International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT2063103

Comparative Study of Datasets used in Cyber Security Intrusion Detection

Rahul Yadav, Phalguni Pathak , Saumya Saraswat
Department of Computer Science Application, ITM University, Gwalior, Madhya Pradesh, India

Article Info
Volume 6, Issue 5
Page Number: 302-312
Publication Issue :
September-October-2020

Article History
Accepted : 05 Sep 2020
Published : 30 Oct 2020

ABSTRACT

In recent years, deep learning frameworks are applied in various domains and achieved shows potential performance that includes malware detection software, self-driving cars, identity recognition cameras, adversarial attacks became one crucial security threat to several deep learning applications in today's world Deep learning techniques became the core part for several cyber security applications like intrusion detection, android malware detection, spam, malware classification, binary analysis and phishing detection. . One of the major research challenges in this field is the insufficiency of a comprehensive data set which reflects contemporary network traffic scenarios, broad range of low footprint intrusions and in depth structured information about the network traffic. For Evaluation of network intrusion detection systems, many benchmark data sets were developed a decade ago. In this paper, we provides a focused literature survey of data sets used for network based intrusion detection and characterize the underlying packet and flow-based network data in detail used for intrusion detection in cyber security. The datasets plays incredibly vital role in intrusion detection; as a result we illustrate cyber datasets and provide a categorization of those datasets.

Keywords : Architecture , Attack , Detection , IDS , Datasets , Prevention

I. INTRODUCTION

With the progressive incorporation of the Internet and social life, the Internet is changing how people become skilled and work, but it also exposes us to ever more severe security threats. The world is becoming more reliant on connected actuators and sensors, changing the life of millions of people. Therefore, it is critical to build robust tools to protect networks against security threats. The foremost challenge is to identify unidentified and obfuscated malware attacks. Malwares are produced deliberately to compromise computer systems and acquire benefits of shortcomings in intrusion detection systems. In addition, there has been significant growth in security threats such as zero-day attacks which are intended to mainly target internet users [1]. There are outsized numbers of cybercriminals round the world aggravated to steal information, unlawfully collect revenues, and hit upon new targets. So it is required to develop efficient IDS which detect sophisticated malware. The purpose of

IDS is to recognize different forms of malware as early as possible, which cannot be achieved by a conventional firewall. In the last few decades, machine learning has been accustomed improve intrusion detection, and currently there is a necessity for an up-to-date, thorough taxonomy and survey of this recent work.

DL technology has enabled people to benefit from more data, obtain better results, and develop more potential. It has dramatically changed people's lives and reshaped traditional AI technology. AI has a wide range of applications, such as face recognition, speech recognition, and robotics, but its application scope goes far beyond the three aspects of image, voice, and behavior. It also has many other outstanding applications within the field of cyber security, such as malware monitoring and intrusion detection. In the early development of AI technology, machine learning (ML) technology played an important role in addressing cyberspace threats [2]. Although ML is extremely powerful, it relies to a great extent on feature extraction. This flaw is especially glaring when it is applied to the field of cyber security. ML algorithms work according to the pre-defined specific feature, which means that features which are not pre-defined will escape detection and cannot be discovered. It can be concluded that the performance of most ML algorithms depends on the accuracy of feature recognition and extraction [3]. In the view of obvious flaws in traditional ML, researchers began to study deep neural network (DNN), also known as DL, which is a sub-domain of ML.

There are a large number of related studies are performed using either the KDD-Cup 99 or DARPA 1999 dataset to substantiate the development of IDSs; yet there is no noticeable answer to the question of which deep learning techniques are more efficient. Secondly, the time taken for building IDS is not considered within the assessment of assorted IDSs techniques, in spite of being a significant factor for the effectiveness of 'on-line' IDSs. The efficiency of an IDS is directly associated with the selected learning model and therefore the quality of the dataset used. An excellent quality dataset can be defined as a dataset that improves better performance metrics in real-world transactions. As mentioned in [4,5] imbalanced datasets present a problem to researchers. A dataset is supposed to be imbalanced when the distribution of classes is not uniform [6]. This can be common problem in many of the classification problems due to the used datasets. Imbalanced dataset results the used classifier biases towards the majority class; however, in most of them, the aim is trying to detect the minority class [7, 8]. This result in a large classification error over the minority class samples and main targets can be missed. Datasets should be balanced according to the data types, to enhance the quality of datasets. The primary purpose of this work is to compile recent works that are oriented to comprehensive overview of data sets which points out the peculiarities of each data set. Thereby a particular focus was placed on attack scenarios within the data sets and their interrelationships. In addition, each data set assessed with respect to the properties of the categorization scheme developed in the first step. This, comprehensive survey intent to support researchers to recognize data sets for their objective. The review of data sets shows that the research community has noticed a lack of publicly available network-based data sets and tries to overcome this insufficiency by publishing a significant amount of data sets in the last few years. This paper presents a literature review on the different datasets that were used for cyber security applications. The purpose of this paper is for those researchers who want to study network intrusion detection.

