall security of the organizational network. The SIEM system can be agent-based, agentless, or hybrid. By this agent-based SIEM system, an agent can be deployed on the data source for security events to be automatically collected while in progress and to bring themselves to the SIEM system. In an agentless SIEM system, no agent is needed for the data source. Instead, the data source communicates security events directly to the SIEM system or by an intermediate logging server, such as a syslog server. The hybrid SIEM system takes advantage of both above techniques [10].

On top of the SIEM system, the security orchestration, automation, and response (SOAR) system is also becoming popular among security professionals. The SOAR system resides at a higher level than a SIEM system. Therefore, a SOAR system has an even broader security view of an organization than the SIEM system, which directs its input to the SOAR system [11]. Over the past few years, block chain-based systems have flourished rapidly, while the SIEM system is especially suitable for secure smart logistics systems [12]. Among other reasons, the following two reasons are the most important for deploying a SIEM system.

- A massive amount of security log data is generated by today's computer networks, so it is impossible to process and analyze this security log data without the use of a SIEM system.
- The SIEM system can perform contextual data analysis, which is crucial for detecting today's malware, as such attacks are not detected by other security devices working independently.

Every organization should select a SIEM system in accordance with its own requirements. The network monitoring solution should be simple, user friendly, attractive, and intuitive. In this way, the administrator can do other useful tasks. It should also provide support for the most common protocols like SNMP, NetFlow, sFlow, and jFlow. Owing to today's growing fields of cloud computing and virtualization, the

network monitoring solution should also support them. Before making a final selection, the organization should review relevant terms and conditions associated with monitoring solutions, and its future support [1]. The SIEM system might cost several hundred thousand dollars [13]. Some SIEM systems provide limited capabilities with lower cost or at free of charge as well, with multiple product capabilities of real-time security monitoring, threat intelligence, behavior profiling, data and user monitoring, application monitoring, analytics, and log management and reporting.

The SIEM system considered suitable for one organization is not necessarily best suited for another organization. As such, every organization should take its environment and different selection criteria into account, and then select an optimal SIEM system. Two selection criteria categories are proposed as functional selection and technical criteria. Functional selection criteria entails whether the SIEM system does what it is supposed to do. Several SIEM systems provide default common functions. Meanwhile, technical criteria involve the vendor, solution integration level, simplicity of deployment, and product evolution. With respect to the selection of a good solution for a specific client, the user's technical and organizational needs should be thoroughly understood, and then the supplier's solutions should be evaluated considering relevant criteria. Quantitative and qualitative methods are needed to be adopted for this process [14], and we have addressed the following key points in our research work.

- Although numerous SIEM systems (both open-source and proprietary) are available in the market today, but their respective structure and architecture remain unknown.
- Currently, there is no clear and well-defined SIEM system architecture available in literature, which identifies a research gap in this area.

In this context, main contributions of this research work are given as follows.

- To the best of our knowledge, none of the existing SIEM systems have described and elaborated their system in detail, thereby making them all difficult to understand and follow. In this study, we have used open-source tools and techniques, and provided a complete description of each module of our proposed SIEM system so that both security experts and novices can better understand and follow it.
- The novelty of this research work is depicted by proposing a well-defined, comprehensive, and modular SIEM architecture that bridges the knowledge gap present in current literature.
- The proposed architecture is a very good foundation for the security personnel who intend to deploy a SIEM system in their organization.
- Our proposed architecture also provides the basis for those software developers who intend to develop a new SIEM system from scratch. The modular architecture proves to be very beneficial for them to comprehend the overall SIEM system and divide their work through a modular approach.

The rest of the paper is organized as follows. In Section 2, the related work is mainly presented to provide information about SIEM systems, either proprietary or open-source ones. This is because they are currently being used in the market for their relatively good functions. In Section 3, discussions on the detailed architecture of the proposed SIEM system are highlighted. In Section 4, some key aspects and limitations related to the proposed SIEM system architecture are dealt with. Lastly, the conclusions are drawn and a recommendation for future work provided in Section 5.

2. Literature Review

2.1 Current SIEM Systems

Organizations have realized the importance of security of their digital assets. Therefore, the demand for a complete security solution is increasing day by day, with many players and vendors coming up with their SIEM systems. We discuss some of the leading SIEM systems that fall in the top-tier category of the latest Gartner report [15]. Besides that, we shall discuss different open-source SIEM systems and present a comparison of different features among them.

2.1.1 Proprietary SIEM systems

Although proprietary SIEM systems are available for use, they incur a considerable cost, especially in the long run. However, proprietary SIEM systems are often simple to use and offer an all-in-one solution. Choosing a proprietary SIEM system for a certain firm should be done with caution. IBM QRadar is a complete SIEM system and is supported by world-leading X-Force research and development. The system has multiple modules, including vulnerability manager, network insights, risk manager, user behavior analytics, incident forensics, etc. Complete solutions can serve as a full-fledged appliance or a virtual appliance. Besides, it can also be used as software-as-a-service (SaaS). The solution provides real-time threat detection along with the complete visibility of entire IT infrastructure [13]. The solution is very famous among medium to large scale enterprises [16].

Splunk provides multiple security solutions, with Splunk enterprise security (ES) being the main SIEM solution. Moreover, they offer Splunk user behavior analytics (UBA) and Splunk Phantom. ES is primarily used for monitoring the overall infrastructure, enabling this system to provide incident response management with dashboards equipped for good visualizations. UBA adds to the flavor of machine learning and performs advance analytics. Splunk Phantom is used to provide security orchestration, automation, and response capabilities. They also offer a customizable SIEM platform that includes third-party components integration according to the requirements of organizations [13, 17].

The Securonix solution has multiple components including Securonix SIEM, security data lake, user and entity behavior analytics (UEBA), SOAR, network traffic analysis (NTA), threat intelligence and apps. Their standard deployment model is SaaS, which is based on Amazon Web Services (AWS) [18]. Most of the users deploy Remote Ingester Nodes (RINs) for data collection and transport to the cloud. The solution is known for detecting behavioral anomalies and insider threats, and the solution is popular among users due to its ease of implementation.

