ADV MATH SCI JOURNAL
Advances in Mathematics: Scientific Journal **9** (2020), no.6, 3849–3854
ISSN: 1857-8365 (printed); 1857-8438 (electronic)
https://doi.org/10.37418/amsj.9.6.64 Spec Issiue on ICAML-2020

REVIEW OF EXISTING DATA SETS FOR NETWORK INTRUSION DETECTION SYSTEM

J. VERMA[1], A. BHANDARI, AND G. SINGH

ABSTRACT. Enormous amounts of data have placed a number of data confidentiality, integrity and availability security challenges and opened gates to malicious activities. Hence, there is a need to address new research challenges. Network Intrusion Detection system (NIDS) is a safeguard technology for ensuring safe and trusted information flow making network secure for modern network-based business and network administration. In this paper we have studied various datasets and presented a review of NIDS related Datasets.

1. INTRODUCTION

Intrusion Detection System is a system or a tool that works with network and monitors traffic for suspicious or unusual malicious activity or policy violations and report it to administrator using security information and event management system. Network if penetrated with malicious activity can lead to loss of potential vital information, data breaches and loss of user trust. There is a need of security of private resources from inside and outside attacks to the organization by exploiting its vulnerabilities like password cracking, network sniffing traffic and doing other malicious activities by masquerader, Misfeasor or Clandestine Users. Firewalls and anti-malware software alone do not provide enough protection of entire network from attacks, [1].

[1]*corresponding author*

2010 *Mathematics Subject Classification.* 68P15, 68M10.

Key words and phrases. NIDS-Network Intrusion Detection System, dataset, IDS-Intrusion Detection System.

2. Classification of Intrusion Detection System

Intrusion detection systems track and detect networks for potential malicious operation, and are prone to false alarms. It involves keeping check on evasion techniques, sending fragmented packets to keep the attacker under radar, avoid-ing default port to avoid port-by-attacker reconfiguration, organizing a search among attackers, address spoofing and server proxy, and Pattern change evasion helps to keep malicious attacks under control. There are two kinds of schemes for detecting intrusion. Host intrusion detection systems only monitor client or independent host network packets and alert the administrator if irregular or mali-cious activity is detected by taking snapshots of current device files and compar-ing them to the previous snapshot, [2]. Network intrusion detection systems are intelligently distributed within networks through hardware or software-based de-vices, depending on the IDS device supplier, which can be connected to Ethernet, FDDI and other network media.

3. Datasets for Network Intrusion Detection Systems

Network intrusion detection systems are designed in a manner that they monitor and analyze network traffic to mitigate security risks and network inva-sion. The KDD Cup 1999 is a dataset consisting of 41 features, classified into basic features, traffic characteristics and content features, is most commonly used to test intrusion detection models comprising a specific collection of data to be audited, including a wide range of intrusions. Lee et al. worked with KDD Cup 99 dataset for simulation in a military network environment in [3]. Tavallaee and et al. evaluated the level of complexity in KDD data set information. Mohammad Khubeb Siddiqui et al. tried to build the relation between attack over the network and protocols used by the hacker in [4]. Revathi and Malathi et al. suggested the re-dundant and overlapping information in KDD Cup 99 datasets, [5]. Tavallaee et al. proposed the NSL-KDD dataset to fix inherent KDD'99 data sets issues. Dhanabal et al. in [3,6] analysed NSL KDD and examined the efficacy of different classification algorithms in detecting anomalies in network traffic patterns and re-vealed the link between protocols and network attacks, [6]. Sarathi Partha et al. in [8] proposed Fuzzy Vectorized GA and Weighted Vectorized GA to detect network attacks for the NSL-KDD data set, [7]. IXIA Perfect creates UNSW-NB15 dataset which consists of creating certain types of attacks.

