B

Al-Azhar Bulletin of Science
Basic Science Sector

THE OFFICIAL PUBLICATION OF
THE FACULTY OF SCINCE
(MEN , GIRLS &ASSUIT BRANCH)
AL-AZHAR UNIVERSITY
CAIRO, EGYPT

ABSB





Al-Azhar Bulletin of Science: Section B, Vol. 32, No. 1, (June) 2021, pp. 27-44
http://doi. 10.21608/absb.2021.63810.1100

Al-Azhar Bulletin of Science: Section B

DISTRIBUTED INTRUSION DETECTION SYSTEMS IN BIG DATA: A SURVEY

Bashar I. Hameed[1]*, AbdAllah A. AlHabshy[1], Kamal A. ElDahshan[1]

1.*Mathematics Department, Faculty of Science, Al-Azhar University, Cairo, Egypt.*
* Corresponding Author: basharibh78@gmail.com
Received: 17 Feb 2021; Revised: 19 Apr 2021; Accepted: 25 Apr 2021; Published: 29 Sep 2021

ABSTRACT

We live in a time where data stream by the second, which makes intrusion detection a more difficult and tiresome task, and in turn intrusion detection systems require an efficient and improved detection mechanism to detect the intrusive activities. Moreover, handling the size, complexity, and availability of big data requires techniques that can create beneficial knowledge from huge streams of the information, which imposes the challenges on the process of both designing and management of both *Intrusion Detection System (IDS)* and *Intrusion Prevention System (IPS)* in terms of performance, sustainability, security, reliability, privacy, energy consumption, fault tolerance, scalability, and flexibility. IDSs and IPSs utilize various methodologies to guarantee security, accessibility and reliability of enterprise computer networks. This paper presents a comprehensive study of the Distributed Intrusion Detection Systems in Big Data, and presents intrusion detection and prevention techniques that utilize machine learning, big data analytics techniques in distributed systems of the intrusion detection.

Keywords: *Intrusion detection; Signature-based detection; Anomaly-based detection; Machine learning; Big data, Distributed systems.*

1. Introduction

The distributed environment field has been significantly advancing in the last two decades, providing services of software and hardware to different users, which has changed the computing environment of data sharing, cycle sharing, and other services in terms of distributed resources [1].

Intrusion is defined as every group of actions attempting to threaten the confidentiality, integrity, or the availability of resources [2]. In an information system, intrusion represents any activity that violates a security policy of a system. While intrusion detection represents the process utilized to identify the intrusions, intrusion detection is based on the behavior of the intruder which is – understandably – expected to be different from legitimate, hence, the significance of information systems as comprehensive assets for organizations. Despite the subsystems of complex intrusion detection are not applications typically, they have been incorporated as elements of the operating systems. Almost all intrusion detection systems attempt to detect the suspected intrusions, and then alert the system administrator. Technologies for automated reactions to intrusion are just in their initial phases of development. The original systems of intrusion detection have assumed that the system of the single processor of stand–alone and detection techniques consisted of post-facto processing of the auditing records, while the current systems consist of various nodes executing multiple operating systems that are linked together forming the single distributed system. Furthermore, the intrusions could be done by multiple intruders, which could substantially

increase the complexity, with no fundamental problems [3].

Usually, the intrusion detection systems (IDSs) are deployed with other mechanisms of preventive security, like authentication and authorization. IDSs act as the second line of defense to protect the information systems. However, there are many reasons which can make the intrusion detection represent as the necessary part for a defense system. For instance, numerous classical applications and traditional systems were developed and expanded without taking security into consideration during the design and implementation phases, and both systems and applications were developed to work within diverse environments, so it might be vulnerable when it deployed within the existing environment. The system can be secure when isolated, however it becomes vulnerable when connected to the Internet. So, Intrusion detection can identify and consequently allow the responses to attacks against those systems. Besides, lack of information security and the software engineering practices might lead to design bugs or flaws which could be utilized via the intruder to attack the systems. Consequently, prevention mechanisms, such as firewalls, might not be as effective as expected [4]. Therefore, the IDSs were designed to reveal the intrusion before discovering the secured system resources. IDSs have always been considered as the second defense line in terms of security, and the cyber–space equivalent to the burglar alarms used today in physical security [5].

The rest of the paper is organized as follows. Section 2 provides an overview of the intrusion detection systems. Section 3 discusses big data, followed by intrusion detection in the distributed big data in Section 4, and finally, Section 5 provides the conclusion.

2. Intrusion detection systems

Intrusion detection systems (IDSs) in secured systems provide most if not all information for other supportive systems: the intruder identification, intrusion time, intruder location, intrusion activities (passive or active), intrusion type (such as the wormhole, sinkhole, black hole, selective forwarding, and so on), the OSI layer at which intrusion could be done (such as physical, data link, or network layer) [4]. Such information is beneficial for handling the result of the attacks since it obtained specific information about the intruder, and so is considered as the third defense line. Hence the importance of intrusion detection systems to the information systems [5].

IDS is defined as the combination of tools, methods, and resources utilized to identity, locate, and report the intrusions. Typically, intrusion detection is considered as a part of the protection system installed around a system or a device, and cannot be a stand–alone protection [6].

2.1. Intrusion Techniques

Intrusion is defined as any attempt to expose, alter, destroy, steal, disable, or gain unauthorized access to make an unauthorized utilization of assets. It can also be defined as an unauthorized or unwanted activity in a network [7]. Generally, there are two types of intrusion; active or passive. Active intrusion attempts to modify the resources of the system or affect their operations. In fact, the intruder impacts the operation in the attacked network, therefore such impact is the objective of the attack and can be detected. Passive intrusion, on the other hand, attempts to utilize the information from the system but does not impact the resources of a system, and the attackers typically are hidden taping a communication link in order to either collect data or destroy the functioning elements for a network [5]. Intruders can be either internal or external. Internal intrusion is initiated from inside the security perimeter (the insider), such as the entity authorized to access some system resources, but utilizes such granted resources to access some other prohibited ones.

External intrusion is – on the contrary – initiated from outside the system via an unauthorized or illegitimate intruder (the outsider) [8]. Over the Internet, the potential of outside intrusions extents from amateur pranksters to organized crimes, hostile

governments, and international terrorists. The intrusion can be done at any time and from anywhere on the host network, furthermore, the intrusion is always achieved on services provided via the internet.

The intrusion can be associated with any protocol type. So, this section presents the common types of network intrusions [9-11] :

Distributed denial of the services (DDoS): An intrusion attempting to disrupt the server/network resources and make those resources unavailable to the intended users.

Denial of the services (DOS): A known cyber–attack where the attacker seeks to overwhelm the machine with many requests of illegitimate connections in order to make resources of the network unavailable for its intended users permanently or temporarily. These may be too difficult to distinguish from the legitimate activity of a network.

The Exploits: Attacks which are achieved via targeting and detecting the known vulnerabilities existing in the operating systems, and exploit software in order to automate the attacks as soon as the potential vulnerability is detected.

Reconnaissance: This kind of intrusion collects the preliminary information about public networks or target hosts, is utilized via the exploit techniques in order to penetrate the target hosts or the networks through leveraging collected information, uses the free information available to the public, and the social media searches assist in the reconnaissance attacks. It is also known as passive reconnaissance.

Worm: A malicious software that spreads over the network propagations. This kind of attack performs larger rather quicker in terms of networks. It also affects computers, turning them into zombies or bots, together with an intent to utilize them in the distributed attacks via the formation of the botnets.

Fuzzers: The attack which utilizes the massive and huge amounts of randomized data recognized as the “Fuzz” in order to trigger network failure or crash the important network servers.

