International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 02 February 2024
Page No: 254 - 260
https://irjaeh.com

A Comprehensive Study on CIC-IDS2017 Dataset for Intrusion Detection Systems

Zafar Iqbal Khan[1], Mohammad Mazhar Afzal[2], Khurram Naim Shamsi[3]
[1]Research Scholar, Department of Computer Science, Glocal University, Saharanpur, Uttar Pradesh, India.
[2]Assistant Professor, Department of Computer Science, Glocal University, Saharanpur, Uttar Pradesh, India.
[3]Research Scholar, Department of Computer Science, Glocal University, Saharanpur, Uttar Pradesh, India.
Email id: *mzafar.ikhan@gmail.com[1], mazhar.afzal@gmail.com[2], knshamsi@gmail.com[3]*

Abstract
In the present era of digital technology, electronic assaults result in the compromise of confidential information and substantial financial ramifications for individuals, organizations, and nations. Hence, the role of cybersecurity resources is essential in safeguarding data from any Cybersecurity event. Researchers are prioritizing the use of anomaly-based intrusion detection systems for the identification of cybersecurity threats. Machine learning algorithms are crucial in this endeavor since they possess the ability to identify such attacks reliably. It's a dataset that enhances the efficiency of the machine learning algorithms. The datasets currently employed in intrusion detection systems exhibit a notable deficiency in accurately representing actual network threats and attacks. They also contain a significant number of concealed threats, thereby restricting the precision of detection within existing machine-learning intrusion detection system approaches. Consequently, these systems are unable to effectively cope with the growing number of novel attacks in the real word scenarios, in cloud environments. The objective of this study is to integrate the categorization and analysis of current datasets to enhance the generation of future datasets that accurately replicate actual network data. This will enhance the efficacy of the next generation of intrusion detection systems and correctly mirror network threats.

Keywords: *Intrusion Detection Systems, Machine Learning, Cyber Attack, UNSW-NB15 dataset, CICIDS2018 dataset, CICIDS2017 dataset, DARPA dataset, UNSW-NB15 dataset, NSL-KDD dataset, KDD99 dataset, ADFA-IDS data.*

1. Introduction
Researchers have relied on standard datasets to evaluate their results when using anomaly-based intrusion detection systems against attacks. However, the currently available datasets lack the actual characteristics of network traffic, which is why most anomaly-based intrusion detection systems are not applicable in current business environments [1]. Furthermore, intrusion detection systems are unable to adapt to constant changes in networks (i.e., new nodes, changes in traffic loads, changing topology, etc.) [2]. These changes make it difficult to rely on old data sets only It does not help in the development of intrusion detection systems. The process of creating new datasets must take into account this fact of constant change. For example, proposing to generate a standard dataset with extensible functionality would remove the burden of creating datasets from scratch [3]. Data sets are either real (i.e., recorded from network settings) or synthetic (i.e., simulated or injected traffic) [4]. Artificial attack injection can be used to either introduce attacks onto an existing data set or parallelize existing attack classes in the data set [5]. The researcher mentioned in [3] that in order to deal with a data set, the set must cover the following matters

- **Network Configuration:** Network Configuration indicates that one has complete knowledge of the network topology on how networking devices are connected in

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 02 February 2024
Page No: 254 - 260
https://irjaeh.com

the test environment so that real-life attack scenarios are captured [6].

- **Network Traffic:** Network Traffic refers to the capture of all network packets from the host, router, firewall, and web applications to download the flow and generate the data set [7].
- **Labeled Dataset**: Refers to labeling instances of data captured from network traffic to get a complete understanding of network interaction [8].
- **Network Interaction:** Network Interaction refers to the existence of a complete record of network communications inside and outside the network [9].
- **Capturing the Traffic**: Capturing the Traffic refers to capturing functional and non-functional network traffic to measure DR and FPR from IDS [10].
- **Protocols:** An ideal data set should include all communications using different protocols, whether normal or malicious [11].
- **Attacks:** The data set should consist of broad and up-to-date attack categories [12].
- **Features:** The dataset must maintain a complete set of well-defined features to classify the stars [13].
- **Heterogeneity:** The dataset must be collected from different sources to cover all details of the attack detection procedure [14].
- **Metadata:** The dataset must contain appropriate documentation describing the test environment, the infrastructure of the attack system, the infrastructure of the victim system, and the scenarios used in the attacks [15, 16]. Table 1 summarizes the available datasets and has been classified based on the domain to which they belong.