The Exabeam Security Management Platform (SMP) mainly consists of multiple products [19] as follows. The SMP can be employed as a complete package on customer premises, and also be used as a cloud-based SIEM that is managed and hosted by Exabeam. The solution includes Exabeam Data Lake, cloud connectors, threat hunter, entity analytics, case manager, advanced analytics, and Exabeam Incident Responder. Exabeam is built on multiple big data platforms to handle data logging without limitations.

The LogRhythm named its SIEM solution as the LogRhythm NextGen SIEM system [20]. XDR stack is the essential component of the SIEM solution that contains DetectX, AnalytiX, and RespondX modules. For network traffic analysis, NetworkXDR and SysMon for system monitoring is included in the package. The SIEM can be deployed as physical hardware or an application. Its cloud-based solution called LogRhythm Cloud is also available. In addition to SIEM features, the system incorporates endpoint monitoring, forensics, and management capabilities to facilitate large-scale deployment.

2.1.2 Open-source SIEM systems

Open-source SIEM software is becoming very popular, and thus used by many public and private institutions. Its cost is the main factor that small to medium-sized organizations should consider in opting for open-source solutions. An open-source solution allows them to explore and assess various capabilities before pursuing proprietary solutions.

Open-source SIEM (OSSIM) is the most used software the world over [21]. It was provided by AT&T Cybersecurity. The key components of OSSIM manly include event acquisition, asset identification and inventory management, vulnerability analysis, normalization, intrusion detection, behavioral monitoring, and event correlation. Although OSSIM is very popular among the open-source community, it also has some inherent limitations, for example, being difficult to set up and configure specifically in a Windows environment. Its log management is not very rich, and thus can only provide limited functionality in terms of application and database monitoring.

The Elasticsearch-Logstash-Kibana (ELK) stack is a very famous log management platform [22]. It is used as a building block for open-source SIEM. It mainly comprises of a robust platform that acquires and analyzes data from many sources. Scalable and centralized storage is employed for data storage, and

several open-source tools are incorporated for data analysis. Elasticsearch is a time-series data analytics engine that stores and indexes data. Kibana is the visualization layer that allows one to investigate and display data. The agents deployed on edge hosts are known as beats, and their primary role is to collect and transport the data via Logstash.

Wazuh is an open-source utility formally defined as a host-based intrusion detection system (HIDS) [23]. Wazuh is a complete security monitoring solution and provides threat detection, incident monitoring, incident response, and compliance. Major components of Wazuh are agent, server, and the elastic stack, with all of these components being applied to collect, aggregate, analyze, and correlate all events and logs and finally respond to them with defined actions.

In 2014, the Mozilla foundation established the Mozilla Defense Platform (MozDef) as a collection of various utilities and services that have been merged to form an open-source SIEM solution [24]. The primary purpose of doing so is to automate the entire security incident handling process. MozDef resides in between the elastic search and the log management modules to allow it to interact directly. In addition to event aggregation and correlation it also uses machine learning techniques to provide some advanced features. However, it lacks the reporting and compliance capabilities.

The community edition of SIEMonster is a collection of various open-source security solutions, and can support up to 100 endpoints and reports in real-time risk information [25]. It can be deployed in physical, virtual, and cloud contexts. Although it provides all the basic features of a SIEM, it does not support machine learning techniques, without any available upgrades and support.

2.1.3 Open-source vs. proprietary SIEM systems

It is always recommended to start with an open-source SIEM system. Most small to medium-sized organizations prefer to use open-source SIEM systems because such systems fulfill most of their requirements, while large enterprises are recommended to go with a proprietary SIEM system. However, they must begin with an open-source SIEM system and thoroughly understand their objectives and requirements before selecting a suitable proprietary SIEM solution. Although open-source SIEM systems are available and free of charge for deployment, but they also warrant a considerable amount of time and manpower inputs for deployment. On the other hand, proprietary SIEM systems are costly but are relatively easy to deploy, thereby requiring less time and manpower to expend.

Table 1. Comparative analysis of SIEM systems

Feature	Open-Source SIEM					Proprietary SIEM					Proposed SIEM
	OSSIM	ELK	Wazuh	MozDef	SIEMonster	QRadar	Splunk	Securonix	Exabeam	LogRhythm	
Real-time monitoring	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓
Threat intelligence	×	✓	✓	×	✓	✓	✓	✓	✓	✓	✓
Behavior profiling	✓	✓	×	✓	×	✓	✓	✓	✓	✓	✓
Data monitoring	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
User monitoring	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Application monitoring	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Analytics	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Log management	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Updates	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓
Reporting	×	✓	✓	×	✓	✓	✓	✓	✓	✓	✓
GUI	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Detailed system description	×	×	×	×	×	×	×	×	×	×	✓
Database	MySQL	ES	MySQL	ES	ES	Ariel	GZip-files	A.Hadoop	ES	SQL-server	MySQL

ES=Elasticsearch.

No matter which open-source or proprietary SIEM system is selected, some key parameters must be carefully considered [13], which include real-time monitoring, threat intelligence, behavior profiling, data and user monitoring, application monitoring, analytics, log management, updates, reporting, graphical user interface (GUI), detailed system description, and a database. Table 1 provides a short description of various SIEM systems.

3. Proposed SIEM System Architecture

For a better understanding and development of a new SIEM system, its low-level modular architecture should be viewed and analyzed. Despite this, current SIEM systems do not unfortunately unveil their detailed architecture for this purpose. As a result, their exact low-level architecture remains furtive. The proposed SIEM system architecture bridges this gap by proposing a well-defined, detailed, and modular SIEM system architecture. It is modular in design, allowing more modules to be added to this architecture in the future according to the requirement, if necessary. The SIEM system is actually a collection of a finite number of components or modules that work together to achieve the overall system goals. Fig. 1 presents the modules of the proposed SIEM system.

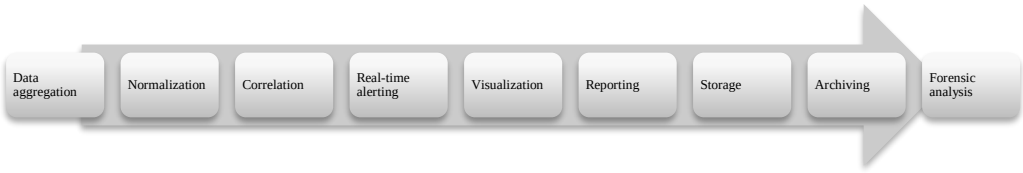

Fig. 1. Modules of proposed SIEM system.

