Data Mining with Big Data in Intrusion Detection Systems: A Systematic Literature Review

Fadi Salo*, MohammadNoor Injadat†, Ali Bou Nassif‡, Aleksander Essex*

*Department of Electrical and Computer Engineering, The University of Western Ontario, London, ON, Canada
Email: {faldo, minjadat, aessex}@uwo.ca
†Department of Electrical and Computer Engineering, University of Sharjah, Sharjah, UAE
Email: anassif@sharjah.ac.ae

Abstract—Cloud computing has become a powerful and indispensable technology for complex, high performance and scalable computation. The exponential expansion in the deployment of cloud technology has produced a massive amount of data from a variety of applications, resources and platforms. In turn, the rapid rate and volume of data creation has begun to pose significant challenges for data management and security. The design and deployment of intrusion detection systems (IDS) in the big data setting has, therefore, become a topic of importance. In this paper, we conduct a systematic literature review (SLR) of data mining techniques (DMT) used in IDS-based solutions through the period 2013-2018. We employed criterion-based, purposive sampling identifying 32 articles, which constitute the primary source of the present survey. After a careful investigation of these articles, we identified 17 separate DMTs deployed in an IDS context. This paper also presents the merits and disadvantages of the various works of current research that implemented DMTs and distributed streaming frameworks (DSF) to detect and/or prevent malicious attacks in a big data environment.

Index Terms—Intrusion Detection System, Real-Time Detection, Data Mining, Big Data, Cloud Computing, Network Security.

I. INTRODUCTION

Enterprises, like government, finance, industry, and health care, need strong, reliable cybersecurity solutions [1]. Undoubtedly, the challenge has increased dramatically with the big data era from the current implemented solutions, such as traditional intrusion detection systems (IDS) [2]. The substantial recent growth of technological capabilities paves the way to cyber threats that target individuals or enterprises with various malicious attacks. In this threat environment, traditional security protections like firewalls, anti-virus software, and virtual private networks (VPNs) are not always sufficient, and robust mechanisms for detecting intrusions must also be deployed. For that reason, deploying an IDS along with the other traditional security systems is an essential strategy [3]. An IDS is automated software used to analyze the network traffic to detect and/or prevent malicious attacks [4]. It can be classified by two common methods namely, misuse-based detection (MD) and anomaly-based detection (AD). In the first approach, classification of the attack is based specifically on known patterns, called signatures [5]. AD, on the other side, seeks to detect abnormal behavior or signatures with the key challenge of distinguishing between normal and abnormal patterns with low error [5]. There are different studies discussing the challenges facing IDS especially with the rise of big data era, which brings up different types of malicious attacks. Accordingly, special tools are required to accommodate the scalability of big data to detect cyber threats. The notion of big data is an umbrella term that encompasses the recent trends toward collection, storage and analysis of enormous volumes of data. Popular big data management tools include Hadoop, Spark, and Shark.

To the best of our knowledge, no prior systematic literature review addresses research questions (RQs) about implemented DMTs in the context of IDS, which has motivated this work. In related work, Ashraf and Habaebi [6] discussed the threat approaches in the Internet of Things using an autonomic taxonomy. Ali et al. [7] summarized the security challenges and vulnerabilities in the mobile cloud computing. Ahmed et al. [8] presented the major categories of anomaly detection including statistical, classification, and clustering. They also discussed the dataset challenges of network IDS. Zuech et al. [9] reviewed the problem of big heterogeneous data. Injadat et al. [10] summarized the data mining techniques implemented in social media. Alguliyev and Imamverdiyev [11] analyzed the big data applications related to information security challenges and illustrated the research directions at the same domain. In this study, we conduct a systematic literature review (SLR) following the framework of Kitchenham and Charters, [12] exploring research into DMTs in the big data setting for IDS. The study covers the period between Jan 2013 and Dec 2018, evaluates, and analyzes research conducted in this space. This paper is divided into four sections. Section II describes our review methodology. Section III presents the results and discussion. Section IV includes the conclusion and recommendations for future work.

