SURVEY

UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Log File Analysis Based on Machine Learning: A Survey

Rawand Raouf Abdalla, Alaa Khalil Jumaa

Department of Information Technology, Technical College of Informatics, Sulaimani Polytechnic University, Sulaimani, Kurdistan Region, Iraq

ABSTRACT

In the past few years, software monitoring and log analysis become very interesting topics because it supports developers during software developing, identify problems with software systems and solving some of security issues. A log file is a computer-generated data file which provides information on use patterns, activities, and processes occurring within an operating system, application, server, or other devices. The traditional manual log inspection and analysis became impractical and almost impossible due logs' nature as unstructured, to address this challenge, Machine Learning (ML) is regarded as a reliable solution to analyze log files automatically. This survey tries to explore the existing ML approaches and techniques which are utilized in analyzing log file types. It retrieves and presents the existing relevant studies from different scholar databases, then delivers a detailed comparison among them. It also thoroughly reviews utilized ML techniques in inspecting log files and defines the existing challenges and obstacles for this domain that requires further improvements.

Index Terms: Log Files, Log Analysis, Machine Learning, Anomaly Detection, User Behavior, Log File Maintenance

1. INTRODUCTION

In the context of computing, logs are bits of data that give insight into numerous events that occur during the execution of a computer program [1]. Information technology utilization has increased at an unparalleled rate during the previous two decades. Data of various types are shared through a broad range of networks, from company-wide LAN networks to public hub wireless access networks. As the transmission and consumption of data through these networks grows, so does number of breaches and network intrusion efforts aimed at obtaining secret and personal information. As a consequence of this, security for networks and data has become a highly significant topic in both the academic and practical computing communities [2].

Log data comprised more than 1.4 billion logs each day is used to detect suspicious business-specific activities and user profile behavior [3].

A series of devices and software generate log files in dissimilar formats. Log files are used by software systems to retain track of their activities. Different system part, like OS, may record its events to a remote log server. An OS is the machine software that controls computer hardware and software resources and permits the execution of multiple applications [4]. The start or end of occurrences or activities of software system, status information, and error information are all captured in the log files. User information, application information, date and time information, and event information are normally included in each log line. When these files are properly analyzed, they may provide important information about numerous characteristics every system. For monitoring, troubleshooting, and problem detection, logs are often gathered [5].

Access this article online

DOI: 10.21928/uhdjst.v6n2y2022.77-84

E-ISSN: 2521-4217

P-ISSN: 2521-4209

Corresponding author's e-mail: rawand.raouf.a@spu.edu.iq

Received: 16-07-2022 Accepted: 07-09-2022 Published: 07-10-2022

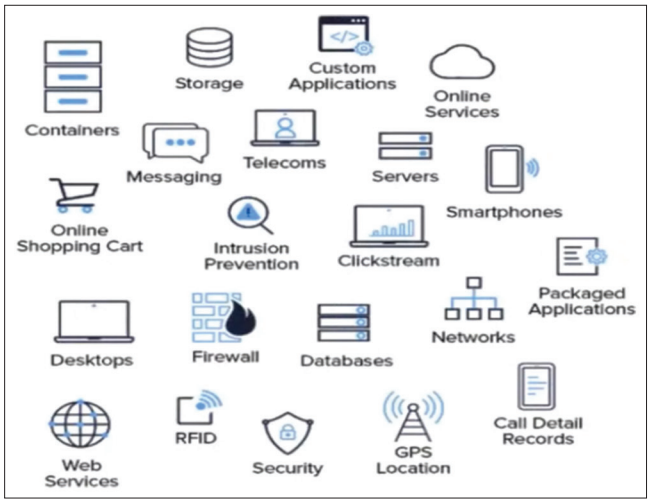

Fig. 1. Some of log file sources [11].