2. CIC-IDS2017 Dataset

Published by the Canadian Institute for Cybersecurity in the year 2017, this data set is now a benchmark for research in anomaly detection and intrusion detection research [17]. The data set is a collection of packets captured during the interval of 5 days in 8 separate sessions. Subsequently, the data has been released into 8 files each from one session; for machine learning research, it is preformatted into Comma Separated Values (.csv). Individual Files in CIC-IDS2017 Data Set with Instances are shown in Figure 1.

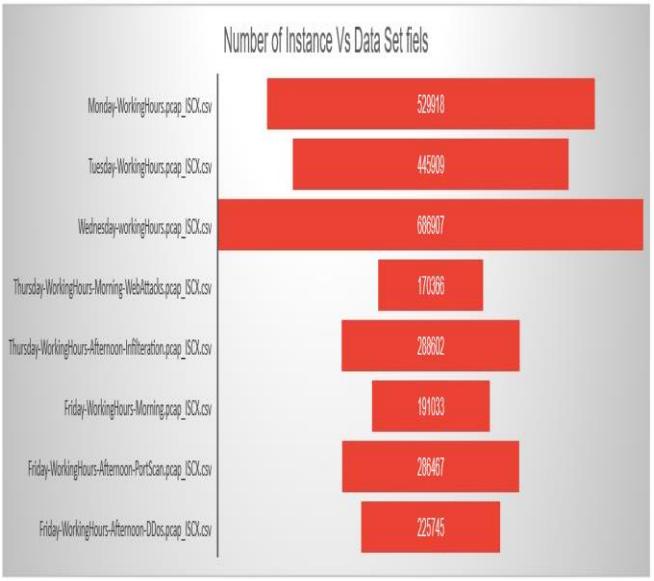

Figure 1 Individual Files In CIC-IDS2017 Data Set with Instances

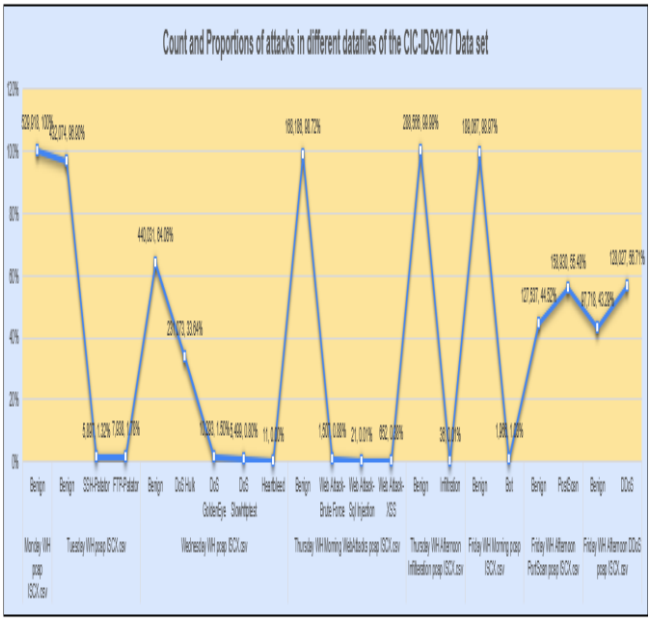

Figure 2 Count and Proportions of Different Attack Types in CIC-IDS2017 Data Set

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 02 February 2024
Page No: 254 - 260
https://irjaeh.com

Table 1 Summary of CIC-IDS2017 Data set

File Name	Attacks	Attack Counts	Proportion	Total Incidents
Monday-WorkingHours.pcap_ISCX.csv	Benign	529,918	100%	529918
Tuesday-WorkingHours.pcap_ISCX.csv	Benign	432,074	96.90%	445,909
	SSH-Patator	5,897	1.32%	
	FTP-Patator	7,938	1.78%	
Wednesday-workingHours.pcap_ISCX.csv	Benign	440,031	64.06%	686,907
	DoS Hulk	231,073	33.64%	
	DoS GoldenEye	10,293	1.50%	
	DoS Slowhttptest	5,499	0.80%	
	Heartbleed	11	0.00%	
Thursday-WorkingHours-Morning-WebAttacks.pcap_ISCX.csv	Benign	168,186	98.72%	170,366
	Web Attack-Brute Force	1,507	0.88%	
	Web Attack-Sql Injection	21	0.01%	
	Web Attack-XSS	652	0.38%	
Thursday-WorkingHours-Afternoon-Infilteration.pcap_ISCX.csv	Benign	288,566	99.99%	288,602
	Infiltration	36	0.01%	
Friday-WorkingHours-Morning.pcap_ISCX.csv	Benign	189,067	98.97%	191,033
	Bot	1,966	1.03%	
Friday-WorkingHours-Afternoon-PortScan.pcap_ISCX.csv	Benign	127,537	44.52%	286,467
	PortScan	158,930	55.48%	
Friday-WorkingHours-Afternoon-DDos.pcap_ISCX.csv	Benign	97,718	43.29%	225,745
	DDoS	128,027	56.71%	