Different data sources send logs, events, and contextual data to the SIEM system. The data aggregation module of a SIEM system receives these data sent by the data sources. After receiving data, the data aggregation module sends them to the normalization module. A normalization module converts these data given in multiple formats into JavaScript Object Notation (JSON) format. For this reason, the other modules of the proposed SIEM system can work on it. It utilizes the concept of parallelism to normalize the data speedily, and thus stores them in the database using the storage module. The correlation module performs a correlation on the normalized data, and if data is matched to some correlation rule, an alert is generated by the correlation module. If correlation module generates an alert, it is displayed on the visualization module (GUI) in the most apposite method through the real-time alerting module. According to the user requirements, customized reports can be generated through the reporting module. As the SIEM system receives a massive amount of data daily, after some predefined time, the log archiving module archives data from the active database to be used later. The forensic analysis and incident response module applies different analysis techniques on the logged data, thereby helping administrators and security experts in incident response.

The data aggregation module must be capable enough to receive and handle data coming at a fast pace, since typical or standard network cards do not support the capacity to process data at the line rate. Thus, for this purpose, the Napatech network acceleration card was used in our proposed SIEM system architecture. It is a hardware-based solution to achieve the capturing and processing of data coming at the line rate. Fig. 2 presents the detailed architecture of the proposed SIEM system.

Ten distinct modules have been shown, having each individual module responsible for performing a subset of the total number of tasks performed by the SIEM system. The mathematical formula of the SIEM system with all its modules is denoted in Equation (1):

$$S = \sum_{i=1}^{N} Mi \quad (1)$$

In the above Equation (1), “S” represents SIEM system, “N” represents number of modules and its value is 10, whereas “M” represents each module of our proposed SIEM system architecture. It is clearly visible that each module “M” of SIEM system “S” performs its designated task or set of tasks. The tasks performed by all the modules constitute the overall functionalities of the SIEM system. The following section provides the detailed and comprehensive architecture of individual modules of the proposed SIEM system.

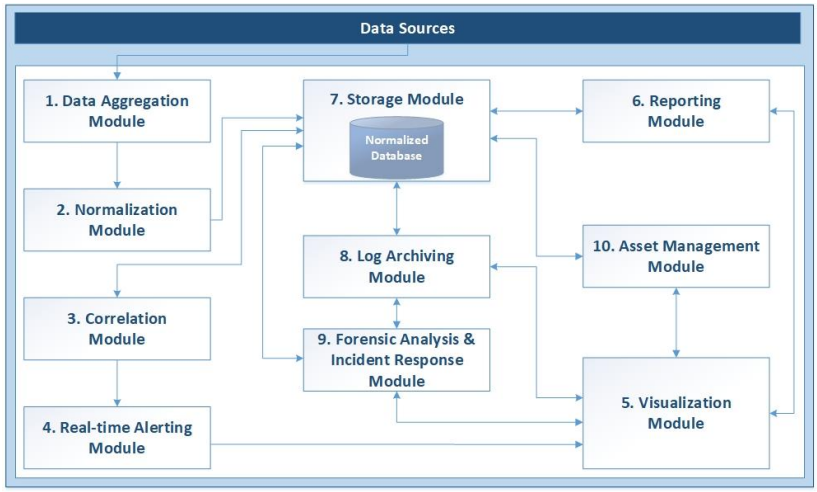

Fig. 2. Architecture of proposed SIEM system.

3.1 Data Aggregation Module

Data aggregation module collects logs and alerts generated from different data sources, which include PC, server, switch, router, firewall, IDS, IPS, etc. Normally, all these data sources send data in different formats. On PCs and servers, Linux sends log data in the syslog format, whereas Windows sends log data in its own format. The other network and security devices send data in some customized format, and this format changes from vendor to vendor as well. Therefore, as depicted in Fig. 3, data aggregation module, must be able to collect and handle any sort of log data format that data sources send.

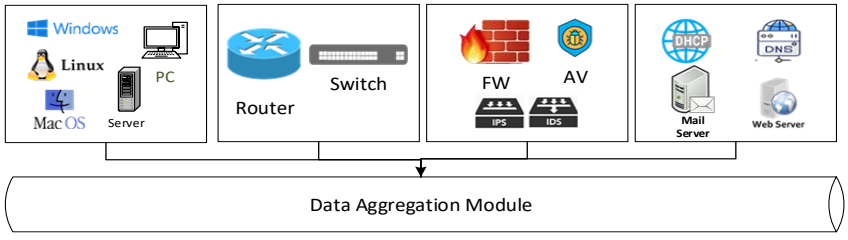

Fig. 3. Data aggregation module.

3.1.1 Data transfer approaches

As depicted in Fig. 4, data is transferred from data sources to the SIEM system by two approaches of push and pull. The push technique involves the data source pushing data to the data aggregation module. Alternatively, in a pull strategy, the data is pulled in the SIEM system by an agent and sent to the data aggregation module.

An example of the “push” approach is the syslog client and the server message exchange mechanism. The syslog client (data sources) sends log messages to the syslog server (SIEM system) whose IP address is configured in the configuration file of the syslog client. An example of the “pull” approach is a secured

database with an appropriate password. In this case, the SIEM system will have to establish a connection to the database through some AAA certified method. Although this approach is safer, it incurs a higher overhead cost. Table 2 provides a summary of data transfer approaches.

Fig. 4. (a) Pull approach and (b) push approach.

Table 2. Comparison of data transfer approaches

	Security	Cost	Configuration	Efficiency
Push approach	Low[a)]	Low	Easy	High
Pull approach	High	High	Difficult	Low

[a)]Older version used UDP, newer version uses TCP.

3.1.2 Operating system (OS) logging services

Logging is an important feature of any operating systems today. Table 3 lists the logging services in various operating systems.