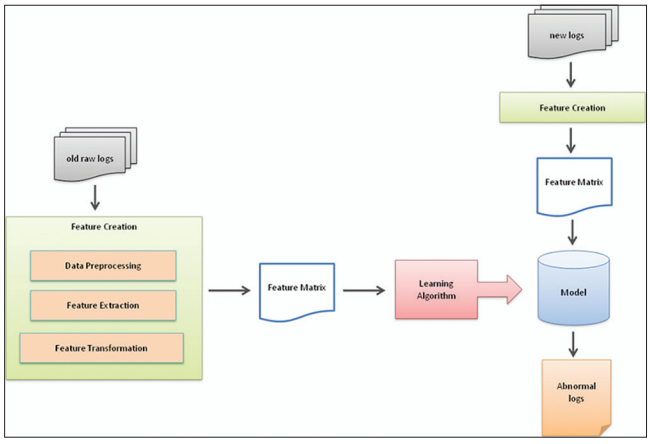

Fig. 2. Overview of the learning system [14].

Normally, log files are saved as text, compressed, or binary files. The most commonly used format is text files, which have the gains of utilizing fewer CPU and I/O resources when producing files, allowing for long-term storage and maintenance, and being easy to read and use. The binary format is a machine-readable log file format created by a platform that requires a particular tool to view, making it unsuitable for long-term team storage. While compressing log files, use an appropriate compression format standard with multi-platform compatibility for efficient log storage and usage [6].

Log files record activity on different kinds of servers, including data servers and application servers. Because log files record a range of server actions and include a huge data in the form of server-produced messages, they are often relatively big files. These messages include valuable knowledge, like what apps were operating on the server, when they were run, and by whom a log file contains a lot of messages that indicate various server actions. In addition, each unique message type will have hundreds, if not thousands, of entries in the log file, each with slight variances in the general structure. The system administrator may utilize these messages for a lot of reasons, such as intrusion detection, reporting, performance monitoring, anomaly detection, and so on.

2. STATE OF THE ART

This survey tries to explore the most recent existing studies on analysis log files by using both of supervised and unsupervised machine learning algorithms. Different models and techniques have been proposed by researchers to different aspects. Several studies have utilized techniques and models to predict attack, user behavior, and system failure to increase server security and systems, marketing, and decrease failure time. This study revealed that there are many gaps which require further improvements such as: Using real dataset in creating models, more log analysis, or mining must be done to obtain meaningful information and minimize the false positive and negative results and the maintenance aspect requires further improvements compared to the other mentioned aspects.

3. COMMON LOG FILE TYPES

Almost every network device produces a unique form of data, and each component logs those data in its own log. As a result, there are several types of logs, such as [7]:

3.1. Application Logs

Developers have a strong grip on application logs. It may include any kind of event, error message, or warning that the program generates. Application logs provide information to system administrators concerning the status of an application running on a server. Application logs should be well-structured, event-driven, and include pertinent data to assist as the center for higher-level abstraction, visualization, and aggregation. The application logs' event stream is required for viewing and filtering data from numerous instances the programs.

3.2. Web Server Logs

Every user communication with the web is saved as a record in a log file called a "web log file" that is a text file with the extension ".txt." The data created automatically of users' interactions with the web, and will be saved in a variance log files, including server access logs, error logs, referrer logs, and client-side cookies. In the style of text file, this web log

TABLE 1: A list of studies for the purpose of (identifying user behavior)

Reference No.	Classification Algorithms	Performance
[16]	NN	Prediction Accuracy is 90%
[17]	(Parzen) (Gauss) (PCA) and (KMC)	Prediction Accuracy for daily activity dataset is 90% e-mail content dataset is 65% e-mail communication network dataset 75%
[18]	Modified Span Algorithm and the Personalization Algorithm	Provides high predication accuracy

TABLE 2: A list of studies for the purpose of (Security)

Reference No.	Classification Algorithms	Performance
[20]	K-Means Clustering	Direction Accuracy ratio is 83%
[21]	SVR, LR, and KNR	Offers excellent security protection
[22]	LR, NN, RF, and XG	Direction Accuracy ratio is 85% with 0.78% false positive rate
[23]	SVM	Direction Accuracy ratio is 99%
[24]	K-Means Clustering	Direction Accuracy ratio for SOM#34 is (84.37%) AAU is (90.01%)

TABLE 3: A list of studies for the purpose of (Maintenance)

Reference No.	Classification Algorithms	Performance
[25]	Discovering Patterns from Temporal Sequences	Provides high system performance
[26]	Principal Component Analysis (PCA)	Provides high system performance

saves all and every web request executed by the client to the servers. Each line or record in the web log file links to a user's request to the servers. The logs file for the web data are: Web Server Logs, Proxy Server Logs, and Browser Logs [8].

