Frontiers in Health Informatics ISSN-Online:2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 3
Open Access

Enhancing Cybersecurity with Tailored Datasets: Building an Intrusion Detection Dataset

Maduri Madhavi[1], Dr. N P Nethravathi[2]
Research scholar, Reva University, Department of CSE, India
Professor, Reva University, Department of CSE, India
[1] madurimadhavi@gmail.com, [2] nethravathi.np@reva.edu.in

Cite this paper as: Maduri Madhavi, N P Nethravathi (2024) Enhancing Cybersecurity with Tailored Datasets: Building an Intrusion Detection Dataset. *Frontiers in Health Informatics, 13* (3), 8788-8795

Abstract: *Datasets play a crucial role in developing and evaluating intrusion detection systems (IDS). These datasets typically contain network traffic data, including both normal and malicious activities, to train and test IDS algorithms. Some widely used datasets for IDS research include KDD Cup 99, NSL-KDD, UNSW-NB15, and CICIDS2017. These datasets provide a diverse range of network attacks, such as denial-of-service (DoS), probe, user-to-root (U2R), and remote-to-local (R2L) attacks. However, it is important to note that many of these datasets are outdated and may not accurately represent current network threats. To address this issue, researchers are continuously developing new datasets that incorporate emerging attack patterns and reflect modern network environments. When selecting a dataset for IDS research, it is crucial to consider factors such as its relevance to current threats, the balance between normal and malicious traffic, and the presence of diverse attack types to ensure the development of robust and effective intrusion detection systems. In this paper different datasets for IDS were discussed along with their characteristics and creation of dataset by simulation of attacks in virtual environment*

I Introduction

Intrusion detection systems play a crucial role in safeguarding computer networks and systems from unauthorized access, malicious activities, and cyber threats. In order to develop effective intrusion detection models, researchers and practitioners often rely on various datasets that serve as benchmarks for evaluating the performance of their detection algorithms. (Prasad et al., 2015) (MaqboolBeigh et al., 2013) The primary

objective of this paper is to present a comprehensive review of the datasets commonly utilized in the field of intrusion detection systems, with a focus on highlighting
their distinct characteristics, advantages, and limitations.
Intrusion detection datasets can be broadly categorized based on their source, network environment, and the types of attacks they encompass. (Jindal & Anwar, 2021) This paper aims to provide a thorough analysis of the existing datasets, their evolution, and the challenges associated with their usage in the context of intrusion detection research.

The key contributions of this paper are as follows:

We provide a taxonomy and comprehensive survey of the commonly used intrusion detection datasets, delving into their sources, features, and the diverse range of attack types they represent, thereby offering researchers and practitioners a consolidated reference on the available benchmark datasets in this domain.
While the existing datasets have been instrumental in advancing intrusion detection research, they often fail to capture the evolving nature of cyber threats and intrusion patterns. Therefore, we discuss the limitations of the current datasets and emphasize the need for the development of more realistic and up to date datasets that can better reflect the dynamic threat landscape.

II Overview of Intrusion Detection Systems

Intrusion detection systems have become an integral component of modern cybersecurity infrastructure, playing a crucial role in monitoring, analyzing, and responding to potential security threats within computer networks and systems. These systems are designed to detect a wide range of malicious activities, including unauthorized access attempts, network based attacks, insider threats, and various other forms of cyber threats, with the primary goal of alerting system administrators or security teams to take appropriate countermeasures.

Intrusion detection systems can be broadly classified into two main categories: (Thakkar & Lohiya, 2020) (Hindy et al., 2018) (MaqboolBeigh et al., 2013)

Signature based detection systems, which rely on predefined patterns or signatures to identify known attacks, and anomaly based detection systems, which aim to detect deviations from normal or expected behavior, potentially uncovering previously unknown or zero day attacks. The effective deployment of intrusion detection systems is contingent upon the availability of comprehensive and representative datasets that can be used to train, test, and evaluate the performance of these systems.

