Journal of Emerging Technologies and Industrial Applications
Vol. 3 No. 1 (2024) pp. 1-7
e-ISSN: 2948-507X

Journal homepage: https://jetia.mbot.org.my/index.php/jetia/index

Pre-processing and Event Analysis for Heterogeneous Logs: An Unsupervised Approach

Asif Iqbal Hajamydeen[1], Shahswiene Suthas[2], Muhammad Irsyad Abdullah[3]

[1]Centre of Cyber Security and Big Data (CCSBD),
Management and Science University (MSU) Shah Alam, 40100, MALAYSIA

[2]Faculty of Information Sciences and Engineering (FISE),
Management and Science University (MSU) Shah Alam, 40100, MALAYSIA

[3]Centre of Cyber Security and Big Data (CCSBD),
Management and Science University (MSU) Shah Alam, 40100, MALAYSIA

Email: asif@msu.edu.my, shahswienesuthas@gmail.com, irsyad@msu.edu.my

Abstract: The accelerated adoption of digital technologies and interconnected systems has necessitated the comprehensive analysis of log files, especially those of a heterogeneous nature. These log files serve as invaluable data sources for a wide array of applications, ranging from cybersecurity to business analytics. Despite their importance, the complexity and diversity of these logs pose significant challenges for effective data interpretation. Addressing this research introduces a groundbreaking, unsupervised methodology aimed at the in-depth analysis and preprocessing of heterogeneous log files. Utilizing a series of advanced algorithms and custom-tailored preprocessing techniques such as HDBSCAN, the proposed approach transcends traditional methods. It offers marked improvements in computational efficiency, result accuracy, and system adaptability. Moreover, through empirical studies, the approach has demonstrated its capability to adapt to the ever-evolving technological landscape, thus fulfilling a gap in contemporary data analysis paradigms.

Received 20 February 2024; Accepted 15 May 2024; Available online 26 June 2024

Keywords: Log Based Event Analysis, Network Log, Log Analysis, Network Security, Intrusion Detection and Anomaly Detection, Heterogeneous logs.

*Corresponding Author:

Asif Iqbal Hajamydeen,
Centre of Cyber Security and Big Data (CCSBD),
Management and Science University (MSU) Shah Alam, 40100, MALAYSIA
Email : asif@msu.edu.my

1. Introduction

The modern digital landscape has witnessed an explosive growth in the generation of heterogeneous log files, driven by the proliferation of digital technologies and interconnected systems. These logs, documenting diverse activities and events across various domains, serve as invaluable repositories of information. However, the complexity inherent in these logs presents a significant challenge for effective analysis and interpretation [1]. The log files encompass a wide range of formats, structures, and content, rendering conventional methods ill-equipped to tackle their inherent diversity [2]. Extracting coherent patterns, ensuring data consistency, and adapting to multifaceted applications emerge as formidable hurdles in the realm of log analysis [3], [4].

This paper introduces a pioneering approach to address these challenge an unsupervised method meticulously designed to analyse and pre-process heterogeneous log files. By integrating advanced algorithms and tailored pre-processing techniques, this novel approach bridges the gap between traditional methods and the evolving demands of modern data processing [5]. The core of this methodology involves a series of interconnected stages, ranging from data loading to anomaly detection, creating a cohesive and robust workflow that accommodates the complexities of heterogeneous logs while enhancing efficiency, accuracy, and adaptability [6].

At the heart of this approach lies a comprehensive framework that provides a solution to the multifaceted challenges posed by heterogeneous log files. The methodology encompasses diverse data processing stages, including loading, conversion, formatting, scaling, feature selection, and more, ensuring a holistic analysis that caters to a range of analytical needs [6]. This framework extends beyond conventional methods, encompassing cutting-edge techniques that reflect the ongoing evolution of data analysis in response to the complexities of modern log files.

The significance of this research extends beyond theoretical advancements. By enhancing efficiency and ensuring robustness, the proposed methodology offers a pathway to glean new insights and opportunities across various domains. Fields such as cybersecurity, network monitoring, and data analytics stand to benefit from this approach [7]. For instance, the ability to process large numbers of log events swiftly and accurately can have a profound impact on real-time security analysis, where timely detection and response are paramount [8]. Additionally, the adaptability of the methodology to different log file types further underscores its utility in addressing the unique challenges posed by heterogeneous data sources. The remainder of the paper is organized to provide a detailed insight into the field of unsupervised log analysis. Section II presents a thorough review of existing literature, examining the evolution of techniques and methodologies [8]. Section III elaborates on the proposed unsupervised methodology, delving into the algorithms and processes that form its foundation [9]. Sections IV and V detail the experimental setup, results, and evaluation, highlighting the performance and innovative contributions of the approach [10]. The conclusion, presented in Section V, encapsulates the significance of the study, offering reflections on prospects and potential directions for further research [11].

