Distributed Denial of Service Defense on Cloud Computing Based on Network Intrusion Detection System: Survey

Esraa Samkari[1†] and Hatim Alsuwat[1†],
S44380084@st.uqu.edu.sa Hssuwat@uqu.edu.sa
[1] Department of Computer Science, College of Computer and Information Systems, Umm Al-Qura University, Saudi Arabia

Summary
One type of network security breach is the availability breach, which deprives legitimate users of their right to access services. The Denial of Service (DoS) attack is one way to have this breach, whereas using the Intrusion Detection System (IDS) is the trending way to detect a DoS attack. However, building IDS has two challenges: reducing the false alert and picking up the right dataset to train the IDS model. The survey concluded, in the end, that using a real dataset such as MAWILab or some tools like ID2T that give the researcher the ability to create a custom dataset may enhance the IDS model to handle the network threats, including DoS attacks. In addition to minimizing the rate of the false alert.

Keywords:
DDoS, DoS, cloud computing attack, network intrusion detection, system detection.

1. Introduction

With the development of technology, the speed and efficiency of the network has become high in transferring data from one device to another. As a result of this progress, we get a variety of network's types, such as Local Area Networks (LAN), Wide Area Networks (WAN), cloud computing, Internet of Things (IoT), etc. Side by side, many government and commercial industries, such as healthcare, defense, banking, and trade, are converting their manual transactions into digital and relied more on computers in their work, in terms of creating, storing, editing, and deleting the data [1].

Some services of trade industry, e.g., displaying and purchasing products, is needed to be available and accessed anytime by the end-user. These industries tend to use cloud computing to provide their services, which makes their data centralized and more vulnerable to attacks intrusions such as network viruses, eavesdropping, etc. [2]. Therefore, there are several lines of defense for networks, including: antivirus, firewall, prevent detection system, and honeypots [3]. However, in the last decade, some famous companies, Amazon and Google, which use cloud computing to offer their services to others, ware vulnerable to Denial of Service (DoS) attacks [4,5].

The DoS attack exploits the nature of the network in dealing with sending and receiving data by using protocols, such as HTTP, TCP, UDP, etc., as it crowds the data traffic and distracts the target from legitimate users. Thus, the legitimate user cannot be accessed to the service that provided from a server [6]. However, the capability of the server is high and performing the DoS attack from one computer is not enough. Therefore, the attacker starts its attack by controlling as many computers as possible, by spread viruses and worms through the internet, to slow the flow of data in the target network. This attack known as Distribute Denial of Service Attack (DDoS).

To mitigate and prevent such attacks, Intrusion Detection System (IDS) is used as one of the defense methods [1,7]. IDS is used to monitoring the traffic network in order to detect the abnormal transmit of the data [3]. It can classify either based on a Network Intrusion Detection System (NIDS) or Host Intrusion Detection System (HIDS). The most important difference between them is the location of IDS that start monitoring the network traffic. The HIDS had a single point to monitor the incoming packet on the single system, where the NIDS monitor the full network [8]. The goal in both of them is to produce an alert if they detect a suspicious transmission of the packet. The implement of the detection can be happen using two techniques: analyzing based on signature or based on behavior [3,8,9].

The IDS can detect the attack based on signature method, also known as knowledge method, which use a several pattern, that already exist on the database, to expose the attack [3]. However, this method failed on detection if incoming packet had a pattern which the method had never seen yet. To solve this problem, the behavior-based IDS method is used. Unlike the signature based, the behavior based does not require to store any pattern to detect the attack, because the method depend on the analyze the behavior of sending the packet. However, unfortunately, this method has high rate of false alert [9].

To minimize the false rating alert of behavior-based IDS, the Machine Learning (ML) and the Deep Learning (DL) have been integrated with it [10]. Different models algorithms such as Random Forest (RF), Naive Bayesian network (NB), Support Vector Machine (SVM), etc. have been used to classify the attack automatically, hence

Manuscript received June 5, 2022
Manuscript revised June 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.6.12

reduce the dependence on human effort [7]. Yet, the false alert rating still has a high value due to the traffic data labeling issue of the dataset that used in training the ML model.

The types of the dataset that available on the internet can be synthetic, emulate, or real [5]. Most of the research, nowadays, used one or more types of existing dataset to train or evaluate their model of IDS [9,11]. Where the other research builds a custom dataset using ML to increase the types of the dataset [12]. Overall, the developing model to enhance the accuracy of IDS is challenging in the research field, which many of surveys had illustrated the reasons, and we will highlight them in the next section.