Web server logs typically include the IP address, the date and time of request, the exact request line providing by users, the URL, and the requested file type.

3.3. System Logs

The OS records specified events in System log. In addition, these logs are an excellent resource for obtaining information about external events. Typically, a system log includes entries generated by the OS, such as system failures, warnings, and errors. Individual programs may generate log files related with user sessions that include data on the user's login time, interactions with the application, authentication result, and so on. While an operating system-generated log file is mentioned to as a system log, files produced by particular programs or users are related to as audit data. Examples include records of successful and unsuccessful login attempts, system calls, and user command executions [9].

3.4. Security Logs

Security logs are utilized to give enough capabilities for identifying harmful actions after their occurrence with the intention of prevent them from recurrence. Security logs preserve track of a range information that has been pre-defined by system administrators. For example, firewall logs contain information about packets routed from their sources, rejected IP addresses, outbound activity from internal systems, and failed logins. Security logs contain detailed information that security administrators must manage, regulate, and evaluate in conformity with their requirements [7].

3.5. Network Logs

Network Logs offer different information on various events that occurred on the networks. Among the events are the recording of malicious activity, a rise in network traffic, packet losses, and bandwidth delays. Network logs may be gathered from a range of network devices, including switches, routers, and firewalls. By monitoring network logs for various attack attempts, network administrators can monitor and troubleshoot normal networking.

3.6. Audit Logs

Record any network or system activity that is not allowed in a sequential order. It aids security managers in analyzing malicious activity during an attack. The source and destination addresses, timestamp and user login information are usually the most significant parts of information in audit log files.

3.7. Virtual Machine Logs

This log files include details on the instances that are executing on the VM, such as their startup configuration, operations, also the time they complete their execution. The VM logs retain track of many processes, including the number of instances operating on the VM, the execution duration of each application, and application migration, which assists the CSP in identifying malicious activity that happened while the attack [10]. Figure 1 show some of log file sources.

4. LOG MINING

Log mining is a technique which employs statistics, data mining, and ML to automatically explore and analyze vast amounts of log data in essence to find useful patterns and trends. The data and tendencies gleaned might assist in the monitoring, administration, and troubleshooting of software systems [12]. Web Usage Mining (WUM) as an example because it the most frequently utilized log file types.

Web mining is separated into three categories: Web Usage Mining, Web Content Mining and Web Structure Mining. Web usage mining is a procedure for capturing web page access data. The pathways leading to viewed web sites are providing by this use data. This data are often collected automatically by the web server then stored in access logs. Other important information provided by Common Gateway Interface CGI scripts includes referral logs, survey log data, and user subscription information. This area is significant to the entire usage of data mining by businesses, institutions, and their data access and web-based applications. Three steps necessity be taken in Web Usage Mining which are [13]:

4.1. Data Preprocessing

The web logs contain raw data that cannot be utilized to generate information. During this step, engineers use techniques to transform original data for a usable format. Typically, real-world data are incomplete, unexpected, and lacking of behavior or patterns, in addition including many mistakes. Data pre-processing is a tried-and-true way of solving these issues.

4.2. Pattern Discovery

Those pre-processing results are then used to determine a pattern of frequent user access. To identify significant information, several data mining methods for instance association rules, clustering, classification, and sequential pattern approach will be used in pattern discovery. The information that has been obtained could be presented in a number of ways including graphs, charts, tables, and so on.