Table 3. OS logging services

Operating system	Logging service	Log viewing applications
Windows	Windows logging service (WLS)	Event viewer
MacOS (10.12 and later)	Unified logging system (ULS)	Console, log, OSLog
Linux	RSyslog	grep, tail, head
iOS (10 and later)	Unified logging system (ULS)	Console
Android	Native, application and JVM logging	Logcat, dmesg

Linux logging service

The Linux operating system uses a “rsyslog” service for event logging purposes. This service can not only log event data on the local system but also transfer them to another syslog server for storage purposes. The rsyslog service uses 514 as the default port in Linux, which can be changed, of course.

The rsyslog service can use both TCP and UDP protocols for the transfer of event log data, with Fig. 5 depicting the workflow of the Linux logging service. Many systems or application-level components like DHCP client, NTPD, Linux PAM, user applications, etc., send their log messages to the syslog service for logging. The rsyslog service uses a standard log message format which is defined in RFC 5424 [26].

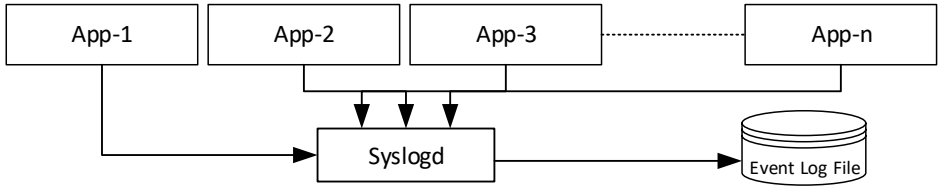

Fig. 5. Linux logging service.

Windows logging service

When logging of events was first introduced in Windows NT 3.5 in 1994, only three kinds of logs were supported in Windows NT 3.5 i.e., system, application, and security logs. This log format is known as the Windows event log (EVT) format. The EVT log format is the default format in Windows 2000,

Windows XP, and Windows 2003 server. The event log is composed of a header, description of the event and an additional optional data. Through the release of Windows Vista, Microsoft introduced a new logging mechanism known as the Windows XML event log (EVTX) format. This new logging mechanism was released to overcome issues present in the older versions of the logging mechanism [27]. Fig. 6 provides the Windows event logging formats (EVT and EVTX).

Microsoft Vista XML event log (EVTX) format				
Header	Chunk 1	Chunk 2	. . .	Chunk N

Microsoft Windows NT 3.5 event log (EVT) format		
Header	Description	Additional optional data

Fig. 6. Windows logging formats.

macOS logging service

The macOS can store logs as plain text files. These text files can be viewed by "console" utility provided by macOS. The apple system logger (ASL) was the logging service in older versions of macOS. Currently, the unified logging system (ULS) is the logging service in macOS version 10.12 and later. In ULS, logs are stored in compressed binary format.

3.1.3 Other logging service

Web server is the most common and widely used server nowadays. These most widely used web servers include Nginx, Apache, IIS, etc.—Apache which is open source. It has been widely used as a web server that can by default create log entries in common log format (CLF), but a custom log format can also be specified. The contents of a web server log file may contain information, such as username, visiting path, user agent, timestamp, success rate, page last visited, URL, and request type. Web server logs are mainly divided into four types of agent log, transfer log, error log and referrer log. In all likelihood, DNS server, mail server, DHCP server, SNMP server, NTP server, etc., also store logs for analysis purposes. Switches and routers (such as CISCO that are widely used worldwide) send logs in the syslog format. However, the log format can be changed to extensible markup language (XML) or some customized formats as well. Firewall, IDS, IPS, etc., can also be configured to send logs to a log server in syslog or some other format depending on type and vendor of device.

3.2 Normalization Module

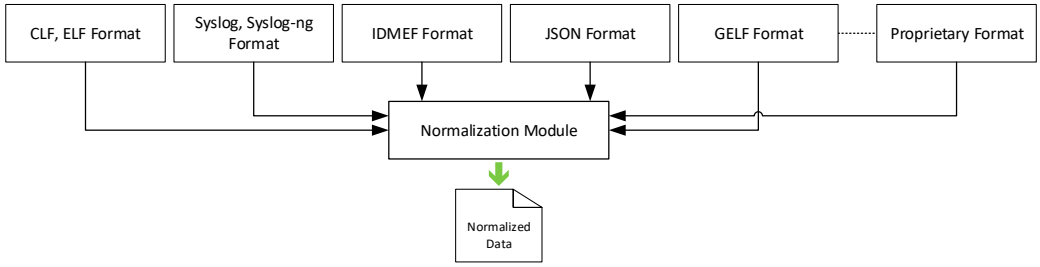

Fig. 7. Normalization module.

The data received by the data aggregation module is presented in different formats, and the SIEM system can perform various operations on it only when this data is transformed into a single format. The diversity of data sources exists both in terms of the type of system or vendor. Almost every type of vendor

has its own log format. The normalization module, which is depicted in Fig. 7, converts all these different formats into a single format, and this can be understood by other modules of the SIEM system [28], with the normalization process being performed by parsing or syntactic analysis. Nowadays, numerous log formats are used, and some of the most common ones are discussed below.

3.2.1 Common log format

Web servers commonly use the Common Log Format (CLF) to log requests from source computers in order to get finer statistics, and this standardized log format helps analytic programs to work on the log data more conveniently. As shown in Fig. 8, the CLF format of a log file entry is basically a text-based log format.

HIP	UAuth	UID	DateTime	Req	Code	Bytes

Fig. 8. Log file entry's CLF format.

where HIP is the host IP address of the client, UAuth is user authentication (a process that owns the TCP/IP connection); UID is the user ID of the requesting person for a document as determined by HTTP authentication; DateTime is to display the date, time, and time zone when the server request is fully completed; Req is to contain the request method, the requested information or resource and the protocol used by the client; Code is status code sent by the server to the client; and Bytes is the size of the response object, excluding the response header.

3.2.2 Extended log format

The extended log format (ELF) is generally used by web servers, and is more flexible because it contains more information than the CLF format. The ELF format is basically an extended version of the CLF format with some additional information, which includes a user agent, cookie, and referrer to be briefly described below.

User agent: This field contains information about the browser and operating system of client.

Cookie: It is a persistent token that is transmitted to a client.