II. RELATED STUDIES

Buczak et al. [9] published a study which addresses machine learning approaches utilized by the intrusion detection systems. This study focuses on Machine Learning and Data Mining techniques used in cyber security, with a prominence on the ML/DM techniques and their descriptions. This survey classified the datasets into three types, namely, Packet-Level Data, Net Flow data and public datasets.

Diro and Chilamkurti [10] approached deep learning as a novel intrusion detection technique with promising results. The authors also reported that thousands of zero-day attacks appear because of the addition of various protocols, mainly from IoT and that most of them are small variants of previously known cyber-attacks. Such a situation indicated that even advanced mechanisms such as traditional machine-learning systems face the difficulty of detecting these small mutants of attacks over time.

Milenkoski et al. [11] provided IDS evaluation design space practices in cyber security intrusion detection by analyzing existing systems by evaluating standard parameters, namely, workloads There are three types of workloads with respect to workload content: pure benign (workloads that don't contain attacks), pure malicious (workloads that contain only attacks), and mixed, metrics There are two types of metrics with respect to the aspect of IDS behavior they compute: security related measures accuracy of attack detection in Intrusion Detection System and performance related assess nonfunctional properties of Intrusion Detection System .

Lopez-Martin et al. [12] proposed a new network intrusion detection method specifically developed for an IoT network. This method is based on a Conditional Variational Autoencoder (CVAE) which integrates the intrusion labels inside the decoder layers. The proposed model is also able to perform feature reconstruction, and it also can be used in the current Network Intrusion Detection System, which is part of network monitoring systems, and particularly in IoT networks. The proposed approach operates in a single training step, therefore saving computational resources.

Another recent survey by Zarpelao et al. [13] discusses the problems to security, especially concerning IoT, and therefore the integration of real-world devices with the web since cyber security threats are delivered to most daily activities. Attacks against critical infrastructures, like power plants and public transit, can have severe consequences for cities and whole countries. .The authors presented a research about intrusion detection approaches specifically used for internet of things (IoT) and they also proposed taxonomy to classify IDSs for IoT which was based on the subsequent attributes: Intrusion Detection System (IDS) placement strategy, security threat, detection method, and validation strategy

Giovanni Apruzzese [14] presents an analysis which addressed machine learning techniques applied to three relevant cyber security problems: intrusion detection, malware analysis and spam detection. Initially they proposed an original taxonomy of the most popular categories of ML algorithms and show which of them are currently applied to which problem. This study explores several issues that influence the application of ML to cyber security. All approaches are susceptible to adversarial attacks and require constant re-training and cautious parameter tuning that cannot be automatized. When the similar classifier is applied to recognize different threats, the detection performance is unacceptably low; a possible alleviation can be achieved by using different ML classifiers for detecting specific threats.

Ramos et al. [15] presented a review that focused on model- based quantitative security metrics that address to evaluate comprehensive network flexibility against intrusions. In this survey, an in-depth review of the state-of-the-art of Network

Security Metrics (NSMs) has been presented focused in the Common Vulnerability Scoring System (CVSS) framework, which is used as input by several security metric models. The differences between the security metrics field and other correlate areas have also been conducted. This study carried out a comprehensive and detailed review of the main metric proposals and has been presented more specifically in the realm of model-based quantitative NSMs; a complete and thorough review of the main metric proposals has also been presented. The main pros and cons of each reviewed work have also been described. Eventually, an in-depth investigation of the main properties of the reviewed security metrics has been presented, along with open issues and suggestions for future research directions, followed by a discussion on past related work. According to what has been presented in this review, it is reasonable to assume that the field of model-based quantitative NSMs is still in development and significant more progress still needs to be done.

