УДК 004.896
МРНТИ 50.05.09

10.51889/2959-5894.2023.83.3.020

IMPLEMENTATION OF THE ARTIFICIAL IMMUNE SYSTEM ALGORITHM FOR SECURITY INFORMATION AND EVENT MANAGEMENT SYSTEMS

Çelik Y.[1], Fındık O.[1,2], Alaca Y.[3], Assanova B.[2], Sharmukhanbet S.[2]*

[1] Karabuk University, Karabuk, Türkiye
[2] Kh.Dosmukhamedov Atyrau University, Atyrau, Kazakhstan
[3] Hitit University, Çorum, Türkiye
[*]e-mail: oguzfindik@karabuk.edu.tr

Abstract

One of the most used technologies in computer and network security is Security Information and Event Management (SIEM) systems. A SIEM system is a tool that collects logs of all software and hardware connected to the network, detects security violations using these logs, and generates alarm notifications. The SIEM system generates several alerts during operation. It is an essential problem because of the abundance and the correctness of these alerts generated. In this paper, we implement the Artificial Immune System (AIS) algorithm to generate more stable alerts and to increase the verification rates of the alerts produced by SIEM systems. The results show that the adaptation of the AIS algorithm to SIEM systems is successful. When we apply the SIEM systems powered by AIS, then we had got more successful result than the traditional SIEM systems.

Keywords: Security information and event management (SIEM) systems, Information security, Cyber security, Artificial immune systems, Intrusion detection systems.

Аннотация

РЕАЛИЗАЦИЯ АЛГОРИТМА ИСКУССТВЕННхОЙ ИММУННОЙ СИСТЕМЫ ДЛЯ ИНФОРМАЦИОННЫХ СИСТЕМ БЕЗОПАСНОСТИ И УПРАВЛЕНИЯ СОБЫТИЯМИ

Ю. Челик[1], О. Финдик[1,2], Ю. Аладжа[3], Б. Асанова[2], С. Шармұханбет[2]
[1]Карабукский университет, г. Карабук, Турция
[2]Атырауский университет им.Х.Досмухамедова, г. Атырау,Казахстан
[3]Университет Хитит, г. Чорум,Турция

Одной из наиболее часто используемых технологий в области компьютерной и сетевой безопасности являются системы управления информацией о безопасности и событиями (SIEM). Система SIEM — это инструмент, который собирает журналы всего программного и аппаратного обеспечения, подключенного к сети, обнаруживает нарушения безопасности с помощью этих журналов и генерирует тревожные уведомления. Система SIEM генерирует несколько предупреждений во время работы. Это существенная проблема из-за обилия и корректности генерируемых предупреждений. В этой статье мы реализуем алгоритм искусственной иммунной системы (AIS) для генерации более стабильных оповещений и повышения скорости проверки оповещений, генерируемых системами SIEM. Результаты показывают, что адаптация алгоритма AIS к SIEM-системам прошла успешно. Когда мы применили системы SIEM на базе AIS, мы получили более успешный результат, чем традиционные системы SIEM.

Ключевые слова: системы безопасности информации и управления событиями (SIEM), информационная безопасность, кибербезопасность, Искусственные иммунные системы, системы обнаружения вторжений.

Аңдатпа

АҚПАРАТТЫҚ ҚАУІПСІЗДІК ЖӘНЕ ОҚИҒАЛАРДЫ БАСҚАРУ ЖҮЙЕЛЕРІ ҮШІН ЖАСАНДЫ ИММУНДЫҚ ЖҮЙЕ АЛГОРИТМІН ЕНГІЗУ

Ю. Челик[1], О. Финдик[1,2], Ю. Аладжа[3], Б. Асанова[2], С. Шармұханбет[2]
[1]Карабук университеті, Карабюк қ., Түркия
[2]Х.Досмұхамедов атындағы Атырау университеті, Атырау қ., Қазахстан
[3]Хитит университеті, Чорум қ., Түркия

Компьютерлік және желілік қауіпсіздік саласындағы ең көп қолданылатын технологиялардың бірі-қауіпсіздік және оқиғаларды басқару жүйелері (SIEM). SIEM жүйесі-бұл желіге қосылған барлық бағдарламалық жасақтама мен аппараттық құралдардың журналдарын жинайтын, осы журналдар арқылы қауіпсіздіктің бұзылуын анықтайтын және дабыл хабарламаларын жасайтын құрал. Siem жүйесі жұмыс кезінде бірнеше ескерту жасайды. Бұл жасалған ескертулердің көптігі мен дұрыстығына байланысты маңызды мәселе. Бұл мақалада біз тұрақты

ескертулер жасау және SIEM жүйелері жасаған ескертулерді тексеру жылдамдығын арттыру үшін жасанды иммундық жүйе (AIS) алгоритмін енгіземіз. Нәтижелер AIS алгоритмін SIEM жүйелеріне бейімдеу сәтті болғанын көрсетеді. AIS арқылы жұмыс істейтін SIEM жүйелерін қолданғанда, біз дәстүрлі SIEM жүйелеріне қарағанда табысты нәтижеге қол жеткіздік.