Analysis: The assortment intrusion which penetrates the web applications via ports (such as port scans), the emails (such as spam), and the scripts of web (such as files of HTML).

Backdoors: The technique of by–passing or transiting the normal authentication and remote access control to the device, then locating an entrance to the plaintext while it is struggling to continue without being observed.

Shellcode: An attack where the attacker perforates the slight piece of the code, then begins from the shell in order to control a compromised machine.

Generic: A technique established against block–ciphers via utilizing the hash function to perform the collision attack regardless of the block–cipher structure.

2.2. Intrusion Detection Techniques

In a secured information system, if Intrusion Prevention (the first defense line) does not prevent intrusions, then Intrusion Detection (which is the second defense line) comes into play. Detection is the process of detecting any suspicious behavior in networks performed through members of the network [5]. that it has been also utilized to identify intrusions [4]. Intrusion detection determines any unauthorized access to a system or any intruder attempt to gain system resource.

The approaches of automated detection do not identify the intrusion before the intruder initiates the interaction with a system, but rather the system administrators routinely take actions to prevent the intrusions. Those actions require access control to the system before the user can gain any access to a system, and these techniques identify the system vulnerabilities which the intruder can exploit or gain unauthorized access to the system through, and blocking some or all access points to the network, as restricting system access to physical access [3].

IDSs is also defined as the security services controlling and examining all system activities, challenges to access and identify the system resources, and unauthorized activities.

Typically, IDSs are utilized with other techniques of protective security which are called authentication and access control. IDS is the most significant part of the full defense systems [12].

The design of an IDS should fulfil the following requirements [5]:

- Does not introduce new weaknesses for the system,
- Consume minimal resources from the system and does not degrade the overall system performance with overhead operations,
- Runs continuously and remains hidden from the system and users.
- Uses the standards to be cooperative and open.
- Reliable and reduces the false positives and the false negatives in the detection phase.

So, the IDS aims to identify different types of malware as early as possible, which cannot be identified by the traditional firewall. Developing IDSs has become extremely important due to the increasing volume of computer malware [13].

IDSs can be classified to two main categories which are host–based IDSs and network–based IDSs according to data sources utilized by the IDSs [4]. Host–based IDS (HIDS) monitors all activities of the single host and detects if there is any malicious activity. Mainly, HIDS can be used to monitor the processes activities and to ensure that the policies of system files security, registry keys, and system logs are applied [14]. Host–based intrusion detection can run on the individual systems that include mechanisms of analyzing and collecting information on the particular system. Network–based intrusion detection (NIDS) is utilized to monitor and analyze the data from the traffic of a network in order to protect the system from the network–based attacks [15]. NIDS are installed in the network points such as gateways and routers to check the intrusions in network traffic [16]. The host-based IDS utilized the logs-based detection as a data source. Log detection techniques are primarily hybrids dependent on the rule and machine learning rely on log features and utilize text analysis-based techniques. So, logs-based detection methods could be divided into features engineering, text analysis, and combine of a rule-based system. Network-based IDS methods could be divided into three sub-classes; which are packet-based detection, flow-based detection, and session-based detection. A network-based IDS utilizes the network traffic as the data source which is typically the packets. These packets represent the primary units of communication of the network. The session is a sequence of packets combined into a piece of information in the network. Packets contain packet headers and payloads. So, detection of the packet includes packet parsing and the analysis of a payload. Depended on the feature extraction, flow-based detection methods could be divided for feature engineering and deep learning. Moreover, traffic grouping is a unique approach to flow detection. The session-based detection methods could be divided depending on whether the sequence information is utilized for the statistical feature or sequence feature [17]. The following Fig. 1 shows the Classification of IDS depending on the source of data.

The major advantage of the host-based IDSs is that they can precisely locate the intrusions and also initiate the responses. Host-based IDSs can monitors the behaviours of significant objects (such as sensitive files, programs, and ports). In contrast, the disadvantages are that the host-based IDSs occupy the host resources, depend on the host reliability, and unable to detect the network attacks. A network-based IDS usually is deployed on major hosts or switches. A majority of network-based IDSs are independent of the operating system (OS). Consequently, it could be utilized in a different environment than the OS. Moreover, the network-based IDSs have the ability to detect a specific kind of protocol and network attacks. The major drawback is that it is only monitoring the traffic that passes via the specific segment of a network [17].

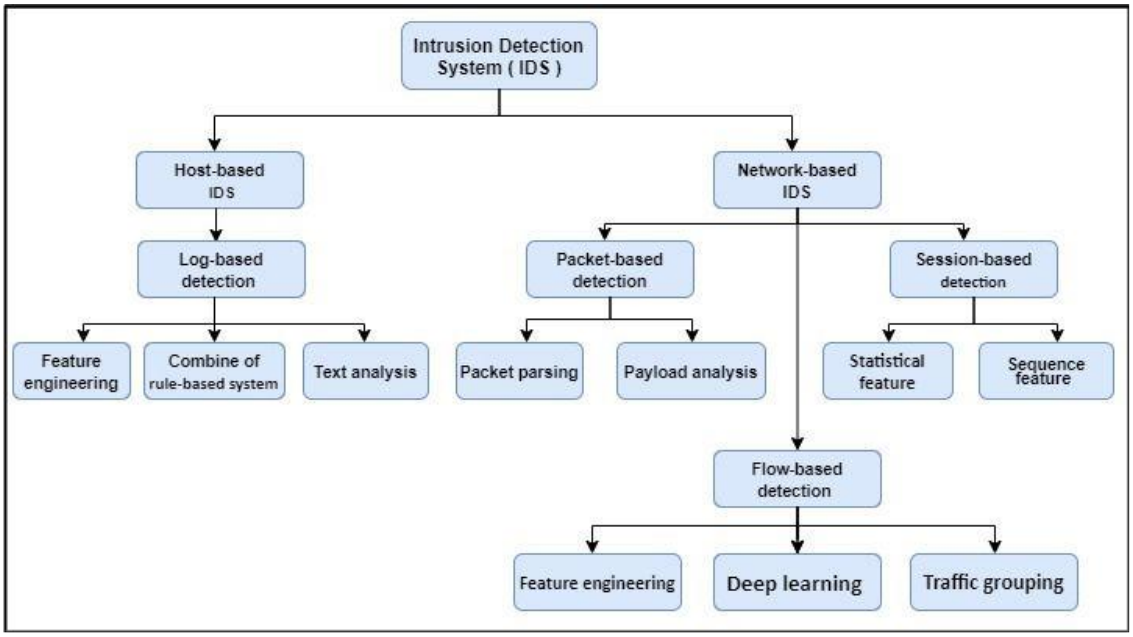

Fig. 1. Classification of IDS depending on the source of data

Table 1. Explaining a comparison between HIDS and NIDS [17, 18]

	HIDS	NIDS
Source of a data	Logs of the operating system or programs of application	Traffic of the network
The Deployment	Every host. Dependent on an operating system. Difficult in deploy.	Key nodes of the network Easy in order to deploy.
Efficiency of detection	Low, it must process various logs.	High, it can detect the attacks in the real time.
Intrusion traceability	Trace the process of intrusion according to the system call paths.	Trace a position and the time of intrusion depending to IP the addresses and the timestamps.
Threat anticipation	Good in trending and detecting suspicious behavior patterns	Good in trending and detecting suspicious behavior patterns
Deterrence of the intruder	Strong deterrence for the inside intruders	Strong deterrence for outside intruders
Assessing damage	Excellent in determining the extent of a damage	So weak in determining the extent of a damage
Intruder prevention	Good in preventing inside intruders	Good in preventing outside intruders
The limitation	It cannot analyze network behaviors	Monitor a traffic passing only via the specific segment of a network.