II. METHODOLOGY

The systematic methodology of this review consists of three stages [12]: planning, conducting, and reporting, with each stage consisting of several sub-stages. Detailed information of this methodology is presented in the following subsections. Fig. 1 demonstrates this methodology.

A. Research Questions (RQs)

The key element of this study is to analyze and report the empirical research about the DMT with big data in the

Fig. 1 demonstrates this methodology.
field of IDS during the desired period. Towards this goal, four RQs were identified: RQ1: Which data mining techniques with big data have been used in IDS? RQ2: What are the strengths and weaknesses of the DMT and DSF employed in the literature? RQ3: What are the evaluation metrics that have been implemented to validate the DMT in IDS? and RQ4: What type of attacks does the IDS in the literature designed for? What are the characteristics of the utilized datasets used to achieve this goal?

B. Study Selection

This research requires comparative investigation across the relevant articles. The digital libraries used in our review included: IEEE Xplore, ACM, Web of Science, Elsevier, Springer, and Wiley. Our initial search returned 545 articles. Tangential and duplicate articles were then identified during a subsequent filtration step. Next, we applied our inclusion and exclusion criteria to exclude the irrelevant articles. Next, we applied our Quality Assessment Rules (QARs) as a benchmark to evaluate the quality of the candidate articles. The inclusion criteria consist of the following rules: 1) Include articles that use data mining techniques with big data in IDS. 2) Include comparative studies of data mining techniques with big data in IDS. 3) Include the most recent version of articles (if any). 4) Consider the articles during the publication period between Jan 2013 and Dec 2018. On the other hand, exclusion criteria consisted of these rules: 1) Exclude articles that involve data mining with big data but are not related to IDS. 2) Exclude articles that involve IDS but are not related to data mining with big data. 3) Exclude articles that involve data mining in IDS but are not related to big data. 4) Exclude articles not categorized as peer-reviewed journal or conference papers. After applying the inclusion and exclusion criteria, we obtained 53 journal and conference articles as the primary source of the quality assessment stage.

C. Quality Assessment Rules (QAR)

As a step to guarantee the quality of the articles, 10 quality assessment questions were identified as follows: QAR1: Is the research purpose/problem clearly defined? QAR2: Are the DMT utilized with big data clearly defined? QAR3: Is the IDS principle clearly described? QAR4: Is the data preprocessing in the experiment well suited for the intended research? QAR5: Is the employed dataset clearly defined? QAR6: Are the big data mining tools clearly pinpointed? QAR7: Are the merits and shortcomings of the employed big data analysis tool, software, or algorithm addressed? QAR8: Is the background of attacks clearly depicted? QAR9: Does the research apply accurate evaluation metrics? and QAR10: Are the experiment results and findings are clearly reported?

Based on these QARs, we evaluated the 53 articles to assure the richness and suitability of the final outcome. Each question was graded from 0 to 1 mark, where each QAR was scored as follows: “Excellent” = 1, “Very good” = 0.75, “Good” = 0.5, “Poor” = 0.25, “not answered” = 0. Accordingly, article with score 5 or more was considered for the data extraction stage. Overall, a total of 32 articles were considered in the data extraction stage. Table I list the selected articles group by the final score.

D. Extract and Synthesize Data

The aim of this stage is to extract information that will answer the RQs. For the sake of quality assurance, each author plays the role of extractor and checker of others’ work. To ensure the consistency of the analytical results, common names were used for the same DMT, software, or big-data tool. For instance, Bayesian Network replaced the other related names such as Naive Bayes, Naive Bayes, or Naïve Bayesian. Different methods were adopted from [12] to provide evidence
TABLE II
SELECTED ARTICLES