Түйін сөздер: ақпараттық қауіпсіздік және оқиғаларды басқару жүйелері (SIEM), ақпараттық қауіпсіздік, киберқауіпсіздік, жасанды иммундық жүйелер, интрузияны анықтау жүйелері.

Introduction

Network infrastructures consist of many devices and software tools such as computers, clients, servers, etc. All machines and software in the network produce logs every day. Due to a large amount of log data, it isn't easy to filter the data that can be helpful to take from network traffic. Network movements need to be reported to analyze incoming and outgoing traffic data and obtain meaningful results for the network.

Identifying the events occurring in information security or network security and taking precautions against them can be provided using these reports and log analysis [1]. If attacks on the network can not be detected correctly, then the security of all components connected to the network is compromised. Since the network is a multi-user system and many connected systems, it is challenging to ensure network security. Log records of all users and system activity must be collected to ensure information security [2]. One of the goals of log management is collecting log records automatically from systems, databases, and users. One of the information security systems developed especially for analyzing logs is a SIEM system.

A SIEM system can become an automated and self-adaptive system using artificial intelligence techniques. A traditional SIEM system has difficulty with newly encountered and multi-step attacks by combining data from many heterogeneous systems. To overcome these difficulties, a self-adaptive and automatic SIEM system is introduced using genetic programming and deploying artificial neural networks [3].

SIEM systems have two classic tasks. The first is to collect log data from software and hardware such as network servers, routers, firewalls, and intrusion detection systems (IDS). Another task is to listen to the network traffic in real-time and apply a series of correlation rules to detect suspicious events and analyze the encrypted network. Since the whole network is complex and large, SIEM log records should be collected in a specific location [4]

In large-scale networks with a large number of security incidents, SIEM systems face problems such as the normalization of data, reducing the number of false positives, and the length of data analysis. Effective data processing plays a vital role in eliminating these problems [5].

A cyber kill chain model was first developed by Lockheed Martin and is known as the Lockheed Martin Kill-Chain model, which is currently used by the National Institute of Standards and Technology (NIST) as a component of the cyber security framework. The Kill-Chain model was used in SIEM systems to generate fewer false positive and false negative alerts [4]. Supervisory Control and Data Acquisition (SCADA) systems are integrated with SIEM systems to be more reliable [4].

One of the methods developed against cyber-attacks is the Artificial Immune System, and it is used to detect abnormal activities in the network for intrusion detection [6].

Artificial Immune System (AIS) has attractive features such as self-configuration, self-learning, and self-adaptation. Thanks to these features, AIS is generally used when anomalies are identified as non-self. Thus, a proactive system is developed to identify and prevent new and unseen anomalies [7].

A self-adaptive SIEM system is improved by deploying artificial neural networks. It is possible to identify the attacks in the existing system by generating the rules and correlations using Artificial Neural Networks (ANN) [4].

Denial of service attacks (DoS) is constantly growing and threatening. Due to their sophistication, ease of application, and improvements in hiding fingerprints, they have recently been the most common types of attacks. When dealing with DoS flooding attacks, the AIS approach has yielded successful results in reducing such attacks [8]. In this approach, various agents are distributed to the devices and software in the network. These agents can identify threats and perform behaviors similar to the biological defense mechanisms of human beings, such as creating quarantine areas and constructing immune memory.

To ensure the security of computers and networks, keep unauthorized users away from the system, or prevent them from logging into the system and capturing information, first, unique authentication and access are provided to users. Such phases form the lowest level of security [9].

The necessary signatures should be well defined, and the size of the large-scale signature databases should be adjusted by this method. There is a high probability of false alarms in a signature database that is not well-sized [10]. If the signature database is more minor than necessary, it is ineffective in detecting attacks and

negatively affects security. A profile of events is created to monitor user behaviors using an anomaly intrusion detection method [10].

The remainder of the paper is organized as follows. A review of SIEM systems is given in Section 2. The design of the dataset for cyberattacks on SIEM systems and the use of SIEM together with AIS are introduced in Section 3. In Section 4, the experimental test results obtained from the proposed SIEM and AIS methods are presented, and finally, some concluding remarks are given in Section 5.

1. Security information and event management system (SIEM)

SIEM systems are software tools that provide the log records produced by all network software and hardware resources to be collected and reported on a central system [11]. A SIEM system contains and stores logs. If the identical records are generated repeatedly, then SIEM will show them as a single event. It is a tool that makes the collected logs readable and categorizes them, manages the events created by generating reports and alarms, and establishes relationships between many events [4].

There are three main challenges in conventional SIEM systems.

1. The existing SIEM systems are highly dependent on the configuration of multiple sensors (agents) deployed over the network. Numerous techniques have been proposed to combine data derived from different sources to help identify attacks.
2. Existing correlation engines require collecting relevant warnings of operators and making a minor effort to choose the most appropriate precaution.
3. SIEM systems attach great importance to the use of automated procedures for real-time analysis of security events. SIEM systems are equipped with autonomous up-to-date information on attacks that prevent further damage to zero-day attacks by self-learning and adaptation to the event association.