Information security researchers study the intrusion detection approaches from two main perspectives; anomaly detection, and misuse detection [19]. So, the intrusion detection techniques are classified into two major categories; Signature–based Detection (SD), and Anomaly–based Detection (AD) [12].

Anomaly–based Detection IDSs depend in the assumption that an attacker behavior differs from the normal user behavior, which helps to detect evolving attacks. Anomaly–based Detection IDSs also identify the normal behavior of a system and keep updating it [16]. One primary drawback of Anomaly–based Detection IDSs is defining their sets of rules. The efficiency of a system depends on implementing the AD and testing it on all protocols, and on the rule that defining the process is affected by different protocols utilized by several vendors. Moreover, custom protocols make it more difficult to define the rules, as detailed knowledge of the accepted behavior of a given network needs to be developed by the administrators for detecting the intrusion correctly, however, once the rules are defined and the protocol is built then the anomaly detection system will work well.

Signature–based Detection, in the contrary, involves searching in network traffic for malicious bytes series or sequences of packets [15]. It references a stored collection of the previous signatures of attacks like the specific patterns, the sequences of known malicious instructions, byte sequences in the network traffic, and the known vulnerabilities of a system [14].

The signature engines have main limitations such as detecting only the attacks of which signatures are stored previously in the database; the signature have to be generated for each attack, hence new attacks cannot be detected, a technique easily deceived as it is dependent on the string matching and regular expressions. Moreover, these techniques look for only the strings with the packets that transmit over the network. Further signatures work very well only against fixed and specific behavioral patterns. So, it fails when dealing with attacks created by humans or the worm within a behavioral characteristic of self–modifying [15]. The following Fig. 2 shows the types of IDS depending on intrusion technique.

Systems of signature-based detection are

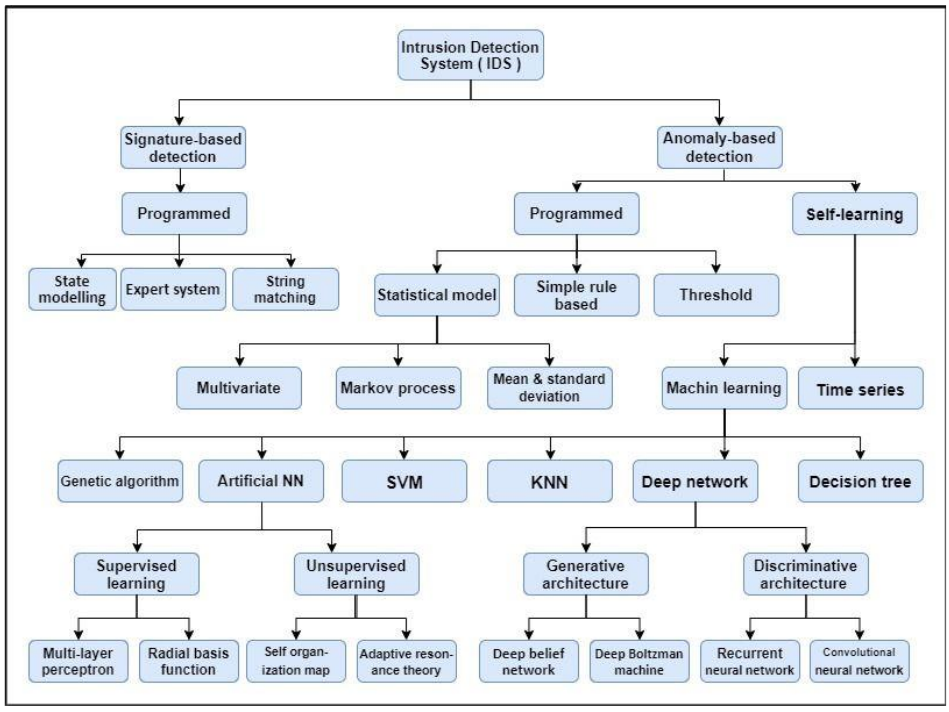

Fig. 2. Types of IDS depending on intrusion technique

programmed together within distinct decision rules. These rules retain the detection code in a straight and forward manner to detect the intrusion. It is programmed in three categories: state modelling, an expert system, and the string matching [18]:

- **State modelling** is encoding of the attacks as number of various states in a finite automaton. Also, each one of these attacks has observed in the traffic profile to be considered as an intrusion.
- **Expert systems** contain a group of rules that utilized to characterize the scenarios of attacks which known to the system.
- **String matching** is the process of knowledge acquisition just as the expert system but it has diverse approach in knowledge exploiting [20].

Anomaly detection categorized into two types depending on the way that normal profile of the system specified [18]:

- **Programmed:** The programmed technique is when the system needs the user or the external person to teach a system how to detect the changes in behavior. The programmed techniques also are divided into three major classes which are; the threshold, the simple rule based, and the statistical models. Statistical models are sub-categorized into three types which are; the deviation of mean(standard), the multivariate, and Markov process.
- **Self-learning:** system of self-learning operates through the example together with the baseline set of the normal operation. This system can be sub-divided into the major categories which are the model of time series and the machine learning.
    - **Time series:** The time series clearly indicates that the data pattern is stationary but the time of packets arrival differs from cycle to cycle. So, the time period of every cycle is different. The cycles can be distinguished from one another using the send-receive packets method [21].
    - **Machine learning:** The techniques of machine learning can be classified into:

1. Genetic Algorithm (GA) is a method of the adaptive search in the class of an evolutionary computation utilizing techniques inspired from the convolutional biological process [22].
2. Artificial NN: It consists of the elements of information processing known to mimic neurons of a brain. ANN is categorized into supervised and unsupervised learning. In supervised learning, a neural network is provided together within a labelled set of the training which learns the mapping from the inputs to the outputs and given the labelled set of the pairs of inputs-output. Also, there are several kinds of algorithms of supervised learning such as Multilayer Perceptron (MLP) which is the feedforward neural network, and Radial Basis Function (RBF) which is another neural network of a feed-forward. In unsupervised learning, a neural network is provided only together with the input data without conceptualizing an output. It discovers the patterns within data autonomously. There are several typical kinds of unsupervised learning such as Self Organization Maps (SOMs) and Adaptive Resonance Theory (ART).
3. Support Vector Machines (SVM) is the algorithm of machine learning which learns to classify the data by utilizing the points labelled training examples which fall into one or two classes.
4. K-Nearest Neighbor (K-NN): It is known to be the non-parametric and the highly efficient in a classification.
5. Deep Network: It is conducted via the training data together with various layers in the hierarchical networks with unsupervised learning. Deep networks IDS could be categories depended on how architectures and the techniques are being utilized for generative architecture and discriminative architecture. In the generative architecture, there are several sub-classes such as Deep Boltzmann Machine (DBM) and Deep Believe

Networks (DBN). While in the discriminative architecture, the recurrent neural network and the convolutional neural network represents the sub-classes of discriminative architecture [23].

6. Decision Tree (DT): This algorithm learns and also models the set of data in problems of the classification [18].

The major advantages of Signature-based detection systems are; good detection performance with a low rate of false alarm, high detection efficiency, and strong interpretative ability. While it has many disadvantages such as; it only detects the known attacks, all detections almost based on the domain knowledge, high rate of missed alarm, and efficiency of the detection decreases within a scale of the signature database. On other hand, anomaly-based detection has several advantages which are; it detects known and unknown attacks. Also, it depends on the complexity of the model regarding the detection efficiency. While the disadvantages are; high false alarm rate of detection performance and low missed alarm rate, only feature design depends on the domain knowledge, the outputs are only detection results, and a weak interpretative ability [17].

