International Journal of Engineering and Innovative Research

http://dergipark.gov.tr/ijeir

MACHINE LEARNING BASED SECURITY ANALYSIS: ALARM GENERATION AND THREAT FORECASTING

Fatma Bozyiğit[1*], Okan Türksever[2], Ozan Türksever[2],Deniz Kılınç [1]

[1] İzmir Bakırçay University, Faculty of Engineering and Architecture, Computer Engineering, Izmir, Turkey.
[2] Software Engineer, Logsign, İstanbul, Turkey.

[*]Corresponding Author: fatma.bozyigit@bakircay.edu.tr

(**Received:** 01.01.2020; **Revised:** 05.03.2020 **Accepted:** 13.04.2020)

ABSTRACT: T Log files keep activity records of each process performed have an important place in terms of security. Systems that provide infrastructure for applications such as network security mainly work on log management. Recently, when the security mechanisms of popular applications are examined, it has been observed that they aim to strengthen their infrastructures with machine learning (ML) methods, but in some respects, they have shortcomings. In this study, we aim to develop an alarm and security reporting system using ML methods. Our study differs from the others since it considers five separate feature (IP reputation, web reputation, malware destination access, botnet) and includes them into ML model.

Keywords: Log Analysis, Security Management, Alarm System, Machine Learning.

1. INTRODUCTION

Log data is the recordings produced by information systems for various purposes (security, error recovery, performance, control, etc.) [1]. Log management, which is performed by system administrators for fault detection, have become widely used for security issues, information analysis, and compliance with standards in recent times. Moreover, storing the log data is obligated by the law of 5651, “İnternet ortamında yapılan yayınların düzenlenmesi ve bu yayınlar yoluyla işlenen suçlarla mücadele edilmesi” and ISO 27001 which is the standards of the IT sector.

Logging is critical task for detection of cyber-attack and event management. Cyber security specialists reveal the security flaws in the system by analysing the log records. More recently, Security Information and Event Management (SIEM) has been joined by the broad use of log management technology that focuses on collecting an extensive variety of logs. SIEM approach is seen as a more advanced system than logging, since it offers more detailed and real time configuration and reporting options. One of the most important features of SIEM is the correlation technique that helps to identify possible attacks by establishing meaningful connections between independent events with the help of the security policies and rules [2]. SIEM has composed of two main functions; security information management (SIM) and security event management (SEM) [3]. SIM provides the collection, reporting and analysis of log data; primarily from host systems and applications, and secondarily from network and security devices, to support regulatory compliance reporting, internal threat management and resource access monitoring. SEM processes log and event data from security devices, network devices, systems and applications in real-time to provide security monitoring, event correlation and incident responses.

Traditional SIEM systems push security notifications by conducting static rule-based model through previous safety threats. However, this approach causes missing different types of attack which have not appealed before. Generating specific rule sets is necessary to detect dissimilar cyber-attacks considering the factors such as change in network characteristics, usage behavior's, and users' tendency. Another point to be noticed in current SIEM systems is that they do not use the analysis results of log file to explore future threats.

In this study, we propose to detect users having unexpected actions on a network by using a novel hybrid approach consisting of non-static rules and ML methods. We also aim prediction of the network attacks which may occur in the future. For this aim, we trainee base models with the use of dataset including log files of many applications to detect the current network anomalies and unpredictable problems which are revealed using current security issues. The proposed ML approaches are compared in terms of Root Mean Square Error (RMSE) and it is concluded that our study is promising in future network attack predictions.

This paper is organized as follows. In Section 2, related works are presented. Section 3 gives information about the methods used for security analysis of log files. Section 3 gives information about proposed ML methods in our study. In Section 4, some experiments are implemented on the data set and the results are discussed. Section 5 concludes the paper and gives information about our future work.

2. RELATED WORKS

One of the most noticeable points derived from existing literature is that current SIEM infrastructures contain some limitations. A contribution to the existing literature to eliminate these limitations is realized by AlSabbagh and Kowalski [4]. They develop a new network analysis and incident management system to support the companies where socio-technical security operations are performed [4]. To eliminate the drawbacks and constraints of the SIEM infrastructure the researchers calculate the risk management maturity level by considering different metrics.