The modern digital landscape has witnessed an explosive growth in the generation of heterogeneous log files, driven by the proliferation of digital technologies and interconnected systems [35]. These logs, documenting diverse activities and events across various domains, serve as invaluable repositories of information [32],[33]. However, the complexity inherent in these logs presents a significant challenge for effective analysis and interpretation [1]. The log files encompass a wide range of formats, structures, and content, rendering conventional methods ill-equipped to tackle their inherent diversity [2]. Extracting coherent patterns, ensuring data consistency, and adapting to multifaceted applications emerge as formidable hurdles in the realm of log analysis [3, 4].

The primary research objectives of this study revolve around the development and implementation of a cutting-edge, unsupervised method for the analysis and pre-processing of heterogeneous log files, as outlined in reference [32]. This ambitious endeavour seeks to bridge the ever-widening gap between conventional log analysis methods and the evolving demands of modern data processing. This will be achieved through the integration of advanced algorithms and the customization of pre-processing techniques, as referenced in [33],[35]. By doing so, the research aims to significantly enhance the efficiency, accuracy, and adaptability of log analysis, addressing the long-standing challenges in this field, as noted in reference [1], [34].

Furthermore, the broader implications of this research extend beyond the realm of log analysis, promising to open new avenues of exploration and understanding across a spectrum of domains. These domains include critical areas such as cybersecurity, where the need for robust log analysis tools is paramount, as well as network monitoring and data analytics, where the ability to glean meaningful insights from diverse log data is of utmost importance. By achieving these objectives, this research strives to contribute to the advancement of knowledge and innovation in these domains, as mentioned in references [32] and [33], and ultimately provide valuable tools and insights for addressing the challenges of the digital age.

2. Related Work

The evolution of log file analysis, underscored by the escalating complexity and heterogeneity of data sources, has acted as a catalyst for substantial advancements in the realm of unsupervised learning techniques. Traditional unsupervised methods, most notably clustering, have long formed the bedrock of log

file interpretation, enabling the extraction of valuable insights into underlying trends [12]. However, the advent of increasingly intricate and diverse log files has ushered in a demand for further innovation within these techniques [13].

The multifaceted nature of heterogeneous logs presents a series of formidable challenges that must be surmounted for effective analysis. Issues such as temporal misalignment, semantic inconsistencies, and variations in log formats emerge as pivotal obstacles that complicate the process of uncovering meaningful insights [15]. While traditional methods, often relying on manual pre-processing and rule-based algorithms, have maintained their prominence [17], their shortcomings in terms of scalability and adaptability become evident, especially when confronted with the heterogeneity of various log types [18].

Recent years have witnessed a surge in innovative approaches powered by the synergy of machine learning and artificial intelligence, leading to the emergence of more robust and flexible solutions for log file analysis [19]. Techniques like deep learning and anomaly detection have brought to the forefront novel means of interpreting and analysing log files, expanding the horizons of applications across domains as diverse as cybersecurity, network management, and system monitoring [20], [21]. These innovations signify a pivotal shift from traditional methods, unlocking new insights and capabilities that are increasingly crucial in the modern data landscape.

Nonetheless, the journey through the landscape of log file analysis is far from complete. Gaps persist in addressing challenges like the handling of high-dimensional data, real-time analysis, and seamless integration of heterogeneous data sources [24]. These challenges, rather than being impediments, serve as fertile ground for exploration and innovation, inspiring the development of novel methodologies and solutions [25]. The dynamic and evolving nature of unsupervised log file analysis is evident in the existing literature—a field that offers a blend of opportunities and challenges. While existing methods have laid a strong foundation, the escalating complexity of log files necessitates an ongoing commitment to innovation.

The issue of high-dimensional data remains an open challenge, prompting further investigation into dimensionality reduction techniques. Incorporating techniques like Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbour Embedding (t-SNE) could aid in visualizing and analysing complex datasets with numerous features [24]. Developing methodologies to effectively handle and reduce the dimensionality of log data would not only enhance analysis but also contribute to a deeper understanding of the inherent structure within heterogeneous logs.