Classical management flaws still persist; for example, bringing together many security events reported from multiple heterogeneous systems poses exciting challenges in adapting to new and multi-step attacks.

Methods such as ANN and Genetic Programming have been used to minimize the disadvantages of SIEM systems. ANNs classify all events collected by a SIEM system. The ANN is trained to categorize the attack scenarios whose events are dynamically defined and controlled by the information examined. More precisely, an ANN creates patterns containing the type and number of events to represent a particular multi-step attack. It generates efficient correlation rules by introducing Genetic Programming (GP) into the SIEM correlation engine. It also provides a correlation infrastructure to learn and develop correlation rules for different multi-step attacks automatically. Therefore, it associates events with a specific attack context and makes the response of the event more efficient [12].

A standard SIEM solution must have the components given in Figure 1 to manage and analyze security events at a single place and to follow the incident response processes, if necessary.

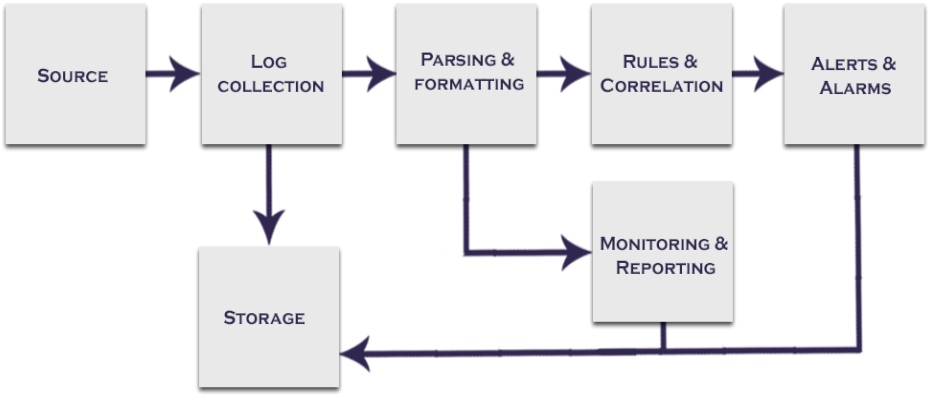

Figure 1. Six basic steps of a SIEM system

1.1 Log Collection Aggregation

SIEM systems collect logs from many different devices connected to the network. Depending on the variability of the systems, before the log collection process is performed the records created by that system in their structure differ [13].

Processing log files collected from software, hardware, and application resources of each system has significant problems [12]. These problems can be listed as follows:

1. The log records being high both in number and size,
2. The recording patterns are different from each other since the logs of each source are collected in its log format,
3. The content of the logs is very different from each other since various events come together to form records.

All devices and resources that are actively connected to the network collect their log records in their format, which allows the patterns of the log records to be designed differently. Due to this difference, it isn't easy to read and analyze the logs. The main reason for the difference in these patterns is that the application areas are not considered quite broad, readability is not prioritized, this process is left to the discretion of the application developers, and most importantly, there is no specific standard. No organization has set a standard regarding this issue, and its deficiency continues. SIEM systems define these events by combining these patterns within specific rules [14].

1.2 Normalization

Normalization correlates by enabling the data to be converted to another format to reveal appropriate information from the log files without disturbing the integrity of the data. Filtering should be applied to log data with specific criteria to reduce massive data. By using this process, unnecessary parts of massive data are removed. The data can be saved in databases so that it may be organized logically. Thus, efficiency will be achieved in terms of time and performance [15].

1.3 Correlation

It is establishing a relationship between logs independent of each other and taking action as a result. In other words, designing of reports from records collected via device or raising alarms from the rules created is called correlation [16]

Correlation is the process of checking millions of data collected every day and storing them in a single directory before reporting. This helps to reduce huge data and creates easy to read by the system administrator. Complex alarms are reported by the correlation system [17].

1.4 Prioritization

Prioritization is the classification process according to critical values if the system is under any attack and this attack is successfully performed. A priority value between 0 and 5 is determined according to the magnitude of the attack [18]. For example, scanning to find vulnerabilities in a network or server port is more important than scanning a printer's port. Administrators can change these priorities in the normalization tables for each event.

1.5 Alarm, Report

These are services for collecting information (logs) of the events that occurred and recorded in the systems and for the analysis and reports of the operations made in the systems. It is the process of creating reports for analysis of all successful and unsuccessful event information or logs in systems having SIEM systems [17]. It is the process of determining that the systems have been attacked by creating alarms and reports and analyzing these attacks in order not to have such attacks again. In contrast, reporting of the log records, graphs, and past analysis is performed. If a problem occurs within certain criteria, an alarm is generated, and the responsible system administrators are warned.

1.6 Post-event Analysis (Incident Response)

A log containing an attack vector or a risky situation caused by multiple threats is called an incident [17]. For example, the reverse TCP connection attempt from the machine where SQL Injection experiment was performed and adding users to that machine can be given as an example. After the event, archiving the logs to analyze logs is called Post-event Analysis (Incident Response).