Some studies in the literature claim that the user information must be analyzed along with data obtained through log analysis. One of these studies is carried out by Deliang. He states that cyber profile creation is the most important step for the web applications to obey privacy and security policies [5]. Accordingly, the user information is also considered beside log analysis for system threats and problem identification.

In case of considerable number of alarms arising, users in the network must focus on the most important alarms and ignore non-critical alarms. This situation (warning the users with inappropriate security alerts) can lead to time consumption and performance decrease. Thus, specification of the appropriate security alerts is the critical sub task in security systems. One of the studies on alarm classification is conducted by Schleburg et al. [6]. Researchers identify the significance level of automatically generated alarms using Market Basket Analysis method.

In the past several years there has been extensive research into detection and information gathering against external threats. However, a few numbers of study have been utilized for insider threat which is widely known as the most dangerous attack. Insider has many advantages since knowing internal organization. Focusing on insider threats may cause the increase in the false generated alarms. One of the studies aiming to find a solution to this issue is carried out by Ambre and Shekokar [7]. The researchers aim to decrease the unnecessary alarms to detect insiders as soon as possible.

3. PROPOSED APPROACHES

3.1. K-Means Algorithm

The K-means clustering is one of the simplest and most popular unsupervised ML algorithms. Unsupervised algorithms reveal information from datasets using only input vectors without referring to tagged attributes [8]. To achieve this goal, the number of clusters in the data set is determined first. Clusters represent the collection of the data points due to certain similarities. Assignment of samples to clusters is done by determining the distances from cluster centers.

The general architecture of the K-means clustering algorithm is demonstrated with the flow diagram in Figure 1.

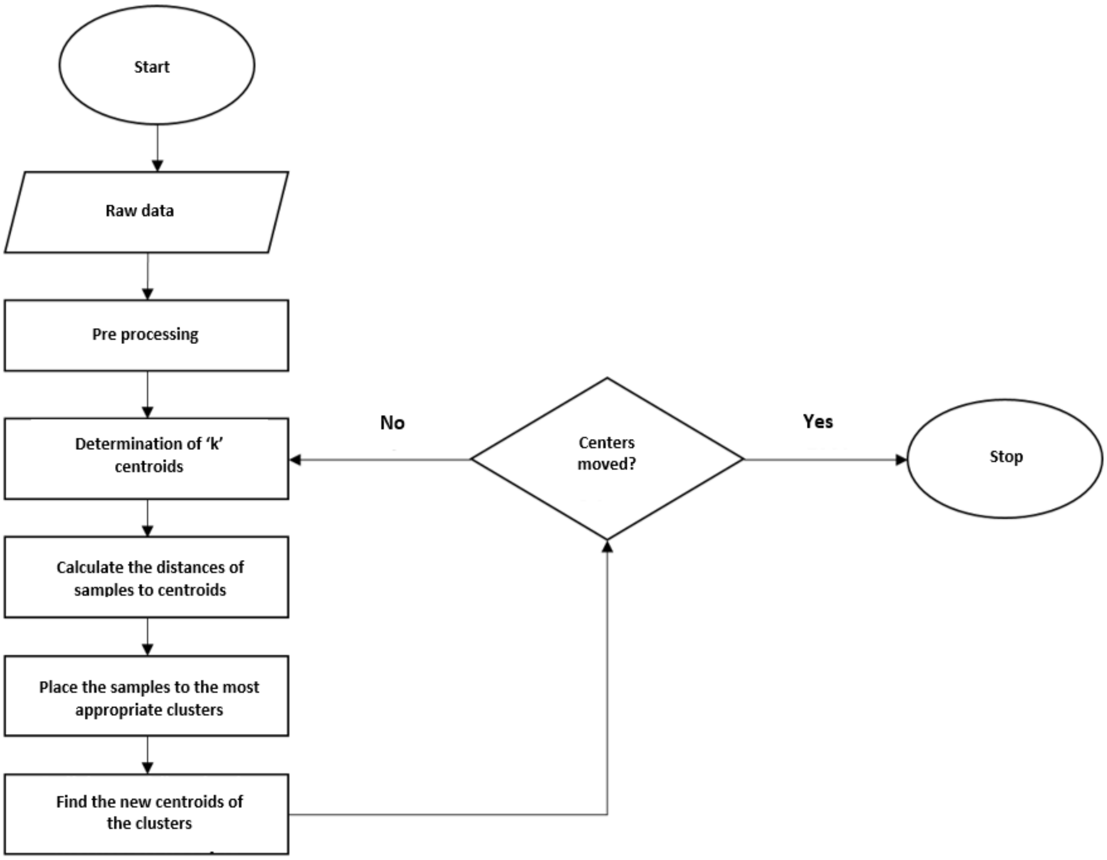

Start
Raw data
Pre processing
Determination of 'k' centroids
Calculate the distances of samples to centroids
Place the samples to the most appropriate clusters
Find the new centroids of the clusters
Centers moved?
No
Yes
Stop