TABLE 1. Popular Datasets for NIDS

S. N	Name of the dataset	Total number of instances	Type of attack	Number of Features	Labelled (Y/N)	Availability	Research done
1	KDD-Cup'99	5000000 Imbalanced classes	Normal DoS, Probe, U2R, R2L	41	Y	http://archive.ics.uci.edu/ml/datasets/kdd+cup+1999+data	Tavallaee and et al [9] Preeti Aggarwal and et al[17] Mohammad Khubeb Siddiqui et.al [4]Revathi & Malathi et. al.[5]
2	NSL-KDD	Training set= 489431 Test set = 311027	Normal DoS, Probe, U2R, R2L	41	Y	http://www.unb.ca/cic/datasets/nsl.html	Tavallaee et al. [3]Dhanabal et. al [6]Partha sarathi et. al[18]
3	Kyoto 2006+	Covers three years of real traffic data (2006-2009)	Multiple	24	N	http://www.takakura.com /Kyoto_data/	Jungsuk SONG et[12] Iman Sharafaldin et. al. [13] Danijela D. Protić et. al. [19]
4	ISCX 2012	Training Dataset = 9 Testing Dataset = 9	Normal Attack	9	Y	http://www.unb.ca/cic/datasets/ids.html	Aldwairi et al M Ring et al. [11] Saeid Soheily-Khah at. el.[16]
5	DARPA Datasets	Multiple datasets	DoS, Probe, U2R, R2L	Multiple Dataset	N	http://ll.mit.edu/IST	Sharafaldin et al.[13] M Ring et. al.[11]
6	CICIDS2017	Contains total 5 days data, i.e. Monday to Friday.	Infiltration, Botnet, Brute Force FTP , DoS, Web attacks.	80	Y	http://www.unb.ca/cic/datasets/ids.html	Razan Abdulhammed et. al.[15] Sharafaldin et al[13] Mohamed Hamid Abdul Raheem et. al. [20] Benjamin J. Radford et al.[21]
7	DEFCON	Contain only attack traffic during DEFCON competition	Port Scan, Buffer-Flow attacks	None	N	http://cctf.shmoo.com	Ali Shiravi et. Al.[22] Markus Ring et al[11]
8.	UNSW-NB15	Contains 175,341 Training sets and 82,332 Testing Set	Fuzzers Analysis, Backdoors, DoS, Exploits, Generic	49	Y	https://cloudstor.aarnet.edu.au/plus/index.php/s/2DhnLGDdEECo4ys	Khan et al.
9.	CICDDoS2019	--	LDAP, MSSQL, SNMP, SSDP, NTP, DNS, UDP	80	Y	http://www.unb.ca/cic/datasets/ids.html	Iman Sharafaldin et. al. [23]

The UNSW-NB15 dataset includes around two million and 540,044 vectors with 49 features. Nour Moustafa and et. al. (2019) introduced existing and novel methods used to produce the UNSWNB15 data set, [8]. Tharmini Janarthanan et. al. analysed the features included in the UNSW-NB15 dataset, [9]. Mukrimah Nawair conducted multi-classification of the UNSW-NB15 network anomaly detection method, [10]. DEFCON was created by collecting normal and abnormal traffic while performing hacking and anti-hacking competitions in a restrictive environment collection. Ali Shiravi et al. suggested use of DEFCON for alarm correlation assessment techniques, [20]. Markus Ring et al.

presented a survey of data sets for intrusion detection on a network basis, [11]. According to Sharafaldin et al. [14] DARPA 1998 dataset is based on network traffic and analysis and it con-tains seven weeks of network-based attacks while the test data contains two weeks of network-based attacks. This dataset isn't real-world network traffic. There are three collections of data available for testing, DARPA 98, 99 and 2000. Ring et al. reviewed data sets, and considered DARPA 1998 to be one of the most popular intrusion detection data sets, [13].

Honeypots, darknet sensors, e-mail servers, and web crawler tools were used to produce KYOTO dataset. The Kyoto 2006 + dataset is compiled using honeypots, darknet sensors, email servers and web crawler. Jungsuk Album et al. provided detailed overview of Benchmark Data at the University of Kyoto [12]. Table 1, [3–6,9,11–22] describes popular NIDS datasets with the details of dataset name, number of instances, type of attacks, number of features, labelling flag, availability URL and details of research work done on the specified dataset.

4. Conclusion

An IDS works by gathering snapshots of the entire system and then using the information gathered from pre-established patterns, it provides awareness and insight into how an attack occurred. A NIDS perform network traffic analysis and check the traffic that is transmitted to the database and library for established at-tacks on the subnets. If an attack consisting of an anomalous activity is detected, the alert is sent to the administrator. In our paper, we reviewed common Datasets used for Network Intrusion Detection Systems and their origins along with the researchers who worked on Network Intrusion Detection System with these datasets.

References

[1] V. Das, V. Pathak, S. Sharma, Sreevathsan, M. Srikanth, T. Gireesh Kumar: *Network Intrusion Detection System Based On Machine Learning Algorithms*, International Journal of Computer Science and Information Technology, **2**(6) (2010), 138–151.

[2] A. Khraisat, I. Gondal, P. Vamplew, J. Kamruzzaman: *Survey of intrusion detection systems: techniques, datasets and challenges*, Cybersecurity, **2**(1) (2019), 1–22.

[3] M. Tavallaee, E. Bagheri, W. Lu, A. A. Ghorbani: *A Detailed Analysis of the KDD CUP 99 Data Set*, IEEE Symposium on Computational Intelligence for Security and Defense Applications, CISDA 2009, 1–6.