III. INTRUSION AND INTRUSION DETECTION SYSTEM

Cyber security is a set of technologies and processes designed to protect computers, networks, programs and data from attacks and unauthorized access, alteration, or destruction [16]. A network security system is consisting up of a computer security system and a network security system. Every system includes antivirus software, firewalls and intrusion detection system (IDS). At present, network intrusion detection systems (NIDS) present a enhanced solution to the security crisis compared with other traditional network defense technologies, such as firewall systems. NIDS helps network administrators detect attacks, vulnerabilities, and breaches inside an organization's network [17]. Intrusion detection systems are classified according to several different criteria.

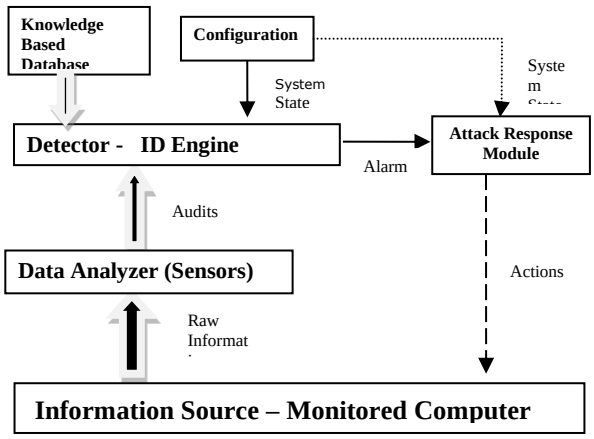

Fig 1. Working of Intrusion Detection System

To classify these kinds of systems, we can have two categories; based on architectural configuration another is based on the data processing time. According to the location intrusion detection systems are classified into two types Host-Based and Network- Based [18]. Also to classify IDSs, it can be done according to their techniques; Signature-Based and Anomaly-Based. There are three main types of network analysis for IDS: misuse-based, also known as signature-based, anomaly-based, and hybrid.

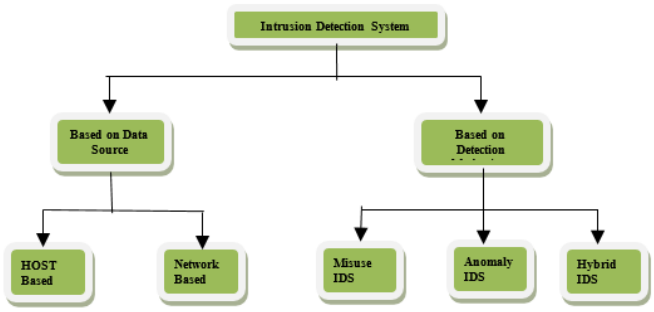

Fig 2. Classification of Intrusion Detection Systems

3.1 Signature-based Method:

Misuse-based detection techniques aim to detect known attacks on the basis of the specific patterns. It detects on the basis of previously identified malicious instruction sequence that is used by the malware. Those detected patterns in the IDS are identified as signatures [19]. They are used for known types of attacks whose pattern (signature) already exists in

system without generating a large number of false alarms. However, it is quite difficult to detect the new malware attacks as their pattern (signature) is not known. To overcome this problem administrator needs to regularly update database rules and signatures manually. New (zero-day) attacks which, which can result in potential damage and present serious security risks to your computer or personal data cannot be detected by the techniques based on misused technologies.

3.2 Anomaly-based Method:

Anomaly-based IDS used to detect the unidentified malware attacks that are developed rapidly [20]. Anomaly-based IDS are appealing due to their capability to detect zero-day attacks. In anomaly-based IDS concept of machine learning is employed to construct a trustful activity model and incoming data is compared with that predefined model and if it is not found in model it is declared as suspicious. Another benefit of using this method is that all profiles of usual activity has been modified for applications, network or each system, which makes it all the more complex for attacker to seek out which behavior can perform undetected. Additionally, the information on which anomaly-based techniques alert (novel attacks) can be used to characterize the signatures for misuse detectors. The main disadvantage of anomaly-based techniques is the potential for high false alarm rates because previously unseen system behaviour can be categorized as anomalies.