2. Design of siem dataset

In 1998, DARPA created datasets for intrusion detection. KDD'99 dataset was created in 1999 by adding new ones to these datasets and changing some of the data in the datasets (KDD'99, http://kdd.ics.uci.edu/, Tarih yok). KDD'99 dataset, which has become standard, is still up to date for academic studies (DARPA). This dataset consists of 41 features. Among them, 32 features are continuous variables, and nine are discrete variables (KDD'99, http://kdd.ics.uci.edu/databases/kddcup99/task.html, Tarih yok). We used the KDD'99 dataset to test the method's success proposed in this study [20].

Training Dataset

The training data consists of 65536 data taken randomly from KDD'99, of which 10% can be used. These data determine the population of AIS, that is, the antibody. These training data are mutated with the Clonal

Selection Algorithm, and the population of antibodies is created with mutated data. Antibody data shows the number of data in four types of attacks. Log records consist of DoS, U2R, R2L, Probe attacks, and non-attack log data. Since there is no log data when there is no attack, there are normal type values. Thus, by looking at these values, false positive and false negative numbers can be determined.

Table 1 Antibody populations

Attack Type	Attack Name	Nr of data	Subtotal of Attacks	All Attacks	Total of dataset
DoS	Land	16	42681	44098	65536
	Neptune	29086			
	Pod	20			
	Smurf	13459			
	Teardrop	100			
U2R	Buffer-overflow	2	4		
	Loadmodule	1			
	Perl	1			
R2L	Imap	11	34		
	Multihop	1			
	Warezmaster	20			
	Phf	2			
Probe	Ipsweep	102	1379		
	Nmap	101			
	Portsweep	238			
	Satan	938			
Normal	Normal	21438	21438	21438	

As seen in Table 1, 10% of KDD log data contains 65536 attack data in total. Among the total log data, most of which is DoS attack data. Although there are four different types of attacks, the number of antibodies varies with the scarcity and distribution of other data.

Test Datasets

It is possible to test multiple antigen populations with the antibody population obtained from KDD'99. In Table 1, three sets of antigens were created by selecting random values from the antibody population of the data. In this way, it is aimed that the antibodies differentiate against different antigens and try to identify these antigens.

Preprocessing

The preprocessing has to be performed since KDD'99 data are in txt format, and these data must be used in AIS. These preprocess are digitization and normalization. A sample of KDD'99 data is shown in Figure 2. All data have 42 features in total and are converted into 42 columns. The data converted into columns are given in Table 2.

0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.00,0.00,1.00,0.00,0.00,255,252,0.99,0.01,0.00,0.00,0.00,0.00,0.00,0.00,snmpgetattack.

1,tcp,smtp,SF,3170,329,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,2,0.00,0.00,0.00,0.00,1.00,0.00,1.00,54,39,0.72,0.11,0.02,0.00,0.02,0.00,0.09,0.13,normal.

0,tcp,http,SF,297,13787,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.00,0.00,1.00,0.00,0.00,177,255,1.00,0.00,0.01,0.01,0.00,0.00,0.00,0.00,normal.

Figure 2. Sample KDD'99 data

Table 2. Data parsed to columns

0	*udp*	*private*	*SF*		*105*	*146*	*...*	*...*	*...*	*0.00*	*0.00*	*0.00*	*0.00*	*snmpgetattack.*
1	*tcp*	*smtp*	*SF*		*3170*	*329*	*...*	*...*	*...*	*0.02*	*0.00*	*0.09*	*0.13*	*normal.*
0	*tcp*	*http*	*SF*		*297*	*13787*	*...*	*...*	*...*	*0.00*	*0.00*	*0.00*	*0.00*	*normal.*
0	*tcp*	*http*	*SF*		*291*	*3542*	*...*	*...*	*...*	*0.00*	*0.00*	*0.00*	*0.00*	*normal.*

Data converted into columns must be transformed into numerical data used in AIS. Due to the size of the data, the data were digitized with the preprocessing phase. In the digitization process, the digitization of the protocol names were carried out as TCP = 0, UDP = 1 and ICMP = 2. Flag names, attack names, and service names are shown in Table 3, Table 4, and Table 5, respectively.

Table 3. Digitization of flag names

Flag	*S0*	*S1*	*S2*	*S3*	*SF*	*SH*	*OTH*	*REJ*	*RSTO*	*RSTOSO*	*RSTR*
Numerical Value	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*

Table 4. Digitization of attack names

Nr	*Attack Names*
1	*ipsweep, nmap, portsweep, spy, warezclient, saint, mscan,*
2	*back, buffer_overflow, land, neptune, pod, satan, smurf, teardrop, warezmaster, apache2, udpstorm, mailbomb, processtable*
3	*loadmodule, perl, rootkit, ps, xterm, sqlattack*
4	*ftp_write, guess_passwd, imap, multihop, phf, named, sendmail, snmpgetattack, xsnoop, httptunnel, snmpguess, worm*