There are many papers and surveys that suggest using methods to find out the anomaly behavior on the network. Since the integration of the ML in IDS is still a trending topic, some research tends to suggest one algorithm above another or advise avoiding some dataset that has a particular characteristic [5]. Therefore, the aim of this paper is to collect and classify the past research to presents a list of suitable datasets for IDS with the list of ML algorithms based on the type of intrusion attack. Thus, the contribution of the paper is as follows:

- Giving a list of the dataset names that were used to train the IDS model, which may assist other researchers in finding the most appropriate dataset for their research.
- Listing the most ML and DL algorithms used for building and enhancing the model of IDS in recent two years.
- Comparing and discussing the difference between previous IDS surveys in the past five years in terms of their methodology and analyzing the research paper.

The rest of the paper is divided as follows: Section 2 presents the recent survey that talking about the IDS. Where Section 3 illustrates the steps of this article in collecting the research papers. The Section 4 focuses on the previous related work that purposed a method to enhance the IDS using different traffic dataset. Section 5 discusses the strength and weaknesses of the previous work section with recommendations for improving the IDS. Finally, the Section 6 summaries and give the conclusion of the paper.

2. PAST SURVEYS

Many surveys have been conducted over the last five years to gather, evaluate, and analyze academic papers that discuss the challenges and solutions of network intrusion detection as shown in table 1. With the huge amount of the transmit packets on the network, most of the research has focused on the use of ML or enhancing the dataset. Therefore, we find the surveys in table 1 focus on ML or DL algorithms.

Hao [13] presented the problem of network security in intrusion detection, including the false alert rate and detection time rate with their solutions by reviewing several papers talking about improving the performance of the detection system using ML. In addition, the paper explained the differences between the three common datasets, which are KDD CUP 99, DARPA 1998, and IDS2018. Zichuan et al. [14] looked into how most of the data mining methods could be used to utilized to detect intrusions. In addition, they classify the papers based on the detection accuracy among of different intrusion types of attacks including U2R R2L, DOS, and Probe. The similarity points of the two papers were in explaining the core types of algorithms that use in ML including Decision Tree, Artificial Neural Network, Fuzzy Clustering, K-means clustering, Support Vector Machine, and so on, with a clarification of the weaknesses point of each of them in the field of IDS. However, Hao's paper forced on Apache Spark framework.

Mohamed et al. [15] divided the studies into two categories, alert processing techniques, and detection techniques. Moreover, they illustrate the four possible alerts of IDS events. After investigating many papers based on specific parameters, the research found the Support Vector Machine algorithm has the best result in reducing the false alert rate. Unlike the previous survey, which taxonomic studies based on custom structure, Arun and Satish [16] collected and analyze the studies between 2012 and 2018 that are related to intrusion detection, then they identified the studies' techniques with its measurement. However, they end up with two metrics, the Manhattan Distance and Euclidean Distance. In the end, both surveys make the agreement that using ML algorithms with carefully selecting the dataset is the solution to improve the false alert rate.

Among the previous surveys, the current survey of Dylan and Meng [5] had covered a wide area of NIDS techniques from 1999 to 2021. They taxonomic the studies based on the intrusion detection challenges that are related to data driven. These challenges illustrated the natural problem in building the models including the labeling issue in most datasets, the instances amount available of data in different network types, and the volume of redundant and noisy data that exists in the dataset. Furthermore, as they mentioned, instead of using the real dataset, most research articles employed synthetic and imitate datasets to create the IDS model. In the end they concluded by saying that future research should focus on real dataset such as UGR and LITNET to get better perform of IDS.

Each of the five previous surveys concentrates either on the ML algorithm with the dataset or on the different IDS defense techniques with their weak points. However,

this article will gather all active real data of datasets as a list then specify the most ML algorithms that had the highest rate in reducing the false alert. After that, we will present some other techniques of IDS and comparing with the ML algorithms in terms of accuracy and performance, which most of these techniques and algorithms from papers in 2021.

