Cybersecurity attacks: which dataset should be used to evaluate an intrusion detection system?

Danijela D. Protić[a], *Miomir* M. Stanković[b]

[a] Serbian Armed Forces, General Staff, Department for Telecommunication and Informatics, Center for Applied Mathematics and Electronics, Belgrade, Republic of Serbia,
e-mail: danijelaprotic318@gmail.com, **corresponding author**,
ORCID iD: https://orcid.org/0000-0003-0827-2863,

[b] Mathematical Institute of the Serbian Academy of Sciences and Arts, Belgrade, Republic of Serbia,
e-mail: miomirdanijela@gmail.com,
ORCID iD: https://orcid.org/0009-0002-8504-6966

DOI: 10.5937/vojtehg71-46524; https://doi.org/10.5937/vojtehg71-46524

FIELD: computer sciences, electronics, telecommunications
ARTICLE TYPE: original scientific paper

Abstract:

Introduction: Analyzing the high-dimensional datasets used for intrusion detection becomes a challenge for researchers. This paper presents the most often used data sets. ADFA contains two data sets containing records from Linux/Unix. AWID is based on actual traces of normal and intrusion activity of an IEEE 802.11 Wi-Fi network. CAIDA collects data types in geographically and topologically diverse regions. In CIC-IDS-2017, HTTP, HTTPS, FTP, SSH, and email protocols are examined. CSE-CIC-2018 includes abstract distribution models for applications, protocols, or lower-level network entities. DARPA contains data of network traffic. ISCX 2012 dataset has profiles on various multi-stage attacks and actual network traffic with background noise. KDD Cup '99 is a collection of data transfer from a virtual environment. Kyoto 2006+ contains records of real network traffic. It is used only for anomaly detection. NSL-KDD corrects flaws in the KDD Cup '99 caused by redundant and duplicate records. UNSW-NB-15 is derived from real normal data and the synthesized contemporary attack activities of the network traffic.

Methods: This study uses both quantitative and qualitative techniques. The scientific references and publicly accessible information about given dataset are used.

Results: Datasets are often simulated to meet objectives required by a particular organization. The number of real datasets are very small compared to simulated dataset. Anomaly detection is rarely used today.

Conclusion: The main characteristics and a comparative analysis of the data sets in terms of the date they were created, the size, the number of features, the traffic types, and the purpose are presented.

Key words: ADFA, AWID, CAIDA, CIC-IDS-2017, CSE-CIC-2018, DARPA, ISCX 2012, KDD Cup '99, Kyoto 2006+, NSL-KDD, UNSW-NB15.

Introduction

With the increase in computer applications and large amounts of data being processed around the world, the need for data protection has multiplied in recent years. Intrusion detection systems (IDSs) are the primary line of defense that protects networks from malicious attacks. The IDS is generally classified into three installation types: host-based, network-based, and hybrid (Protić & Stanković, 2020). The network-based intrusion detection systems can also be divided into signature-based and anomaly-based, both of which are inspired by the human immune system. The signature-based (misuse-based) IDS protects the network by proactively detecting the presence of known attacks by comparing unknown network traffic against a database of known attack signatures. It detects malicious software based on the knowledge gathered through known attacks. The main advantage of signature-based IDSs is their high detection speed. The main disadvantage of signature-based IDSs is the difficulty in detecting unknown attacks. Anomaly-based IDSs detect unusual network behavior by detecting deviations from a statistical model of normal network behavior and by looking for activities that deviate from the created model. The main advantage of anomaly-based IDSs is the detection of unknown attacks. The main challenge in anomaly detection is determining what is identified as normal.