3.3 Hybrid Method

Hybrid detection combines misuse and anomaly detection [21]. It is used to enhance the detection rate of known intrusions and to trim down the false positive rate of unknown attacks. Most DL methods are hybrids.

3.4 Host-Based IDS

Host-based intrusion detection systems are used to analyze the activities on a specific Host commonly called HIDS[18]. They have numerous advantages similar to network-based intrusion detection systems (NIDSes) do, but they have reduced scope of operation [22]. A potential flaw with with host-based intrusion detection systems is that any information that they might gather needs to be communicated over the network. If the machine is being actively attacked, via the same network, this may not be possible. A common implementation of host-based IDS will be established in several antimalware products used today. So many IDSs (Host-Based) depends on string matching or signature to identify potential threat, and can be defeated by easy change of tool which states the signature is not matched.

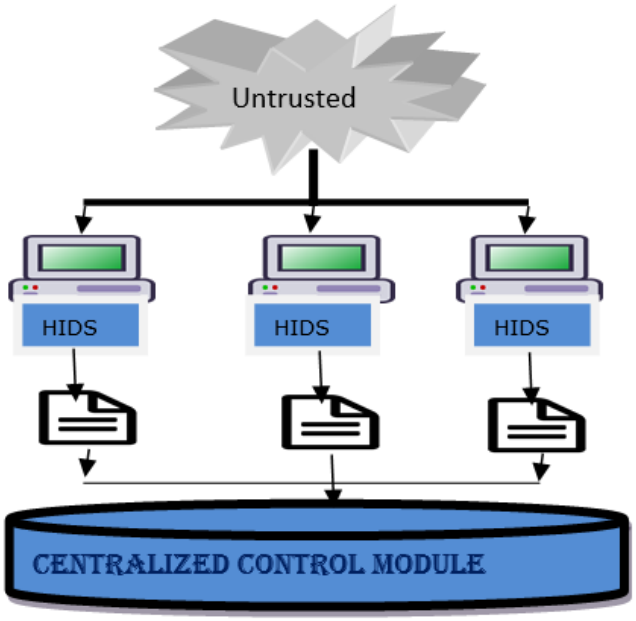

Fig. 3 Block Diagram of Host Based IDS

3.5 Network-Based IDS

A network-based intrusion detection system often known as NIDS is employed to observe and analyze network traffic to safeguard a system from network-

based threats [23]. They are easy to secure and can be more difficult for an attacker to detect. Network-based Intrusion Detection Systems (NIDS) are located at a premeditated point (or points) to supervise the traffic on the network. It NIDS reads all inbound packets analyses the passing traffic on the entire subnet, and searches for any suspicious patterns matches the traffic that is passed on the subnets to the library of known attacks. When threats are is identified, or abnormal behavior is detected based on its severity, the system can send an alert is sent to the administrator or barring the source IP address from accessing the network. For simulation in network intrusion detection systems, OPNET and NetSim, are commonly used tools.

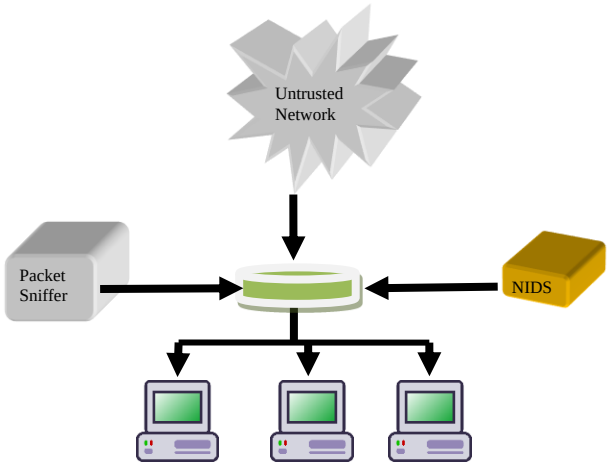

Fig. 4 Block Diagram of Network Based IDS

IV. DATASETS

Biggest challenge is to determine Intrusion Detection System's performance in finding the appropriate data set. The information to be used for the data may be obtained by observing the network. Collecting information from the network is costly; therefore developers want to manage their network or systems using available datasets. In this section, the data sets, which is usually used for attack detection systems are mentioned.