Table 1: Related surveys

Paper	*Title*	*Year*	*Cover*
Arun and Satish [16]	An Extensive Survey on Intrusion Detection- Past, Present, Future	2018	The numerous studies on intrusion detection and their techniques were discussed
Zichuan et al. [14]	Survey of Intrusion Detection Methods Based on Data Mining Algorithms	2019	Measurement of the performance of machine learning algorithms based on the ability to detect the different type of intrusion attack
Mohamed et al. [15]	A survey and taxonomy of techniques used for alerts of Intrusion Detection Systems	2019	Categorize the research based on detect techniques and alert processing techniques
Dylan and Meng [5]	A Survey on Data-driven Network Intrusion Detection	2021	Taxonomy the IDS challenging of cloud environment into eight phases based on anomaly data driven
Hao [13]	A Survey on Machine Learning based Intrusion Detection Systems Using Apache Spark	2021	Illustrate the different ML algorithms that used for IDS and how they were used on the Apache Spark framework

3. STEPS OF COLLECTING THE PAPERS

This article followed a non-systematic method in collecting the other articles that related to network attacks, specifically the DoS attacks. However, the process of collecting articles was not random, but rather followed specific criteria.

First, all papers must be recent research, 1-3 years ago, and have a unified topic with a diverse methodology for solving the network attack. Second, the title of the article must contain one of these words or phrases: DoS, DDoS, cloud defense, network intrusion detection, attack detection, and anomaly detection. Third, the abstract must have a clear statement that defines the type of network attack problem with the proposed solution.

The result shows the shifting in most research in following specific methodology, where they rely on ML and DL to detect network attacks using NIDS. Furthermore, they did not focus on one type of attack, e.g., DoS, due to the ability of NIDS to detect the multi kinds of network attacks. In addition, they also did not specify the network types e.g., IoT or cloud.

Therefore, these articles follow the same steps methodology, starting from selecting the dataset and ending with building and evaluating the models. Figure 1 shows these steps in order.

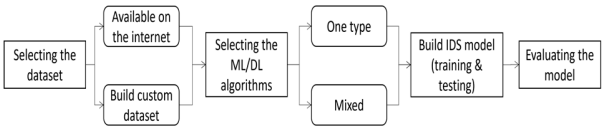

Fig. 1 The steps of build IDS model.

4. NETWORK TRAFFIC DETECTION BY AN INTRUSION DETECTION SYSTEM

4.1. DATASET

Building NIDSs nowadays are heavily relying on the dataset [29]. Some of these datasets are available on the internet with PCAP (Packet Capture) file extension [5]. This type of file contains a collection of network packets, and these packets can be binary classified, normal or abnormal [12]. The abnormal packet is a malicious packet, which can be PROBE, Botnet, DoS, R2L, Port Scan, or other network attacks [10, 17]. Many an open-source datasets can be access through the internet, where in this subsection we illustrated some of them and the table 2 gave the summary about them.

Table 2: Comparison between famous datasets

Dataset name	*Year*	*Is data synthetic?*	*Packet attack type*	*Amount*
DARPA 98-99 [13]	1998 - 1999	Yes	R2L U2R DoS Probe	3.5M
KDD Cup 99 [8]	1999	Yes	DoS Probe R2L U2R	4.8M
NSL-KDD [8]	2009	Yes	DoS Probe R2L U2R	125K
UNSW_NB15 [18]	2015	Yes	Analysis DoS Generic Reconnaissance Fuzzers Backdoor Exploits	2.5M

CICIDS 2017 [5]	2017	No	DoS DDoS Botnet Brute-force Web attacks	2.8M
CSE-CICIDS2018 [5]	2018	No	Not defined	4.5M
MAWILab [12]	1999 – now	No	DoS Port Scanning	Not defined

4.1.1 *DARPA 98 & 99*. The Defence Advanced Research Project Agency (DARPA) dataset was created in 1998 using emulated network environment at the Massachusetts Institute of Technology Lincoln Laboratory [8]. The dataset has 3.5 million instance of data, which some of them are benign and the other are malicious. The malicious packets can be classified into four types, and they are: R2L, U2R, DoS, and Probe. However, the problem with this dataset is that it needed more steps when applied with classic ML models [13].

4.1.2 *KDD Cup 99*. It was created in 1999 and is an extension of the DARPA dataset. The KDD Cup 99 had around 4.8 million records [8]. However, it suffers from duplicate values and has the same packet attack type as the DARPA dataset, i.e., Dos, Probe, R2L, and U2R. Yet, it considers one of the most famous datasets used by researchers in the field of cybersecurity [5,29]. In addition, this dataset is an improved version of the one before it, DARPA, which can be easily used with the ML/DL algorithm [8].