Referrer: This field contains the URL of the site that a client came from.

3.2.3 Intrusion detection message exchange format (IDMEF)

IDMEF is the format used by open-source IDS systems for exchanging incident data among IDS, IPS or security information-gathering software. This format defines the data format and exchange methods among IDS, IPS and security information-gathering software, e.g., SIEM.

3.2.4 Syslog/syslog-ng

The syslog format is generally used by Windows and Unix/Linux servers, switches, and routers. The syslog format is defined in RFC 5424 [26]. The syslog protocol is based on a layered architecture, and can use multiple transport protocols for the transmission of syslog messages. The latest implementation of the syslog message service (namely rsyslog) supports both UDP and TCP protocols for transmitting syslog messages. The header of a syslog message specifies the "Version, Timestamp, Hostname, Application, Process ID, Message ID, and Priority" fields. After the header, the message body comes in as structured data with data blocks in "key=value" format. Fig. 9 presents a sample syslog message.

```
<34>1 2021-07-15T22:00:00.003Z server.machine.com su – ID0 – USER'su root' failed for lonvick on /dev/pts/0
```

Fig. 9. Sample syslog message.

The syslog-ng format reinforces the syslog format, but in fact, it is an advanced version of the syslog format, which helps filter messages according to its contents. The syslog-ng provides reliable delivery of log messages as it uses TCP protocol for transmitting log messages, while also providing the facility of log message encryption. In fact, there are three versions of syslog-ng, i.e., syslog-ng open-source edition (OSE), syslog-ng premium edition (PE) and syslog-ng store box (SSB).

3.2.5 Other log formats

Apart from those formats discussed above, there are also other d formats available, even though they are less commonly used. Such formats include common event expression (CEE), incident object description exchange format (IODEF), ArcSight common event format (CEF), JSON, Graylog extended log format (GELF), etc.

3.2.6 Proprietary log formats

A proprietary format is one whose specs and details are not publicly available. Some security devices (i.e., antivirus, IDS/IPS, firewall, SIEM), web and proxy servers, and business applications have to use such a format. In our case as an example, the normalization module uses the JSON format. It converts every log format it receives to the JSON format for further processing. As depicted in Fig. 10, an event log is received from a Linux system in the syslog format, and converted into the JSON format.

Fig. 10. Event log conversion from syslog format to JSON format.

3.3 Correlation Module

As presented in Fig. 11, the correlation module is responsible for performing the correlation process on the normalized data, and determines the relationship between several events. The main function of the correlation module is to transforms a heap of events into a smaller bunch of more meaningful events.

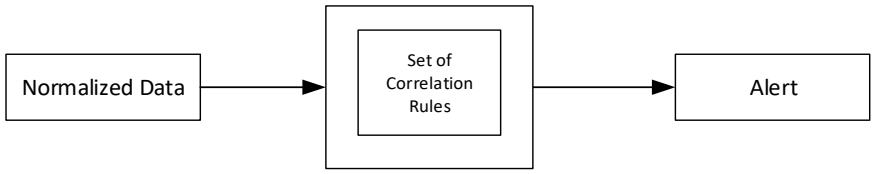

Fig. 11. Correlation module.

As the data is received from various data sources, it is possible that more than one alert are generated for only one anomaly. If an anomaly is detected both by a firewall and IDS, then two alerts are generated. When the SIEM system receives these two alerts, it normalizes them before they are sent or delivered to the correlation module. The correlation module performs the correlation process on these two alerts, and merges them into a single alert to be displayed on the GUI of the SIEM system. Another scenario tends to appear when there is no alert generated by any device in the network. Despite this, there are still some suspicious activities in the logs received by the SIEM system. After performing the correlation process on the received logs, the correlation module generates an alert for a suspicious activity, such as a wrong password attempt. This type of correlation is generally performed by threshold values specified in correlation rules.

3.3.1 Event correlation stages

The process of event correlation can be divided into four stages, as listed in Table 4.

Table 4. Event correlation stages

S. No.	Correlation stage	Description
1	Event filtering	Irrelevant events are discarded to reduce a large number of events.
2	Event aggregation & de-duplication	Closely related events are aggregated in event aggregation. Events that are identical are merged in event de-duplication.
3	Event masking	Events involving systems downstream of a failed system are ignored.
4	Root cause analysis	Dependencies between events are examined using tools, such as dependency graphs to explain events through other events.

3.3.2 Event correlation techniques

There are several event correlation techniques available, such as finite state machine-based, rule-based, case-based, Bayesian-based, artificial neural network-based event correlation, etc. [29]. The rule-based event correlation technique is the most frequent and basic correlation approach because this has been utilized in many current SIEM solutions [29]. The correlation module of the SIEM system is also known as a "correlation engine." Today, several correlation engines are available, and employ some of the above correlation techniques. Some of the most common event correlation engines include simple event correlator (SEC), Esper, Drools, Nodebrain, Prelude, OSSEC, OSSIM, etc. Drools can perform the best (based on high throughput and low execution time) when the number of rules and input events are greater (500 rules and 1,000,000 input events) [3]. For a modest to medium number of rules and input events, Nodebrain can work well. Although SEC tends to exhibit very mediocre performance, it is appropriate for cases with a small number of rules and input events, such as embedded systems, and Esper's performance is mediocre. Table 5 summarizes the event correlation engines.

Table 5. Event correlation engines

Event correlation engine	Developed in	Memory usage
Esper	Java	High
Drools	Java	High
Prelude	C, Python	Less[a)]
Nodebrain	C	Less[a)]
SEC	Perl	Less[a)]
OSSEC	C	Less[a)]
OSSIM[b)]	C	Less[a)]

[a)]Less as compared to Java-based libraries. [b)]Refers to a correlation engine of OSSIM.

3.4 Real-Time Alerting Module

A real-time alerting module can serve as the system to generate an alert after receiving inputs from the correlation module. This alert is sent to the visualization module for further processing immediately after it is generated.