4.1 DARPA1998

The DARPA1998 dataset [24] was the earliest effort to create an IDS dataset and is a widely used benchmark dataset in IDS studies. In 1998, at the Lincoln Labs of MIT DARPA introduced a program to produce a comprehensive and realistic IDS benchmarking conditions and they designed the KDD98 (Knowledge Discovery and Data Mining (KDD)) dataset. The researchers collected Internet traffic over nine weeks to create this datasets; the primary seven weeks form the training set, and the last two weeks form the test set. This test data further categorized as normal or abnormal on the basis of individual features that are extracted from 2 million records and each record has 41 features.

The extracted data contains a large variety of attacks simulated in a military network environment. The dataset contains both raw packets and labels. The labels are categorized into five types and they are: denial of service (DOS), User to Root (U2R) , Probe, normal and Remote to Local (R2L). Because raw packets cannot be directly applied to conventional machine learning models, so to overcome this drawback the KDD99 dataset was created. DARPA1998 dataset accuracy and capability while considering the real-life conditions have been widely criticized, although this dataset was an important contribution to the research on IDS.

4.2 KDD Cup99

The KDD99 [25] dataset is the most widespread data set for testing of intrusion detection systems at present. It had been designed as a simulation data set in 1998. KDD Cup99 is particularly utilized in the field of data mining and machine learning. Its compilers extracted 41 features from DARPA 1998 which can be categorized as basic features, host-based statistical features, time-based statistical features and content features, within the data set [18]. KDD Cup99 includes both training and test data. The data in this dataset can be classified into 5 main categories,

4 of them are attack, and 1 is Normal. Normal; non-attack type data. Attack types: DOS (Denial of Service), R2L (Root to Local), Probe (Probing attacks), and U2R (User to Root). Additionally, this data sets data also contain a feature to show the label of the data whether it is an intrusion or not. Unfortunately, the KDD99 dataset includes many defects. First defect is the data is quite unbalanced, which makes the classification results biased toward the majority classes. Also, v0arious duplicate records and redundant records also exist. Before using the dataset, Researchers have to filter it carefully. As a result of the various experimental results and from different studies KDD data are too outdated to represent the existing network environment.

4.3 NSL-KDD

NSL-KDD [26] was proposed to overcome the shortcomings of the KDD99 dataset. NSL-KDD dataset was created by deleting duplicate and redundant records from KDD99 data set; therefore, it contains only a moderate number of records. To avoid the classification bias problem, records of different classes are balanced within the NSL-KDD. Therefore, the experiments can be implemented on the full dataset, and also the results from different papers are reliable and comparable. The issues of data bias and data redundancy are alleviated by NSL-KDD, to some extent. However, the NSL-KDD does not consist of new data; because of that minority class samples are still lacking, and its samples are still out-of-date [25]. The distinctions NSL-KDD has over the original KDD Cup99 are as follows:

1) The classifier does not give biased results because there are no redundant data within the training set.

2) The reduction ratio is lower because there is no repetitive data in the test set.

3) The growth of records within the KDD dataset is proportional to the number of records selected from individual difficult level group .Each data contains, 41 attributes, which unfolds various features of the flow. Either attack type or normal are the labels, which can assigned to them. Further categorizations of these attacks are DOS, Probe, R2L and U2R.

4.4 CIC IDS 2017

CIC-IDS2017 was created by The Canadian Institute for Cybersecurity (CIC) in 2017. The CIC IDS 2017 dataset includes latest attacks, which similar to the real-world data. It was created by analyzing network traffic using information from the timestamp, CICFlowMeter, source, and destination IPs- ports, protocols and attacks [27]. It includes 86 network related features that also contain IP addresses and attack types Furthermore, the CIC has identified eleven criteria that are necessary for building a reliable benchmark dataset. These criteria are : Complete Traffic, Labeled Dataset, Available Protocols , Complete Interaction, Complete Network configuration/Structure, Complete Capture, Heterogeneity, Feature Set, Attack Diversity and Metadata. CICIDS2017 dataset comprises both benign behavior and also details of recent malware attacks: like Brute Force FTP, Brute Force SSH, DoS, Heartbleed, Web Attack, Infiltration, Botnet and DDoS [28]. Label of this dataset is basically based on the timestamp, source and destination IPs, protocols and attacks, source and destination ports. An entire network topology was configured to bring together this dataset which contains Modem, Firewall, Switches, Routers, and nodes with different operating systems (Microsoft Windows (like Windows 10, Windows 8, Windows 7, Windows Vista , Windows NT and Windows XP), Apple's macOS iOS, and open source operating system such as Linux). The dataset, which is mentioned, contains 80 network flow features from the captured network traffic. Since machine learning techniques are applied in AIDS Comparison of public IDS datasets, are very important to assess these techniques used for realistic evaluation.