Figure 1. Flowchart of K-means algorithm

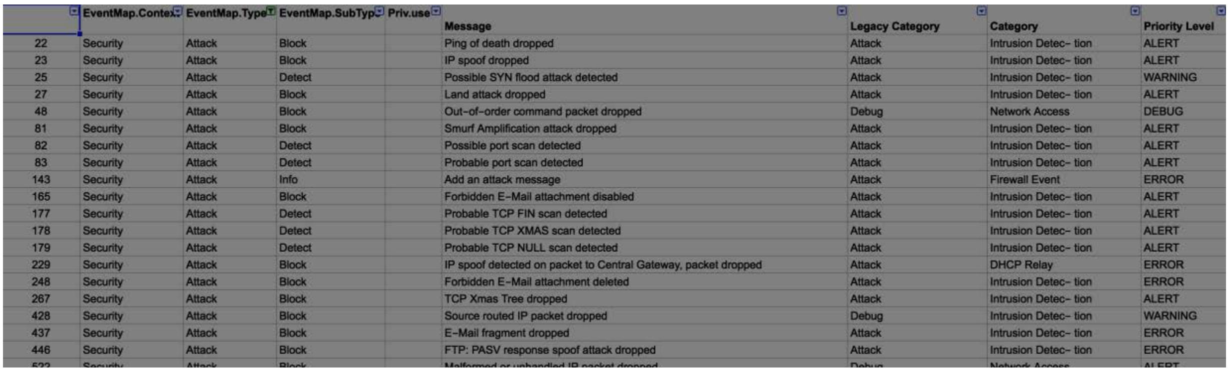
3.2. Gaussian Mixture Model

Gaussian Mixture (GM), which has similar properties with the K-mean, is one of the most widely used clustering methods. According to researches, GM method has some advantages over K-means. Firstly, K-means cluster the data without considering the variance, while GM model uses the variance to reveal the distribution patterns of the data in a clearer manner [9]. To summarize, K-means model is that it places a circle (or, in higher dimensions, a hyper-sphere) at the center of each cluster, with a radius defined by the most distant point in the cluster. This works fine for when the data is circular. However, when data takes on different shape, K-means does not perform well. In contrast, Gaussian mixture models can handle even very oblong clusters.

4. EXPERIMENTAL STUDY

4.1. Experimental Dataset

The experimental dataset of our study is constructed collecting data from more than one hundred information system components (network, operating system, database, and so on). First collected data is normalized and split into columns to make the analyse process easier. During this process, legacy category and priority level of the events are tagged. Figure 2 shows the examples of the categorized samples.



	EventMap.Contex	EventMap.Type	EventMap.SubTyp	Priv.use	Message	Legacy Category	Category	Priority Level
22	Security	Attack	Block		Ping of death dropped	Attack	Intrusion Detec- tion	ALERT
23	Security	Attack	Block		IP spoof dropped	Attack	Intrusion Detec- tion	ALERT
25	Security	Attack	Detect		Possible SYN flood attack detected	Attack	Intrusion Detec- tion	WARNING
27	Security	Attack	Block		Land attack dropped	Attack	Intrusion Detec- tion	ALERT
48	Security	Attack	Block		Out-of-order command packet dropped	Debug	Network Access	DEBUG
81	Security	Attack	Block		Smurf Amplification attack dropped	Attack	Intrusion Detec- tion	ALERT
82	Security	Attack	Detect		Possible port scan detected	Attack	Intrusion Detec- tion	ALERT
83	Security	Attack	Detect		Probable port scan detected	Attack	Intrusion Detec- tion	ALERT
143	Security	Attack	Info		Add an attack message	Attack	Firewall Event	ERROR
165	Security	Attack	Block		Forbidden E-Mail attachment disabled	Attack	Intrusion Detec- tion	ALERT
177	Security	Attack	Detect		Probable TCP FIN scan detected	Attack	Intrusion Detec- tion	ALERT
178	Security	Attack	Detect		Probable TCP XMAS scan detected	Attack	Intrusion Detec- tion	ALERT
179	Security	Attack	Detect		Probable TCP NULL scan detected	Attack	Intrusion Detec- tion	ALERT
229	Security	Attack	Block		IP spoof detected on packet to Central Gateway, packet dropped	Attack	DHCP Relay	ERROR
248	Security	Attack	Block		Forbidden E-Mail attachment deleted	Attack	Intrusion Detec- tion	ERROR
267	Security	Attack	Block		TCP Xmas Tree dropped	Attack	Intrusion Detec- tion	ALERT
428	Security	Attack	Block		Source routed IP packet dropped	Debug	Intrusion Detec- tion	WARNING
437	Security	Attack	Block		E-Mail fragment dropped	Attack	Intrusion Detec- tion	ERROR
446	Security	Attack	Block		FTP: PASV response spoof attack dropped	Attack	Intrusion Detec- tion	ERROR
522	Security	Attack	Block		Malformed or unhandled IP packet dropped	Debug	Network Access	ALERT