4.3. Pattern Analysis

Final outcome of the pattern discovery step are not utilized directly in the analysis. Accordingly, during this phase, a strategy or tool will be developed to assist analysts in comprehending the knowledge which has been gathered. Visualization approaches, Online Analytical Processing OLAP analysis, and this phase might involve the use of tools or methods for instance knowledge query mechanisms. Figure 2 Overview of the learning system.

5. TYPES OF LOG FORMAT

Common servers use one of these three types of log file formats [15]:

5.1. Common Log File Format

Web servers create log files utilizing this standardized text file format. The setup of the standard log file format is provided in the box that follows.

Example of Common Log File Format [15].

5.2. Combined Log Format

It is like the previous log file format with the add of the referral field, user agent field, and cookie field. The setup for this format is shown in the box follows.

```
LogFormat "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{Useragent}i\"" combined CustomLog log/access_log combined eg: 127.0.0.1 - frank [15/Oct/2021:14:59:38 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2328 "https://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I ;Nav)"
```

Example of Combined Log Format [15].

5.3. Multiple Access Logs

It is a hybrid of the common log and the combined log file format, with the ability to establish several directories for access logs. The structure of various access logs is detailed in the box that follows.

```
LogFormat "%h %l %u %t \"%r\" %>s %b" common CustomLog logs/access_log common CustomLog logs/referer_log "%{Referer}i -> %U" CustomLog logs/agent_log "%{User-agent}i"
```

Example of Multiple Access Logs [15].

6. DATA PREPROCESSING AND ML

6.1. Data Preprocessing

It is critical to preprocess data to handle with different flaws in raw gathered data, which may include noise such as mistakes, redundancies, outliers, and other missing values or unclear data. The most prevalent procedures in data preprocessing are [16]:

6.1.1. Data cleaning
Handle data inconsistencies, noise, and missing values.

6.1.2. Data integration
Seeks to integrate data from several sources into a cohesive data storage unit. Which is not an easy operation, since it entails establishing compatibility across several schema types. Weak or inefficient data integration might result in inconsistency and redundancy, while a well-implemented solution would surely improve accuracy and improve subsequent operations. Data integration techniques involve entity identification, correlation analysis, tuple deduplication, redundancy, and along with the discovery and resolution of data value conflicts.

6.1.3. Data transformation
Aims to transform data in a style which is both useable and meaningful format. The reason is for data mining processes to be more efficient. Smoothing, feature building, normalization, discretization, and generalization of nominal data are all examples of data transformation strategies. These subtasks are heavily reliant on the preprocessed data and need human supervision.

6.2. Machine Learning
A science focused with the theory, performance, and features of learning systems and algorithms. ML is a highly interdisciplinary field that depend on techniques from several fields, including artificial intelligence, cognitive science, optimization theory, information theory, statistics, and optimal control. ML has permeated virtually every scientific subject on consequence of its broad use in a various of applications, having a tremendous impact on both research and society. It was used to a range of challenges, such as autonomous control systems, informatics and data mining, recognition systems, and recommendation systems [17].

ML is broadly classified into three subfields [18]:
- Supervised Learning: It needs training on labeled data that contain both inputs and outputs.
- Unsupervised Learning: Not needs labeled training data, as the environment solely offers unlabeled inputs.
- Reinforcement Learning: It permits learning to occur on consequence of feedback obtained from interactions with the external environment.

Analysis of log files relevant to a failed execution may be laborious, particularly if the file contains thousands of lines. Utilizing current advancements in text analysis using deep neural networks (DNN), research [3] presents an approach to decrease the effort required to study the log file by highlighting the most likely informative content in the failed log file, which may aid in troubleshooting the failure's causes. In essence, they decrease the size of the log file by deleting lines deemed to be of less significance to the problem.

7. CONTRIBUTION OF THE LOG ANALYSIS

The log analysis contribution split into four categories, as follows [19]:

7.1. Performance
Used to find the system's performance during the optimization or troubleshooting phase. In the instance of performance, logs assist the administrator in clarifying how a specific system's resource has been utilized.

