INTERNATIONAL JOURNAL OF ADVANCED INNOVATIVE TECHNOLOGY IN ENGINEERING

Published by Global Advanced Research Publication House Journal

Home page: www.ijaite.co.in

Network Intrusion Detection Over the Big Data

[1]Ashish Kokate, [2]Prof. Thamraj Ghorsad

[1,2]Department of Computer Science & Engineering, Dr. Sau. Kamalatai Gawai Institute of Engineering & Technology, Darapur, Maharashtra, India

[1]ashishsk1@gmail.com, [2]raj.ghorsad@gmail.com

Article History

Received on: 25 May 2022

Revised on: 22 April 2022

Accepted on: 24 May 2024

Keywords: Big Data, Intrusion Detection, Stream Data, Big Data Analytics

e-ISSN: 2455-6491

Production and hosted by
www.garph.org

ABSTRACT

Analyzing network flows, logs, and system events has been used for intrusion detection. Network flows, logs, system events, etc. generate big data. Big Data analytics can correlate multiple information sources into a coherent view, identify anomalies and suspicious activities, and finally achieve effective and efficient intrusion detection. This paper presents methods and subsequent evaluation criteria for network intrusion detection, stream data characteristics and stream processing systems, feature extraction and data reduction, conventional data mining and machine learning, deep learning, and Big Data analytics in network intrusion detection. Current challenges of these methods in intrusion detection are also introduced.

1. INTRODUCTION

The concepts of network and information security related to a large-scale cyber-attack in any organization have continued to dominate headlines and surpass any probability of a land-based terror attack in the organization. In the area of cyber security, effective and efficient situational awareness often requires knowledge of current and historical cyber (i.e., host or network) activities to detect and respond to threatening behaviors [2]. The analysis of cyber threats could be improved by correlating security events from numerous heterogeneous sources [3]. Organizations must implement intrusion detection and prevention systems (IDPS) to protect their critical information against various kinds of attacks because anti-virus software and firewalls are not enough to provide full protection for their systems [4].

Intrusion detection systems (IDSs) can be categorized into three types: a network-based intrusion detection system (NIDS), a host-based intrusion detection system (HIDS), and a hybrid-based intrusion detection system (hybrid IDS). A

HIDS detects malicious activities on a single computer while a NIDS identifies intrusions by monitoring multiple hosts and examining network traffic. In a NIDS, sensors are located at choke points of the network to perform monitoring, often in the demilitarized zone (DMZ) or on network borders, and capture all the network traffic. Hybrid-based IDSs detect intrusions by analyzing application logs, system calls, file-system modifications (password files, binaries, access control lists, capability databases, etc.), and other host states and activities [5]. IDSs are often used with other technologies (e.g., routers and firewalls). IDS technologies such as HIDS, NIDS, network behavior anomaly detection (NBAD), and wireless local area network (WLAN) IDS are used together to correlate data from each device and make decisions according to what these IDSs monitor [6].

An intrusion prevention system (IPS) is an advanced combination of anti-virus software, personal firewalls, IDS, etc. The objective of an IPS is not only to detect attacks but also to stop them by responding automatically such as disabling connections, logging users offline, ending processes, shutting the system down, etc. Intrusion prevention systems can be classified into two categories: network-based intrusion prevention systems and host-based intrusion prevention systems [4]. The purpose of this paper is to investigate the state-of-the-art of methods and techniques in network intrusion detection, and the advances and challenges of Big Data analytics in intrusion detection to explore new techniques that aid in intrusion detection analysis.

A. Methods in Network Intrusion Detection and Evaluation Criteria

There are three kinds of intrusion detection methods including signature-based detection (also called misuse detection), anomaly-based detection, and hybrid intrusion detection. The commonly used and accurate of these is signature-based intrusion detection. After a new attack is launched, the attack pattern or signature is defined which can be targeted resources during an attack, the way that the resources are targeted, or a name (in characters) within the body of the attack code. Network security specialists can design a defense against a new assault after the attack signature is studied. Per the proposed defense, the IDS is updated accordingly to detect the new attack pattern and respond to it. This method is very effective in detecting known attacks and produces a small number of false-positive (FP) alarms that classify normal traffic as malicious. However, if the attack pattern is slightly altered, this method will not be able to identify the changed versions of the attack [7].