4.5 CSE-CIC-IDS2018

CSE-CIC-IDS2018 dataset [28] the most recent dataset proposed by The Canadian Institute for Cyber security (CIC) and Communications Security Establishment (CSE). It includes detailed information of attacks with abstract distribution models that can be applied to various network protocols with different topologies for computer systems. Furthermore, the dataset was enhanced, therefore the number of duplicate data is very low and uncertain data is nearly absent,The dataset CSE-CIC-IDS2018 includes seven different attack scenarios, including, Heartbleed, Brute-force, DoS attack, Web attack, Infiltration attack, Botnet attack, DDoS attack, and Heartleech .Similarly to CICDS2017 dataset [27] , the CICFlowMeter tool is used to extract 80 statistical network flow features separately from the generated network traffic in the forward and reverse direction. The CIC team reported the raw data every day along with the network traffic and event logs. The CICFlowMeter-V3 is used in features extraction technique from the raw data and In, this process it extracted more than 80 network traffic features. Finally, they saved them as a CSV file if Artificial Intelligence techniques are to be used.

4.6 ADFA2013 Dataset

Dataset is proposed by Creech and Hu [29] for evaluation by system call based host level intrusion detection system. The ADFA dataset is a host level intrusion detection system provided by the Australian defence academy (ADFA) [27], which is widely employed for testing of intrusion detection Systems. In this dataset, payloads and vectors are used to attack the Ubuntu OS. The data set consist of two OS platforms, ADFA Linux Dataset (ADFA-LD) and ADFA Windows Dataset (ADFA-WD), which logs the order created from the evaluation of system calls based on HIDS. ADFA-LD was developed using Ubuntu Linux version 11.04 which was used as the host operating system. ADFA-LD and ADFA-WD are designed as public datasets that represent the structure and methodology of the modern attacks. Several attack occurrences in ADFA-LD were derived from new zero-day malware, forming this dataset good enough for showing disparity between SIDS and AIDS approaches to intrusion detection. For evaluation of HIDS the ADFA Windows Dataset (ADFA-WD) is presented as a contemporary Windows data set. The payloads consist of password brute-force, java based meterpreter, add latest superuser, C100 Webshell and linux meterpreter payload. The dataset formation consist of three data types, namely, (1) attack Data (2) normal validation data, and (3) normal training data,. The normal training data contains 4373 traces. The normal validation data contains 833 traces. The attack data contains 10 attacks per vector. In this data set, real network traffic footprints were evaluated to recognize normal performance for computers from real traffic of SMTP, HTTP, POP3, SSH, IMAP, and FTP protocols. This dataset contains assorted and labeled attacks scenarios which is based on realistic network traffic.

4.7 UNSW-NB15

The UNSW-NB15 [30] dataset was created by four tools, namely, IXIA Perfect- Storm tool, Tcpdump tool, Argus tool, and Bro-IDS tool, where researchers configured three virtual servers to capture network traffic and extract features using tool. These tools are used to create some types of attacks, including DoS, Reconnaissance, Exploits, Shellcode, Generic, and Worms. This dataset consist of numerous forms of attacks than the KDD99 dataset, and its features are more plentiful. The classification of data consists of standard data and nine types of attacks. These features include flow features, content features, basic features, additional features, time features and labeled features. The UNSW-NB15 dataset comprise of around two million and 540,044 vectors with 49 features. Additionally, Moustafa et al. [31] published a partition from this dataset which consist the

training set (175,341 vectors) and the testing set (82,332 vectors). The UNSW-NB15 is representative of new IDS datasets, and has been used in various modern studies. At present the influence of UNSW-NB15 is inferior to that of KDD99; therefore it is imperative to create new datasets for developing new IDS.