2.1 IC-IDS2017 and Other Data Sets

There is a plethora of Data sets available for research in the field of Network Security and intrusion Detection. Table 2 summarizes the key Characteristics of the CIC-IDS2017 Data set with some of the previously available data sets. Count and Proportions of Different Attack Types in CIC-IDS2017 Data Set are shown in Figure 2.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 02 February 2024
Page No: 254 - 260
https://irjaeh.com

Table 2 Comparison of CIC-IDS with Other Data Sets

Feature	CICIDS2017	NSL-KDD	UNSW-NB15	BoT-IoT	KDDCup 1998	CSE-CIC-IDS2018
Attack Types	Diverse (DoS, DDoS, Botnet, Brute-force, Web, Infiltration)	DoS, DDoS, Probe, U2R, R2L	Diverse (DoS, DDoS, Botnet, Web, Fuzzers, Exploits)	IoT-specific attacks (DoS, DDoS, Botnet, Scanning, Malicious Traffic)	DoS, Probe, U2R, R2L	Botnet, DoS, DDoS, Web, Infiltration, Brute-force
Traffic Capture	PCAP and CSV files	Pre-processed flow records	PCAP files and processed features	PCAP files and processed features	PCAP files and processed features	PCAP, CSV files, and processed features
Benign Traffic	Yes	No	Yes	Yes	No	Yes
Attack Scenarios	Simulated realistic scenarios	Predefined attack types	Diverse real-world attack recordings	Simulated and real-world IoT attacks	Predefined attack types	Real-world attack recordings (malware+network traffic)
Number of Instances	11.8 million	489,843	2.5 million	3.1 million	4.8 million	2.7 million
Number of Features	85	41	66	233	41	85
Complexity	Moderate	Simple	Complex (mix of real-world and simulated)	Moderate (simulated and real-world)	Simple	Complex (real-world malware + network traffic)
Strengths	Realistic scenarios, diverse attacks, labeled data	Large dataset, simple to use	Real-world attacks, rich features, labeled data	IoT-specific, a mix of simulated and real-world	Large dataset, established benchmark	Real-world malware interaction, diverse attacks, labeled data
Weaknesses	Scattered Presence, Huge Volume of Data, Missing Values	Outdated attacks, unrealistic scenarios	Imbalanced classes, complex features	Limited attack types, simulated scenarios	Outdated attacks, unrealistic scenarios	High computational cost, imbalanced classes

3. Research Work done with CIC-IDS Data Set

Plenty of work has already been done with the CIC-IDS2017 data set. It has been explored deeply for various research objectives. We will summarize our observations into various categories of the work done.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 02 February 2024
Page No: 254 - 260
https://irjaeh.com

3.1 Intrusion Detection Systems (IDS)

a. **Machine Learning Algorithm Evaluation:** Researchers compare and evaluate various ML algorithms like SVM, Random Forests, and Neural Networks for intrusion detection using CIC-IDS2017, identifying effective approaches for practical IDS implementations [18].

b. **Deep Learning Applications: Deep Learning Applications:** Studies explore the use of deep learning techniques like CNNs and RNNs for advanced intrusion detection with high accuracy and adaptability to complex attack patterns [19].

c. **Anomaly Detection Techniques**: Research investigates anomaly-based IDS approaches using statistical methods or one-class classifiers to identify abnormal network traffic patterns indicative of potential intrusions [20][21].

3.2 Feature Selection and Engineering

a. **Identifying Relevant Features:** Studies analyze the CICIDS2017 features to determine their effectiveness in intrusion detection, leading to selecting the most features and reducing data complexity for better model performance [22].

b. **Identifying Relevant Features:** Studies analyze the CICIDS2017 features to determine their effectiveness in intrusion detection, leading to selecting the most features and reducing data complexity for better model performance [23,24]

3.3 Network Traffic Analysis and Characterization

a. **Attack Behavior Understanding:** Studies analyze the characteristics of different attack types present in the CIC-IDS2017 dataset to understand their patterns and potential evasion strategies [25].

b. **Botnet Detection and Analysis:** Research focuses on identifying and analyzing botnet activity within the network traffic data, aiming to develop effective defense mechanisms against botnet-based attacks [26].

c. **Emerging Threat Detection:** Studies explore the use of the CIC-IDS2017 dataset for training models to detect novel or zero-day attacks not explicitly labeled in the data, enhancing the adaptability of IDS systems [27].