An anomaly detection system has a profile of normal behavior patterns in the defense system. A coming data pattern is classified as an attack when it is different from the normal pattern. The anomaly detection method can use unsupervised learning techniques to detect new emerging attacks without the need for labeled patterns; however, it can generate false alarms [8]. Anomaly detection techniques demonstrate good accuracy in detecting network-level attacks such as the SYN flood, teardrop, and denial of service (DOS), etc.; but not in recognizing application-level exploits such as Remote to Local (R2L) and User-to-Root (U2R). All anomaly detection schemes consider only the packet header fields such as flags, port numbers, IP addresses, etc.; therefore, they work well if an attack involves only the related fields at the network level. Unfortunately, they have no way to detect attacks if the payload is involved, for example, an attack on Microsoft IIS induces users to download a malicious script file, but because there are no invalid packet header fields involved, header-based techniques will not trigger any alarm. The malicious payload is a type of application-level attack; if the payload is ignored in anomaly-based detection, poor performance in detecting payload-associated attacks is obvious [9]. The hybrid intrusion detection method has been developed to improve the performance and capabilities of intrusion detection and prevention systems (IDPS) by combining the signature-based method (misuse detection) and the anomaly-based method [10].

Some evaluation criteria that can be used to compare the performance of algorithms in an IDS include [11]:

1) Accuracy, 2) false-negative rate (FNR), 3) false-positive rate (FPR), 4) time used, 5) memory consumption and 6) Kappa Statistic. Among the six evaluation criteria, three practical criteria are often used for the IDS [12]:

- Accuracy: measuring the percentage of failure and correct detection as well as the

number of false alarms generated from the IDS.

- FNR: the percentage of the samples that are reported as normal when they are anomalous. In the false-negative situation, the IDS do not detect real attacks.
- FPR: the proportion of normal instances that are incorrectly classified as anomalous.

Table 1. A Comparison between Anomaly Detection and Misuse Detection

Aspects	Anomaly Detection	Misuse Detection
Characteristics	Uses the deviation from normal usage patterns to identify intrusions.	Uses the patterns of known attacks (signatures) to identify intrusions.
Drawbacks	- Has to study the sequential interrelation between transactions - False positives.	-Known attacks have to be hand-coded -Unable to detect new attacks -Need signatures update -False negatives

Table 1 shows a comparison of characteristics and drawbacks between the anomaly method and the misuse method for intrusion detection [10]. Table 2 compares three detection methods based on different performance criteria [13].

B. Stream Data on the Communication Network

Stream data and its processing have the following characteristics: it is dynamic and constantly changing, has a large volume, allows only one or a small number of scans, flows in and out in a fixed order, and requires quick responses. Web clickstreams and network traffic are typical examples of stream data. Effective analysis and management of stream data is a huge challenge since stream data is generally not stored in any type of data repository. The continuous query model is typical in a stream data management system where predefined queries evaluate incoming streams constantly, collect aggregate data, respond to the changes of data streams, and report their status. Stream data mining involves dynamic changes and efficient discovery of general patterns within the stream data. People are interested in identifying intrusions based on the anomaly of message flow that can be discovered by dynamically constructing stream models and clustering stream data, or comparing the current frequent patterns with those at specific previous times. Most stream data reside at a low level of abstraction, but analysts are usually more interested in higher as well as multiple levels of abstraction. Therefore, multidimensional and multi-level online analysis and mining should be conducted on stream data as well [14].

The main problems associated with stream data mining are concept evolution, concept drift, and infinite length. Concept evolution is defined as the development of novel classes, while concept drift means data changes with time. Infinite length means that stream data requires infinite length storage and training time [15]. Concept drift in the learning model is introduced because of the velocity component of big stream data; specifically, concept drift indicates that statistical properties of the target variable predicted by a model change with time in an unforeseen manner. This is a major problem since the prediction will be less accurate with time [12]. Real-time intrusion detection is a tedious task because of the large volume of data involved. Data imbalance is also a major hurdle. If the imbalance level in the data is high, classifiers will be lower in accuracy and reliability. Imbalance is an inevitable problem in real-time data due to the large size and low frequency of certain transactions. Sampling techniques are common approaches to reducing the impact of imbalance on classifiers [16].