Table 2. Digitization of service names

Service	**Nr**	**Service**	**Nr**	**Service**	**Nr**	**Service**	**Nr**	**Service**	**Nr**
http	0	exec	33	ftp_data	13	csnet_ns	46	daytime	26
smtp	1	printer	34	rje	14	pop_2	47	ctf	27
finger	2	efs	35	time	15	sunrpc	48	nntp	28
domain_u	3	courier	36	mtp	16	uucp_path	49	shell	29
auth	4	uucp	37	link	17	netbios_ns	50	IRC	30
telnet	5	klogin	38	remote_job	18	netbios_ssn	51	nnsp	31
ftp	6	kshell	39	gopher	19	netbios_dgm	52	http_443	32
eco_i	7	echo	40	ssh	20	sql_net	53	urh_i	59
ntp_u	8	discard	41	name	21	vmnet	54	X11	60
ecr_i	9	systat	42	whois	22	bgp	55	urp_i	61
other	10	supdup	43	domain	23	Z39_50	56	pm_dump	62
private	11	iso_tsap	44	login	24	ldap	57	tftp_u	63
pop_3	12	hostnames	45	imap4	25	netstat	58	tim_i	64
								red_i	65

The Use of the Clonal Selection Algorithm

When creatures are exposed to an antigen, this antigen is responded by producing antibodies through B lymphocytes in the bone marrow. As shown in Figure 3, antibodies are on the surface of the B cell and bind into incoming antigens to recognize them. A uniform antibody against each antigen is produced by the B cell. If any antibody recognizes an antigen, then B cells divide and reproduce with the help of the T cells and then grow. This reproduction occurs by cloning a single cell or cluster of cells.

Burnet first introduced the clonal selection theory in 1976 [19]. This theory consists of three basic features, as shown in Figure 3. First, the most suitable antibody is selected among n antibodies. The antibody is proliferated and divided into a set of subpopulation groups. The proliferating antibodies to detect different antigens differ.

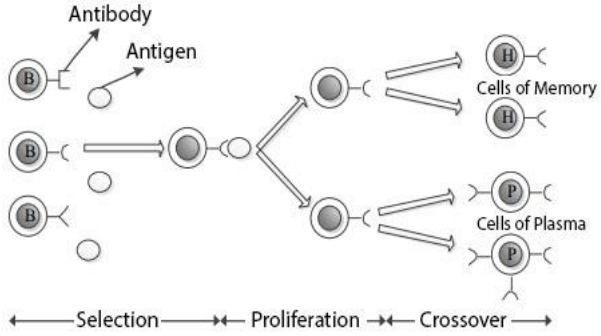

Figure 3. Clonal selection principle

The learning mechanism in the immune system also has an important place. Clonal selection algorithm (CSA) designed to perform machine learning and pattern recognition tasks, where an explicit antigen population represents a set of input patterns at the beginning of 2000 [19]. As shown in Figure 4, the antibody population is first created. The training data that has been preprocessed are read, and an antibody population is created. Then the threshold value that controls the specificity level of the antibodies is determined. Antibodies in the population are cloned to produce new antibodies. Antibodies whose fitness values are lower than the threshold are eliminated. Otherwise, they are added to the population. The test data after the preprocessing are read, and an alarm is generated by comparing the test data with the antibody. The pseudocode of the algorithm is given in Figure 5.

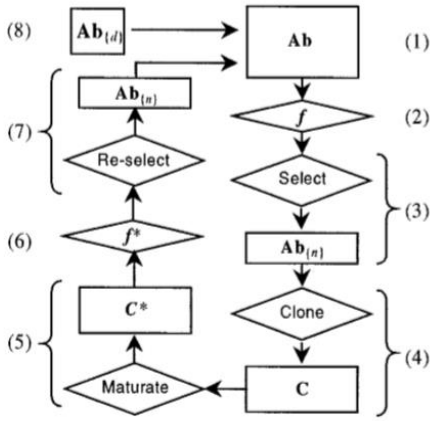

Figure 4. The flowchart of the clonal selection algorithm [1]

```
Input: Ab,Ag; iter,n,d,L;
Output:Abm
for t=1 to iter
        for j=1 to M,
                f: =coding(Ab);
                Abn: =selection(Ab,f;n);
                C: =cloning(Abn;f);
                C: = hypermutation(C,f);
                f: =coding(C);
                Abn: = selection(C;f;n);
                Ab: = add(Ab,Abn);
                Abd: = produce(d,L);
                Ab: =  crossover(Ab;Abd;f))
        end
end
```

Figure 5. The pseudocode of the implementation of CSA on SIEM systems

The threshold value is determined using the Euclidean distance formula. With this formula, the distance between the antibody and antigen is found. Antibody data consists of 42 different data. The test data consists of 42 different data, as well. The distance between the first data of the antibody and the first data of the test data is calculated. By repeating this calculation to all features, all values are added together. The threshold value is determined by taking the average of all these values.

The antibodies and antigens shown in Table 6 and Table 7 are calculated using the Euclidean distance formula starting from the first column to the last column in each row. The threshold value is computed as the average of these values. Antibodies whose values are lower than the threshold are eliminated.

Euclidean distance between Ab (Antibody) and Ag (Antigen) is calculated by Equation (1).

$$D=\sqrt{\sum_{i=1}^{L}(Ab_i - Ag_i)^2} \tag{1}$$

Cloning process:

In this process, all antibodies in Table 6 are cloned using Equation (2).

$$C_i = round\left(\frac{\beta \cdot s(\mathrm{Ab})}{i}\right), \tag{2}$$

where i is index of the antibody in the population, β is an input parameter and $s(Ab)$ is the size of Ab population. C_i determines the number of antibodies to be cloned.