| ID | RQ1 | RQ2 | RQ3 | RQ4 | Year | Ref |
|----|-----|-----|-----|-----|------|-----|
| P1 | 1   | 1   | 1   | 0   | 2013 | [13]|
| P2 | 0   | 1   | 0   | 1   | 2014 | [14]|
| P3 | 1   | 1   | 1   | 0   | 2014 | [15]|
| P4 | 1   | 1   | 0   | 0   | 2015 | [16]|
| P5 | 1   | 1   | 1   | 1   | 2015 | [17]|
| P6 | 1   | 0   | 1   | 1   | 2015 | [18]|
| P7 | 1   | 0   | 1   | 1   | 2015 | [19]|
| P8 | 1   | 0   | 1   | 1   | 2015 | [20]|
| P9 | 1   | 0   | 1   | 1   | 2015 | [21]|
| P10| 1   | 1   | 1   | 0   | 2015 | [22]|
| P11| 1   | 0   | 1   | 1   | 2015 | [23]|
| P12| 1   | 0   | 1   | 1   | 2015 | [24]|
| P13| 1   | 0   | 1   | 0   | 2015 | [25]|
| P14| 1   | 1   | 0   | 1   | 2016 | [26]|
| P15| 1   | 0   | 1   | 1   | 2016 | [27]|
| P16| 1   | 1   | 1   | 0   | 2016 | [28]|
| P17| 1   | 1   | 1   | 0   | 2016 | [29]|
| P18| 1   | 1   | 1   | 1   | 2016 | [30]|
| P19| 1   | 1   | 1   | 0   | 2016 | [31]|
| P20| 1   | 1   | 1   | 1   | 2016 | [32]|
| P21| 1   | 0   | 1   | 1   | 2016 | [33]|
| P22| 1   | 0   | 1   | 1   | 2016 | [34]|
| P23| 1   | 1   | 1   | 0   | 2017 | [35]|
| P24| 1   | 1   | 0   | 1   | 2017 | [36]|
| P25| 1   | 1   | 1   | 1   | 2017 | [37]|
| P26| 1   | 1   | 1   | 1   | 2017 | [38]|
| P27| 1   | 1   | 1   | 1   | 2018 | [39]|
| P28| 1   | 1   | 1   | 1   | 2018 | [40]|
| P29| 1   | 1   | 1   | 1   | 2018 | [41]|
| P30| 1   | 1   | 1   | 1   | 2018 | [42]|
| P31| 1   | 1   | 1   | 1   | 2018 | [43]|
| P32| 1   | 1   | 1   | 1   | 2018 | [44]|

of the collected data that are sought to provide answers to the present RQs. We used narrative synthesis (RQ1, RQ3, and RQ4), binary outcomes (All RQs), and reciprocal translation (RQ2) methods since those methods outfit the answers of our RQs. Table II shows the RQs that were answered by each selected article.

III. RESULTS AND DISCUSSION

A. RQ1: Which data mining techniques with big data have been used in IDS?

Throughout the selected articles, we have identified 16 DMTs in addition to one novel technique named Unit Ring Machine (URM). The techniques are: Adaptive Incremental Clustering (AIC), Artificial Neural Network (ANN), Association Rule (AR), Bayesian Networks (BN), Conjunctive Rule (CR), Density Estimation (DE), Decision Tree (DT), Expectation-Maximization (EM), FP-Growth (FPG), Genetic Algorithms (GA), k-Means, k-Nearest Neighbors (kNN), Local Outlier Factor (LOF), Logistic Regression (LR), PageRank, Support Vector Machines (SVM), and URM. As shown in Fig. 2, DT and BN are by far the most widely used techniques in the IDS that suit the nature of big data with a total of 13 papers each. Some experiments propose enhancement features in the traditional DMT to acquire the desirable scalability in machine learning module. A novel technique (URM) was proposed in [13] which uses the geometric properties of both normal and intrusion network traffic. The experimental results revealed that URM shows high suitability with big data in IDS.