Malicious attacks and intrusions are dynamic and are needed to perform intrusion detection in a real-time environment with data streams. An event may be normal on its own, but it is malicious if it is considered as part of a sequence of events. Stream data analysis is used to help identify intrusions in this kind of situation. It can be very useful in identifying sequences of events that frequently occur together, discovering sequential patterns, and recognizing outliers or anomalies [10]. There are three main types of anomalies [17]: 1) point anomalies data samples that are detected as anomalies concerning the rest of a dataset; 2) collective anomalies collections of data samples that are anomalous altogether, and 3) contextual (conditional) anomalies being anomalous only in certain contexts. A live network situational awareness system was developed based on streaming algorithms for determining important stream features and identifying anomalous behavior. In addition, the performance of this system was improved through refining and

enhancing a well-known streaming algorithm. This system was used in a live high-speed mid-scale enterprise network and the performance and detection results of the system were presented [2].

Massive online analysis (MOA) is a framework for stream data mining including tools for collecting and evaluating machine learning algorithms. It can implement clustering, classification, regression, frequent graph mining, and frequent pattern mining. It contains online and offline collections for clustering and classification as well as evaluation tools. MOA is currently one of the best tools for stream data mining [15]. Stream data may be collected from various sources and processed in a stream processing engine so that the results are written to a destination system. Flink, Storm, and Spark Streaming are the three main open-source platforms for distributed stream processing. Flink and Storm are up to 15 times higher in throughput efficiency than Spark Streaming, a micro-batch processing system. The storm is better in throughput efficiency than the others, but Spark Streaming is robust in node failures and provides a recovery without losses [18]. The storm is one of the main distributed stream processing engines. It has had a lot of applications such as continuous computation, real-time analytics, distributed remote procedure calls (RPC), online machine learning, and the ETL (extract, transform, and load) process. The storm is fault-tolerant and scalable and guarantees that data will be processed [19]. Table 2 [20] compares the three-stream processing systems.

Table 2. A comparison of Stream Processing Systems

Aspects	Flink	Storm	Spark Streaming
Build Language	Java/Scala	Java/Closure	Java/Scala
Failure Mechanism	Check-point	Upstream Backup	Parallel Recovery
Messages Semantic	Exactly one	At least once	Exactly one
Application Program Interface (API)	Declarative	Compositional	Declarative
Throughput Efficiency	Medium	Highest	Lowest
Failures Subsystem	No	Nimbus, Zookeeper	No

A major challenge of anomaly detection is the large volume of data. Anomaly detection techniques need to be computationally efficient in dealing with inputs (data) in large size and the inputs are stream data that require online analysis. Another problem is false alarms due to the large volume of inputs. Because the input data may contain millions of data objects, a low percentage of false alarms can make analysis overwhelming. Labeled data corresponding to normal behavior is often available, but labels for intrusions are often not. Therefore, unsupervised or semi-supervised anomaly detection methods are often preferred [21]. Dealing with huge amounts of stream data, ranging from structured data to unstructured data, numerical data to text streams in micro-blogs is a challenge because the stream data is dynamic and may be very heterogeneous [22].

II. Data Mining and Machine Learning Applications in Network Intrusion Detection

A. Big Data, Feature Extraction, and Data Reduction for Network Intrusion Detection

Feature selection is important in machine learning and data mining. Removing redundant or irrelevant features and performing the principal component analysis (PCA) results in data dimension reduction. Feature selection can improve the prediction performance of models by reducing the data dimension and speeding up the learning process. Feature selection has many applications that are involved with high-dimensional data. Most studies have been conducted in an off-line learning approach in which the features of training samples were given a priori. However, such an assumption is not suitable in some real-world applications where training instances may arrive in an online manner or it is costly to collect all the data. Online feature selection (OFS) was studied. The purpose of online feature selection is to develop online classifiers that only use a small and fixed number of features [23]; therefore, making online feature selection for mining big data an important research topic.