4.1.3 *NSL-KDD 2009*. The NSL-KDD dataset was created in 2009, which is the better version of KDD Cup 99 that remove the redundant records from the training and testing samples [5]. Thus, the size of the dataset gets reduced to 125 thousand records. The dataset has not any new type of packet attack and has the same attack types as the KDD Cup 99 dataset (DoS, Probe, U2R, and R2L). Therefore, it is suitable to use it as a whole dataset instead of taking part in it. However, the data is old and does not fit the modern attack types [8].

4.1.4 *UNSW_NB15*. In 2015, the *UNSW_NB15* dataset has been created in the Cyber Range Lab by Dr. Nour Moustafa [8, 5]. He used a traffic generator, named IXIA, to fill the dataset with normal and abnormal packets. He classified the abnormal packets into seven classes, which are DoS, Analysis, Generic, Reconnaissance, Fuzzers, Backdoor, and Exploits [18]. The total of the data was above two million.

4.1.5 *CICIDS 2017*. The Canadian Institute of Cybersecurity (CIC) build the dataset of CIC based on record the behavior of 25 users on the network using B-profile system [5]. The dataset contains a normal behavior of application network protocol including HTTPS, HTTP, FTP, SSH, and the email. In addition, five different scenarios have been used to act as abnormal behavior. The abnormal behavior is happened by performing several kinds of attacks on each specific day. These attacks are DoS, DDoS, Port Scan, Brute force, and Bot net.

4.1.6 *CSE-CICIDS2018*. In 2018, the collaboration was happened between CIC (Canadian Institute of Cybersecurity) and CSE (Communications Security Establishment) in building CSE-CICIDS2018 dataset [5]. Beside the using B profile to record the user behavior, they also used the M profiles to check the scenarios of the network traffic. The capturing and analyzing are happen in AWS (Amazon Web Services) cloud computing [8]. The data contain 8 types of modern attacks; however, the dataset faced some problem such as the size and the distribution of the data.

4.1.7 *MAWILab*. The network data of the MAWILab dataset was collected from 1999 until now, where the dataset every day must gather the network packets for 15 minutes [12]. In general, each packet can have one of these labels, which are notice, suspicious, benign, or anomalous. Since the data network is recorded every day, the size of the dataset is huge and cannot be defined. However, the capture of anomalies packets can be estimated, where the minimum rate of capturing anomalies packets is around 50 and the maximum rate of capturing the anomalies packets is approximate 500 packets [5]. Unfortunately, the tracing network data is happened between two endpoints, U.S. and Japan, and the dataset is not available to access [12].

4.2. MACHINE LEARNING ALGORITHMS

Capturing the malicious packets on the network based on their signature is tedious work. Therefore, most research advised using ML algorithms in order to detect abnormal behavior automatically. The ML can be classified as supervised learning, data with the label, or unsupervised learning, data alone, and the model trying to figure out the pattern. Some of the supervised machine learning algorithms that are used in building the NIDS model will be defined in this section, including SVM, NB, DT, and RF.

4.2.1 *Support Vector Machine (SVM).* When a line separate entirely between two different classes is known as linear two-dimension space, which also is one type of SVM algorithm. Another type of SVM known as non-linear p-dimension space, where p is the number of different classes. In this type, an arc (or radius) it used instead of line to classify between the various of data by Kernel trick [14].

4.2.2 *Naive Bayes (NB).* The name of Naive Bayes (NB) came from Bayes' theorem, which heavily depends on prior probability. NB is one of the pattern classification methods of supervised ML. In addition, the Bayesian Networks, Gaussian Naive Bayes, Hidden Naive Bayes,

and Multinomial Naive Bayes are considered other models of NB [13, 5, 19].

4.2.3 *Decision Tree (DT).* It is one of the classification methods of supervised ML that is used to reduce the number of features [1]. The DT algorithm is represented as node and leaf or as nested if/else statements [20, 13]. However, it has a problem overfitting the training data, which can be fixed by pruning some leaf nodes using the Random Forest (RF) algorithm. Some DT algorithms are C4.5 and CART [17, 20].

4.2.4 *Random Forest (RF).* It used to eliminate the overfitting problem of DT and get an accurate result. The general idea behand RF algorithm is to implement multiple DT then get the average tree in this forest [20]. However, the building time is high compared to the previous algorithm, i.e., DT.