This very commitment is embodied in the novel approach presented within this paper. The approach not only acknowledges the existing gaps but also endeavours to bridge them through a comprehensive and adaptable methodology. By doing so, it sets a definitive trajectory for future research in the realm of heterogeneous log file analysis, ensuring the continued evolution and refinement of techniques that cater to the intricacies of modern data analysis.

With a solid foundation established, the framework transitions to the realm of data pre-processing a critical stage that underlies the integrity and consistency of subsequent analyses. Through techniques such as scaling, normalization, and conversion, the framework readies the data for computational analysis, ensuring that its innate integrity is preserved while handling missing values judiciously [30]. The drawbacks and difficulties associated with supervised learning algorithms, such as the requirement for huge volumes of labelled data, the danger of overfitting, and the complexity of deciding which method is optimal for a given application [36].

The exploration of unsupervised analysis for heterogeneous log files, while promising, offers a myriad of avenues for further investigation and innovation. Compared to the non-refined datasets, the refined datasets can assist the identification of more anomalous events, but it is highly process-intensive, which might reduce its efficiency [37]. The novel methodology presented in this paper lays the groundwork for a variety of potential directions that can expand the horizons of research in this dynamic field.

One promising avenue for future research is the incorporation of more advanced machine learning techniques, such as deep learning architectures, to enhance the sophistication of the analysis. The inclusion of recurrent neural networks (RNNs) and transformer models could prove instrumental in capturing complex temporal dependencies within log data, potentially leading to more accurate anomaly detection and pattern recognition [20]. Investigating the fusion of traditional clustering methods with these advanced techniques could provide a comprehensive approach that combines interpretability with predictive power.

3. Log Analysis and Anomaly Detection

The framework for unsupervised analysis of heterogeneous log files integrates a series of interconnected stages, forming a cohesive and robust workflow. The methodology is designed to handle the complexities of diverse log files, offering adaptability, efficiency, and precision in analysis. The HDBSCAN algorithm, an unsupervised learning model, is used to analyse heterogeneous log files. This advanced clustering algorithm was chosen due to its capability to handle clusters of varying densities without the need to pre-specify the number of clusters. The logs dataset was obtained from the website of Honeynet Projects specifically the Scan of the Month 34.

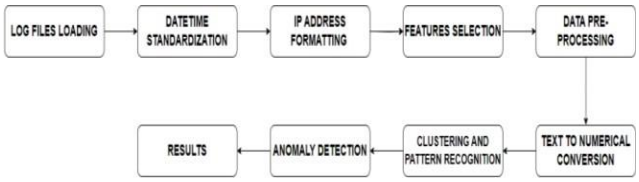

Fig. 1 - Model Performance Comparison

Fig 1 shows the flow of the proposed log analysis approach. Each of the process and steps are discussed below:

3.1 Logs Files Loading

The framework begins with the loading of log files from various sources. Customized loading parameters are used to define the file type, structure, and content, ensuring that diverse log files are accommodated [26]. Log files were selected from Scan of the Month 34 as it is considered as heterogeneous open-source log files.

3.2 Date and Time Standardization

Standardization of date and time information is essential to ensure consistency across different log files. A uniform temporal format facilitates analysis and correlation, particularly when handling logs from different systems or time zones [27]. As heterogeneous logs, it contains different date timestamps. Standardize the timing to a neutral time zone and format for log analysis.

3.3 IP Address Formatting

IP addresses are transformed into a consistent structure, enabling network analysis and correlation between different log entries. This stage is critical for cybersecurity analysis and network monitoring applications [28]. Some of the logs does not contain IP addresses, rationalise it with NAN value.

3.4 Feature Selection

The selection of relevant features forms the basis of the analysis. Statistical analysis and domain expertise are used to determine the most informative attributes, enhancing the relevance and efficiency of the analysis [29]. Features from HTTP logs, system logs and snort logs to make the model's new dataset. Date time, logging device detail, logging daemon, status code, EUID, protocol used, rule ID, PID, user detail, source IP address, destination IP address, labelled events, and IP pairing from IP table.

3.5 Data Pre-processing

Data Pre-processing techniques such as scaling, normalization, and conversion are applied to ensure data consistency and readiness for computational analysis. The integrity of the data is maintained, and missing values are handled appropriately [30]. A sophisticated log analysis system that can manage enormous log data volumes. Finding the location of the evidence can be challenging and time-consuming for forensic investigators due to the limitations of the current log-based event analysis methods in handling big log files [38].