V. CONCLUSION

In this paper a survey of intrusion detection system technologies, have been reviewed to deal with the challenges in intrusion detection. The scope of the work on classifying intrusion detection systems, reviewing the various datasets used for training and testing systems. Datasets used for training and testing systems are very important in network intrusion detection. Therefore, testing is done using publicly available datasets for IDS research. These datasets have been explored and their data collection techniques, evaluation results and limitations have been discussed. While widely accepted as benchmarks, there are many problems with the existing public dataset, such as uneven data, imbalanced ratio and outdated content .Due to which biased results are generated. However, these datasets no longer represent contemporary zero-day attacks. There exists a need for newer and more comprehensive datasets, as most of the existing machine learning techniques are trained and evaluated on the knowledge provided by the old dataset such as DARPA/ KDD99. Though ADFA dataset contains many new attacks, still it is not adequate for training and testing. While various algorithms present outstanding results, but these results are obtained on outdated datasets CNN and DBN have not been exploited in this field and experimental works are still in progress to conclude the fidelity of these algorithms to detect cyber attacks. Unfortunately, the efficient technique of intrusion detection has not yet been recognized. Comprehensively, there won't be an ideal data set, but there are numerous satisfactory data sets existing and the community could assistance from closer association.

VI. REFERENCES

[1]. S. Aftergood, "Cybersecurity: The cold war online," Nature, vol. 547, no. 7661, p. 30, 2017.

[2]. A. Milenkoski, M. Vieira, S. Kounev, A. Avritzer, and B. D. Payne, "Evaluating Computer Intrusion Detection Systems:A Survey of Common Practices," Acm Comput. Surv., vol. 48, no. 1, pp. 1–41, 2015.

[3]. C. N. Modi and K. Acha, "Virtualization layer security challenges and intrusion detection/prevention systems in cloud computing: a comprehensive review," J. Supercomput., vol. 73, no. 3, pp. 1–43, 2016.

[4]. J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," J. Big Data, vol. 6, no. 1, p. 27, 2019.

[5]. A. Ali, S. M. Shamsuddin, and A. L. Ralescu, "Classi_cation with class imbalance problem:Areview," Int. J. Adv. Soft Comput. Appl., vol. 7, no. 3, pp. 176_204, 2015.

[6]. F. Provost, "Machine learning from imbalanced data sets 101," in Proc. AAAI Workshop Imbalanced Data Sets, Menlo Park, CA, USA: AAAI Press, 2000.

[7]. S. Barua, M. M. Islam, X. Yao, and K. Murase, "MWMOTE-majority weighted minority oversampling technique for imbalanced data set learning," IEEE Trans. Knowl. Data Eng., vol. 26, no. 2, pp. 405_425, Feb. 2014.

[8]. X. Gao, C. Shan, C. Hu, Z. Niu, and Z. Liu, "An adaptive ensemble machine learning model for intrusion detection," IEEE Access, vol. 7, pp. 82512_82521, 2019.

[9]. Buczak AL , Guven E . A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Commun. Surv. Tut. 2015;18(2):1153–76 .

[10].A .A . Diro, N. Chilamkurti, Distributed attack detection scheme using deep learning approach for internet of things, Future Gener. Comput. Syst. 82 (2018) 761–768 . Online]. Available: http://www.sciencedirect.com/science/article/pii / S0167739X17308488 .

[11].Milenkoski A , Vieira M , Kounev S , Avritzer A , Payne BD . Evaluating computer intrusion detection systems: a survey of common practices. ACM Comput. Surv. 2015;48(1):12 .

[12].M. Lopez-Martin, B. Carro, A. Sanchez-Esguevillas, J. Lloret, Conditional vari- ational autoencoder for prediction and feature recovery applied to intrusion detection in IoT, Sensors 17 (9) (2017) 1967 Online]. Available, doi: 10.3390/ s17091967 .

[13].Zarpelao BB , Miani RS , Kawakani CT , de Alvarenga SC . A survey of intrusion detection in internet of things. J. Netw. Comput. Appl. 2017;84:25–37 .