Machine learning plays a crucial role in IDS, especially that it takes less training time compared to the algorithms of deep learning and in turn the data feed to the device in algorithms of machine learning, so the user train and evaluate the constructed system [24].

Table (2): Illustrate the Advantages and Disadvantages of Machine Learning Techniques [17, 18].

ML Technique	Advantages	Disadvantages
Genetic Algorithm	Utilizes the technique that is inspired via the process of a convolutional biological. Also, it has a capability in order to solving the optimization problems through the classification.	Gets a stuck in the local optimum (the overfitting).
Artificial NN	Ability to deal within the nonlinear data. Also, the ability of strong fitting	The overfitting. Also, it prone to become a stuck in the local optimum. The model training is a time consuming.
Support Vector Machine	The algorithm is simple to analyze mathematically. Also, all computations performed in the space by utilizing the kernels which giving it the edge to be utilized practically.	Selection of the kernel function is not straight forward. Also, slow in the training and requires more space of memory.
K-Nearest Neighbor (K-NN)	Easy to implement and it can also solve the problems of multi-class.	Slow in the training and requires a big space of memory. Computationally, it complex due to classify the test sample which involve a consideration of all the training samples.
Deep Network	Improve the performance.	The running time are often so long to meet the requirements.
Decision Tree	It has the unique structure so, it easy in order to interpret. It has no limitation to handling the high dimensional of sets of the data.	If the trees are not pruned back, then it causes the overfitting. The kind of a data should be considered when it constructing the tree (such as a categorical or as a numerical).

Machine learning is defined as the method of data analysis where systems gather and learn the knowledge from the performed tasks, and then improve their performance through utilizing that learned knowledge, which usually make machine learning provide a system with the ability to enhance the strategy of an execution. Furthermore, systems using machine learning are useful for different applications, however those systems are expensive. In a diverse application context, machine learning utilizes methods similar to the statistical data mining [14]. Traditional machine learning techniques are useful for IDS, such as Support Vector Machines, Hidden Markov Models, Neural Networks, and Fuzzy Logic [25]. The following Table (3) represents a summary of the existing approaches of IDS depended on machine learning.

Table (3): Summary of an Existing Approaches of IDS Depended on the Machine Learning

Reference	Date	ML Approach	Dataset	Features	Attacks to be Detected	Evaluation measure	Feature Selection Approach	Problem Domain
Hassan et al. [25]	2020	CNN, WDLSTM	UNSW-NB15	undetermined	DoS, Probe, U2R, R2L	Accuracy, precision, Recall, and F1-score	CNN	Hybrid Detection
Alqahtani et al. [26]	2020	BN, NB, RF, DT, RT, DTb, and ANN	KDD'99 cup	undetermined	DoS, Probe, U2R,R2L	Accuracy, precision, Recall, and F1-score	undetermined	Signature Detection
Vinayakumar et al. [27]	2019	DNNs	Benchmark IDS datasets	undetermined	Normal, Dos, Probe, R2L, U2R	Pr, F1, TPR, FPR, ROC	BoW, N-grams, Keras Embedding,	Hybrid Detection
Faker and Dogdu [28]	2019	DNN, RF, GBT	UNSW NB15, CICIDS2017	undetermined	DOS/DDOS, Probing, U2R, and R2L	Accuracy	K-means clustering	Signature Detection
Abdulhammed et al. [29]	2019	RF, Bayesian Network, LDA, and QDA	CICIDS2017	59	Brute Force Attack, HeartBleed Attack, Botnet, DoS, DDoS, Web Attack, Infiltration Attack	DR, F-score, FAR, Accuracy	Feature Selection/Extraction reduction (Auto-Encoder, PCA)	Signature Detection
Gao et al. [30]	2019	DT, RF, kNN, DNN	NSL-KDD	undetermined	DoS, Probe, U2R, R2L	Accuracy, Recall, Precision, F1-score	undetermined	Signature Detection
Belouch et al. [31]	2018	SVM, Naïve Bayes, DT, Random Forest	UNSW-NB15	All	Fuzzers, Analysis, DoS, Exploits, Backdoors, Generic, Reconnaissance, Shellcode, Worms	TP, TN, FP	SVM, Naïve Bayes,	Signature Detection and Anomaly Detection
Shah et al. [32]	2018	SVM, DT, Fuzzy Logic, BN and NB	NSA Snort IDS Alert Logs, DARPA IDS, NSL-KDD IDS	undetermined	MAC Spoofing, DNS Poisoning, IP Spoofing, SSH, FTP, Scanning Attacks, DoS, U2R, R2L, and Probing Attack	DR, FPR, DA	undetermined	Signature Detection
Othman et al. [33]	2018	SVM, SGD	KDD99	17	DoS, Probe, U2R, R2L	AUROC, AUPR	Feature selection (ChiSqSelector)	Signature Detection
Sharafaldin, et al. [34]	2018	KNN, ID3, Adaboost, Naïve-Bayes, RF, MLP, QDA	Generating a New Dataset of Intrusion Detection	undetermined	All	Precision, Recall, F-measure	undetermined	Anomaly Detection
Almseidin et al. [35]	2017	J48, RF, RT, DT, MLP, Naive Bayes and Bayes Network	KDD	undetermined	DoS, Probe, U2R,R2L	Accuracy, precision, FN, FP, TP, TN	undetermined	Signature Detection

Reference	Date	ML Approach	Dataset	Features	Attacks to be Detected	Evaluation measure	Feature Selection Approach	Problem Domain
Chowdhury et al. [36]	2016	SVM	UNSW-NB15	3	DoS, Fuzzer, Analysis, etc.	Accuracy, FPR, FNR	Simulated Annealing	Signature Detection
Amoli et al. [37]	2016	DBSCAN	ISCX	All	DoS, DDoS Probe	Accuracy, Precision, Recall, FPR, TNR,	undetermined	Signature Detection
Moustafa et al. [11]	2015	Decision Tree	UNSW-15	undetermined	9 attacks	Accuracy	undetermined	Hybrid Detection
Lin et al. [38]	2015	CANN, k-NN	KDD-Cup 99	6, 19	DoS, Probe, U2R, R2L	Accuracy, DR, False alarm	FEx(CANN)/Feature Selection (PCA)	Anomaly detection
Yassin et al. [39]	2013	K-means, NB	ISCX	All	DoS, DDoS, brute force SSH	Accuracy, detection rate false alarms	undetermined	Hybrid Detection

Although the machine learning techniques made great moves in the intrusion detection field, there are several challenges that have been still existed such as[17]:

1. The lack of ready-made datasets where the diffuse dataset is currently KDD99, which has many problems, and new datasets are required. However, constructing new datasets depends on expert knowledge, while the cost of labor is still high. Furthermore, variability of the Internet environment intensifies a shortage of dataset. Also, there are new kinds of attacks that are emerging and some of the existing datasets are so old to reflect the new attacks.
2. Low detection accuracy in the actual environments which means that the methods of machine learning have a certain ability to detect the intrusions, but often they do not perform completely good on the unfamiliar data. In consequence, when the dataset does not cover all the typical samples of a real-world, the perfect performance in an actual environment is not guaranteed, even that if models are achieving high accuracy on the test sets.
3. A low efficiency: utmost studies are emphasizing a detection result. So, it commonly employs complicated models and extensive methods of data preprocessing which leading to low efficiency. However, to reduce the harm as much as possible, the IDSs need to detect the attacks in real-time. So, a trade-off exists between the effectiveness and the efficiency.