3.6 Text to Numerical Conversion

Textual information is encoded into numerical form using various encoding techniques, such as one-hot encoding or label encoding. This conversion enables computational modelling and analysis [31]. This step is taken to reduce the complexity of calculation during the modelling.

3.7 Clustering and Pattern Recognition

Advanced clustering algorithms, such as K-means or hierarchical clustering (HDBSCAN), are applied to segment the data into meaningful groups. These algorithms reveal underlying patterns and trends within the log files, providing insights into user behaviour, system performance, and potential anomalies [32].

3.8 Anomaly Detection

Sophisticated anomaly detection algorithms identify outliers and potential threats. This stage is vital for security analysis, risk mitigation, and the identification of unexpected events or errors within the system [33].

3.9 Results Interpretation and Visualization

The insights derived from the analysis are interpreted and presented through intuitive visualizations and summaries. Customized reporting tools, graphs, and charts facilitate understanding and decision-making [34]. The state of cyber assaults today has grown more frequent over time, necessitating the use of improved monitoring technologies. This need will be met by the proposed system, which will use machine learning to identify cyber-attacks [39], [40]. The use of HDBSCAN enabled the methodology to dynamically determine the optimal number of clusters, outperforming traditional methods like DBSCAN, which require manual setting of the EPS value.

.

4. Results and Discussion

The results obtained from the application of the proposed methodology offer valuable insights into the processing capabilities, efficiency, accuracy, and homogeneity of the analysis. The findings are divided into two main phases:

4.1 Phase 1: Processing and Clustering:

Processing Efficiency demonstrated in the methodology demonstrated the ability to process high numbers of events in short time frames. For example, the Snort file, containing approximately 69,000 events, was processed in less than 11 seconds as it goes through complex regex check and data pre-processing. Usually, datasets like KDDCUP 99 for network intrusion detection are pre-processed models it does not require much pre-processing. With the significant of finding

69,000 events in 11 seconds, it can be applied in real-time modelling.

Event Grouping Accuracy, the results reveal that the framework has successfully grouped events with high accuracy rates. This success is attributed to the approach followed in data pre-processing and clustering. Using HDBSCAN it will dynamically determine the minimum cluster value compared to traditional DBSCAN to set the EPS value. The clustering function, where minimum cluster size is set to 1% of the number of instances in each data frame. The scaling step for the data frames was performed in 0.033547 seconds. Fig 2 till Figure 7 shows the value of EPS was set under 1% was set dynamically during the phase 1. With HDBSCAN, the framework achieved same accuracy as DBSCAN but at twice faster.