7.2. Security
Security logs are a lot used to detect security breaches or misconduct and to conduct postmortem investigations into security occurrences. For example, intrusion detection requires reconstructing sessions from logs that identify illegal system access.

7.3. Prediction
In addition, logs have ability of producing predictive information. There are predictive analytic systems that utilize log data to assist with marketing plan development, advertising placement, and inventory management.

7.4. Reporting and Profiling
Furthermore, log analysis is required for analyzing resource usage, workload, and user activity. For instance, logs will capture the attributes of jobs inside a cluster's workloads to profile resources utilize within large data center.

8. SURVEY METHODOLOGY

Collect articles for this research have been done in a systematic manner comprehensive database for the research on automated log analysis was utilized. Relevant articles were identified in online digital libraries, and the repository was extended manually by evaluating the references to these articles. The libraries can now be accessed online. To begin, looked through a range of well-known online digital repositories (e.g., ACM Digital Library, Elsevier Online, ScienceDirect, IEEE Xplore, Springer Online, and Wiley Online). According to these studies, most prevalent uses of log files with ML algorithms is classified into numerous categories, on which we based our study:

8.1. Identify User Behavior

User activity analysis using logs may provide significant information about users. User clustering based on logs enables the gathering of clients considering their activity and subsequent analysis of user access patterns, making it an excellent option for problem solving [20].

Xu *et al.* [21] examine use of HTTP traffic to find the identities of users. Techniques presuppose access to a proxy server's log. Thus, it is likely to develop web use profiles for people who utilize devices with a static IP address. They demonstrated that given a web use profile, it is feasible to identify users on any other device or to monitor when another user uses a device. Technically, they divide web traffic across sessions that link to the traffic of a distinct IP address over a definite time period. They reduce every session to a frequency vector distributed over the vector space of accessible domain. They used a set of methods for instance-based user identification centered on this representation. Experiments showed that centered on gathered web usage profiles using Nearest Neighbor classification, user identification is achievable with a prediction accuracy of greater than 90%. This paper needs to examine the usage of more sophisticated identification and obfuscation methods integrating the time series of URLs more closely.

Kim *et al.*[22], based on user behavior modeling and anomaly detection techniques, the authors offered a framework for detecting insider threats throughout the user behavior modeling process. They constructed three datasets depending on the CERT database: Users daily action dataset, an e-mail content dataset, and an e-mail communication dataset depending on the user account and sending and receiving information. They proposed insider-threat identification models using those datasets, applying ML set anomaly detection methods to imitate real-world companies with just a few potentially harmful insiders' activities. In this work, the authors employed classification algorithms for insider-threat detection. The findings in this study recommend that the suggested framework is capable of detecting malicious insider behaviors relatively effectively. On the basis of the daily activity summaries dataset, the anomaly detection achieved a maximum detection 90% percent by monitoring top 30% of anomaly. According to the e-mail content datasets, the detection 65.64 % detected while 30% of sceptical e-mails have been monitored. The paper's limitation is that, although the dataset (CERT) used to building the system was carefully developed and contains a variety of threat scenarios, it stills an artificially and simulated produced dataset.

Prakash *et al.* [23] investigated for the scope of analyzing user prediction behavior based on users personalization obtained from web logs. A web log records the user's navigation patterns when visiting websites. The user navigation pattern could be analyzed using the user's recent weblog navigation. The weblog has several posts with data such as the status code, IP address, and amount of bytes sent, along with categories and a time stamp. User interests could be categorized according to categories and attributes, which aids in determining user behavior. The goal of this research is to differentiate between interested and uninterested user behavior through classification. The Modified Span Algorithm and the Personalization Algorithm are used to identify the user's interest. Table 1 provides a summary list of studies we reviewed for the purpose of identifying user behavior.

8.2. Security Issues

Recently, some researchers and programmers utilizing data mining methods to log-based Intrusion Detection Systems (IDS) resulted in a powerful anomaly detection-based (IDS) which depended solely on the inflowing stream of logs to discern what may be normal and what is not (possibly an attack) [24].