2.3. Intrusion Prevention Techniques

Prevention is stopping all intrusions before happening, which means that any intrusion technique can defend against the targeted intrusion [5]. Intrusion prevention is defined as the process of intercepting the detected attacks in real–time and preventing them to continue sending data to their intended destinations. Intrusion (attack) is either intentional or illegal access to the data and information, modifying it, and/or making the system unusable. Computer security techniques are useful in detecting and preventing any unauthorized access to the computer system. Both hardware and software data should be protected from destruction and interruption as well [40]. The intrusion prevention system (IPS) is the software or hardware with the capabilities of IDS, and stopping possible incidents. Different organizations use IPS to protect data and network against attacks, embedding encryption and machine learning intrusions, and programmable logic controllers [41]. IPS responds to the detected threat as follows [42]:

- Reconfiguration of additional security controls in the system such as firewalls or routers blocking the future attacks.
- Removing malicious content of the attack in network traffic filtering out the packets of threat.
- Configuring or reconfiguring additional security and privacy controls in the browser settings preventing future attacks.

There are serious breaches of security and privacy that still occur on daily basis, which create the absolute necessity to provide the systems with information security through firewalls, intrusion detection systems and intrusion prevention systems (IDSs/IPSs), authentication, encryption, and other hardware and software solutions [42].

Additional accurate and further comprehensive performance in the operation of detection and a prevention of the malicious acts use technologies such as intrusion detection/prevention system (IDPS). There are four fundamental categories of IDPS; host, network, network behavior analysis (NBA–based), and wireless–based, and each category offers certain capabilities of logging, information gathering, intrusion detection, and intrusion prevention [43]

Host-Based monitors properties of a single host and events occurring within the host for suspicious activity. Host-based IDPs provide a variety of security capabilities [5]. Also, it typically performs extensive logging of the data related to the detected events. In addition, it has the capability to detect many types of malicious activities. But the limitations of the host-based are; alert the generation delays, the centralized reporting delays, and conflicts with existing security controls. Network-Based monitors network traffic for particular network segments or devices and analyze the network and application protocol activity to identify suspicious activity. Network Behavior Analysis (NBA) examines the network traffic to identify the threats which generate unusual flows of traffic, such as the distributed denial of the service (DDoS) attacks. NBA technologies have the capability of detecting many types of malicious activity. Also, it offers extensive information-gathering capabilities. Furthermore, they are so accurate in detecting the attacks that generate large amounts of network activity in a short time period and attacks that have unusual flow pattern. However, a few products of NBA offer a limited capability of signature-based detection. Besides, the delay in detecting the attacks [43]. Wireless monitors the wireless network traffic and analyzes its wireless networking protocols to identify suspicious activities that involving protocols themselves [5]. Wireless IDPs can detect attacks, misconfigurations, and policy violations. Also, wireless IDPS generally is accurate because of the limited scope (an analyzing wireless networking protocols). Although the wireless IDPSs offer robust capabilities of the detection, they do have several significant limitations such as they cannot detect certain kinds of attacks against the wireless networks, they utilize evasion techniques, and the wireless sensors of IDPs are susceptible to an attack [43].

To prevent the intrusion, there are distinct techniques of intrusion prevention such as encryption, access control, authentication, secure routing, which are presented as the first cross of defense against intrusions [5]. Compared to signature–based and anomaly–based, IDPS provide faster execution although with the high false positive rate. Patterns of intrusion and normal patterns do not always comply with distributions; they are not linearly separable, which becomes problematic when IDPS applies statistical learning methods for intrusion detection. Lack of knowledge of packet payload causes poor performance of anomaly–based detection on application level, while in signature–based, an intense and strong analysis has been done on the payload of packet to extract unique signatures. Signature–based approach provides a high accuracy on existing attacks [44].

In general, there are various challenges that faced the IDPs such as detection accuracy which is more challenging in the host-based IDPs due to the variety of the possible detection techniques; Such as, log analysis and file-system monitoring which do not have the knowledge of a context under a detected event that occurred. Also, the analysis of the huge amount of data could be challenging in IDPs. Moreover, the performance is extremely subject to the configuration and the tuning of each kind of product. Despite the testing can be performed via utilizing the default settings of every product. Several products are designed within an assumption which they will need

extensive customization and tuning. So, the performance of IDPs products could be challenging. Quantifying the costs of the life cycle for the IDPs solutions may be difficult due to the distinct factors of environment-specific which impact the cost, and because usually, it is challenging to capture benefits of the cost provided via the IDPs. Moreover, IDPs have challenges in performing IDPs product testing as a part of the evaluation [43].

3. Big data

Big Data is the huge and heterogeneous data being structured, unstructured, and semi-structured data [24]. Big data focuses on technological advancements related to gathering, storing, analyzing, and application of data in real life. Big data is usually characterized by volume, variety, veracity, value, volatility, and complexity. Big data huge; terabytes and petabytes, and its production is extremely fast from multiple resources which collect data from sensors, social media and/or the website's metadata [45]. Big data needs to be accurate and of high quality to create business value [46].

Big data complexity comes from the dynamic relationship among datasets, where a change in the dataset changes the other datasets. The increasing data availability has resulted in innovative advanced technologies used to analyze and provide business data value; such as, advanced analytics, machine learning, and artificial intelligence tools [45]. Since the scope of big data increases the available attack points, it is more difficult to detect the intrusions in big data environment [25]. Big data systems should have the techniques, tools, and mechanisms to load data, deal with data, and extract knowledge from data. Big data systems use the power of parallel computing to complete the analysis of data and composite operations and in turn improve data processing.

A complex big data system – for storing, processing, and analyzing big data – is technically challenging. It is difficult to collect and integrate data from distributed locations because of various autonomous data resources and heterogeneous and massive capacities, hence the need for a system which can handle processing big data. This system should support accountability, reliability, accessibility, retrievability, scalability, and confidentiality. It should mine prudence from huge datasets efficiently at various stages in real or near to real time. These systems characteristics are improved to enhance decision making and gain more advantages [47]. Technologies and tools store, manage, deal, and analyze big data, and provide support to big data techniques. These technologies include – but it not limited to [48]:

- ***Hadoop****:* is the open–source software framework to process big data on the distributed systems for a given problem. This framework stores, deals, manages, and also analyzes hundreds of terabytes and even the petabytes of data [48].
- ***Spark****:* is an open–source framework for big data that is faster than Hadoop (it is 100 times faster in memory and 10 times faster in access to disk). It reduces the iterations of the read/write from a disk and stores the intermediate data within a memory [49].
- ***Python and R****:* is the open–source programming language and software environment which supports the statistical computing and graphics. This software is significantly important as a tool for computational statistics, visualization, and data science [48].

3.1. Relational database

The relationship database (RDB) is a collective set of multiple data sets organized by tables, records and columns, forming a relationship between database tables. Tables communicate and share information that facilitates the search, organization and reporting of data. Structured Query Language (SQL) is a standard user application that offers an easy database interaction programming interface, and it is a data query language originally developed by IBM in 1970 to define and manipulate the data. This language is embedded in all relational database management systems (RDBMS). The SQL language manages authorization, integrity, constraints, views, and concurrency control. It

handles embedded SQL and dynamic SQL as well. The standard set of database system properties includes atomicity, consistency, isolation, and durability; often referred to as ACID which assures the database recovery from potential failures occurring during processing the transaction [50]. SQL databases are developed for data integrity and concurrency control, representing the best option for banking and insurance systems. Data integrity consumes high processing power, reaching the limits of relational databases when it comes to handling data of large volumes [45].