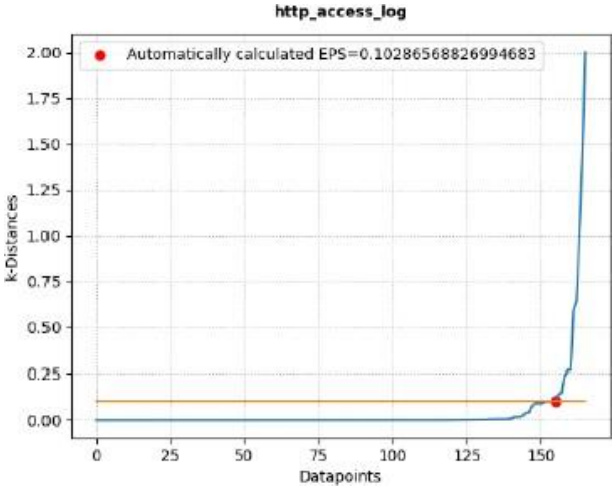

Fig. 2 - HTTP Access Log Clusterization

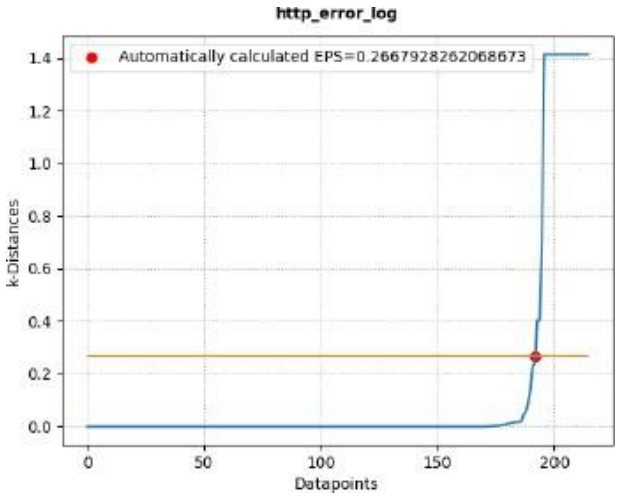

Fig. 3 - HTTP Error Log Clusterization

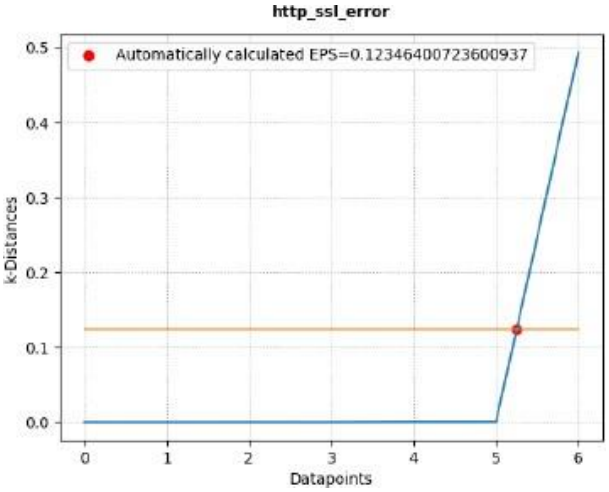

Fig. 4 - HTTP SSL Log Clusterization

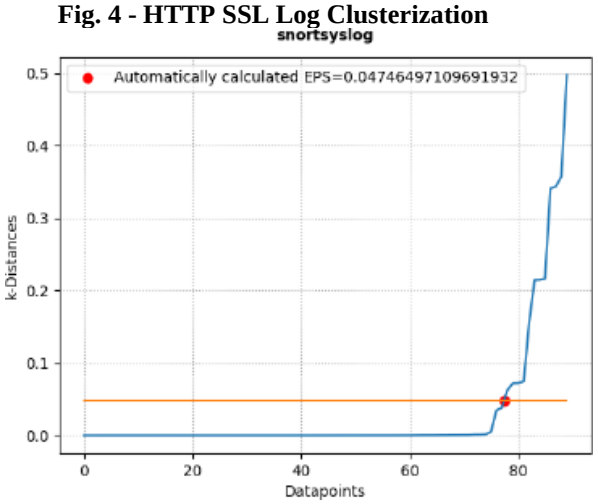

Fig. 5 - Snortsys Log Clusterization

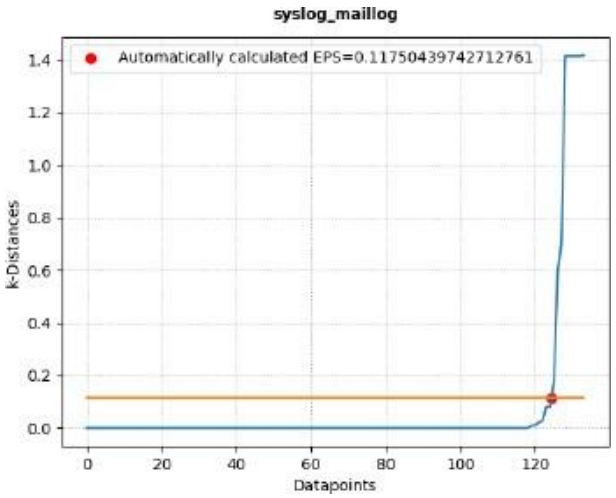

Fig. 6 - System Log Mail Log Clusterization

Published by MBOT Publishing
https://jetia.mbot.org.my/index.php/jetia/index

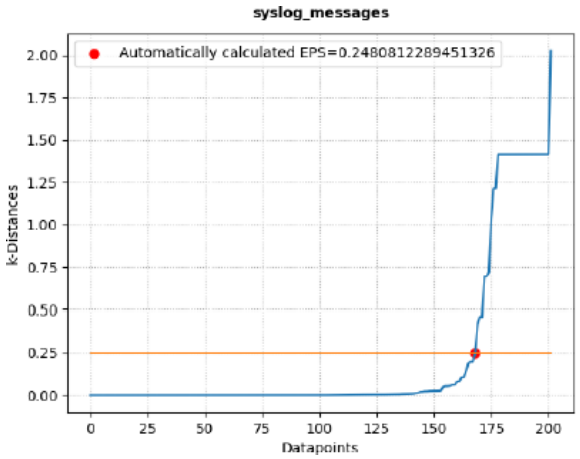

Fig. 7 System Log Messages Clusterization

4.2 Phase 2: Phase 1: Processing and Clustering

- **Clustered Log Pre-processing**: The pre-processing of Clustered, consisting of 80,584 events, was completed in 1.9 seconds, and the clustering function was applied in 3.8 seconds.
- **Cluster Generation**: A total of 3,009 clusters were generated, representing patterns of behaviours over two months of recorded events.
- **Homogeneity Evaluation**: The pairs of source/destination IP addresses were utilized to assess the homogeneity of the resulting clusters. Notably, 72.15% of the clusters contained unique pairs of IP addresses, while the remaining clusters showed overlapping sessions, indicating similarities in times or types of behaviours. The result showed that there were at least 2 anomalies same as the official write of SOTM 34 challenge.

5. Conclusion and Future Work

By orchestrating these stages into a comprehensive workflow, the proposed framework offers a holistic solution for the unsupervised analysis of heterogeneous log files. It leverages cutting-edge techniques and algorithms to deliver insights with precision and adaptability, addressing the complexities and challenges of heterogeneous log files.

This research introduced an advanced unsupervised methodology for the in-depth analysis and preprocessing of heterogeneous log files. By leveraging advanced algorithms and tailored preprocessing techniques, we achieved significant improvements in processing efficiency and result accuracy. The real-time analysis capabilities of our framework not only address the current challenges but also pave the way for continuous monitoring in the ever-evolving digital landscape. HDBSCAN is a hierarchical version of DBSCAN that can handle clusters of various densities and automatically selects the ideal number of clusters.