In order to address this requirement, researchers have developed and utilized a wide range of intrusion detection datasets, each with its own unique characteristics, such as the source of the data, the network environment in which it was collected, and the types of attacks they attempt to capture (Hindy et al., 2018)(Thakkar & Lohiya, 2020)(MaqboolBeigh et al., 2013)(Gharib et al., 2016).(Thakkar & Lohiya, 2020)Researchers have developed and utilized a wide range of intrusion detection datasets, each with its own unique characteristics, such as the source of the data, the network environment in which it was collected, and the types of attacks they attempt to capture.

III Literature Review

The literature review presents a comprehensive overview of the various datasets that have been used in the field of intrusion detection systems. The existing research has explored the limitations and challenges associated with the currently available intrusion detection datasets, highlighting the need for more realistic and up to date datasets that can better reflect the evolving nature of cyber threats (MaqboolBeigh et al., 2013) (Thakkar & Lohiya, 2020) (Gharib et al., 2016)
Intrusion Detection Systems (IDS) rely heavily on high quality datasets for development, training, and evaluation. This review examines the most significant datasets that have shaped IDS research over the past decades, analyzing their characteristics, applications, and limitations.

Traditional Benchmark Datasets

KDD Cup 99 Dataset

The KDD Cup 99 dataset, derived from DARPA'98 IDS evaluation program, has been one of the most widely used datasets in the field of network intrusion detection [1]. It contains approximately 4.9 million instances with 41 features and represents network connection records with labeled normal and attack traffic.

This dataset, created by DARPA, has been a benchmark for many years in IDS research. It contains a wide range of simulated intrusions in a military network environment. However, it has been criticized for being outdated and not representative of modern network traffic patterns. The KDD Cup 1999 dataset has been extensively used for detecting and classifying anomalies in computer networks[8]. It was generated by domain experts at MIT Lincoln Lab and consists of five million records, each containing 41 features that classify malicious attacks into four categories: Probe, DoS, U2R, and R2L (Protić, 2018; Shah & Trivedi, 2015). This dataset has been a cornerstone for research in intrusion detection systems, with various studies focusing on improving its effectiveness and efficiency. Interestingly, while the KDD Cup 1999 dataset has been widely used, it has some limitations. It cannot

reflect real traffic data as it was generated by simulation over a virtual computer network (Protić, 2018). To address this, the NSL-KDD dataset was created by removing redundant and duplicate records from the KDD Cup 1999 dataset's training and test sets (Protić, 2018). Additionally, the Kyoto 2006+ dataset was developed using real three-year network traffic data, incorporating 14 statistical features derived from the KDD Cup 1999 dataset and 10 additional features [9][10].

Key characteristics:

41 features per connection record
4 main categories of attacks: DoS, Probe, R2L, and U2R
Includes basic features (e.g., duration, protocol_type) and derived features

Limitations:

Outdated attack patterns
Redundant records
Unrealistic attack to normal ratio

NSL KDD Dataset

The NSL KDD dataset was proposed to address the inherent problems of the KDD Cup 99 dataset [2]. It eliminates redundant records and provides a more balanced distribution of attack types.

The NSL-KDD dataset is a widely used benchmark in intrusion detection research, addressing several shortcomings of its predecessor, the KDD Cup '99 dataset. It was created by removing redundant and duplicate records from the KDD Cup '99 dataset, making it more suitable for realistic network traffic analysis (Protić, 018; Sahli, 2022; Yadav & Kalpana, 2019). One of the key advantages of the NSL-KDD dataset is its improved Representation of real-world network traffic. Unlike the KDD Cup '99 dataset, which was generated through simulation over a virtual computer network, the NSL-KDD dataset provides a more accurate reflection of actual network behavior[11]. This enhancement makes it a valuable resource for developing and testing intrusion detection systems (IDS). Interestingly, the NSL-KDD dataset has been used in various machine learning approaches for intrusion detection[12]. For instance, a study implementing Deep Generative Adversarial Networks (GANs) for data augmentation achieved an impressive accuracy of 99.78% using the XGBoost model (Madany et al., 2023). Another research utilized decision trees with Correlation Feature Selection (CFS) subset evaluation, demonstrating high detection rates and accuracy for both binary and multi-class classification tasks (Ingre et al., 2017). However, it's worth noting that a study focusing on manual rule development highlighted a mislabelling problem in the NSL-KDD dataset, particularly for R2L and U2R attacks [13][14].