3.5 Visualization Module

The end user uses visualization module (GUI) after authentication, with his personal credentials to be kept secret by some mechanism [30]. After receiving an alert from the real-time alerting module, a visualization module displays it on the GUI. There are several components in the visualization module, such as text animations, graphs, pop-ups, and so on. All these elements make the GUI more attractive, informative, and intuitive, and convey the message early in an intuitive manner to the end user using different visualization techniques. The visualization module displays data on the screen according to various priority levels. The data that needs urgent attention from the end user is given the highest priority.

Aggregation plot, histogram, bar plot, spine plot, box plot, scatter plot, matrix plot, connected scatter plot, tile bars, value bars, pop-ups are some of the visualization tools for visualization of data on the SIEM system's GUI. The visualization module presents the front end of SIEM, thereby hiding the design and implementation complexities of the SIEM system from its end user. Augmented reality is another latest technology that improves attack perception and presentation to the end user [31].

3.6 Reporting Module

The key purpose of a reporting module is to generate reports on stored SIEM data based on different filters and search criteria. Most of the SIEM solutions give multiple pre-defined report formats for the ease of users. Fig. 12 describes the five most common types of reports provided by different SIEM solutions.

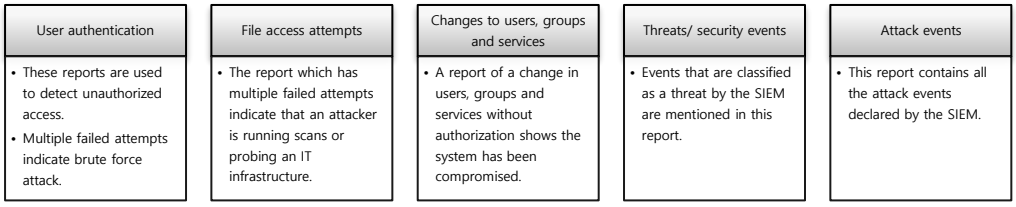

Fig. 12. Most common types of reports provided by different SIEM solutions.

3.7 Storage Module

The storage module is mainly used to keep normalized data as well as alerts generated by the correlation module. This data can remain active in the database, and whenever a report is required, the reporting module applies various filters and search criteria to generate the report. The storage module must be able to receive and save data in real time. Normally, it is a better approach because it can use some big data solution to achieve this purpose. A non-big data solution is only recommended for a small organization whose size is unlikely to grow to a great extent. However, senior management of the organization makes the final decision after analyzing their requirements. There are some other areas or factors that are also important to be considered for the storage module of the SIEM system [32], as summarized in Table 6.

Table 6. Storage module

	Feature	Detail
Storage module	Storage management	Centralized management, fully indexed, cost, auto backups
	Cloud support	Amazon S3, iDrive, Dropbox, NextCloud, etc.
	Security	Confidentiality, integrity, availability
	Size	Scalability, log filtering, compression, archiving
	Regulatory compliance	PCI DSS, HIPAA, SOX, GDPR

3.8 Log Archiving Module

The key purpose of the archiving module is the retention of data for long time. The data stored by the storage module remains in an active state in the database for a predefined amount of time. After this, the time-archiving module shifts the stored data from the active database to an archive, as a database in a dormant state, for later use. There are various types of data archiving available, which can be selected according to the requirement of accessibility of the desired data. Table 7 lists different types of archiving media and features for the selection of the media.

Table 7. Archiving media and its features

Type	Example	Size	Reliability	Speed	Cost
Optical storage	CD, DVD and Blu-Ray disc	Low	Low	Medium	High
Cloud	Dropbox, Google drive, iCloud, etc.	High	High	Low	Medium
Tape drive	Magnetic tape, IBM linear tape-open	High	High	Low	Medium
Portable HD	Portable hard drive connected via USB	Medium	Low	High	Low
External SSD	Solid-state drives employ integrated circuits to store data.	Medium	High	High	High

3.9 Forensic Analysis and Incident Response Module

This module has recently received a great deal of attention and significance. However, standardization is also required for incident response utilization [33]. This module is mainly divided into two sections of host forensic and network forensic. The former is all about collecting the evidence from a computer or machine, whereas the latter deals with monitoring and analysis of network traffic for investigation. Monitoring data for file integrity, network connections, layer-7 flow, processes, registry, layer-3 packets, and application IDs are required to execute a bare minimum forensic analysis. The module performs comprehensive forensic analysis especially related to an operating system, disk and file system, live memory, web, email, network, multimedia [34]. Machine learning and deep learning-based approaches can be very useful for analysis purpose [35]. Immediately after an incident or anomaly is detected and reported, its mitigation is involved in the next step. This module provides an overall forensic analysis and response to the incident that has occurred.

3.10 Asset Management Module

Asset management has become a very important task in the organizational environment. Asset management through manual means is no longer an option now. An asset management module can manage all sort of assets in an organization in an automated way. An asset includes everything that can enable an organization to better perform its business operations. Hardware and software both types of assets are monitored and managed by this module. The modules discussed above provide different functionalities, and thus can work together to achieve the core functionalities of the SIEM system. In addition to the core functionality of the SIEM system, new requirements for additional functionalities arise from time to time. This modular design of the SIEM system is flexible enough to best accommodate more modules if desired in the future.

4. Discussion

In this research work, the detailed structure of a SIEM system was proposed. All the minute details like log formats, event correlation engines, storage module media and technology, archiving strategies etc. are made available. However, due to limited space, only the most important and relevant details have been included. Furthermore, some technologies are proprietary, e.g., proprietary log formats and their details are not available. Therefore, such proprietary technologies have only been mentioned, and their details have not been included yet. Moreover, another important focus of this research work is placed on modularity, and the proposed architecture is deliberately divided into 10 modules for efficient design. Although the number of modules can be changed according to the end user requirements, e.g., data aggregation and normalization modules can be merged into a single module per requirement.

The correlation module of our proposed SIEM system uses rule a based approach. However, a rule-based approach has its own limitations of requiring a continuous updating of rules and failing to detect

zero-day attacks. Another approach is to use machine learning-based correlation engines. However, such models might not perform well in a live network environment. Furthermore, these models are heavily computational and time-intensive, which can generate large number of false positives. Nevertheless, a better tuned machine learning-based model or hybrid method between machine learning and metaheuristics can prove to be very effective and efficient for detection mechanisms [36–41]. A near-miss event is an event that is not detected as a malicious event due to the rule-based approach used. When it comes to certain rules, threshold values are used for detection. In the case of a rule using a threshold value of 50 and the event having a value of 49, the rule will not be matched, and thus, the event will be treated as a benign event. However, this will most probably serve as a malicious activity event, and as a result, it is not detected under the threshold-based rule. An efficient visualization module can overcome this limitation and help security administrators to achieve better detection and analysis.