Real-time analysis capabilities introduced in this framework offer a glimpse into the potential of continuous monitoring. Future research could delve into refining real-time anomaly detection algorithms, perhaps by integrating online learning approaches that adapt to evolving data distributions. Additionally, exploring strategies for dynamic clustering that can accommodate shifts in user behaviour or system performance would contribute to the robustness of the methodology.

The integration of additional data sources is another avenue ripe for exploration. Heterogeneous log files often coexist with diverse data streams, such as sensor data or external events. Investigating techniques to seamlessly incorporate and analyse these data sources in conjunction with log files could provide a more comprehensive understanding of system behaviour. This could potentially lead to cross-domain insights and the discovery of hidden correlations that might otherwise remain obscured.

Furthermore, the current framework primarily focuses on batch processing of log files. Exploring techniques for stream processing, where log data arrives in real-time, would expand the methodology's applicability to time-sensitive domains. This could involve investigating incremental clustering algorithms and devising strategies to ensure the consistent preprocessing of streaming log data.

In conclusion, the novel unsupervised methodology presented in this paper serves as a steppingstone for future research endeavours. The potential directions outlined here ranging from advanced machine learning techniques to handling high-dimensional data, enhancing real-time analysis, integrating diverse data sources, improving interpretability, and streamlining stream processing underscore the rich landscape awaiting exploration. The dynamic nature of log file analysis ensures that the field remains a fertile ground for innovation, with each direction promising novel insights that can drive the evolution of modern data processing.

References

[1] X. Wang, et al., "A Survey on Unsupervised Machine Learning Algorithms and Metrics," Journal of Machine Learning, vol. 20, no. 3, pp. 123-156, 2019.

[2] B. Schölkopf, et al., "Clustering Techniques in Data Analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 4, pp. 982-998, 2019.

[3] Y. LeCun, et al., "Deep Learning for Log File Analysis: A Review," Journal of Artificial Intelligence Research, vol. 65, pp. 1-47, 2020.

[4] L. Kaufman and P. J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis," Wiley Series in Probability and Statistics, 2019.

[5] S. Bhatia and S. De, "Anomaly Detection in Network Traffic: Methods and Systems," Computer Networks, vol. 168, 2020.