3.2. NoSQL databases

NoSQL databases are developed to enable data insertion without predefined schema. NoSQL systems are non–relational distributed databases designed for storing large data and processing them in parallel with multiple commodities. These systems use the SQL and other mechanisms to interact with data, usually through APIs which interpret SQL statements to systems native query language. NoSQL systems support heterogeneous data processing and analysis activities such as predictive and exploratory analytics, data ETL (Extraction, Transform, and Load), and OLTP (Online Transactional Processing). NoSQL systems are – contrary to traditional DBMS and Data Warehouses (DWHs) –highly scalable and can scale thousands, or even millions of users perform read/write operations. They are the best fit for organizations collecting large amounts of data, and offer increased stability through commodity hardware and focusing on analytical processing of large quantity datasets. NoSQL databases do not require consistency to fulfill higher partitioning and availability, which are grouped in property known as the BASE [51].

BASE ensures NoSQL database reliability despite the loss of consistency. Various applications use NoSQL databases such as embedded IR, parallel data processing through distributed systems, explorative analytics on unstructured and semi–structured data, large–scale data storage, social networks, search engines, nuclear modeling, geospatial analysis, data warehouses, and caching [45].

4. Distributed intrusion detections systems in big data

In distributed systems, both network and host systems with multiple devices, software, platforms, sensors, and other sources deliver services to users and organizations at any time and everywhere. Furthermore, these systems monitor organizations demands through analytical mechanisms and big data tools to carefully support and distinguish between normal and anomalous incidences, hence, capturing and processing big data are increasing dramatically in terms of volume, velocity, and variety [52].

Analyzing flows of the network, the logs, and the system events which have been utilized to detect any intrusion. Network flows, logs, and system events generate big data [53]. Conventional technologies do not support long–term and large–scale analytics because retaining operation of large data volume is usually neither economical nor feasible. Event logs and other recorded activities are generally deleted after periods of fixed retention. It is not efficient to conduct analysis and complex queries on unstructured and large datasets within noisy and deficient data. Big data applications assist in preparing, cleaning, and querying the heterogeneous data with incomplete and/or the noisy records, and in turn a part of the security management software, since they [54]. Big data analytics correlates sources of multiple information into the coherent view, identifies anomalies and suspicious activities, then achieves intrusion detection effectively and efficiently. Big data analytics sifts big data quicker for more efficient and effective diverse systems [53].

Big data analytics has been gaining more attention in terms of intrusion detection and mitigating network security problems, through; studying data of high complexity and volume, diverse formats, and from diverse resources; anomaly detection, and combating cyber–attacks. Ultra–high dimension models of data create an accurate profile of online data streams which helps to predict and detect intrusions and attacks in real–time. Big data technologies such as Hadoop eco–system and stream processing

store and analyze large, diverse and heterogeneous datasets at a high speed, transforming the security analytics via: (1) capturing large data scales from different number of internal and the external resources such as vulnerability databases; (2) conducting deep data analytics; (3) presenting the integrated view of security–related information; (4) real–time analysis of the data stream. The tools of big data analytics need to be properly configured, and system analysts and the architects should have accurate knowledge of their systems [55]. Big data classical techniques are traditionally referred to as the field of end–to–end or request–response solutions. While intrusion detection systems require online analysis, they use methods of data–flow computing of incoming data [56].

In distributed systems, the components of IDS are distributed amongst diverse physical locations. For example, in wireless sensor networks (WSN) and internet of things (IoT), components of the IDS system can be placed in many architectures such as a hierarchal architecture, since the processed data moved up in layers. IDS can be independently run on every node through stand–alone architecture. IDS operations are carried out in diverse processing platforms including PC, the cloud, network device, server, mobile device, and IoT device [7]. The following Table (4) presents a summary of existing intrusion detection techniques in distributed systems.

5. CONCLUSIONS

Security represents the most important parameter in the distributed and big data systems, and it cannot be ignored. Currently, intrusion detection represents one of the most important security issues in the cyber–world.

Table (4): Summary for Existing Techniques of Intrusion Detection in Distributed Systems

Reference	Date of publication	Dataset	Technique	Advantage	Limitation
Zhong et al. [57]	2020	DARPA1998, ISCX2012 and CICIDS2017	DT, SVM, CNN, RNN-CNN, BDHDLS	It shows that big data techniques and the parallel strategies for feature selection, clustering, and training significantly reduce the model construction time	BDHDLS uses more computational resources to achieve performance gains
GAO et al. [58]	2019	NSL-KDD and UNSW-NB15	Random Forest (RF)	good results in accuracy and false positive rate (FPR)	Lower rate of PAR because KDD dataset contains some unknown DoS attacks that increases detection difficulty
Idhammad et al. [59]	2018	CIDDS-001	Naive Bayes model	High detection performances for several attack types	Did not making decisions regarding to the specific intrusions
Dahiya, and Devesh [60]	2018	UNSW-NB 15	Canonical Correlation Analysis (CCA) and Linear Discriminant Analysis (LDA), Naïve Bayes, Random Forest	High accuracy using Random Tree and Random Committee, Accuracy is also improved by using feature reduction methods	Accuracy of Naïve Bayes was not good
Moustafa et al. [52]	2017	NSL-KDD and UNSW-NB15	DMM-based ADS technique	Help in choosing best model to identify the attacks as outliers	No decision making regarding to specific intrusions
Vimalkumar and Radhika [61]	2017	Synchrophasor dataset	Deep neural networks, support vector machines, decision trees, naive bayes algorithm and random forest	Improves detection time rate	Evaluation using a small collected dataset
Pan et al. [62]	2014	synchrophasor dataset	Bayesian networks	Provides better accuracy than the support vector machines classifier	support only for the static dataset

So, there is a significant number of techniques that have been developed depending on machine learning approaches. However, they did not succeed in identifying all intrusions types, but there are many attempts to detect and prevent intrusions from causing damages in the distributed systems.

This paper presents a detailed investigation of intrusion detection and machine learning techniques. Moreover, it presents the problems associated with various machine learning techniques for intrusion detection activities in distributed systems and big data.

References

[1] Ahmed W, Wu YW. A survey on reliability in distributed systems. Journal of Computer and System Sciences. 2013;79(8):1243-55. DOI: https://doi.org/10.1016/j.jcss.2013.02.006

[2] Abraham A, Jain R, Thomas J, Han SY. D-SCIDS: Distributed soft computing intrusion detection system. Journal of Network and Computer Applications. 2007;30(1):81-98. DOI: https://doi.org/10.1016/j.jnca.2005.06.001

[3] Jones AK, Sielken RS. Computer system intrusion detection: A survey. Citeseer; 2000. DOI: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.24.7802&rep=rep1&type=pdf

[4] Ning P, Jajodia S. Intrusion detection techniques. The Internet Encyclopedia. 2004;2:355-67.DOI: https://doi.org/10.1002/047148296x.tie097

[5] Butun I, Morgera SD, Sankar R. A survey of intrusion detection systems in wireless sensor networks. IEEE communications surveys & tutorials. 2013;16(1):266-82. DOI: https://doi.org/10.1109/SURV.2013.050113.00191

[6] andala S, Ngadi MA, Abdullah AH. A survey on MANET intrusion detection. International Journal of Computer Science and Security. 2007;2(1):1-11. DOI: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.224.3523&rep=rep1&type=pdf

[7] Aldweesh A, Derhab A, Emam AZ. Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues. Knowledge-Based Systems. 2020;189:105124.DOI: https://doi.org/10.1016/j.knosys.2019.105124

[8] Lunt TF. A survey of intrusion detection techniques. Computers & Security. 1993;12(4):405-18.DOI: https://doi.org/10.1016/0167-4048(93)90029-5