Table 3 Digitized Ab data

1	2	3	4	5	6	…	…	…	38	39	40	41	42
1	0	5	0	3170	329	…	…	…	0.02	0.00	0.09	0.13	0
0	0	0	0	297	13787	…	…	…	0.00	0.00	0.00	0.00	0
0	0	0	0	291	3542	…	…	…	0.00	0.00	0.00	0.00	0

Table 4 Digitized Ag data

1	2	3	4	5	6	…	…	…	38	39	40	41	42
1	0	0	4	215	626	…	…	…	0.02	0.00	0.09	0.65	0
1	0	13	4	232	2019	…	…	…	0.00	0.00	0.00	1.00	2
0	0	13	4	314	315	…	…	…	0.00	0.00	0.00	0.20	2

Each antibody is cloned n times. Each row and column is cloned with a certain similarity rate. The best n of cloned antibodies are selected and added to the antibody population. The K-means clustering algorithm conducts selection. The best k antibodies are selected for the cloned n antibodies and added to the population. Adding new antibodies provides to detect different antigens. Using three different antigen sets, the algorithm detects false-negative and false-positive alarms.

As shown in Figure 5, the data is first read from the file to create an antibody population. New antibodies are generated by determining the similarity ratios and cloning the population of antibodies. The test data read from the file are compared with the population, and attacks are detected. Steps of the clonal selection algorithm is shown in Figure 6.

3. Computational test results

Experiments on the development of SIEM systems using AIS are introduced in this section. False-positive and false-negative rates of SIEM systems have been significantly reduced by using AIS. Three experiments were carried out using the KDD'99 dataset. KDD'99 dataset includes approximately 4,900,000 instances. Each instance has 42 columns. The last column represents the class of the instance. Each instance is either labeled as normal or as an attack. Table 8 shows the attack types.

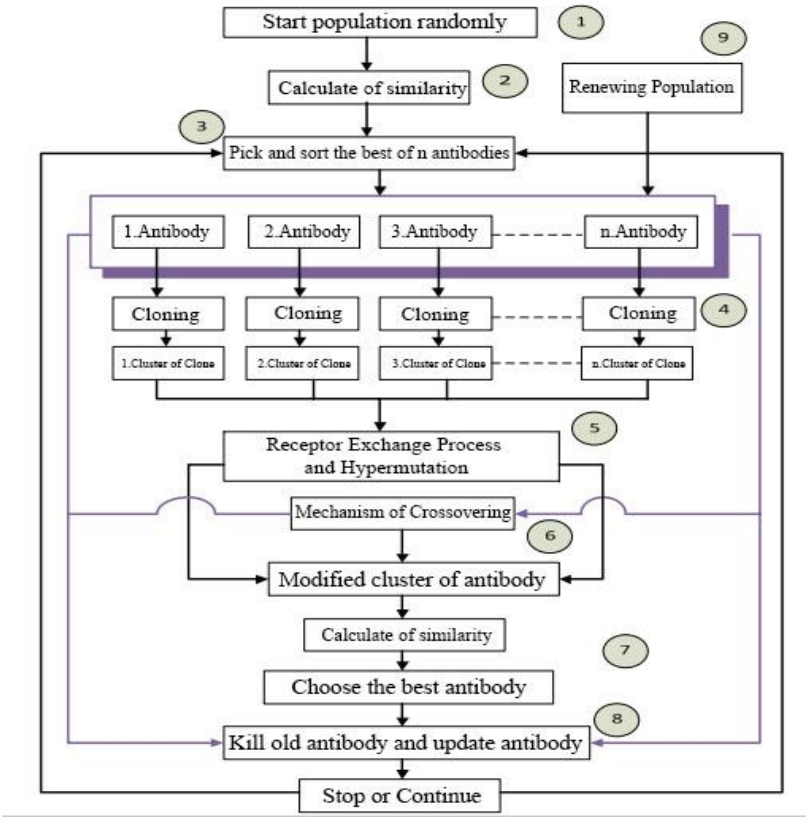

Figure 6. Steps of the clonal selection algorithm

Table 8. Instance and attack types

Class	*Type of attack*
0	*Normal*
1	*Probe*
2	*DoS – Denial of Service*
3	*U2R – User to Root*
4	*R2L – Remote to Local*

The antibody population used in the experiments is shown in Table 9. It is seen that the number of DoS attacks is high in the training data given to AIS. However, since the CSA algorithm is used in the experiments and this algorithm clones itself, there is no problem detecting the low number of attacks.

Table 9. Training population size

Class	*Number of attacks*
0	*21438*
1	*1379*
2	*42681*
3	*4*
4	*34*
Total	*65536*

Three experiments were carried out in total. The test data used in Experiment 1 are shown in Table 10. The number of DoS attacks is higher than the others in the attacks given as antigen data in Experiment 1. This way measures how much AIS strengthens SIEM in determining the antigens with a different number of attacks.