Big data is defined as “datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.” The volume of big data in many sectors ranges from a few dozen terabytes (TB: approximately 1012 bytes) to multiple petabytes (PB: approximately 1015 bytes) [24]. Big data can be too large in volume, moves too fast, or may not fit the strictures of conventional database architectures [25]. Technologies for big data include machine learning, data mining, crowdsourcing, natural

language processing, stream processing, time series analysis, cluster computing, cloud computing, parallel computing, visualization, and graphics processing unit (GPU) computing, etc. [26]-[27]. Distributed file systems, cluster file systems, and parallel file systems are the main tools used in big data [24].

Unfortunately, redundant attributes and records make intrusion detection in Big Data analytics an especially complicated and challenging task [28]. PCA has been used in extracting features from the attributes of high dimension datasets, especially datasets with redundant attributes. Experts have used this method in feature selection for IDSs [29]. Unsupervised anomaly detection in unstructured system logs was studied using PCA and its effects, which were also presented in the study. Dimension reduction is highly important for both performance and efficiency and helps reduce the computational complexity in anomaly identification and improves the performance of the classifiers [17].

B. Conventional Data Mining and Machine Learning in Network Intrusion Detection

Data mining methods such as clustering, classification, and association rule mining are often used to obtain valuable information about network intrusion through analyzing the network data. Clustering can be used in both misuse detection and anomaly detection while classification is mainly used in anomaly detection and is a supervised learning method. An IDS based on classification can classify all the network traffic as either malicious or normal. Association rule mining searches for frequently occurring items from a large dataset and identifies correlation relationships or association rules between data items of the large dataset [30]. Specifically, there is a difference between the applications of classification and clustering in IDSs [31]:

1) Clustering: An advantage of clustering over the classification method is that it does not need to use labeled data set. Clustering is useful in intrusion detection because malicious activity should cluster together, separating itself from normal activity.

2) Classification: A classification-based IDS tries to classify all the traffic as either malicious or normal. The challenge is how to minimize the number of false negatives and false positives. Five general techniques have been used to perform classification for intrusion detection including support vector machine (SVM), fuzzy logic, inductive rule generation, neural networks, and genetic algorithms.

Some data mining algorithms are good for attack or intrusion detection. For example, the decision tree (DT) is thought one of the most effective and efficient techniques for detecting attacks in anomaly detection. Other data mining techniques such as support vector machines (SVM), Naive Bayes, Random Forest, and neural networks have been used in analyzing the traffic of network systems. Attack signatures are needed in misuse detection which uses techniques such as SVM, neural networks, and DT. A hybrid classifier, on the other hand, is a mixed-method based on algorithms such as neural networks, k-nearest neighbors (k-NN), DT, and fuzzy logic. It can detect both misuse attacks and anomaly attacks. Network events have been treated as a data stream and various data stream-based learning models have been used in presenting a new insight into intrusion detection [32]. The applications of data mining in communication network control include [33]:

- Classification of network status, application types, user types, etc. to make the most suitable action for each situation.
- Clustering of users or nodes into groups based on commonalities such as applications, channel conditions, hardware resources, and position to improve network
- load distribution, network performance, and user satisfaction.
- Developing regression models to predict service duration, traffic load, channel quality, the number of service requests, and desired service quality.
- Identify frequent sequential patterns to find typical network traffic trends, typical user behavior, network security attacks, etc.
- Discovering association rules between network control parameters, user satisfaction, and network performance.

Intrusion detection based on text processing has been a research topic in information and network security. An intrusion detection mechanism using the text-processing-based k-nearest neighbor (k-NN) classifier was presented. The classifier was

used on the database DARPA 1998 and was shown to produce better results than those of other algorithms [34]. Summarizing research in attack detection is presented here as follows [35]: 1) emphasis on trying to find hybrid solutions and detection classification from the year 2001 to 2008; 2) emphasis more on data mining and machine learning, as well as hybrid solutions of anomaly detection and misuse detection from 2010 to 2015. Machine learning methods have been studied for the design of IDS including Bayesian networks, linear genetic programming (LGP), neural networks, support vector machines (SVM), fuzzy inference systems (FISs), multivariate adaptive regression splines (MARS), etc. Neural networks especially have been used in both misuse detection and anomaly detection [36]. Machine learning methods like SVM are also under the umbrella of data mining and each of the data mining and machine learning methods has its pros and cons in intrusion detection. Table 4 [37] describes and compares these methods.