Figure 2. Attributes of sample data.

Finally, the attributes of the samples in the dataset are combined with each other with respect to generate new features (Figure 3).

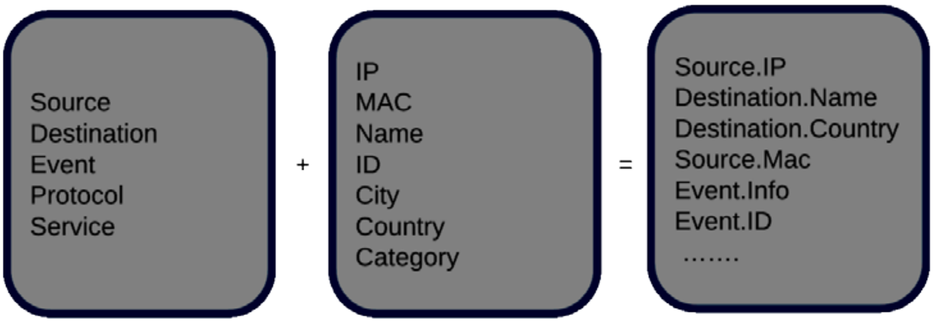


Figure 3. Combination of attribute names.

4.2. Evaluation Results

In this study, our goal is to reveal possible threats on the system and to minimize the number of false alarms received during the day. Static rules (not specific for company's network architecture) can cause excessive alarm triggers, performance loss occurs accordingly. Considering this situation, False Positive alarms, generated by using static rules, is examined. The results are visualized as in Figure 4.

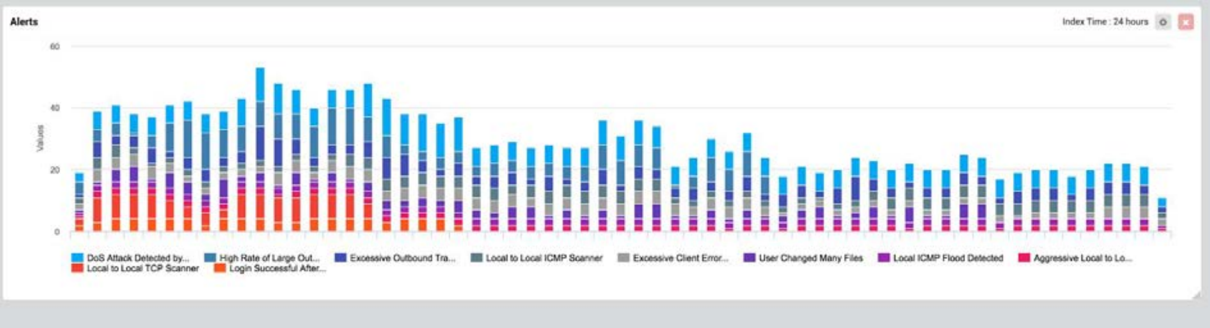


Figure 4. Analysis of False Positive alarms

The threshold value for the alarms tagged as False Positive is dynamically set to minimize the number of security threats. This dynamic architecture determines the alarms to be triggered with respect to change of the user actions within 24 or 48 hours. As a result, the number of the false alarms decreases by 70% to ensure more accurate notification system.