[9] Shah AA, Hayat MS, Awan MD. Analysis of machine learning techniques for intrusion detection system: a review. 2015. DOI: https://doi.org/10.5120/21047-3678

[10] Bagui S, Kalaimannan E, Bagui S, Nandi D, Pinto A. Using machine learning techniques to identify rare cyber‐attacks on the UNSW‐NB15 dataset. Security and Privacy. 2019;2(6):e91. DOI: https://doi.org/10.1002/spy2.91

[11] Moustafa N, Slay J, editors. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). 2015 military communications and information systems conference (MilCIS); 2015: IEEE.DOI: https://doi.org/10.1109/MilCIS.2015.7348942

[12] 1. Samrin R, Vasumathi D, editors. Review on anomaly based network intrusion detection system. 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT); 2017: IEEE. DOI: https://doi.org/10.1109/ICEECCOT.2017.8284655

[13] Khraisat A, Gondal I, Vamplew P, Kamruzzaman J. Survey of intrusion detection systems: techniques, datasets and challenges. Cybersecurity. 2019;2(1):1-22. DOI: https://doi.org/10.1186/s42400-019-0038-7

[14] Jose S, Malathi D, Reddy B, Jayaseeli D, editors. A survey on anomaly based host intrusion detection system. Journal of Physics: Conference Series; 2018: IOP Publishing. DOI: https://doi.org/10.1088/1742-6596/1000/1/012049

[15] Jyothsna V, Prasad R, Prasad KM. A review of anomaly based intrusion detection systems. International Journal of Computer Applications. 2011;28(7):26-35.DOI: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.259.1390&rep=rep1&type=pdf

[16] Mishra P, Varadharajan V, Tupakula U, Pilli ES. A detailed investigation and analysis of using machine learning techniques for intrusion detection. IEEE Communications Surveys & Tutorials. 2018;21(1):686-728. DOI: https://doi.org/10.1109/COMST.2018.2847722

[17] Mishra P, Varadharajan V, Tupakula U, Pilli ES. A detailed investigation and analysis of using machine learning techniques for intrusion detection. IEEE Communications Surveys & Tutorials. 2018;21(1):686-728. DOI: https://doi.org/10.1109/COMST.2018.2847722

[18] Mishra P, Varadharajan V, Tupakula U, Pilli ES. A detailed investigation and analysis of using machine learning techniques for intrusion detection. IEEE Communications Surveys & Tutorials. 2018;21(1):686-728. DOI: https://doi.org/10.1109/COMST.2018.2847722

[19]Liao H-J, Lin C-HR, Lin Y-C, Tung K-Y. Intrusion detection system: A comprehensive review. Journal of Network and Computer Applications. 2013;36(1):16-24. DOI: https://doi.org/10.1016/j.jnca.2012.09.004

[20] Debar H, Dacier M, Wespi A, editors. A revised taxonomy for intrusion-detection systems. Annales des télécommunications; 2000: Springer.DOI: https://doi.org/10.1007/bf02994844

[21] Khatkhate AM. Symbol time series analysis (STSA) for network event/intrusion detection.

[22] Eid HFAM. Computational Intelligence in Intrusion Detection System: MSc Thesis, Al-Azhar University; 2013. DOI: https://scholar.cu.edu.eg/sites/default/files/abo/files/phd_thesis_computational_intelligence_in_intrusion_detection_system_2013.pdf

[23] Deng L. A tutorial survey of architectures, algorithms, and applications for deep learning. APSIPA Transactions on Signal and Information Processing. 2014;3. DOI: https://doi.org/10.1017/atsip.2013.9

[24] Saravanan S, editor Performance evaluation of classification algorithms in the design of Apache Spark based intrusion detection system. 2020 5th International Conference on Communication and Electronics Systems (ICCES); 2020: IEEE. DOI: https://doi.org/10.1109/ICCES48766.2020.9138066

[25]Hassan MM, Gumaei A, Alsanad A, Alrubaian M, Fortino G. A hybrid deep learning model for efficient intrusion detection in big data environment. Information Sciences. 2020;513:386-96. DOI: https://doi.org/10.1016/j.ins.2019.10.069

[26] Alqahtani H, Sarker IH, Kalim A, Hossain SMM, Ikhlaq S, Hossain S, editors. Cyber intrusion detection using machine learning classification techniques. International Conference on Computing Science, Communication and Security; 2020: Springer. DOI: https://doi.org/10.1007/978-981-15-6648-6_10

[27] Vinayakumar R, Alazab M, Soman K, Poornachandran P, Al-Nemrat A, Venkatraman S. Deep learning approach for intelligent intrusion detection system. IEEE Access. 2019;7:41525-50. DOI: https://doi.org/10.1109/ACCESS.2019.2895334

[28] Faker O, Dogdu E, editors. Intrusion detection using big data and deep learning techniques. Proceedings of the 2019 ACM Southeast Conference; 2019. DOI: https://doi.org/10.1145/3299815.3314439

[29] Abdulhammed R, Musafer H, Alessa A, Faezipour M, Abuzneid A. Features dimensionality reduction approaches for machine learning based network intrusion detection. Electronics. 2019;8(3):322. DOI: https://doi.org/10.3390/electronics8030322

[30] Gao X, Shan C, Hu C, Niu Z, Liu Z. An adaptive ensemble machine learning model for intrusion detection. IEEE Access. 2019;7:82512-21. DOI: https://doi.org/10.1109/ACCESS.2019.2923640

[31] Belouch M, El Hadaj S, Idhammad M. Performance evaluation of intrusion detection based on machine learning using Apache Spark. Procedia Computer Science. 2018;127:1-6. DOI: https://doi.org/10.1016/j.procs.2018.01.091

[32] Shah SAR, Issac B. Performance comparison of intrusion detection systems and application of machine learning to Snort system. Future Generation Computer Systems. 2018;80:157-70. DOI: https://doi.org/10.1016/j.future.2017.10.016

[33] Othman SM, Ba-Alwi FM, Alsohybe NT, Al-Hashida AY. Intrusion detection model using machine learning algorithm on Big Data environment. Journal of Big Data. 2018;5(1):1-12. DOI: https://doi.org/10.1186/s40537-018-0145-4

[34] Sharafaldin I, Lashkari AH, Ghorbani AA. Toward generating a new intrusion detection dataset and intrusion traffic characterization. ICISSp. 2018;1:108-16. DOI: https://doi.org/10.5220/0006639801080116

[35] Almseidin M, Alzubi M, Kovacs S, Alkasassbeh M, editors. Evaluation of machine learning algorithms for intrusion detection system. 2017 IEEE 15th International

Symposium on Intelligent Systems and Informatics (SISY); 2017: IEEE. DOI: https://doi.org/10.1109/SISY.2017.8080566

[36] Chowdhury MN, Ferens K, Ferens M, editors. Network intrusion detection using machine learning. Proceedings of the International Conference on Security and Management (SAM); 2016: The Steering Committee of The World Congress in Computer Science, Computer. DOI: https://www.proquest.com/conference-papers-proceedings/network-intrusion-detection-using-machine/docview/1807002945/se-2?accountid=201395

[37] Amoli PV, Hamalainen T, David G, Zolotukhin M, Mirzamohammad M. Unsupervised network intrusion detection systems for zero-day fast-spreading attacks and botnets. JDCTA (International Journal of Digital Content Technology and its Applications. 2016;10(2):1-13. DOI: http://users.jyu.fi/~pavahdan/Unsupervised%20Network%20Intrusion%20Detection%20Systems%20for%20Zero-Day%20Fast-Spreading%20Attacks%20and%20Botnets.pdf