3.4 Dataset Analysis and Improvement

a. **Data Quality Assessment:** The research investigates the quality and potential limitations of the CIC-IDS2017 dataset, identifying issues like class imbalance or data biases that might affect research results [28].

b. **Dataset Augmentation Techniques:** Studies explore methods to artificially generate additional data points based on the CIC-IDS2017 dataset, addressing class imbalance, and potentially improving the generalizability of trained models [29].

c. **Comparative Analysis with Other Datasets:** The research compares the CIC-IDS2017 dataset with other intrusion detection datasets to evaluate its strengths and weaknesses for different research objectives [30].

Conclusion

Most of the datasets do not include authentic network traffic. The majority of organizations refrain from revealing their network traffic as a result of concerns over confidentiality. Consequently, there is a substantial need for up-to-the-minute network traffic statistics. The aforementioned data sets are inadequate to keep pace with the ongoing advancements in new attacks that have adopted novel and unfamiliar methods. Moreover, these datasets fail to address or incorporate attacks and threats associated with software containers, which have recently gained rapid adoption. This situation poses significant challenges for intrusion detection systems and the existing datasets, necessitating the development of a new behavioral representation mechanism to detect unknown threats. Effective development of an Intrusion Detection System (IDS) necessitates the provision of a real-time assault scenario, including new attacks.

References

[1] K. Kendall, “A Database of Computer Attacks for the Evaluation of Intrusion Detection Systems,” Master Thesis, pp. 12–26, 1999, doi: citeulike-article-id:9077111.

[2] R. P. Lippmann et al., “Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation,” Proc. - DARPA Inf. Surviv. Conf. Expo. DISCEX 2000, vol. 2, 2000, pp. 12–26, doi: 10.1109/DISCEX.2000.821506.

[3] Lee, W., &Stolfo, S. J. A Framework for Constructing Features and Models for Intrusion Detection Systems (Vol. 3), 2001.

[4] M. Tavallaee, E. Bagheri, W. Lu and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, Ottawa, ON, Canada, 2009, pp. 1-6, doi: 10.1109/CISDA.2009.5356528.

[5] O. Atilla and E. Hamit, “A review of KDD99 dataset usage in intrusion detection and machine learning between 2010 and 2015,” PeerJ, 2016, pp. 0–21, doi: 10.7287/PEERJ.PREPRINTS.1954V1.

[6] “Generation of a new IDS test dataset: Time to retire the KDD collection.” in Proceedings of the Wireless Communications and Networking Conference (WCNC). IEEE, 2013, pp. 4487–4492.

[7] M. J. M. Turcotte, A. D. Kent, and C. Hash, “Unified Host and Network Data Set.” arXiv e-prints, pp. 1–16, August 2017.

[8] S. Behal and K. Kumar, “Measuring the impact of DDoS attacks on web services-a realtime experimentation.” International Journal of Computer Science and Information Security, vol. 14, no. 9, p. 323, 2016.

[9] J. J. Santanna, R. van Rijswijk-Deij, R. Hofstede, A. Sperotto, M. Wierbosch, L. Z. Granville, and A. Pras, “Booters â˘A ˘T an analysis of DDoSas-a-service attacks.” in 2015 IFIP/IEEE International Symposium on Integrated Network Management (IM), 2015, pp. 243–251.

[10] B. A. Khalaf, S. A. Mostafa, A. Mustapha, M. A. Mohammed and W. M. Abduallah, "Comprehensive review of artificial intelligence and statistical approaches in distributed denial of service attack and defense methods", IEEE Access, vol. 7, pp. 51691-51713, 2019.

[11] Kurniabudi, D. Stiawan, Darmawijoyo, M. Y. B. Idris, A. M. Bamhdi and R. Budiarto, "CICIDS-2017 dataset feature analysis with information gain for anomaly detection", IEEE Access, vol. 8, pp. 132911-132921, 2020.

[12] Ahmad and E. Harjula, "Evaluation of machine learning techniques for security in SDN", Proc. IEEE Globecom Workshops, pp. 1-6, Sep. 2020.