4.3. DEEP LEARNING ALGORITHMS

The basic idea of the DL (Deep Learning) is to build a model by training and passing the data through multilayers, which adjusts the accuracy automatically. However, problems like overfitting and underfitting had a chance to happen. Therefore, it is important to consider the size and quality of the data. Some of the DL algorithms are used in building the NIDS model [21] including FNN, CNN, RNN, and AE, which will be illustrated in the following paragraphs.

4.3.1 Forward *Neural Network (FNN).* It classifies the data by passing them through multi layers, input, hidden, and output. The input layer is used to feed the neural network by passing the initial data to the hidden layer. The hidden layer can be one or more layers, which use some logistic function with activation function like sigmoid or ReLu to simplify the classification process [20]. The output layer can be two or more neurons, and the final result is ended with choosing one of them.

4.3.2 *Convolutional Neural Networks (CNN).* Usually, the CNN algorithm is used to detect and extract the pattern from several images [2]. However, it can be used to extract the spatial features from the network dataset [21]. The CNN contain an input, hidden, and output layers. It can be used to classify the data into binary classification, e.g., the NIDS model that build using CNN algorithm, can classify the network packet eighter to benign or malicious. Yet, some research combines the CNN algorithm with other DL algorithms, e.g., RNN, to utilize the best result and get spatial-temporal features [21, 2].

4.3.3 *Recurrent Neural Network (RNN).* It is a normal neural network with a short memory cell. With the memory in each neuron of RNN, the RNN process can perform in sequence in order to recognize a pattern of the input data. However, it is suffered from ”Vanishing Gradient problem”, the backpropagation value is either far above or below the zero [3]. Therefore, the Long Short-Term Memory (LSTM), one form of RNN, is used to solve the previous problem by adding another memory, a long memory.

4.3.4 *Auto Encoder (AE).* It is used to compress and reduce the data size and dimensionality by encoding function, then it used a decode function to convert the low dimensionality to high dimensionality and retrieve the original data. The data is trained as a result of that process, and the neural network's parameters are adjusted automatically using the backpropagation algorithm [17]. As a result, the NIDS model can detect an anomaly packet of the network.

4.4. INTRUSION DETECTION SYSTEM IN CLOUD COMPUTING

Nowadays, most companies, Amazon, Google, eBay, etc., rely on cloud computing to provide their services to the customer. Cloud computing offers its customers flexibility in performance expansion as their needs under pay-as-you-go service. However, the advantage of sharing the same set of resources with different clients creates a vulnerability when the attacker sends a lot of dummy requests to crowd the clients' network. And thus, the DoS attack occurred by exploiting such a feature.

In order to detect suspicious transmit in the network, like DoS, two general methods are available to use, signature-based or behavior-based. The signature-based, or sometimes called knowledge-based, is a less efficient method due to the requirement to save and memorize every attack signature on the database [24]. On the counter side, the detection based on behavior, known as detection of misuse, is an efficient method due to its ability to analyze the network transmit behavior and then detect an anomalous attitude.

Threats like data breaches, account hijacking, misuse of cloud usage, data loss, malicious insiders, and Denial of Service (DoS) are some vulnerabilities that may happen in the cloud, and there are different types of IDS that can handle them [24]. NIDS (Network Intrusion Detection System), SIDS (Signature Intrusion Detection System), HIDS (Host Intrusion Detection System), and DIDS (Distributed Intrusion Detection System) are some types of the detection systems, where all of them have the same objective, detect anomalous behavior, and send alarm [8].

The NIDS is behavior-based that analyzes the behavior of send/receive packets by monitoring the network transmit [25]. Many researchers engaged the AI (Artificial Intelligence) algorithms to build such a detection system. With the power of AI algorithms, the machine can detect the new attack network by learning and analyzing the network behavior automatically [26]. However, issues such as a high error rate or low detection accuracy may

occur. Therefore, different solutions have been proposed by different researchers to reduce these problems as explained in the next subsection.

4.5. NETWORK INTRUSION DETECTION SYSTEM BASED ON MACHINE AND DEEP LEARNING ALGORITHMS

There are a lot of papers reduce the DoS attack by building a NIDS using ML/DL algorithms. However, some articles still proposed other methods. For example, [4] proposed using the game theory as a defense mechanism against DDoS attacks on the cloud. The theory tries to help the cloud service provider make the right decision if their infrastructure is under attack and the cloud resource is under the limit.

Where some others used a particular network tool. For instance, [27] reduced the impact of the TCP SYN flood attack, type of DoS attack, by implementing the OpenFlow switch infrastructure, which analyzed incoming TCP users' packets and gave the right authorization to access the server. Nevertheless, the number of articles that are interested in ML/DL algorithms with NIDS are increased.