Zeufack *et al.* [25] offered a fully unsupervised framework for real-time detection of abnormalities. This concept is separated into two phases: A knowledge base development stage, that use clustering to identify common patterns, and A streaming anomaly detection stage that detects abnormal occurrences in real time. They test their framework on (Hadoop Distributed File System) log files and it successfully detects anomalies with an F-1 score of 83%. This framework ought to be improved to get advantages for other features that are embedding in a log file and has positive impact on anomalies detection.

The authors of [26] presented a Dempster-Shafer (D-S) evidence theory-based host security analysis technique. They acquire information of monitoring logs and use it to design security analysis model. They utilize three regression models as sensors for multi-source information fusion: Logistic regression, support vector regression, and K-nearest neighbor regression. The suggested technique offers excellent strong security for host. Improved ML approaches may increase accuracy of evidence in this research, resulting in more accurate probability values for host security analysis.

Study [27] a ML-based system for identifying insider threats in organizations' networked systems is provided. The research discussed four ML algorithms: Neural Networks (NN), Random Forest (RF), Logistic Regression (LR), and

XGBoost (XG) across multiple data granularities, limited ground truth, and training scenarios to assist cyber security analysts in detecting malicious insider behaviors in unseen data. Evaluation results showed that the proposed system can successfully learning from the limited training data and generalize to detect new users with malicious behaviors. The system has a great detection rate and precision, mainly when user-generated findings are considered. The downside: Will examine the utilization of temporal information in user activities. Specifically, all the systems in this research gave labels based on a single exemplar's state description. Allowing models to view many exemplars or to maintain state (recurrent connections) can allow models to make non-Markovian decisions.

Shah *et al.* [28] offered an expanded risk management strategy for Bring Your Own Device (BYOD) to increase the safety of the device environment. The proposed system makes usage mobile device management system, system logs, and risk management systems to detect malicious activities using machine learning. They can state that the result achieved 99% detection rate with the practice of Support Vector Machine algorithm.

Tadesse *et al.* [29] employed multilayer log analysis to discover assaults at several stages of the datacenter. Thus, identifying distinct assaults requires considering the heterogeneity of log entries as an initial point for analysis. The logs were integrated in a common format and examined based on characteristics. Clustering and Correlation are the root of the log analyzer in the center engine, which operate alongside the attack knowledge base to detect attacks. To calculate the quantity of clusters and filter events according on the filtering threshold, clustering methods for instance Expectation Maximization and K-means were utilized. On the furthermore, correlation establishes a connection or link between log events and provides new attack concepts. Then, they analyzed the developed system's log analyzer prototype and discovered that the average accuracy of SOM #34 and AAU is 84.37% and 90.01%, respectively. The downside: More log analysis or mining must be done to obtain meaningful information and minimize the false positive and negative results. Table 2 provides a summary list of studies we reviewed for the purpose of Security.

8.3. System Maintenance

Log analysis is typically required during system maintenance because to the intricacy of network structure.

Chen *et al.* [30] studied the issue of extracting useful patterns using temporal log data. They present a new algorithm Discovering Patterns from Temporal Sequences (DTS) algorithm for extracting sequential patterns from temporally regular sequence data. Engineers can utilize the patterns that find to well know how a network-based distributed system behaves. They apply the Minimum Description Length (MDL) concept to well-known issue of pattern implosion and take another step forward in summarizing the temporal links between neighboring events in a pattern. Tests on actual log datasets showed the method's effectiveness. Extensive tests on real-world datasets show that the suggested methodologies are capable of swiftly discovering high-quality patterns.

Cheng *et al.* [31] suggested a method for detecting anomalies using log file analysis. They extract normal patterns from log data and then do anomaly detection using Principal Component Analysis (PCA). Depending on the experimental results, they concluded that the proposed technique is a great success; this enables the technique to be devised and implemented to the real log file analysis, which makes the work of the system auditor easier. With a minimum of 66% and a high of 92.3%, the average accuracy in detecting anomalies is about 80%. Table 3 provides a summary list of studies we reviewed for the purpose of Maintenance.