Table 10. The number of antigen attacks used in Experiment 1

Class	*Number of attacks*
0	*6367*
1	*809*
2	*57256*
3	*3*
4	*1100*
Total	*65536*

The number of attacks used in Experiment 2 is given in Table 11. Normal values given as antigens in Experiment 2 are higher than others. Thus, when test data that does not contain a normal attack is given to AIS, the system's success has been measured. In particular, the number of false negatives was measured in this experiment. AIS powered by SIEM was tested on test data with a high number of normal data, and as a result, the false-negative number was measured.

Table 11. The number of antigen attacks used in Experiment 2

Class	*Number of attacks*
0	*16698*
1	*2654*
2	*41791*
3	*4257*
4	*37*
Total	*65536*

The number of attacks used as test data in Experiment 3 is shown in Table 12. The number of antibodies and antigens in the other two experiments is equal. In Experiment 3, the number of attacks in the training data used as the antibody population was defined to be greater than the number of attacks in the test data used as the antigen population. In this way, the number of false positives was measured. By giving SIEM log data to AIS, it was tested how many of these attacks the system can measure in case of an attack.

Table 12. The number of antigen attacks used in Experiment 3

Class	*Number of attacks*
0	*8212*
1	*15*
2	*38235*
3	*1*
4	*2422*
Total	*48885*

According to the results of the experiments, it is seen that the success rates of SIEM systems, which AIS powers, are quite high. Three experiments were conducted with software developed using the Clonal Selection Algorithm (CSA), one of the main algorithms of AIS, and the experimental test results are given in Table 13. The experimental results show that the use of AIS together with SIEM systems yields high success rates.

Table 13. Experimental results

Experiments	*Accuracy Rate*	*False Positive Rate*	*False Negative Rate*
Experiment 1	*%95,13*	*%5,14*	*%3,91*
Experiment 2	*%94,75*	*%7,73*	*%8,23*
Experiment 3	*%97,68*	*%3,45*	*%4,52*
Average	*%95,85*	*%6,44*	*%5,55*

The antibody population was trained with the training data in AIS, and experiments were carried out with three different test data containing different attacks. The experimental results given in Table 13 show that the success rate of AIS is quite satisfactory and high. Another result of this system is that it produces some values using AIS in SIEM systems. These values are as follows. Detection Rate: Detecting an attack if it exists or identifying a normal situation as abnormal is called detection rate. False Positive: If there is an attack, then the system indicates it as normal. False Negative: Although there is no attack on the system, it indicates an attack and generates an alarm. Table 13 presents false positive, false negative, and detection rates obtained in three experiments. The difference in the experimental results is due to the different test data used in the experiments. It is seen that Experiment 3 is more successful, and the detection rate is higher than the other two experiments.

120 iterations are performed for all experiments, and the success percentages are obtained. A summary of the regression statistics based on the percentage of success and the number of iterations are presented in Table 14. Standard error rates are given according to the number of observations made with this statistical information.

Table 14. Regression statistics

Regression statistics	
Multiple R	*0,965952403*
R Square	*0,933064045*
Adjusted R Square	*0,93302643*
Standard Error	*10,20824399*
Observation	*65536*

In addition to the regression statistics in Table 14, the regression graph was obtained in Figure 7. With this regression graph, the percentage of success increased due to the number of iterations. The regression graph changes with the number of observations of the test data used and the number of iterations.

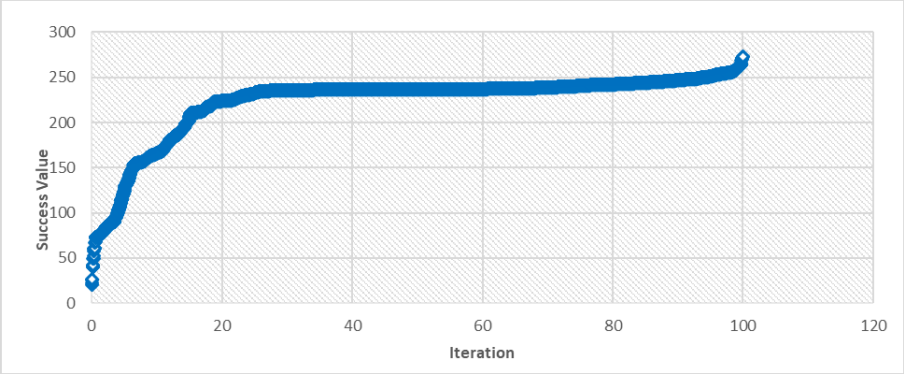

Figure 7. Regression graph

To obtain the regression graph, the iteration number and the percentage of success were chosen as dependent and independent variables, respectively. The regression graph increases as the iteration number increases. As a result, the system was tested on three test data containing different log records and was successful. It is seen that traditional SIEM systems provide significant information security. However, traditional SIEM systems can sometimes give false positive and false negative alarms due to many log data and collecting from multiple sources. Artificial Immune Systems have been used to reduce the number of these alarms, and SIEM systems have been made more robust.