C. Deep Learning in Network Intrusion Detection

Deep learning is also called deep machine learning, hierarchical learning, or deep structured learning. It can be unsupervised or supervised learning from the collected data based on multiple layered models. Deep learning algorithms are very useful for analyzing large amounts of unsupervised data with high variety, which gives it potential in analyzing network data for intrusion detection. However, deep learning has some challenges in big data [38]:

- Large-scale deep learning models are appropriate for handling a large volume of inputs associated with big data. However, how determining the optimal number of model parameters and how to improve the computational practicality is a challenge in deep learning for big data.
- High-dimensional data sources can result in large volumes of raw data, making it a challenge to develop deep learning algorithms for big data with high dimensions.
- It is necessary to adapt deep learning to stream data because within incremental learning for non-stationary data, one of the challenges is handling streaming and fast-moving input data.

Table 3: Various Data Mining and Machine Learning Methods Used in Intrusion Detection

Methods	Pros	Cons
Support Vector Machine	Insensitivity to input data dimension High training rate and decision rate better learning ability for small samples	Limited to binary classifiers which cannot give additional information about detected attack type Training time is lengthy
Decision Tree	High accuracy in detection Works well with huge data sets	It is computationally intensive to build
Bayesian Network	Can incorporate both prior knowledge and data Can encode probabilistic relationships among the variables of interest	If prior knowledge is incorrect, it is possible not to contain any good classifiers Hard to handle continuous features
Genetic Algorithm	Biologically inspired and uses evolutionary algorithms Can derive best classification rules and select optimal parameters	Can be over-fitted Constant optimization response times are not assured
Neural Networks	Do not need expert knowledge and can find novel or unknown intrusions Generalization capable of limited, incomplete, and noisy data	Possible to over-fit during training Not suitable for real-time detection because of a slow training process
Fuzzy Logic	Effective, especially for port scans and probes	Difficult to identify reduced, relevant rule subsets and to update dynamic rules at runtime High resource consumption

D. Big Data Analytics in Network Intrusion Detection

Conventional technologies often cannot support long-term and large-scale analytics primarily because retaining a large volume of data is often not economical or feasible. Therefore, most event logs and other recorded activities are deleted after a fixed retention period. Secondly, it is too inefficient to conduct analysis and complex queries on unstructured and large datasets with noisy and incomplete data. Big data applications are becoming a part of security management software because they help prepare, clean, and query heterogeneous data with incomplete and/or noisy records [39]. In addition, issues related to data fusion, security information and event management (SIEM) systems, and heterogeneous

intrusion detection architectures have been studied [3].

Big Data analytics for intrusion detection and mitigating network security problems has been attracting more and more attention because it promotes the study of large volumes of complex and disparate data with various formats from heterogeneous sources, detects anomalies, and combats cyber-attacks. Ultra-high-dimensional data models can be created to profile stream data accurately online, which helps predict and detect intrusion and attacks in real-time [40]. Big Data technologies like the Hadoop ecosystem and stream processing can store and analyze large heterogeneous datasets at a high speed, transforming security analytics by:

(a) capturing large-scale data from numerous internal and external sources such as vulnerability databases; (b) conducting deep analytics on the data; (c) achieving real-time analysis of stream data, and (d) presenting an integrated view of security-related information. Because Big Data analytics tools need to be configured properly, system analysts and architects are required to have an intimate knowledge of their systems [41].

By correlating the security events from heterogeneous sources, a holistic view and excellent situational awareness of intrusion or attacks can be achieved. Only a single event source like network traffic can generate big data and introduce challenges; however, the more heterogeneous data sources available, the more challenging and complicated big data will become. On the other hand, integrating security events from heterogeneous sources (such as NetFlow, firewalls, IDS, and host log files) for better situational awareness is another major challenge. Furthermore, a comprehensive approach is required to develop tools, techniques, and infrastructure that adapts to permit continuous querying over stream data and spans across the areas of statistical analysis, deductive reasoning (inference), inductive reasoning (machine learning), and high-performance computing (parallelization). Automated (or at least partially automated) distribution of tasks over clusters and big data-specific parallelization techniques are also necessary for effective stream processing [22].