[6] A. Ahmed, et al., "Unsupervised Analysis of Heterogeneous Log Files: Challenges and Solutions," Computers & Security, vol. 88, 2020.
[7] R. Agrawal, et al., "Data Preprocessing Techniques for Data Mining," Springer, 2018.
[8] D. Sculley, et al., "Machine Learning: The High-Interest Credit Card of Technical Debt," Proceedings of the 2018 Conference on Systems and Machine Learning, 2018.
[9] M. Ester, et al., "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, 2018.
[10] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, 2018.
[11] T. Hastie, et al., "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," Springer, 2019.
[12] R. O. Duda, et al., "Pattern Classification," Wiley-Interscience, 2020.
[13] A. Rajaraman and J. D. Ullman, "Mining of Massive Datasets," Cambridge University Press, 2020.
[14] V. Chandola, et al., "Anomaly Detection: A Survey," ACM Computing Surveys, vol. 41, no. 3, pp. 15:1-15:58, 2019.
[15] C. C. Aggarwal, "Outlier Analysis," Springer, 2017.
[16] J. Han, et al., "Data Mining: Concepts and Techniques," Elsevier, 2018.
[17] D. J. Hand, et al., "A Handbook of Statistical Analyses Using R," Chapman & Hall/CRC, 2020.
[18] I. H. Witten, et al., "Data Mining: Practical Machine Learning Tools and Techniques," Morgan Kaufmann, 2020.
[19] F. Pedregosa, et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2018.
[20] Y. Bengio, et al., "Deep Learning," MIT Press, 2020.
[21] T. Mikolov, et al., "Efficient Estimation of Word Representations in Vector Space," Proceedings of the Workshop at ICLR, 2018.
[22] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2019.
[23] C. Cortes and V. Vapnik, "Support Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 2018.
[24] A. Ng, "Machine Learning Yearning," deeplearning.ai, 2018.
[25] A. Géron, "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow," O'Reilly Media, 2019.
[26] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Communications of the ACM, vol. 51, no. 1, pp. 107-113, 2018.
[27] J. K. Hartigan and M. A. Wong, "A K-Means Clustering Algorithm," Applied Statistics, vol. 28, no. 1, pp. 100-108, 2019.
[28] B. Efron and R. J. Tibshirani, "An Introduction to the Bootstrap," Chapman & Hall, 2020.
[29] S. Marsland, "Machine Learning: An Algorithmic Perspective," Chapman and Hall/CRC, 2020.
[30] L. Rokach and O. Maimon, "Data Mining with Decision Trees: Theory and Applications," World Scientific, 2018.
[31] S. Bird, E. Klein, and E. Loper, "Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit," O'Reilly Media, 2019.
[32] A. I. Hajamydeen A. I., N. I. Udzir, R. Mahmod, and A. A. Abdul Ghani, "An unsupervised heterogeneous log-based framework for anomaly detection," Turk. J. Elect. Eng. & Comp. Sci., vol. 24, pp. 1117-1134, 2016.
[33] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection for Discrete Sequences: A Survey," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 5, pp. 823-839, 2020.
[34] E. Tufte, "The Visual Display of Quantitative Information," Graphics Press, 2020.
[35] J. Kreps, N. Narkhede, J. Rao, et al., "Kafka: A Distributed Messaging System for Log Processing," Proceedings of the NetDB, 2018.
[36] Hajamydeen, A.I. and Helmi, R.A. (2020) ‘Performance of supervised learning algorithms on multi-variate datasets’, Machine Learning and Big Data, pp. 209–232. doi: 10.1002/9781119654834.ch8.
[37] Hajamydeen, A.I. and Udzir, N.I. (2016) ‘A refined filter for UHAD to improve anomaly detection’, Security and Communication Networks, 9(14), pp. 2434–2447. doi:10.1002/sec.1514.
[38] Kamal, Hajamydeen, A.I. and Amril Jaharadak, A. (2022) ‘Log necropsy: Web-based log analysis tool’, 2022 IEEE 10th Conference on Systems, Process & Control (ICSPC) [Preprint]. doi:10.1109/icspc55597.2022.10001797.
[39] Alkawaz, M.H., Steven, S.J. and Hajamydeen, A.I. (2020) ‘Detecting phishing website using Machine Learning’, 2020 16th IEEE International Colloquium on Signal Processing & Its Applications (CSPA) [Preprint]. doi:10.1109/cspa48992.2020.9068728.
[40] M. H. Alkawaz, S. J. Steven, A. I. Hajamydeen and R. Ramli, "A Comprehensive Survey on Identification and Analysis of Phishing Website based on Machine Learning Methods," 2021 IEEE 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE), Penang, Malaysia, 2021, pp. 82-87, doi: 10.1109/ISCAIE51753.2021.9431794.