[13] J. Meira, R. Andrade, I. Praça, J. Carneiro, V. Bolón-Canedo, A. Alonso-Betanzos, et al., "Performance evaluation of unsupervised techniques in cyber-attack anomaly detection", J. Ambient Intell. Humanized Compute. vol. 11, no. 11, pp. 4477-4489, Nov. 2020.

[14] Sharafaldin, A. Gharib, A. H. Lashkari and A. A. Ghorbani, "Towards a reliable intrusion detection benchmark dataset", Softw. Netw. vol. 2017, no. 1, pp. 177-200, 2017.

[15] Q. Dang, "Studying machine learning techniques for intrusion detection systems", Stud. Mach. Learn. Tech., 2019.

[16] J. Kim, J. Kim, H. Kim, M. Shim and E. Choi, "CNN-based network intrusion detection against denial-of-service attacks", Electron. vol. 9, no. 6, pp. 1-21, 2020.

[17] Canadian Institute for Cybersecurity, “Intrusion detection evaluation dataset (CICIDS2017).” 2017. Available: http://www.unb.ca/cic/datasets/ids-2017.html

[18] A. H. Hassan, W. M. Shah, M. F. I. Othman and H. A. H. Hassan, "Evaluate the performance of K-means and the fuzzy C-means algorithms to formation balanced clusters in wireless sensor networks", Int. J. Electr. Comput. Eng., vol. 10, no. 2, pp. 1515-1523, 2020.

[19] Zafar Iqbal Khan and Mohammad Mazhar Afzal, “Security in Wireless Sensor Networks: DoS Perspective,” International Journal of Engineering Research and, vol. V6, no. 01, Jan. 2017, doi: 10.17577/ijertv6is010250.

[20] M. Nawir, A. Amir, N. Yaakob and O. B. Lynn, "Multi-classification of UNSW-NB15 dataset for", J. Theor. Appl. Inf. Technol., vol. 96, no. 15, pp. 5094-5104, 2018.

[21] Z. Pelletier and M. Abualkibash, "Evaluating the CIC IDS-2017 dataset using machine learning methods and creating multiple predictive models in the statistical computing language R", Science, vol. 5, no. 2, pp. 187-191, 2020

[22] H. Yao, D. Fu, P. Zhang, M. Li and Y. Liu, "MSML: A novel multilevel semi-supervised machine learning framework for intrusion detection system", IEEE Internet Things J., vol. 6, no. 2, pp. 1949-1959, Apr. 2018.

[23] K. Peng, V. C. M. Leung, L. Zheng, S. Wang, C. Huang and T. Lin, "Intrusion detection system based on decision tree over big data in fog environment", Wireless Commun. Mobile Compute. vol. 2018, pp. 1-10, Mar. 2018.

[24] B. Chakrabarty, O. Chanda and M. Saiful, "Anomaly based intrusion detection system using genetic algorithm and K-centroid clustering", Int. J. Comput. Appl., vol. 163, no. 11, pp. 13-17, Apr. 2017.

[25] V. Chahar, R. Chhikara, Y. Gigras and L. Singh, "Significance of hybrid feature selection technique for intrusion detection systems", Indian J. Sci. Technol., vol. 9, no. 48, Jan. 2017.

[26] R. A. R. Ashfaq, X.-Z. Wang, J. Z. Huang, H. Abbas and Y.-L. He, "Fuzziness based semi-supervised learning approach for intrusion detection system", Inf. Sci., vol. 378, pp. 484-497, Feb. 2017.

[27] M. Ring, S. Wunderlich, D. Grüdl, D. Landes and A. Hotho, Big data analytics for intrusion detection system: Statistical decision-making using finite Dirichlet mixture models, pp. 3-31, 2017.

[28] M. Idhammad, K. Afdel and M. Belouch, "Semi-supervised machine learning approach for DDoS detection", Appl. Intell., vol. 48, no. 10, pp. 3193-3208, 2018.

[29] W. He, H. Li and J. Li, "Ensemble feature selection for improving intrusion detection classification accuracy", Proc. Int. Conf. Artif. Intell. Comput. Sci., pp. 28-33, Jul. 2019.

[30] V. Kumar, D. Sinha, A. K. Das, S. C. Pandey and R. T. Goswami, "An integrated rule based intrusion detection system: Analysis on UNSW-NB15 data set and the real time online dataset", Cluster Comput., vol. 23, no. 2, pp. 1397-1418, Jun. 2020.