A model with a three-layered architecture has been used to describe big data systems, including an application layer, a computing layer, and an infrastructure layer. The application layer includes software or other application resources for implementing data analytics while the computing layer is also called the middleware layer, including the software tools for integrating and managing data. And the infrastructure layer, which includes a network of storage systems based on virtualization and cloud computing is also a system of distributed hardware for storing big data [42]. Big data technologies can be categorized into stream processing and batch processing. Stream processing is performed on data in motion; while batch processing is performed on data at rest. A typical technology with the batch processing of big data is Hadoop. Using a Big Data analysis system that can conduct both stream processing (real-time analytics) and batch processing is the best defense strategy that will efficiently perform intrusion detection and protect critical information infrastructures (CIIs) [42]. Network intrusion prediction and detection are time-sensitive and need highly efficient Big Data technologies to deal with problems on the fly [20].

An advanced persistent threat (APT) is a targeted attack against a high-value asset or a physical system and is one of the most serious information security problems. Big Data analysis is an appropriate approach to APT identification [41] because of the assistance in facilitating APT identification by supporting [43]:

- Anomaly detection is based on the correlation of historical and recent events. For example, an increased volume of Domain Name System (DNS) traffic from a system during a short time can be due to legitimate users' behaviors. But such a pattern indicates covert data exfiltration if it is also detected in historical traffic over days. Furthermore, this kind of correlation helps reduce the false positive rate (FPR) of alerts. Big Data analytics increases the scope and quantity of data that can compute the correlation.
- Managed and dynamic capture, integration, and correlation of data from heterogeneous data sources like network traffic, event data (e.g., IDS, network devices), and operating system artifacts to help defenders correlate sporadic low-severity events as the result of ongoing intrusion or attack behavior. Compared with SIEM systems, Big Data analytics

does not have a limited-time window within which the correlation is performed.

CONCLUSION

Anomaly detection methods are good in detecting network-level attacks, but not in detecting application-level exploits. Accuracy, FNR, and FPR are three practical evaluation criteria for the IDS. Storm, Flink, and Spark Streaming are primarily open-source platforms for distributed stream processing. Removing redundant and irrelevant features helps feature selection and PCA is an approach to extracting features from high dimension data. Intrusion detection can be improved by a comprehensive approach to monitoring security events from various heterogeneous sources. Conventional data mining and machine learning methods are useful in intrusion detection, but they have limitations in dealing with big data on the network. Deep learning algorithms also have some challenges in big data but have the potential in analyzing large amounts of unsupervised network data with a high variety for intrusion detection.

Big Data analytics can be used to analyze network traffic, virus signatures, user behaviors, website profiles, the timing of incident events, etc. It can also compute the correlation of attributes from heterogeneous sources and improve the agency's security through the continuous monitoring of stream data. Streaming data analysis in real-time is becoming the fastest and most efficient way to obtain useful knowledge. One Big Data challenge in stream processing is the proper storage, processing, and management of big volumes of stream data. Big Data analytics has the potential to solve the problems and challenges in cleaning and querying data with incomplete and/or noisy records and in a variety of data structures due to heterogeneous data sources.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

FUNDING SUPPORT

The author declares that they have no funding support for this study.

REFERENCES

[1] [1] Conteh, N. Y., & Schmick, P. J. (2016). Cybersecurity: risks, vulnerabilities, and countermeasures to prevent social engineering attacks. International Journal of Advanced Computer Research, 6(23), 31-38.

[2] [2] Streilein, W. W., Truelove, J., Meiners, C. R., & Eakman, G. (2011, November). Cyber situational awareness through operational streaming analysis. In Military Communications Conference, 2011-MILCOM 2011 (pp. 1152-1157). IEEE.