5. Conclusion and Future Work

SIEM systems have become an integral part of any organizations, and provide a holistic view of an organization’s security. The SIEM system makes the organization's security management convenient and manageable, while also monitoring the security status of all the devices connected in an organizational network by analyzing the logs sent from these devices. There are many SIEM systems available today, and the discussions about these systems and their features are also provided above. The method of choosing the right and suitable SIEM systems (either open-source or proprietary) for a certain organization has been discussed. Currently, there is no clearly defined SIEM architecture that an organization can follow to develop or deploy its own SIEM systems. In this research, a comprehensive and modular architecture for an efficient SIEM system was proposed. The modules of the SIEM system were also discussed in detail, while the overall functionality of the SIEM system was elaborated. The modular structure of the SIEM system allows additional modules to be added and integrated depending on organizational requirements. Our proposed SIEM system architecture will help organizations make development or deployment of a SIEM system more efficient and effective and easier. As for the future work, we plan to extend and integrate the SIEM system with a SOAR system. Also, additional modules added to the file integrity module (FIM) are considered because this can possibly help improve the overall security of the enterprise network. Similarly, in the visualization module, the research community can capitalize on augmented reality (AR) to better improve the visualization and presentation of attacks to the end user.

Author’s Contributions

Conceptualization, MHD. Funding acquisition, MHD. Investigation and methodology, MS, MAP, MUH. Formal analysis, SMM, SSB, AM. Resources, SSB, AM. Writing of the original draft, MS, MAP, MUH. Writing of the review and editing, MS, SMM. Supervision, MHD.

Funding

This research work was supported by the IT&T Endowment Fund PIEAS and funded by the Ministry of Planning, Development and Special Initiatives through the Higher Education Commission of Pakistan under the National Center for Cyber Security (NCCS).

Competing Interests

The authors declare that they have no competing interests.

References

[1] G. Gonzalez-Granadillo, S. Gonzalez-Zarzosa, and R. Diaz, "Security information and event management (SIEM): analysis, trends, and usage in critical infrastructures," *Sensors*, vol. 21, no. 14, article no. 4759, 2021. https://doi.org/10.3390/s21144759

[2] S. M. M. Hossain, R. Couturier, J. Rusk, and K. B. Kent, "Automatic event categorizer for SIEM," in *Proceedings of the 31st Annual International Conference on Computer Science and Software Engineering*, Riverton, NJ, 2021, pp. 104-112.

[3] L. Rosa, P. Alves, T. Cruz, P. Simoes, and E. Monteiro, "A comparative study of correlation engines for security event management," in *Proceedings of the 10th International Conference on Cyber Warfare and Securit (ICCWS)*, Kruger National Park, South Africa, 2015, pp. 277-285.

[4] S. Saleem, M. Sheeraz, M. Hanif, and U. Farooq, "Web server attack detection using machine learning," in *Proceedings of 2020 International Conference on Cyber Warfare and Security (ICCWS)*, Islamabad, Pakistan, 2020, pp. 1-7.

[5] W. S. Hwang, J. G. Shon, and J. S. Park, "Web session hijacking defense technique using user information," *Human-centric Computing and Information Sciences*, vol. 12, article no. 16, 2022. https://doi.org/10.22967/HCIS.2022.12.016

[6] D. Tang, R. Dai, L. Tang, and X. Li, "Low-rate DoS attack detection based on two-step cluster analysis and UTR analysis," *Human-centric Computing and Information Sciences*, vol. 10, article no. 6, 2020. https://doi.org/10.1186/S13673-020-0210-9/FIGURES/24

[7] A. Alqahtani and F. T. Sheldon, "A survey of crypto ransomware attack detection methodologies: an evolving outlook," *Sensors*, vol. 22, no. 5, article no. 1837, 2022. https://doi.org/10.3390/s22051837

[8] N. Kshetri and J. Voas, "Ransomware: pay to play?," *Computer*, vol. 55, no. 3, pp. 11-13, 2022. https://doi.org/10.1109/MC.2021.3126529

[9] E. Tcydenova, T. W. Kim, C. Lee, and J. H. Park, "Detection of adversarial attacks in AI-based intrusion detection systems using explainable AI," *Human-centric Computing and Information Sciences*, vol. 11, article no. 35, 2021. https://doi.org/10.22967/HCIS.2021.11.035

[10] H. Mokalled, R. Catelli, V. Casola, D. Debertol, E. Meda, and R. Zunino, "The applicability of a SIEM solution: requirements and evaluation," in *Proceedings of 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, Napoli, Italy, 2019, pp. 132-137.

[11] R. F. Gibadullin and V. V. Nikonorov, "Development of the system for automated incident management based on open-source software," in *Proceedings of 2021 International Russian Automation Conference (RusAutoCon)*, Sochi, Russian Federation, 2021, pp. 521-525.

[12] A. E. Azzaoui, T. W. Kim, Y. Pan, and J. H. Park, "A quantum approximate optimization algorithm based on blockchain heuristic approach for scalable and secure smart logistics systems," *Human-centric Computing and Information Sciences*, vol. 11, article no. 46, 2021. https://doi.org/10.22967/HCIS.2021.11.046

[13] S. S. Sekharan and K. Kandasamy, "Profiling SIEM tools and correlation engines for security analytics," in *Proceedings of 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, Chennai, India, 2017, pp. 717-721.

[14] H. Mokalled, R. Catelli, V. Casola, D. Debertol, E. Meda, and R. Zunino, "The guidelines to adopt an applicable SIEM solution," *Journal of Information Security*, vol. 11, no. 1, pp. 46-70, 2019.

[15] K. Kavanagh, T. Bussa, and J. Collins, "Gartner magic quadrant for security information and event management," 2020 [Online]. Available: https://www.gartner.com/en/documents/3981040?msclkid=59680934c41611ec9375f609d75c35bb.