9. CONCLUSION

Log files are records and track of computing events across different kinds of servers and systems. ML is a reliable solution for automatically analyzing log files. Log analysis as a ML application is a fast-emerging technique for extracting information from unstructured text log data files. This study analyzed several studies from various academic databases. They each utilized a different ML method for a different objective. We have summarized the importance, methodologies, and algorithms utilized for each element we have studied. Many of the recent publications provided models intended to forecast assaults, user behavior, and system failure to improve server and system security, marketing, and failure times. The disadvantage is that methods that discriminate between normal and abnormal data require a threshold. Selecting a correct threshold is challenging and involves prior knowledge; utilizing actual datasets in model creation; many log analyses or mining must be performed to gain significant information; and minimizing false positive and negative findings. Furthermore, due to the lack of studies on, the maintenance component requires further improvements compared to the other specified features; nonetheless, interested scholars can study it further.

REFERENCES

[1] E. Shirzad and H. Saadatfar. "Job failure prediction in hadoop based on log file analysis". *International Journal of Computers and Applications*, vol. 44, no. 3, pp. 260-269, 2022.

[2] A. U. Memon, J. R. Cordy and T. Dean. "*Log File Categorization and Anomaly Analysis Using Grammar Inference*". Queen's University, Canada, 2008.

[3] M. Siwach and S. Mann. "Anomaly detection for web log data analysis: A review". *Journal of Algebraic Statistics,* vol. 13, no. 1, pp. 129-148, 2022.

[4] H. S. Malallah, S. R. Zeebaree, R. R. Zebari, M. A. Sadeeq, Z. S. Ageed, I. M. Ibrahim, H. M. Yasin and K. J. Merceedi. "A comprehensive study of kernel (issues and concepts) in different operating systems". *Asian Journal of Research in Computer Science*, vol. 8, no. 3, pp.16-31, 2021.

[5] I. Mavridis, I and H. Karatza. "Performance evaluation of cloud-based log file analysis with apache hadoop and apache spark". *Journal of Systems and Software*, vol. 125, pp. 133-151, 2017.

[6] T. Yang and V. Agrawal. "Log file anomaly detection". *CS224d Fall*, vol. 2016, pp. 1-7, 2016.

[7] S. Khan, A. Gani, A. W. A. Wahab, M. A. Bagiwa, M. Shiraz, S. U. Khan, R. Buyya and R. Y. Zomaya. "Cloud log forensics: Foundations, state of the art, and future directions". *ACM Computing Surveys (CSUR)*, vol. 49, no. 1, pp. 1-42, 2016.

[8] V. Chitraa and A. S. Davamani. "A survey on preprocessing methods for web usage data". *International Journal of Computer Science and Information Security*, Vol. 7, no. 3, p. 1257. 2010.

[9] R. A. Bridges, T. R. Glass-Vanderlan, M. D. Iannacone, M. S. Vincent and Q. Chen. "A survey of intrusion detection systems leveraging host data". *ACM Computing Surveys (CSUR)*, vol. 52, no. 6, pp. 1-35, 2019.

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

[11] Available from: https://www.humio.com/glossary/log-file [Last accessed on 2022 Sep 01].

[12] S. He, P. He, Z. Chen, T. Yang, Y. Su and M. R. Lyu. "A survey on automated log analysis for reliability engineering". *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1-37, 2020.

[13] M. Kumar, M. Meenu. "Analysis of visitor's behavior from web log using web log expert tool". In 2017 *International conference of Electronics, Communication and Aerospace Technology (ICECA)*. vol. 2, Institute of Electrical and Electronics Engineers, Manhattan, New York, pp. 296-301, 2017.

[14] W. Li. "*Automatic Log Analysis Using Machine Learning: Awesome Automatic Log Analysis Version 2.0*". 2013.