4. Concluding remarks

In detecting cyber security attacks, SIEM systems play an important role. Reviewing the studies carried out on the SIEM systems, it is noticed that they produce false positive and false negative alarms. In this study, to reduce the false-negative alarms generated by SIEM systems, the traditional SIEM system's data analysis phase is adjusted to the CSA of AIS methods, so the SIEM systems' success is increased.

According to the experimental test results obtained from the proposed study:

1. The SIEM systems powered by AIS are more successful than the traditional SIEM systems.

2. The log data used in the SIEM systems should be preprocessed to be used in AIS.

3. The heterogeneous data collected from different sources should pass certain criteria to filter the unnecessary data.

4. Log data contains a large amount of event data. To increase the system's performance, similar log and event data should be recorded as a single log.

5. It has been observed that the variety and amount of training and test data did not affect the success rate.

Considering all these results, it has been observed that implementation of AIS on the SIEM systems reduced false positive false negative alarm rates and worked more efficiently. As a suggestion, the artificial immune algorithm applied in this paper to the SIEM system can be adjusted to other intrusion detection and prevention systems to increase their performance.

References:

1 Katsikas S. and Anastopoulos V. (2019), A Methodology for the Dynamic Design of Adaptive Log Management Infrastructures, EAI Endorsed Transactions on Security and Safety,6(19):1-14.

2 Sun L., Zhang H., Fang C.(2021), Data security governance in the era of big data: status, challenges, and prospects, Data Science and Management, 2:41-44.

3 Coppolino, L., D'Antonio, S., Nardone, R. et al(2023), A self-adaptation-based approach to resilience improvement of complex internets of utility systems, Environment Systems and Decisions, https://doi.org/10.1007/s10669-023-09937-8.

4 Mauro M. and Sarnob C.(2018), Improving SIEM capabilities through an enhanced probe for encrypted Skype traffic detection, Journal of Information Security and Applications, 85-95.

5 Gunduz M.Z., Das R.(2020), Cyber-security on smart grid: Threats and potential solutions, Computer Networks, 169:1-14.

6 Aldhaheri S., Alghazzawi D., et al(2020), Artificial Immune Systems approaches to secure the internet of things: A systematic review of the literature and recommendations for future research, Journal of Network and Computer Applications, 157:1-24.

7 Hajisalem V., Babaie S.(2018), A hybrid intrusion detection system based on ABC-AFS algorithm for misuse and anomaly detection, Computer Networks, 136:37-50.

8 Vidal J. M., Orozco A. L. S. and Villalba L. J. G. (2018), Adaptive artificial immune networks for mitigating DoS flooding attacks,Swarm and Evolutionary Computation, 38:94-108.

9 Singh A. P., Kumar S., Kumar A. and Usama M.(2022), Machine Learning based Intrusion Detection System for Minority Attacks Classification, 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), Greater Noida, India, , 256-261, doi: 10.1109/CISES54857.2022.9844381.

10 Ma C., Du X., Cao L. (2019), Analysis of Multi-Types of Flow Features Based on Hybrid Neural Network for Improving Network Anomaly Detection, in IEEE Access, 7:148363-148380.

11 Ali A., Khan A., Ahmed M., Jeon G. (2021), BCALS: Blockchain-based secure log management system for cloud computing, Transactions on Emerging Telecommunications Technologies, 33(4).

12 González-Granadillo G., González-Zarzosa S., Diaz R. (2021), Security Information and Event Management (SIEM): Analysis, Trends, and Usage in Critical Infrastructures, Sensors, 21(14),1-28.

13 Kenaza T. (2021). An ontology-based modelling and reasoning for alerts correlation. International Journal of Data Mining, Modelling and Management, 13(1-2), 65-80.

14 Mercl L., & Horalek J. (2020), SIEM Implementation for Small and Mid-Sized Business Environments, Journal of Engineering and Applied Sciences, 14:10497-10501.

15 Al-Duwairi B., Al-Kahla W., AlRefai M. A., Abedalqader Y., Rawash A., & Fahmawi R. (2020), SIEM-based detection and mitigation of IoT-botnet DDoS attacks, International Journal of Electrical and Computer Engineering, 10(2):2182.

16 Bezas K., & Filippidou F. (2023), Comparative Analysis of Open Source Security Information & Event Management Systems (SIEMs), Indonesian Journal of Computer Science, 12(2):443-468.

17 Albasheer H., Md Siraj M., Mubarakali A., Elsier Tayfour O., Salih S., Hamdan M., ... & Kamarudeen S. (2022), Cyber-attack prediction based on network intrusion detection systems for alert correlation techniques: a survey, Sensors, 22(4):1494.

18 Ahmad A., Desouza K. C., Maynard S. B., Naseer H., & Baskerville R. L. (2020), How integration of cyber security management and incident response enables organizational learning, Journal of the Association for Information Science and Technology, 71(8):939-953.

19 Singh K., Kaur L., & Maini R. (2022), A survey of intrusion detection techniques based on negative selection algorithm, International Journal of System Assurance Engineering and Management, 1-11.

20 KDD'99, "http://kdd.ics.uci.edu/," [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html. [Accessed 10 08 2023].