[16] IBM, "Capabilities in your IBM QRadar product," 2023 [Online]. Available: https://www.ibm.com/docs/en/qradar-on-cloud?topic=administration-capabilities-in-your-qradar-product.

[17] Splunk, "Splunk Enterprise Security," 2023 [Online]. Available: https://docs.splunk.com/Documentation/ES.

[18] Securonix, "Next-Gen Security Information and Event Management," 2023 [Online]. Available: https://www.securonix.com/.

[19] Exabeam, "Exabeam Documentation Portal," 2023 [Online]. Available: https://docs.exabeam.com/.

[20] LogRhythm, "SIEM platform & security operations center services," 2023 [Online]. Available: https://logrhythm.com/.

[21] AT&T Cybersecurity, "AlienVault OSSIM: an open-source SIEM," 2023 [Online]. Available: https://cybersecurity.att.com/products/ossim.

[22] Elastic, "Elastic Stack: Elasticsearch, Kibana, Beats, and Logstash," 2023 [Online]. Available: https://www.elastic.co/elastic-stack/.

[23] Wazuh Inc., "Wazuh documentation," 2023 [Online]. Available: https://documentation.wazuh.com/current/index.html.

[24] Mozilla Enterprise Defense Platform, "Overview of the MozDef," 2021 [Online]. Available: https://mozdef.readthedocs.io/en/latest/overview.html.

[25] SIEMonster, "Introducing SIEMonster," 2023 [Online]. Available: https://www.siemonster.com/solution/.

[26] R. Gerhards, "The syslog protocol," Internet Engineering Task Force, Fremont, CA, *RFC 5424*, 2009.

[27] H. Studiawan, F. Sohel, and C. Payne, "A survey on forensic investigation of operating system logs," *Digital Investigation*, vol. 29, pp. 1-20, 2019.

[28] A. Iskhakov and S. Iskhakov, "Data normalization models in the security event management systems," in *Proceedings of 2020 13th International Conference on Management of Large-Scale System Development (MLSD)*, Moscow, Russia, 2020, pp. 1-5.

[29] I. Kotenko, A. Fedorchenko, and E. Doynikova, "Data analytics for security management of complex heterogeneous systems: event correlation and security assessment tasks," in *Advances in Cyber Security Analytics and Decision Systems*. Cham, Switzerland: Springer, 2020, pp. 79-116. https://doi.org/10.1007/978-3-030-19353-9_5

[30] J. Lee, P. Park, S. Ryu, and H. Cha, "2FA-SF: two-factor assessment-based secure framework for clinically distributed multicenter study," *Human-centric Computing and Information Sciences*, vol. 11, article no. 47, 2021. https://doi.org/10.22967/HCIS.2021.11.047

[31] N. J. Ahuja, S. Dutt, S. L. Choudhary, and M. Kumar, "Intelligent tutoring system in education for disabled learners using human-computer interaction and augmented reality," *International Journal of Human–Computer Interaction*, 2022. https://doi.org/10.1080/10447318.2022.2124359

[32] F. Menges, T. Latzo, M. Vielberth, S. Sobola, H. C. Pohls, B. Taubmann, et al., "Towards GDPR-compliant data processing in modern SIEM systems," *Computers & Security*, vol. 103, article no. 102165, 2021. https://doi.org/10.1016/j.cose.2020.102165

[33] D. Schlette, M. Caselli, and G. Pernul, "A comparative study on cyber threat intelligence: the security incident response perspective," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 4, pp. 2525-2556, 2021.

[34] A. R. Javed, W. Ahmed, M. Alazab, Z. Jalil, K. Kifayat, and T. R. Gadekallu, "A comprehensive survey on computer forensics: state-of-the-art, tools, techniques, challenges, and future directions," *IEEE Access*, vol. 10, pp. 11065-11089, 2022. https://doi.org/10.1109/ACCESS.2022.3142508

[35] A. Goswami, M. M. Krishna, J. Vankara, S. M. P. Gangadharan, C. S. Yadav, M. Kumar, and M. Khan, "Sentiment analysis of statements on social media and electronic media using machine and deep learning classifiers," *Computational Intelligence and Neuroscience*, vol. 2022, article no. 9194031, 2022. https://doi.org/10.1155/2022/9194031

[36] M. Zivkovic, M. Tair, K. Venkatachalam, N. Bacanin, S. Hubalovsky, and P. Trojovsky, "Novel hybrid firefly algorithm: an application to enhance XGBoost tuning for intrusion detection classification," *PeerJ Computer Science*, vol. 8, article no. e956, 2022. https://doi.org/10.7717/peerj-cs.956

[37] J. K. Samriya, R. Tiwari, X. Cheng, R. K. Singh, A. Shankar, and M. Kumar, "Network intrusion detection using ACO-DNN model with DVFS based energy optimization in cloud framework," *Sustainable Computing: Informatics and Systems*, vol. 35, article no. 100746, 2022. https://doi.org/10.1016/j.suscom.2022.100746

[38] M. Nawaz, M. A. Paracha, A. Majid, and H. Durad, "Attack detection from network traffic using machine learning," *VFAST Transactions on Software Engineering*, vol. 8, no. 1, pp. 1-7, 2020.

[39] A. Bhardwaj, M. Kumar, T. Stephan, A. Shankar, M. R. Ghalib, and S. Abujar, "IAF: IoT attack framework and unique taxonomy," *Journal of Circuits, Systems and Computers*, vol. 31, no. 2, article no. 2250029, 2022. https://doi.org/10.1142/S0218126622500293

[40] R. Ch, T. R. Gadekallu, M. H. Abidi, and A. Al-Ahmari, "Computational system to classify cyber crime offenses using machine learning," *Sustainability*, vol. 12, no. 10, article no. 4087, 2020. https://doi.org/10.3390/su12104087

[41] C. Choudhary, I. Singh, and M. Kumar, “A real-time fault tolerant and scalable recommender system design based on Kafka,” in *Proceedings of 2022 IEEE 7th International Conference for Convergence in Technology (I2CT)*, Mumbai, India, 2022, pp. 1-6.