[14].Giovanni Apruzzese, Luca Ferretti, Mirco Marchetti, Michele Colajanni, Alessandro Guido On the Effectiveness of Machine and Deep Learning for Cyber Security

[15].A. Ramos, M. Lazar, R.H. Filho, J.J.P.C. Rodrigues, Model-based quantitative network security metrics: a survey, IEEE Commun. Surv. Tut. 19 (4) (2017) 2704–2734 . Online].Available: http://www.sciencedirect.com/science/article/ pii/S0167739X17308488 .

[16].D. Rozenblum, “Understanding Intrusion Detection Systems,” SANS Inst., no. 122, pp. 11–15, 2001.

[17].H. Debar, M. Dacier, and A. Wespi, “A revised taxonomy for intrusion-detection systems,” Ann. Des Télécommunications, vol. 55, no. 7–8, pp. 361–378.

[18].Sans Penetration Testing, “Host- vs. Network-Based Intrusion Detection Systems,” 2001. Online]. Available: https://cyber-defense.sans.org/resources/papers/gsec/host-vs-network-based-intrusion-detection-systems-102574. Accessed: 24-Feb-2016].

[19].Roesch.M, “Snort - Lightweight Intrusion Detection for Networks” 13th USENIX Conference on System Administration, USENIX Association (1999) 229–238

[20].Hossein M. Shirazi,”Anomaly Intrusion Detection System Using Information Theory, K-NN and KMC Algorithms”, Australian Journal of Basic and Applied Sciences, 3(3): 2581-2597, 2009

[21].SANS Institute InfoSec Reading Room, “Application of Neural Networks to Intrusion Detection,” 2001. Online]. Available: https://www.sans.org/reading-room/whitepapers/detection/application-neural-networks-intrusion-detection-336. Accessed: 24-Feb-2016].

[22].S. S. Tirumala, H. Sathu, and A. Sarrafzadeh, “Free and open source intrusion detection systems: A study,” in 2015 International Conference on Machine Learning and Cybernetics (ICMLC), 2015, vol. 1, pp. 205–210.

[23].V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection,” ACM Comput. Surv., vol. 41, no. 3, pp. 1–58, Jul. 2009. 35

[24].G. Karatas, O. Demir, and O. Koray Sahingoz, "Deep learning in intrusion detection systems,'' in Proc. Int. Congr. Big Data, Deep Learn. Fighting Cyber Terrorism (IBIGDELFT), Dec. 2018, pp. 113_116.

[25].G. Karatas and O. K. Sahingoz, "Neural network based intrusion detection systems with different training functions,'' in Proc. 6th Int. Symp. Digit. Forensic Secur. (ISDFS), Mar. 2018, pp. 1_6.

[26].D. Proti¢, "Review of KDD cup '99, NSL-KDD and Kyoto 2006C datasets,'' Vojnotehni£ki Glasnik, vol. 66, no. 3, pp. 580_596, 2018.

[27].M. Tavallaee, E. Bagheri,W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set,'' in Proc. IEEE Symp. Comput. Intell. Secur. Defense Appl., Jul. 2009, pp. 1_6.

[28].A. Gharib, I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "An evaluation framework for intrusion detection dataset,'' in Proc. Int. Conf. Inf. Sci. Secur. (ICISS), Dec. 2016, pp. 1_6.

[29].I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "Toward generating a newintrusion detection dataset and intrusion traf_c characterization,'' in Proc. 4th Int. Conf. Inf. Syst. Secur. Privacy, 2018, pp. 108_116.

[30].R. K. Sharma, H. K. Kalita, and P. Borah, “Analysis of Machine Learning Techniques Based Intrusion Detection Systems,” in International Conference on Advanced Computing, Networking, and Informatics, 2016, pp. 485–493.

[31].Moustafa, N.; Slay, J. UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In Proceedings of the 2015 Military Communications And Information Systems Conference (MilCIS), Canberra, Australia, 10–12 November 2015; pp. 1–6.

Cite this article as :

Rahul Yadav, Phalguni Pathak , Saumya Saraswat, "Comparative Study of Datasets used in Cyber Security Intrusion Detection", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 5, pp. 302-312, September-October 2020. Available at doi : https://doi.org/10.32628/CSEIT2063103
Journal URL : http://ijsrcseit.com/CSEIT2063103