The papers [23,28] study the performance of reducing the false rate by implementing different ML algorithms. The [23], however, expanded its study to include the DL algorithms, which, both the ML and DL, applied to the same dataset, CICIDS 2017 and CICIDS 2018. At the end of the article's experiment, the most accurate result was obtained by the two algorithms, RF and RNN. However, the RNN gets a more accurate result.

Side by side, the papers [2,10,3] did the same experiment with the LSTM algorithm, a better version of the RNN model; however, they combined the LSTM algorithm with other types of neural network algorithms to reduce the false rate alert and extract the spatial-temporal features.

Some researchers shifting their focus on the problems of the dataset instead of improving the NIDS models. Most of the dataset that available to download are suffered from labeling the data, high volume of abnormal packets exists in the dataset without labeling. Therefore, some research, [7,18,30], proposed to use algorithms, such as ADASYN (Adaptive Synthetic Sampling), CBRS (Class Balancing Reservoir Sampling),and DPLAN (Deep Q-learning with Partially Labeled ANomalies), to solve this issue by balancing the dataset.

Another proposed solution is to focus only on analyzing the labeling of normal packets using the AE algorithm. This idea was adopted by several studies to solve these kinds of issues related to unlabeled data [20,22,17,1]. Table 3 summarizes all these researches by notifying the types of ML/DL algorithms and datasets used in their articles.

Table 3: Summary of NIDS models of different article

Ref	Year	Dataset	Algorithm	Summary
[20]	2020	NSL-KDD	AE GMM	Implementation the AE algorithm to train the IDS model using only benign data from NSL-KDD database
[2]	2020	NSL-KDD & UNSW-NB15	LSTM CNN	Extracting the temporal and spatial feature using CNN followed by RNN to enhance the accuracy of abnormal detection
[22]	2021	CICIDS 2017	AE GMM	Using two layers, GMM and AE, to detect then identify the abnormal traffic in real time
[17]	2021	KDD Cup 99	AE	Reduce the dependence on having a big dataset by only analyze the normal data to detect the abnormal data using one-dimensional AE
[1]	2021	NSL-KDD	AE	Using the AE to select a feature and polish the data instead of using ML algorithm
[23]	2021	CICIDS 2017 & CICIDS 2018	MLP SVM DT RF KNN RNN	Comparing the performance of four ML algorithms in detecting the network attacks versus the DL algorithm
[28]	2021	CIDDS-002	LR xgboost RF catboost	Applying different ML algorithms on the CIDDS part two dataset to measure their performance in intrusion detection
[10]	2021	CICIDS 2017	LSTM FNN	Indicating the pattern of abnormal data by extracting the temporal feature using LSTM and FNN
[7]	2021	CICIDS 2017	ADASYN RF	Enhancing the accuracy of IDS by balancing the dataset using ADASYN first, then apply the RF algorithm to classify the data
[3]	2021	Kyoto	RNN LSTM	Detecting the suspicious of different network attack types by exploiting the advantage of sequential of LSTM algorithm
[18]	2021	UNSW_ NB15	DPLAN	Proposing to use DPLAN algorithm to solve unbalance dataset that have a high unlabeled dataset
[30]	2021	CICIDS 2017	CNN MLP CBRS	Proposing CBRS algorithm to enhance the detection performance by balancing the dataset sample

5. DISCUSSION AND RESULT

There is no doubt that the use of Artificial Intelligence (AI) to detect malicious data in the network is an active area of research than the use of traditional techniques. With some datasets available on the internet, the algorithms of ML/DL have been applied to build a NIDS model by numerous researchers. However, each researcher faced a problem either in getting a high rate of false notifications or facing issues with the datasets that they selected.

The general purpose of IDS is to send a notification when an attack is detected. However, the NIDS model suffers from false-positive alerts, in which normal packets are classified as a threat. Therefore, several researchers suggested using one of the ML/ DL algorithms over another to reduce the false alert and enhance the performance of the detection model, as illustrated previously in Section 4.5.

Since datasets are essential for building an intrusion detection model, their data should be redundancy-free, label-related, and have a balanced distribution. However, most datasets are old, redundant, unbalanced, synthetic, and unlabeled data. Different solutions are provided by different researchers, such as using the AE algorithm to solve the unlabeled problem, as it showed in the earlier section.