In the second part of our study, K-Means and Gaussian Mix clustering algorithms are performed on the experimental data set. Using the obtained results, we identify the users who behave differently from what it should be. K-means algorithm split the samples into two category such as “frequently seen situations” and “rare cases” (Figure 5). If any log that are newly introduced to the system has different characteristics from the rare cases, they are included in the "rarely seen" cluster and a security alarm is triggered.

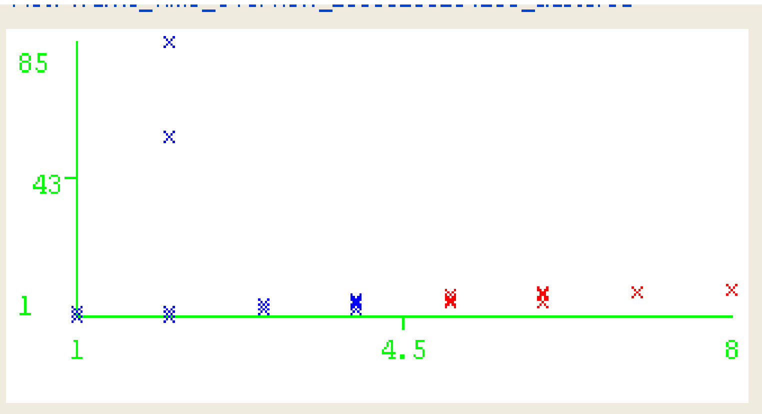


Figure 5. K-means clustering (k=2)

When the results Figure 6 illustrates are evaluated, it is concluded that K-means is not an appropriate alternative for our study, and the GK model performs better on the experimental dataset.

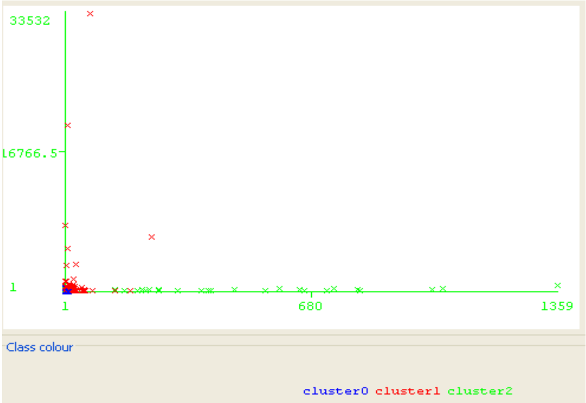
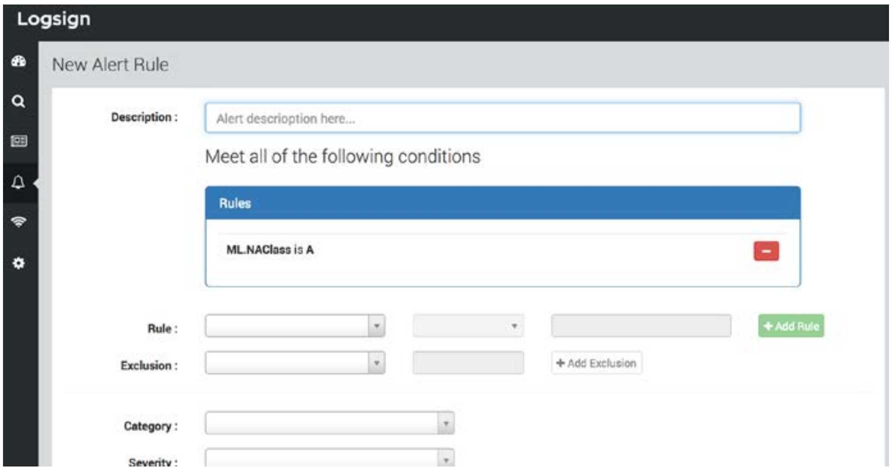


Figure 6. GK clustering (k=3)

Figure 7 and Figure 8 show some of the system interfaces.



Logsign

New Alert Rule

Description : Alert descrioption here...