[38] 1. Lin W-C, Ke S-W, Tsai C-F. CANN: An intrusion detection system based on combining cluster centers and nearest neighbors. Knowledge-based systems. 2015;78:13-21. DOI: https://doi.org/10.1016/j.knosys.2015.01.009

[39] Yassin W, Udzir NI, Muda Z, Sulaiman MN, editors. Anomaly-based intrusion detection through k-means clustering and naives bayes classification. Proc 4th Int Conf Comput Informatics, ICOCI; 2013. DOI: http://icoci.cms.net.my/PROCEEDINGS/2013/PDF/PID49.pdf

[40] Das S, Nene MJ, editors. A survey on types of machine learning techniques in intrusion prevention systems. 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET); 2017: IEEE. DOI: https://doi.org/10.1109/WiSPNET.2017.8300169

[41] 1. Alves T, Das R, Morris T. Embedding encryption and machine learning intrusion prevention systems on programmable logic controllers. IEEE Embedded Systems Letters. 2018;10(3):99-102. DOI: https://doi.org/10.1109/LES.2018.2823906

[42] Patel A, Qassim Q, Wills C. A survey of intrusion detection and prevention systems. Information Management & Computer Security. 2010. DOI: https://doi.org/10.1108/09685221011079199

[43] Scarfone K, Mell P. Guide to intrusion detection and prevention systems (idps). NIST special publication. 2007;800(2007):94. DOI: https://doi.org/10.6028/nist.sp.800-94

[44] Wang L. Big Data in intrusion detection systems and intrusion prevention systems. Journal of Computer Networks. 2017;4(1):48-55. DOI: https://doi.org/10.12691/jcn-4-1-5

[45] ElDahshan KA, AlHabshy AA, Abutaleb GE. Data in the time of COVID-19: a general methodology to select and secure a NoSQL DBMS for medical data. PeerJ Computer Science. 2020;6:e297. DOI: https://doi.org/10.7717/peerj-cs.297

[46] Siddiqui S, Gupta D. Big data process analytics: a survey. Int J Emerg Res Manag Technol. 2014;3(7):117-23.

[47] Bendre MR, Thool VR. Analytics, challenges and applications in big data environment: a survey. Journal of Management Analytics. 2016;3(3):206-39. 8. DOI: https://doi.org/10.1080/23270012.2016.1186578

[48] ElDahshan K, Mancy H, editors. HPC based Modeling, Analyzing and Forecasting of a Century of Climate Big Data. The International Congress for global Science and Technology; 2015. DOI: https://www.researchgate.net/publication/284869677_Artificial_Intelligence_and_Machine_Learning_Journal_Volume_15_Issue_1_ICGST_Delaware_USA_Dec_2015

[49] Howlett RJ, Jain LC, Adelaide MLC. Smart Innovation, Systems and Technologies 022. DOI: https://doi.org/10.1007/978-3-642-27509-8

[50] Kumar KS, Mohanavalli S, editors. A performance comparison of document oriented NoSQL databases. 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP); 2017: IEEE. DOI: https://doi.org/10.1109/ICCCSP.2017.7944071

[51] Abed AH. Recovery and concurrency challenging in big data and NoSQL database systems. International Journal of Advanced Networking and Applications. 2020;11(04):4321-9. DOI: https://doi.org/10.35444/ijana.2020.11041

[52] Moustafa N, Creech G, Slay J. Big data analytics for intrusion detection system:

Statistical decision-making using finite dirichlet mixture models. Data analytics and decision support for cybersecurity: Springer; 2017. p. 127-56. DOI https://doi.org/10.1007/978-3-319-59439-2_5

[53] Wang L, Jones R. Big data analytics for network intrusion detection: A survey. International Journal of Networks and communications. 2017;7(1):24-31. DOI: https://doi.org/10.5923/j.ijnc.20170701.03

[54] Raja MC, Rabbani MA. Big data analytics security issues in data driven information system. Int J Innov Res Comput Commun Eng. 2014;2(10):6132-5. DOI: https://www.researchgate.net/publication/267753116_Big_Data_Analytics_Security_Issues_in_Data_Driven_Information_System

[55] 1. Cardenas AA, Manadhata PK, Rajan SP. Big data analytics for security. IEEE Security & Privacy. 2013;11(6):74-6. DOI: https://doi.org/10.1109/MSP.2013.138

[56] Shterenberg S, Poltavtseva MA. A distributed intrusion detection system with protection from an internal intruder. Automatic Control and Computer Sciences. 2018;52(8):945-53. DOI: https://doi.org/10.3103/S0146411618080230

[57] Zhong W, Yu N, Ai C. Applying big data based deep learning system to intrusion detection. Big Data Mining and Analytics. 2020;3(3):181-95. DOI: https://doi.org/10.26599/BDMA.2020.9020003

[58] Gao Y, Wu H, Song B, Jin Y, Luo X, Zeng X. A distributed network intrusion detection system for distributed denial of service attacks in vehicular ad hoc network. IEEE Access. 2019;7:154560-71. DOI: https://doi.org/10.1109/ACCESS.2019.2948382

[59] Idhammad M, Afdel K, Belouch M. Distributed intrusion detection system for cloud environments based on data mining techniques. Procedia Computer Science. 2018;127:35-41. DOI: https://doi.org/10.1016/j.procs.2018.01.095

[60] Dahiya P, Srivastava DK. Network intrusion detection in big dataset using spark. Procedia computer science. 2018;132:253-62. DOI: https://doi.org/10.1016/j.procs.2018.05.169

[61] Vimalkumar K, Radhika N, editors. A big data framework for intrusion detection in smart grids using apache spark. 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI); 2017: IEEE. DOI: https://doi.org/10.1109/ICACCI.2017.8125840

[62] Adhikari U, Morris TH, Pan S, editors. A causal event graph for cyber-power system events using synchrophasor. 2014 IEEE PES General Meeting| Conference & Exposition; 2014: IEEE. DOI: https://doi.org/10.1109/PESGM.2014.6939285

أنظمة كشف الاختراق الموزعة في البيانات الضخمة: دراسة

بشار ابراهيم حميد[1]، عبدالله عادل محمد الحبشي[1]، كمال عبدالرؤوف الدهشان[1]

1. قسم الرياضيات، كلية العلوم، جامعة الأزهر، القاهرة، مصر

البريد الالكتروني : basharibh78@gmail.com

الملخص

نحن نعيش في وقت تتدفق فيه البيانات في الثانية، مما يجعل اكتشاف الاختراق في أنظمة الحاسوب أو الشبكات مهمة صعبة ومرهقة، وبالتالي تتطلب أنظمة الكشف عن الاختراق آلية كشف فعالة ومحسنة للكشف عن الأنشطة المشبوهة. علاوة على ذلك، فإن التعامل مع حجم وتعقيد وتوافر البيانات الضخمة يتطلب تقنيات يمكن أن تولد معرفة مفيدة من التدفقات الضخمة للمعلومات، مما يفرض تحديات على تصميم وإدارة كل من نظام اكتشاف الاختراق (IDS) ونظام منع التطفل (IPS) من حيث الأداء والاستدامة والأمان والموثوقية والخصوصية واستهلاك الطاقة والتسامح مع الأخطاء وقابلية التوسع والمرونة. البحث يقدم دراسة شاملة لأنظمة اكتشاف الاختراق الموزع في البيانات الضخمة، وتعرض تقنيات كشف الاختراق والوقاية التي تستخدم في التعلم الآلي، وتقنيات تحليل البيانات الضخمة في الأنظمة الموزعة لاكتشاف الاختراق.