[3] [3] Zuech, R., Khoshgoftaar, T. M., & Wald, R. (2015). Intrusion detection and big heterogeneous data: a survey. Journal of Big Data, 2(3), 1-41.

[4] [4] Sandhu, U. A., Haider, S., Naseer, S., & Ateeb, O. U. (2011). A survey of intrusion detection & prevention techniques. In 2011 International Conference on Information Communication and Management, IPCSIT (Vol. 16). 66-71.

[5] [5] Beigh, B. M., & Peer, M. A. (2012). Intrusion Detection and Prevention System: Classification and Quick Review, ARPN Journal of Science and Technology, 2(7), 661-675.

[6] [6] Tyler, G. (2009). Information Assurance Tools Report Intrusion Detection Systems. Information Assurance Technology Analysis Center (IATA).

[7] [7] Kabiri, P., & Ghorbani, A. A. (2005). Research on intrusion detection and response: A survey. IJ Network Security, 1(2), 84-102.

[8] [8] Nieves, J. F., & Jiao, Y. C. (2009). Data clustering for anomaly detection in network intrusion detection. Research Alliance in Math and Science, 1-12.

[9] [9] Zhang, L., & White, G. B. (2007, March). An approach to detect executable content for anomaly-based network intrusion detection. In Parallel and Distributed Processing Symposium, 2007. IPDPS 2007. IEEE International (pp. 1-8). IEEE.

[10] [10] Youssef, A., & Emam, A. (2011). Network intrusion detection using data mining and network behavior analysis. International Journal of Computer Science & Information Technology, 3(6), 87-98.

[11] [11] Faisal, M. A., Aung, Z., Williams, J. R., & Sanchez, A. (2012, May). Securing advanced metering infrastructure using intrusion detection system with data stream mining. In Pacific-Asia Workshop on Intelligence and Security Informatics (pp. 96-111). Springer Berlin Heidelberg.

[12] [12] Najafian, Z., Aghazarian, V., & Hedayati, A. (2015). Signature-Based Method and Stream Data Mining Technique Performance Evaluation for Security and Intrusion Detection in Advanced Metering Infrastructures (AMI). International Journal of Computer and Electrical Engineering, 7(2), 128-139.

[13] [13] Patel, A., Taghavi, M., Bakhtiyari, K., & JúNior, J. C. (2013). An intrusion detection and prevention system in cloud computing: A systematic review. Journal of network and computer applications, 36(1), 25-41.

[14] [14] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

[15] [15] Parikh, D., Tirkha, P. (2013). Data Mining & Data Stream Mining – Open Source Tools, International Journal of Innovative Research in Science, Engineering and Technology, 2(10), 5234-5239.

[16] [16] Balasubramanian, R., Joseph, S.J.S.A. (2016). Intrusion Detection on Highly Imbalanced Big Data using Tree-Based Real-Time Intrusion Detection System: Effects and Solutions, International Journal of Advanced Research in Computer and Communication Engineering, 5(2), 27-32.

[17] [17] Kicanaoglu, B. (2015). Unsupervised Anomaly Detection in Unstructured Log-Data for Root-Cause-Analysis.Master's Thesis, Computing, and Electrical Engineering, Tampere University of Technology.

[18] [18] Lopez, M. A., Lobato, A., & Duarte, O. C. M. B. (2016, December). Performance comparison of Open-Source stream processing platforms. In IEEE Global Communications Conference (Globecom), Washington, USA. 1-6.

[19] [19] Bhattacharya, D., & Mitra, M. (2013). Analytics on big fast data using real-time stream data processing architecture. EMC Corporation. 1-34.

[20] [20] Suthaharan, S. (2014). Big data classification: Problems and challenges in network intrusion prediction with machine learning. ACM SIGMETRICS Performance Evaluation Review, 41(4), 70-73.

[21] [21] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 15.

[22] [22] NESSI (2012). Big Data: A New World of Opportunities, NESSI White Paper, December, 1-25.

[23] [23] Hoi, S. C., Wang, J., Zhao, P., & Jin, R. (2012, August). Online feature selection for mining big data. In Proceedings of the 1st international workshop on big data, streams and heterogeneous source mining: Algorithms, systems, programming models and applications (pp. 93-100). ACM.