[15] N. Singh, A. Jain and R. S. Raw. "Comparison analysis of web usage mining using pattern recognition techniques". *International Journal of Data Mining and Knowledge Management Process*, Vol. 3, no. 4, p. 137, 2013.

[16] M. A. Latib, S. A. Ismail, O. M. Yusop, P. Magalingam and A. Azmi. "Analysing log files for web intrusion investigation using hadoop". In: *Proceedings of the 7th International Conference on Software and Information Engineering*, pp. 12-21, 2018.

[17] J. Qiu, Q. Wu, G. Ding, Y. Xu and S. Feng. "A survey of machine learning for big data processing". *EURASIP Journal on Advances in Signal Processing*, Vol. 2016, no. 1, pp. 1-16, 2016.

[18] N. Jones. "Computer science: The learning machines". *Nature*, vol. 505, no. 7482, pp. 146-148, 2014.

[19] M. A. Latib, S. A. Ismail, H. M. Sarkan and R. C. Yusoff. "Analyzıng log ın bıg data envıronment: A revıew". *ARPN Journal of Engineering and Applied Sciences*, vol. 10, no. 23, pp. 17777-17784, 2015.

[20] H. Xiang. "Research on clustering algorithm based on web log mining". *Journal of Physics Conf Series*, vol. 1607, no. 1, p. 012102, 2020.

[21] J. Xu, F. Xu, F. Ma, L. Zhou, S. Jiang and Z. Rao. "Mining web usage profiles from proxy logs: User identification". In: 2021 *IEEE Conference on Dependable and Secure Computing (DSC)*. Institute of Electrical and Electronics Engineers, Manhattan, New York, pp. pp. 1-6, 2021.

[22] J. Kim, M. Park, H. Kim, S. Cho and P. Kang. "Insider threat detection based on user behavior modeling and anomaly detection algorithms". *Applied Sciences*, vol. 9, no. 19, p. 4018, 2019.

[23] P. G. Prakash and A. Jaya. "Analyzing and predicting user navigation pattern from weblogs using modified classification algorithm". *Indonesian Journal of Electrical Engineering and Computer*, vol. 11, no. 1, pp.333-340, 2018.

[24] A. Abbas, M. A. Khan, S. Latif, M. Ajaz, A. A. Shah and J. Ahmad. "A new ensemble-based intrusion detection system for internet of things". *Arabian Journal for Science and Engineering*, vol. 47, no. 2, pp. 1805-1819, 2022.

[25] V. Zeufack, D. Kim, D. Seo and A. Lee. "An unsupervised anomaly detection framework for detecting anomalies in real time through network system's log files analysis". *High Confidence Computing*, vol. 1, no. 2, pp. 100030, 2021.

[26] Y. Li, S. Yao, R. Zhang and C. Yang. "Analyzing host security using D-S evidence theory and multisource information fusion". *International Journal of Intelligent Systems*, vol. 36, no. 2, pp. 1053-1068, 2021.

[27] D. C. Le, A. N, Zincir-Heywood and M. I. Heywood. "Analyzing data granularity levels for insider threat detection using machine learning". *IEEE Transactions on Network and Service Management*, vol. 17, no. 1, pp. 30-44, 2020.

[28] N. Shah and A. Shankarappa. "Intelligent risk management framework for BYOD". In: 2018 *IEEE 15th International Conference on e-Business Engineering (ICEBE)*. Institute of Electrical and Electronics Engineers, Manhattan, New York, pp. 289-293, 2018.

[29] S. G Tadesse and D. E Dedefa. "Layer based log analysis for enhancing security of enterprise datacenter". *International Journal of Computer Science and Information Security*, vol. 14, no. 7, pp.158, 2016.

[30] J Chen, P Wang, S Du and W Wang. "Log pattern mining for distributed system maintenance". *Complexity*, vol. 2020, no. 2, pp. 1-12, 2020.

[31] X. Cheng and R. Wang. "Communication network anomaly detection based on log file analysis". In: *International Conference on Rough Sets and Knowledge Technology*. Springer, Cham, pp. 240-248, 2014.