After discussing the problems of building the NIDS model, this article suggests focusing on solving the datasets problem more than on selecting the suitable AI algorithms because the result of reducing the false alert by a specific algorithm depends on the quality of the dataset. However, unfortunately, most of the famous datasets are old or suffer from redundant data, such as KDD Cup 99 and CICIDS2017, which may not handle the new threats. Therefore, selecting a dataset like a MAWILab, which has the advantage of collecting a new data each day up to now, or using a tool like an ID2T (Intrusion Detection Dataset Toolkit) [12], enabled to manipulate of the properties of the existing dataset to getting a desirable dataset, is our advice.

6. CONCLUSION AND FUTURE WORK

Detecting the abnormal packets based on behavior saved an effort, however, this technique, i.e., IDS, faced two main problems, issues related to the datasets and a high rate of false notification. This survey dealt with several research that suggested solutions to solve these problems. Some research suggests using certain ML algorithms, which, based on their experience, have reduced the number of false alerts. On the other hand, some research tends to use DL algorithms, such as AE, to avoid the unnamed data problem of some datasets. Where the rest of the research, proposed using their own algorithms, e.g., ADASYN, to balance the dataset. Therefore, the survey recommends solving the datasets problems first, like selecting the datasets containing new and real data rather than old and synthetic data, to decrease the false alert rate. In order to list more names of datasets that need to be selected for building the NIDS model, more research needs to be collected and analyzed as future work.

References

[1] Gaurav, M., Babita, D., Mehul, M., Kamal, H.: *Performance Comparison of Network Intrusion Detection System Based on Different Pre-processing Methods and Deep Neural Network.* In: DSMLAI ’21’, August 9–12, 2021, Windhoek, Namibia (2021)

[2] Jay, S., Manollas, M.: *Efficient Deep CNN-BiLSTM Model for Network Intrusion Detection*. In: AIPR 2020, June 26–28, 2020, Xiamen, China (2020)

[3] Shweta, P., Meenakshi, T., Subham, G.: *Leveraging LSTM-RNN combined with SVM for Network Intrusion*. In: DSMLAI’21, August 9-12, 2021, Windhoek, Namibia (2021)

[4] Kaho, W., Joel, C.: *Game-Theoretic Modeling of DDoS Attacks in Cloud Computing*. In: UCC’21, December 6–9, 2021, Leicester, United Kingdom (2021)

[5] Dylan, C., Meng, J.: *A Survey on Data-driven Network Intrusion Detection.* In: ACM Computing Surveys, Vol. 54, No. 9, Article 182. Publication date: October 2021 (2021)

[6] Kumar, S., Debi, M.: *DDoS Detection and Defense: Client Termination Approach*. In: CUBE 2012, September 3–5, 2012, Pune, Maharashtra, India (2012)

[7] Zhewei, C., Linyue, Z., Wenwen, Y.: *ADASYN−Random Forest Based Intrusion Detection Model*. In: SPML 2021, August 18–20, 2021, Beijing, China (2021)

[8] Aouatif, A., Omar, B., Hicham, B., Abdelmajid, M.: *A Review of Intrusion Detection Systems: Datasets and machine learning methods*. In: NISS2021, April 01, 02, 2021, KENITRA, AA, Morocco (2021)

[9] Long,C., Gao, X., Zhao, J., Wan, W., Shen, H., Gao, P.: *Intrusion Detection Using End-to-End Memory Network*. In: ICCIS 2017, November 7-9, 2017, Wuhan, China (2017)

[10] Andrea, C., Shanchieh, Y., Giovanni, A.: *On the Evaluation of Sequential Machine Learning for Network Intrusion Detection*. In: ARES 2021, August 17–20, 2021, Vienna, Austria (2021)

[11] Florian, W., Felix, O., Steffen, H., Matthias, V., Mathias, F.: *Multi-Stage Attack Detection via Kill Chain State Machines*. In: CYSARM ’21, November 19, 2021, Virtual Event, Republic of Korea (2021)

[12] Carlos, G., Emmanouil, V., Aidmar, W., Max, M., Simin, N.: *On Generating Network Traffic Datasets with SyntheticAttacks for Intrusion Detection*. In: ACM Transactions on Privacy and Security, Vol. 24, No. 2, Article 8. Publication date: December 2020 (2020)

[13] Hao, L.: *A Survey on Machine Learning based Intrusion Detection Systems Using Apache Spark*. In: HPCCT’21, July 02–04, 2021, Qingdao, China (2021)