Improvements over KDD Cup 99:

Removed redundant records
Balanced class distribution
More reasonable record numbers
No duplicate records in train/test sets

IV Modern Datasets

CICIDS2017

The CICIDS2017 dataset represents a significant advancement in IDS datasets, offering modern attack scenarios and benign network traffic [3]. Created by the Canadian Institute for Cybersecurity, it contains network traffic data collected over five days.

The CICIDS2017 dataset is a widely used benchmark for network intrusion detection systems (NIDS) research. It contains network traffic data with various types of attacks and has been employed in numerous studies to develop and evaluate machine learning-based intrusion detection models[18][19]. However, an in-depth analysis of the CICIDS2017 dataset revealed several issues with its data collection pipeline, including problems with traffic generation, flow construction, feature extraction, and labeling (Engelen et al., 2021). These hortcomings

significantly affect the dataset's correctness, validity, and overall utility for learning tasks. As a result, more than 20% of the original traffic traces had to be reconstructed or relabeled, leading to significant improvements in machine learning benchmarks (Engelen et al., 2021). Despite these challenges, the CICIDS2017 dataset remains valuable for NIDS research. Exploratory Data Analysis (EDA) techniques have been used to reveal important relationships between input variables and target classes, providing insights into the dataset's nature, structure, and potential issues (Oyelakin et al., 2023; Oyelakin, 2024). This understanding can help esearchers make informed decisions when using the dataset for future IDS studies and improve the development of machine learning-based intrusion detection system

Key features:
Over 80 network traffic features
Includes modern attack types
Labeled flow based network traffic
Realistic network environment

UNSW NB15

Developed by the Australian Centre for Cyber Security, the UNSW NB15 dataset provides a hybrid of real modern normal activities and synthetic contemporary attack behaviors [4].
The UNSW-NB15 dataset, created by the Australian Cyber Security Centre in 2015, is a widely used benchmark for network anomaly detection. It contains normal data and nine types of attacks with 49 features, extracted using the IXIA tool (Thanh & Lang, 2020). This dataset is particularly valuable for developing and evaluating intrusion detection systems due to its inclusion of modern attack types and a wide variety of real normal activities (Sonule, 2021). Several studies have employed various machine learning techniques to detect anomalies in the UNSW-NB15 dataset(Thanh & Lang, 2020) . For instance, ensemble techniques like AdaBoost and Random Forest have shown promising results, with AdaBoost achieving an F-Measure of 96.76% for detecting Fuzzers attacks (Thanh & Lang, 2020). Other approaches include the use of REPTree classifier, which achieved 99.94% accuracy for Worms class detection with a false alarm rate of 0.000 (Roy & Singh, 2020), and a deep learning-based Convolutional Neural Network (CNN) model that achieved 99.00% testing accuracy (Vibhute et al., 2024). Interestingly, while the UNSW-NB15 dataset is considered more modern and representative of current network traffic, a comparative study found that the older KDD99 dataset performed better in terms of accuracy and false alarm rate (Moustafa & Slay, 2015). Additionally, visual analysis of the UNSW-NB15 dataset revealed issues of class imbalance and class overlap, which researchers should address before using it for classifier model development (Zoghi & Serpen, 2021)
Characteristics:
49 features
Nine types of modern attacks
2.5 million records
Realistic attack scenarios

V Creation of own dataset

In the realm of computer security, the Denial of Service (DoS) is a form of attack strategy that aims to incapacitate the achievement of a computer resource or group of computer networks to its legitimate users by overloading them with requests or traffic that typifies them until the resources available are exhausted making them unable to function normally for the intended users.

Classes of Denial-of-Service Attacks.
Volumetric attacks focus on surpassing the capacity of bandwidth of the targeted network or online service.

UDP Floods An assault of kind involves sending enormous amounts of user datagram protocol packets aimed at

specific ports in the target so as to exhaust the system into searching for these applications making it unable to respond to other ports this depletes the system resources.