[4] A. KASHYAP, A. NAYAK: *Different machine learning models to predict dropouts in MOOCs*, IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI), (2018), 80–85.

[5] D. A. M. S. REVATHI: *A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection*, International Jornal of Engineering Research and Technology, **2**(12) (2013), 1848–1853.

[6] L. DHANABAL, S. P. SHANTHARAJAH: *A Study on NSL-KDD Dataset for Intrusion Detection System Based on L. Dhanabal and S. P. Shantharajah Classification Algorithms*, International Journal of Advanced Research in Computer and Communication Engineering, **4**(6) (2015), 446–452.

[7] Y. DING, Y. ZHAI: *Intrusion detection system for NSL-KDD dataset using convolutional neural network*, ACM International Conference Proceeding Series, (2018), 81-85.

[8] N. MOUSTAFA, J. SLAY: *UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set*, Military Communications and Information Systems Conference, MilCIS 2015 - Proceedings, (2015).

[9] T. JANARTHANAN, S. ZARGARI: *Feature selection in UNSW-NB15 and KDDCUP'99 datasets*, IEEE International Symposium on Industrial Electronics, (2017), 1881–1886.

[10] M. NAWIR, A. AMIR, N. YAAKOB, O. N. G. B. I. LYNN: *Multi-Classification of Unsw-Nb15 Dataset*, Journal of Theoretical and Applied Information Technology, **96**(15) (2018), 5094–5104.

[11] M. RING, S. WUNDERLICH, D. SCHEURING, D. LANDES, A. HOTHO: *Multi-Classification of Unsw-Nb15 Dataset*, Computers and Security, **86** (2019), 147–167.

[12] J. SONG, H. TAKAKURA, Y. OKABE: *Description of Kyoto University Benchmark Data*, Description of Kyoto University Benchmark Data, (2019), 10–12.

[13] I. SHARAFALDIN, A. H. LASHKARI, A. A. GHORBANI: *Toward generating a new intrusion detection dataset and intrusion traffic characterization*, ICISSP 2018 - Proceedings of the 4th International Conference on Information Systems Security and Privacy, (2018), 108–116.

[14] D. PROTIĆ: *Review of KDD Cup '99, NSL-KDD and Kyoto 2006+ datasets*, Vojnotehnicki glasnik, **66**(3) (2018), 580–596.

[15] R. ABDULHAMMED, H. MUSAFER, A. ALESSA, M. FAEZIPOUR, A. ABUZNEID: *Features dimensionality reduction approaches for machine learning based network intrusion detection*, Electronics (Switzerland), **8**(3) (2019), 322.

[16] S. SOHEILY-KHAH, P. F. MARTEAU, N. BECHET: *Intrusion detection in network systems through hybrid supervised and unsupervised machine learning process: A case study on the iscx dataset*, ICDIS 2018 - Proceedings - 2018 1st International Conference on Data Intelligence and Security, (2018), 219–226.

[17] P. AGGARWAL, S. KUMAR: *Analysis of KDD Dataset Attributes - Class wise For Intrusion Detection*, Procedia - Procedia Computer Science, **57** (2015), 842–851.

[18] P. S. BHATTACHARJEE: *Intrusion Detection System for NSL-KDD Data Set using Vectorised Fitness Function in Genetic Algorithm*, Advances in Computational Sciences and Technology, **10**(2) (2017), 235–246.

[19] M. H. ABDULRAHEEM, N. B. IBRAHEEM: *A detailed analysis of new intrusion detection dataset*, Journal of Theoretical and Applied Information Technology, **97**(17) (2019), 4519–4537.

[20] B. J. RADFORD, B. D. RICHARDSON, S. E. DAVIS: *Sequence Aggregation Rules for Anomaly Detection in Computer Network Traffic*, Proceedings of the American Statistical Association 2018 Symposium on Data Science and Statistics, (2018).

[21] A. SHIRAVI, H. SHIRAVI, M. TAVALLAEE, A. A. GHORBANI: *Toward developing a systematic approach to generate benchmark datasets for intrusion detection*, Computers and Security, **31**(3) (2012), 357–374.

[22] I. SHARAFALDIN, A. H. LASHKARI, S. HAKAK, A. A. GHORBANI: *Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy*, Proceedings - International Carnahan Conference on Security Technology, (2019).

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,
PUNJABI UNIVERSITY, PATIALA, PUNJAB, INDIA
Email address: Jyoti.SnehiVerma@gmail.com

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,
PUNJABI UNIVERSITY, PATIALA, PUNJAB, INDIA
Email address: bhandarinitj@gmail.com

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,
PUNJAB INSTITUTE OF TECHNOLOGY, RAJPURA, PUNJAB, INDIA
Email address: myselfgurpreet@gmail.com