[24] [24] Merelli, I., Pérez-Sánchez, H., Gesing, S. and D'Agostino, D. (2014). Managing, Analyzing, and Integrating Big Data in Medical Bioinformatics: Open Problems and Future Perspectives, BioMed Research International, Volume 2014, Article ID 134023, 1-13.

[25] [25] Dumbill E. (2013). Making sense of Big Data. Big Data January Preview Issue, BD1-BD2. Available from: http://www.liebertpub.com/mcontent/files/Big%20Data%20 Preview%20Issue.pdf.

[26] [26] George, G, Haas, M.R., Pentland, A. (2014). Big Data and Management, Academy of Management Journal, 57(2): 321–326.

[27] [27] Zhang, D. (2013). Granularities and Inconsistencies in Big Data Analysis, International Journal of Software Engineering and Knowledge Engineering, 23(6): 887–893.

[28] [28] Manandhar, P. (2014). A Practical Approach to Anomaly-based Intrusion Detection System by Outlier Mining in Network Traffic (Doctoral dissertation, Masdar Institute of Science and Technology).

[29] [29] Guillen, E., Sánchez, J., & Paez, R. (2015). The inefficiency of ids static anomaly detectors in real-world networks. Future Internet, 7(2), 94-109.

[30] [30] Patond, M. K., & Deshmukh, P. (2014). Survey on Data Mining Techniques for Intrusion Detection System. International Journal of Research Studies in Science, Engineering and Technology, 1(1), 93-97.

[31] [31] Lappas, T. and Pelechrinis, K. (2010). Data Mining Techniques for (Network) Intrusion Detection Systems, Technical Report, Department of Computer Science and Engineering, UC Riverside, Riverside CA 92521, May 10, 1-13.

[32] [32] Manandhar, P., & Aung, Z. (2014). Intrusion Detection Based on Outlier Detection Method. INCIDENT '2014), April, 21-22.

[33] [33] De Sanctis, M., Bisio, I., & Araniti, G. (2016). Data mining algorithms for communication networks control concepts, surveys, and guidelines. IEEE Network, 30(1), 24-29.

[34] [34] Kumar, G. R., Mangathayaru, N., & Narsimha, G. (2016). Intrusion Detection-A Text Mining Based Approach. International Journal of Computer Science and Information Security, 14, 76-88.

[35] [35] Stouten, F. (2016). Big data analytics attack detection for Critical Information Infrastructure Protection, Master Thesis, Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology, 1-65.

[36] [36] Peddabachigari, S., Abraham, A., Grosan, C., & Thomas, J. (2007). Modeling intrusion detection systems using hybrid intelligent systems. Journal of network and computer applications, 30(1), 114-132.

[37] [37] Singh, J., & Nene, M. J. (2013). A survey on machine learning techniques for intrusion detection systems. International Journal of Advanced Research in Computer and Communication Engineering, 2(11), 4349-4355.

[38] [38] Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. Journal of Big Data, 2(1), 1-21.

[39] [39] Raja, M. C., & Rabbani, M. A. (2014). Big Data analytics security issues in data-driven information systems. IJIRCCE, 2(10). 6132-6135.

[40] [40] Cui, L., F. Yu, R. and Yan, Q. (2016). When Big Data Meets Software-Defined Networking: SDN for Big Data and Big Data for SDN. IEEE Network, January/February, 58-65.

[41] [41] Cárdenas, A. A., Manadhata, P. K., & Rajan, S. (2013). Big data analytics for security intelligence. The University of Texas at Dallas@ Cloud Security Alliance. 1-22.

[42] [42] Oseku-Afful, T. (2016). The use of Big Data Analytics to protect Critical Information Infrastructures from Cyber-attacks, Information Security, masters level 2016, Luleå University of Technology Department of Computer Science, Electrical and Space Engineering, Master Thesis, 1-64.

[43] [43] Nikos Virvilis, C. I. S. A., CISSP, G., Oscar Serrano, C. I. S. A., & CISM, C. (2014). Big Data Analytics for Sophisticated Attack Detection. ISACA Journal, 3, 1-8