An **ICMP Flood** is a type of attack whereby n the target is usually flooded with made requests ICMP (ping) to which it has to reply or echoes draining the resources of the targeted network.
The term Ping of Death describes the act of transmitting over-sized packets, misformed packets or even just malformed packets. Any machine not designed to handle such would most likely hang or blue screen of death. Definition Protocol Attacks is a case whereby the servers benefit from the exploiting of the weaknesses of the network protocols.

Using **SYN Flood** also known as Stacheldraht Attack wherein a pre determined number of TCP SYN packets are sent to a server before any acknowledgement is received thus keeps the connection partly open and eventually uses up all the available resources.

LAND Attack involves delivery of packets containing the same source and destination ip addresses and ports. This act may confuse the target system, as stated before, it can also cause the system to crash.

A Denial-of-service attack in which an attacker causes an ip address to issue an icmp or ping request from within the target network without actually issuing a request i.e. use of a spoofed ip address is called a Smurf Attack.

To create a dataset Denial of service attack was considered. These are the following attacks that were simulated using a virtual machine. In this kali Linux and windows operating systems were used .Kali is used for launching an attack and windows is the victim. The following attacks were simulated ICMP Flood, TCPSYN flood and UDP flood.

The following are the features collected by simulating the attacks using kali Linux and windows7 operating system

UTC Time, Time, Relative Time,Absolute Time Delta Time, Source Destination, Protocol, Length,Info,Source Port, Dest Port, Cumulativebytes,Hwdestaddr,Hwsrcaddr,Unresolved,Destport,UnresolvedSrcport,NetSrcAddr, NetDestAddr, ExpertInfo

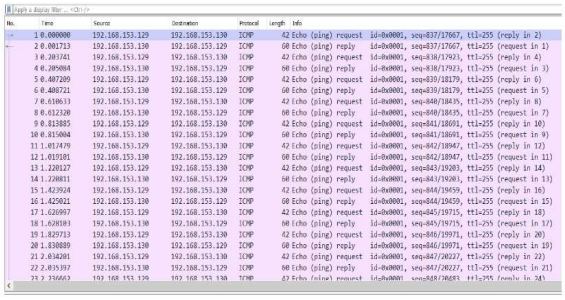
ICMP flood
An ICMP flood, commonly referred to as a ping flood, is a powerful Denial of Service (DoS) attack specifically engineered to overload a target with an excessive number of ICMP Echo Request (ping) packets. This attack exploits the Internet Control Message Protocol (ICMP), which is generally used for connectivity checks and response time measurements. However, in the case of an ICMP flood, it is weaponized to drain the resources of the targeted machine or network.

How an ICMP Flood Works:

1. Packet Transmission: The attacker unleashes a substantial volume of ICMP Echo Request packets (pings) directed at the target.

2. Resource Consumption: The target machine is obligated, according to the ICMP protocol, to respond with an ICMP Echo Reply for each packet it receives.

3. Overwhelming the Target: When the attack traffic reaches a critical level, it decisively consumes the target's bandwidth or processing resources, resulting in:
4. Network Congestion: Legitimate traffic is obstructed due to the massive influx of ping requests.
5. System Overload: The target's CPU and memory are pushed to their limits as they struggle to process the relentless barrage of ICMP requests.

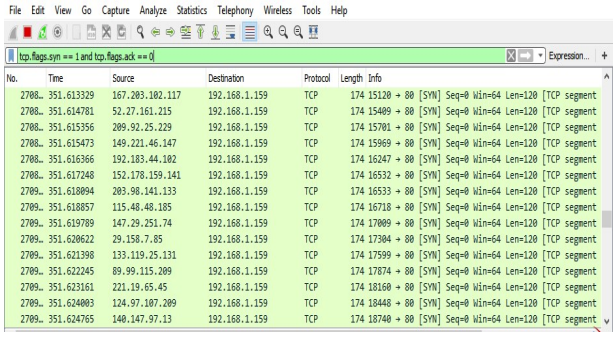


TCPSYN flood

TCP SYN flood is a Denial of Service (DoS) attack that exploits the TCP three-way handshake process to overwhelm a server with half-open connections, exhausting its resources and preventing legitimate users from connecting.