Next, we enumerated the DMTs used in the feature selection/extraction (FS) phase since it is a vital step in making classification computationally feasible and reasonable accurate. Fig. 3 demonstrates 15 DMTs used for FS, which include: Averaging Approach (AA), Automated Branch-and-Bound (ABB), Binary Bat Algorithm (BBA), Backward Elimination Ranking (BER), Conjunctive Approach (CA), Correlation Based (CB), Chi-square, Connection-Window Based Features (CWBF), Distribution-Based Feature Selection (DBFS), Forward Selection Ranking (FSR), Information Gain (IG), Knowledge Engineering Expert (KEE), Minimum Redundancy Maximum Relevance (mRMR), and Principal Component Analysis (PCA). As shown in Fig. 3, PCA is the most popular technique used to reduce the dimensionality. Table III shows the pre-processing techniques applied by the included articles.

It is worth mentioning, as depicted in Table III, that a big number of experiments did not consider the FS step in their approach, although it is important step to enhance the speed and effectiveness of the proposed models, especially in the big data setting. Section III-B includes further discussion about the utilized DMTs.
TABLE III
UTILIZED PRE-PROCESSING TECHNIQUES

| Paper ID          | Pre-processing Technique                 |
|-------------------|------------------------------------------|
| P3, P6, P8, P9, P15, P16, P17, P18, P19, P22, P23, P25, P30 | FS                                       |
| P1, P12, P29      | Normalization, FS                        |
| P13, P26, P31     | Encoding, Normalization, FS              |
| P11, P24          | Encoding, Data cleansing, FS             |
| P14               | Normalization, Data cleansing            |
| P7                | Normalization                            |
| P32               | Encoding                                 |
| P2, P4, P5, P10, P20, P21, P27, P28 | NA                                      |

B. RQ2: What are the strengths and weaknesses of the DMT and DSF employed in the literature?

This RQ provides a rich source of information about the utilized DMTs and big data tools in the literature. We counted 9 different DSFs including: Spark, Hadoop, Kafka, Storm, Azure, Eucalyptus, HAMR, Jubatus, and Mahout. Among the included articles, it has been observed that Apache Spark outperforms Hadoop and other big data distributed streaming frameworks in term of speed, data streaming, and modifying the data in real time. Tables IV & V present the experiments and knowledge outcomes, which form a potential source and guideline for future research in this topic area.

C. RQ3: What are the evaluation metrics that have been implemented to validate the DMT in IDS?

The aim of this RQ is to enumerate the most popular evaluation metrics utilized by the respective IDSs. Overall, 19 approaches were identified in the related literature, detailed explanation about the techniques are discussed in [45]. We categorized our findings into two metrics: efficiency and performance. The efficiency metric includes 16 methods: Accuracy (Acc), Confusion Matrix (CM), Detecting Rate (DR), Error Rate (ER), False Alarm Rate (FAR), F-Measure (FM), False Negative Rate (FNR), False Positive Rate (FPR), Precision, ROC Curve (ROC), True Negative Rate (TNR), True Positive Rate (TPR), Z-score, and Visual method. Performance metrics, on the other side, are: Computational Complexity (CC), CPU Utilization (CPU-U), Memory Utilization (MU), Testing Time (TST), Training Time (TRT). Fig. 4 shows that TPR, Acc, and FPR are the most commonly used for efficiency measurement, while the popular performance measurements are TST, CC, and TRT. Even though evaluation metrics are considered vital to measure and evaluate the performance of IDSs, some papers did not validated their results sufficiently insofar as they relied on a single evaluation metric as a performance indicator, which is known to generate misleading results in some cases [46]. For instance, P2, P9, P11, and P13 reported either Acc, ER, or FPR in their studies. In the same context, we observed 12 articles did not consider the performance metrics utilization at all.

D. RQ4: What type of attacks does the IDS in the literature designed for? What are the characteristics of the utilized datasets used to achieve this goal?