Meet all of the following conditions

Rules

ML.NAClass is A

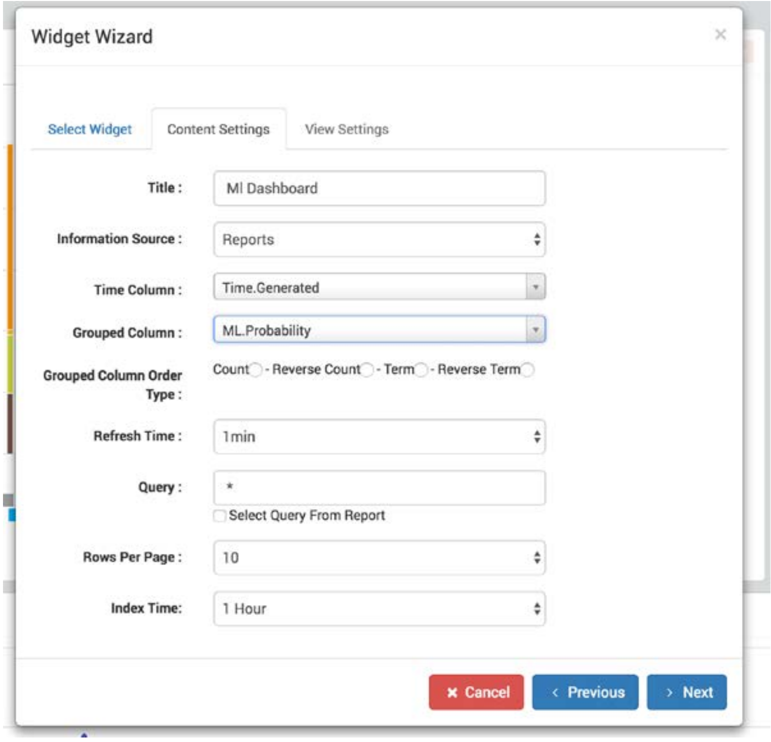
Rule :

Exclusion :

Category :

Severity :

Figure 7. System Interface



Widget Wizard

Select Widget | Content Settings | View Settings

Title : Ml Dashboard

Information Source : Reports

Time Column : Time.Generated

Grouped Column : ML.Probability

Grouped Column Order Type : Count ○ - Reverse Count ○ - Term ○ - Reverse Term ○

Refresh Time : 1min

Query : *

□ Select Query From Report

Rows Per Page : 10

Index Time: 1 Hour

Figure 8. System Interface

5. CONCLUSION

The most important problem in traditional security mechanisms is the use of static rules for the detection of different attacks. However, this common approach can be inefficient due to the change of network characteristics, usage differences, and user tendencies. Considering the deficiencies in the existing systems, we aim to identify the users performing unusual actions on a network by using a novel hybrid approach consisting of non-static rules and ML methods. We also utilize prediction of the network attacks which may occur in the future. In the experimental studies based on the K-means and the GM clustering algorithms, it is observed that GM gives more accurate results to detect of unusual events in the network.

REFERENCES

[1] Jansen, B. J., Spink, A., & Taksai, I. (2009). Handbook of research on web log analysis. London: Information Science Reference.

[2] T.C. Resmi Gazete. Retrieved from https://www.resmigazete.gov.tr/eskiler/2007/11/20071130-6.htm, Aralık, 2019.

[3] Miller, D. (2011). Security information and event management (SIEM) implementation. McGraw-Hill.

[4] AlSabbagh, B., & Kowalski, S. (2016, August). A Framework and Prototype for A Socio-Technical Security Information and Event Management System (ST-SIEM). In 2016 European Intelligence and Security Informatics Conference (EISIC) (pp. 192-195). IEEE.

[5] Deliang, C., Xing, L., & Qianli, Z. (2016, May). A comparative study on user characteristics of fixed and wireless network based on DHCP. In 2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference (pp. 327-330). IEEE.

[6] Schleburg, M., Christiansen, L., Thornhill, N. F., & Fay, A. (2013). A combined analysis of plant connectivity and alarm logs to reduce the number of alerts in an automation system. Journal of process control, 23(6), 839-851.

[7] Ambre, A., & Shekokar, N. (2015). Insider threat detection using log analysis and event correlation. Procedia Computer Science, 45, 436-445.

[8] Li, T., & Yan, L. (2017, June). Siem based on big data analysis. In International Conference on Cloud Computing and Security (pp. 167-175). Springer, Cham.

[9] Rasmussen, C. E. (2000). The infinite Gaussian mixture model. In Advances in neural information processing systems (pp. 554-560).