How a TCP SYN Flood Works:

The attack takes advantage of how a TCP connection is established:
SYN Packet: The client sends a SYN (synchronize) packet to the server, requesting to open a connection.
SYN-ACK Packet: The server responds with a SYN-ACK (synchronize-acknowledgment) packet, acknowledging the request and indicating readiness to complete the handshake.
ACK Packet: The client is supposed to reply with an ACK packet to establish the full connection.
In a **TCP SYN** flood attack:
The attacker sends many SYN packets to the server with spoofed (fake) IP addresses.
The server responds with SYN-ACK packets to these spoofed addresses, waiting for the final ACK that never arrives.
The server's resources become tied up with these half-open connections, which eventually leads to:
Resource Exhaustion: The server's memory and connection pool get exhausted, preventing it from processing new, legitimate connections.
Denial of Service: The server can become unresponsive to legitimate users as it's overwhelmed with incomplete requests.



UDP flood

A UDP flood is a type of Denial of Service (DoS) attack in which an attacker overwhelms a target with a high volume of User Datagram Protocol (UDP) packets. This attack takes advantage of the connectionless nature of UDP, where packets are sent without establishing a formal connection or handshake with the target.

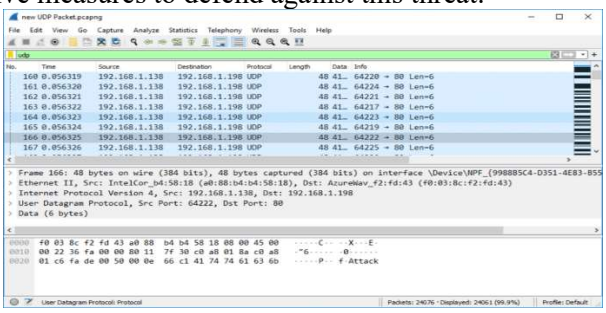
How a UDP Flood Works

1. Packet Transmission : The attacker sends a large number of UDP packets to random or specific ports on the victim's system.

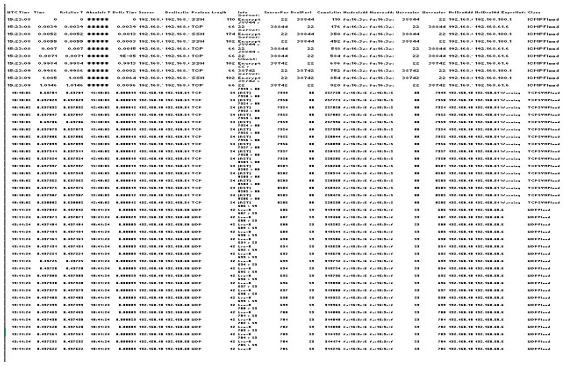
2. Resource Exhaustion : When the target receives a UDP packet on an open port, it attempts to process it as usual. If the packet is sent to a closed port, the target responds with an ICMP "Destination Unreachable" message. In both scenarios, the resources (bandwidth, CPU, memory) of the target are consumed while processing these packets.

3. Denial of Service: When the server or network is flooded with UDP packets, it may experience:
- Lack of Bandwidth: This can cause delays or drops in legitimate traffic.
- Exhaustion of Processing Power: The server's CPU and memory may become overwhelmed by the demand to respond to or handle these packets.

This leads to critical service disruptions, effectively preventing legitimate users from accessing the services offered by the targeted server. The consequences of a UDP flood attack are immediate and damaging, and organizations must take proactive measures to defend against this threat.



The dataset



VI Conclusion

Recent trends in IDS datasets show a move toward: More realistic network environments Integration of multiple data sources Focus on specific domains (IoT, ICS) Inclusion of zero day attacks Better documentation and reproducibilityThe evolution of IDS datasets reflects the changing landscape of cyber threats and network environments. While traditional datasets like KDD Cup 99 and NSL KDD have been instrumental in developing early IDS solutions, modern datasets such as CICIDS2017 and UNSW NB15 provide more realistic and contemporary attack scenarios. Specialized datasets for IoT and ICS environments are becoming increasingly important as these sectors face growing security challenges.