Attacks can be classified into four major types: Denial of Service (DoS), Remote to User (R2L), User to Root (U2R), and Probing. Based on the data extracted, we classified the results either based on realistic datasets captured from blogs, profiles, and network traffic or benchmark datasets including DARPA, KDD Cup 99, NSL-KDD, ISCX-UNB, UNSW-NB15, and CAIDA. Detailed information about these attacks and datasets are discussed in [47]. We noticed, as shown in Fig. 5, the majority of the experiments are designed to provide binary classification of behaviors, for example, as being either normal or malicious. Furthermore, none of the real-time IDS were able to classify intrusions into the four major types. It is worth to mention that a significant number of the experiments are based the DARPA/KDD datasets which they considered obsolete and does not capture contemporary attack types.

E. Further discussion

The discussion in this section pinpoints the contributions and future works of the relevant studies in this SLR. The
comprehensive summary in Table VI draws information from multiple experiments to form a pool of rich source for future researches.

IV. Conclusions

This research explored the relevant empirical studies published in journals and conferences to investigate the utilized DMTs with big data in IDS domain. We manually examined an initial 545 papers, which we eventually reduced to include 32 relevant papers. The findings of this work provide an integrated and unified view of the extracted data to answer the proposed RQs. The conclusion of this SLR is summarized as follows:

- RQ1: The most commonly used DMT with big data in IDS are: DT, BN, k-Means, ANN, and SVM. Likewise, PCA was the most used technique in FS. Furthermore, a large fraction of papers did not consider FS in their approach, which considered an essential step to improve the overall performance.
- RQ2: Hadoop was the most popular DSF with a frequency of 12 experiments. However, other big data tools, such as Spark reported better performance. We also provide an additive summary that highlights strengths and weaknesses of the used big data tools, software, or mining frameworks.
algorithms in the field of IDS.

- **RQ3:** We enumerated the evaluation metrics in terms of efficiency and performance. The most used efficiency metrics are TPR, Acc, and FPR. On the other hand, TST, CC, and TRT were reported more frequently as a performance metrics. In addition, we observed inconsistencies that exist in the evaluation metrics used in IDS.

- **RQ4:** We noticed that most of the experiments have classified attacks as either normal or malicious. In the same scenario, no experiments were able to fully identify the major attack types in real-time mode. We also found 66% of the experiment were conducted based on outdated dataset (DARPA), which may degrade IDS efficiency.

The area of big data analytics in IDS is still in its initial stages, and we anticipate major advances on the horizon. Thus, we recommend that more adaptive DMT be integrated with available big data tools. Moreover, additional effort is required to design real-time IDSs to fully classify several types of attack. Another significant point is that Spark outperforms other popular big data tools such as Hadoop, Storm, and Jubatus. In the light of such a scenario, it is recommended to consider Spark for future research with regards to IDS.

**REFERENCES**

[1] E. Byres and J. Lowe, “The myths and facts behind cyber security risks for industrial control systems,” in *Proceedings of the VDE Kongress*, vol. 116. Citeseer, 2004, pp. 213–218.

[2] O. Osanaiye, K.-K. R. Choo, and M. Dlodlo, “Distributed denial of service (ddos) resilience in cloud: Review and conceptual cloud ddos mitigation framework,” *Journal of Network and Computer Applications*, vol. 67, pp. 147 – 165, 2016.
[43] S. M. Othman, F. M. Ba-Alwi, N. T. Alshybe, and A. Y. Al-Hashida, “Intrusion detection model using machine learning algorithm on big data environment,” Journal of Big Data, vol. 5, no. 1, p. 34, Sep 2018.

[44] E. M. Kurt and Y. Becerikli, “Network intrusion detection on apache spark with machine learning algorithms,” in Engineering Applications of Neural Networks, E. Pimenidis and C. Jayne, Eds. Cham: Springer International Publishing, 2018, pp. 130–141.

[45] N. Munaiah, A. Meneely, R. Wilson, and B. Short, “Are intrusion detection studies evaluated consistently? a systematic literature review,” 2016.

[46] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority over-sampling technique,” Journal of artificial intelligence research, vol. 16, pp. 321–357, 2002.

[47] S. M. Sangve and R. Thool, “A formal assessment of anomaly network intrusion detection methods and techniques using various datasets,” in 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Oct 2015, pp. 267–272.