[14] Zichuan, J., Yanpeng, C., Zheng, Y.: *Survey of Intrusion Detection Methods Based on Data Mining Algorithms*. In:

BDE 2019, June 11–13, 2019, Hong Kong, Hong Kong (2019)
[15] Mohamed, A., Youness, K., El mostapha, C.: *A survey and taxonomy of techniques used for alerts of Intrusion Detection Systems*. In: BDIoT'19, October 23–24, 2019, Rabat, Morocco (2019)
[16] Arun, N., Satish, K.: *An Extensive Survey on Intrusion Detection- Past, Present, Future*. In: ICEMIS '18, June 19–20, 2018, Istanbul, Turkey (2018)
[17] Xue-Chao, S., Hai-Yan, F., Yu-Qing, C.: *Network Intrusion Detection Based on One-dimensional Convolution Layer Autoencoders*. In: ICFEICT 2021, May 21–23, 2021, Changsha, China (2021)
[18] Guansong, P., Chunhua, S., Anton, H., Longbing, C.: *Toward Deep Supervised Anomaly Detection: Reinforcement Learning from Partially Labeled Anomaly Data*. In: KDD '21, August 14–18, 2021, Virtual Event, Singapore (2021)
[19] Eva, P., Sotiris, I.: *A Survey on Encrypted Network Traffic Analysis Applications, Techniques, and Countermeasures*. In: ACM Computing Surveys, Vol. 54, No. 6, Article 123. Publication date: July 2021 (2021)
[20] Radoslava, Š., Christian, L.: *A Semi-Supervised Approach for Network Intrusion Detection*. In: ARES 2020, August 25–28, 2020, Virtual Event, Ireland (2020)
[21] Xiaojie, W., Laisen, N., Zhaolong, N., Lei, G., Guoyin, W., Xinbo., G., Neeraj, K.: *Deep Learning-based Network Trafic Prediction for Secure Backbone Networks in Internet of Vehicles.* In: ACM Trans. Internet Technol (2022)
[22] Yujie, Z., Dezhi, H., Xinming, Y.: *A hierarchical network intrusion detection model based on unsupervised clustering*. In: MEDES '21, November 1–3, 2021, Hammamet, Tunisia (2021)
[23] Hatitye, C., Dane, B.: *Adaptive Machine Learning Based Network Intrusion Detection*. In: icARTi '21, December 9–10, 2021, Virtual Event, Mauritius (2021)
[24] Naga, K., Rajesh, Y., Raghava, S.: *A Painstaking Analysis of Attacks on Hypervisors in Cloud Environment.* In: ICMLT 2021, April 23–25, 2021, Jeju Island, Republic of Korea (2021)
[25] Elisa, B., Imtiaz, K.: *AI-powered Network Security: Approaches and Research Directions*. In: 8th NSysS 2021, December 21–23, 2021, Cox's Bazar, Bangladesh (2021)
[26] Zhao, M., Xiuhua, L., Chuan, S., Qilin, F., Xiaofei, W., Victor, C.: Sleeping Cell Detection for Resiliency Enhancements in 5G/B5G Mobile Edge-Cloud Computing Networks. In: 2022 Association for Computing Machinery (2022)
[27] Nagai, R., Kurihara, W., Higuchi, S., Hirotsu, T.: *Design and Implementation of an OpenFlow based TCP SYN Flood Mitigation*. In: 2573-7562/18/$31.00 ©2018 IEEE DOI 10.1109/MobileCloud.2018.00014 (2018)
[28] Quang, D.: *Evaluating machine learning algorithms for intrusion detection systems using the dataset CIDDS-002*. In: CSSE 2021, October 22–24, 2021, Singapore, Singapore (2021)
[29] Hanan, H., Christos, T., Robert, A.: *Developing a Siamese Network for Intrusion Detection Systems*. In: EuroMLSys '21, April 26, 2021, Online, United Kingdom (2021)
[30] Suresh, A., Thushara, R., Sumohana, C., Bheemarjuna, T.: *On Handling Class Imbalance in Continual Learning based Network Intrusion Detection Systems*. In: AIMLSystems '21, October 21–23, 2021, Bangalore, India (2021)

Esraa Samkari received the B.S. degree in computer science from Umm Al-Qura University, Makkah, Saudi Arabia, in 2019, where she is currently pursuing the M.S. degree. Her research interests include Arabic text recognition and trajectory compression algorithms.
E-mail: s44380084@st.uqu.edu.sa.