References

[1] Stolfo, S. J., et al. (1999). "KDD cup 99 dataset." UCI KDD repository.

[2] Tavallaee, M., et al. (2009). "A detailed analysis of the KDD CUP 99 data set." IEEE CISDA 2009.
[3] Sharafaldin, I., et al. (2018). "Toward generating a new intrusion detection dataset and intrusion traffic characterization." ICISSp 2018.
[4] Moustafa, N., and Slay, J. (2015). "UNSW NB15: a comprehensive data set for network intrusion detection systems." IEEE MilCIS 2015.
[5] Koroniotis, N., et al. (2019). "Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics." Future Generation Computer
Systems.
[6] Alsaedi, A., et al. (2020). "TON_IoT telemetry dataset: a new generation dataset of IoT and IIoT for data driven Intrusion Detection Systems." IEEE Access.
[7] P. Kalpana, P. Srilatha, G. S. Krishna, A. Alkhayyat and D. Mazumder, "Denial of Service (DoS) Attack Detection Using Feed Forward Neural Network in Cloud Environment," *2024 International Conference on Data Science and Network Security (ICDSNS)*, Tiptur, India, 2024, pp. 1-4, https://doi.org/10.1109/ICDSNS62112.2024.10691181.
[8] C. E. Landwehr, A. R. Bull, J. P. McDermott, and W. S. Choi, "A taxonomy of computer program security flaws," ACM Comput. Surv., vol. 26, no. 3, pp. 211–254, 1994.
[9] M. Shyu, S. Chen, K. Sarinnapakorn, and L. Chang, "A novel anomaly detection scheme based on principal component classifier," Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with the Third IEEE International Conference on Data Mining (ICDM03), pp. 172–179, 2003.
[10] KDD Cup 1999. Available on: http://kdd.ics.uci.edu/databases/kddcup 99/kddcup99.html, Ocotber 2007.
[11] Nabi, S. A., Kalpana, P., Chandra, N. S., Smitha, L., Naresh, K., Ezugwu, A. E., & Abualigah, L. (2024). Distributed private preserving learning based chaotic encryption framework for cognitive healthcare IoT systems. *Informatics in Medicine Unlocked*, *49*, 101547. https://doi.org/10.1016/j.imu.2024.101547.
[12] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," pp. 463–484, 2012.
[13] G. Wang, J. Hao, J. Ma, and L. Huang, "A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering," Expert Systems with Applications, vol. 37, no. 9, pp. 6225–6232, 2010.
[14] P. Kalpana, K. Malleboina, M. Nikhitha, P. Saikiran and S. N. Kumar, "Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithm," *2024 International Conference on Data Science and Network Security (ICDSNS)*, Tiptur, India, 2024, pp. 1-7, https://doi.org/10.1109/ICDSNS62112.2024.10691297.
[15] M. S. Abadeh, J. Habibi, Z. Barzegar, and M. Sergi, "A parallel genetic local search algorithm for intrusion detection in computer networks," Engineering Applications of Artificial Intelligence, vol. 20, no. 8, pp. 1058–1069, 2007
[16] R. Longadge,S.S. Dongre,L. Malik, "Class Imbalance Problem in Data Mining: Review", International Journal of Computer Science and Network (IJCSN), Volume 2, Issue 1, February 2013, ISSN 2277-5420
[17] S.M.A. Elrahman and A. Abraham, A Review of Class Imbal-ance Problem, Journal of Network and Innovative Computing, ISSN 2160-2174, Volume 1 (2013) pp. 332-340
[18] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorba-ni, Intrusion Detection Evaluation Dataset (CICIDS2017), Ca-nadian Institute of Cybersecurity, http://www.unb.ca/cic/datasets/ids 2017.html, Accessed on: 13/04/2018

[19] Nallapaneni Manoj Kumar, Pradeep Kumar Mallick, "Blockchain technology for security issues and challenges in IoT", Elsevier Procedia Computer Science Journal, Volume 132, Pages 1815-1823 , 2018, ISSN:1877-0509, UGC Sl No: 46138 and 48229,DOI:https://doi.org/10.1016/j.procs.2018.05.140.