The main problem in intrusion detection is the huge amount of data. Since the type of features and the number of instances determine the applicability of IDSs, analyzing high-dimensional datasets becomes a challenge for researchers. Simulated datasets or datasets obtained from real network traffic differ in size, number of features, purpose, type of attacks, etc. (Omar et al, 2013; Jie et al 2018). A number of authors examine, describe and compare various datasets such as ADFA-LF, ADFA-WD, AWID, CAIDA, CIC-IDS-2017, CSE-CIC-2018, DARPA 98, SCX 2012, KDD Cup '99, Kyoto 2006+, NSL-KDD and UNSW-NB15 data sets, which differ in the number of features, type of attacks and purpose (Protić, 2018; Bohara et al, 2020; Borisniya & Patel, 2015; Thakkar &

Lohiya, 2020; Khraisat et al, 2019; Ferriyan et al, 2021; Serkani et al, 2019; Mighan & Kahani, 2021; Soltani et al, 2021).

In this paper, we present the main characteristics and a comparative analysis of the given data sets in terms of the date they were created, their size, attacks/anomalies, the number of their features, their traffic types, and their purpose.

Data sets

A list of ADFA-LF, ADFA-WD, AWID, CAIDA, CIC-IDS-2017, CSE-CIC-2018, DARPA 98, ISCX 2012, KDD Cup '99, Kyoto 2006+, NSL-KDD, and UNSW-NB15 data sets, with comprehensive descriptions, is given it the text that follows.

ADFA-LD and ADFA-WD datasets

In 2013, the Australian Defense Force Academy (ADFA) developed two data sets containing records from Linux/Unix (ADFA-LD) and Windows (ADFA-WD) systems, respectively. The datasets are free to use for research purposes only. The datasets are evaluated by the host-based IDS (HIDS) (system-call-based). The ADFA-LD consists of system call traces obtained from a temporary local Linux server, and six cyberattacks (Xie et al, 2014). The ADFA-WD is a set of DLL access requests and system calls from a variety of hacking attacks (2015). Both ADFA datasets are the benchmarks for evaluating IDS based on system calls.

ADFA-LD

System call traces are used by HIDS to detect attacks on target systems. ADFA-LD consists of 833 normal training traces, 4372 normal validation traces, and 746 attack traces, all collected under the Linux system, namely: Adduser (91), Hydra_FTP (162), Hydra_SSH (176), Java_Meterpreter (124), Meterpreter (75), and Web Shell (118). Each system call is represented by an integer (Zhang et al, 2020).

ADFA-WD

ADFA-WD is high-quality collection of DLL access requests and system calls for a variety of hacking attacks. The dataset was gathered on a Windows XP SP2 host. The default firewall was enabled, and Norton AV 2013 was installed to detect only sophisticated attacks and ignore low-level attacks. The operating system environment allowed for sharing and the configuration of network printers. It was running applications like webserver, database server, FTP server, streaming media server, PDF reader, and so on. A total of 12 known vulnerabilities for installed applications were exploited using the Metasploit framework and other custom approaches. ADFA-WD is composed of 355 normal training traces, 1827 normal validation traces, and 5542 attack traces (Borisniya & Patel, 2015).

AWID

The Aegean Wi-Fi Intrusion Detection (AWID) data set is a publicly available labeled data set that was created in 2016 and is based on actual traces of normal and intrusion activity of an IEEE 802.11 Wi-Fi network (Natkaniec & Bednarz, 2023). Character data and imbalance between attack and normal data characterize AWID, which may influence IDS evaluation (Chen et al, 2021). The dataset contains 155 distinct features and 14 simulated existing attacks (Sudaroli Vijayakumar & Ganapathy, 2018).

Table 1 – AWID attack classes
Таблица 1 – Классы кислотны атак
Табела 1 – AWID класе напада

Attack class		**Description**
Flooding	Deauthentication	Sending a large number of deauthentication management frames with specific destination MAC address. Results in the connection loss of a client with MAC address or disconnection of all clients that receive the frame.
	Disassociation	Similar to a deauthentication flood, uses disassociation management frames.
	Block Acknowledge	The attacker sends a fake ADDBA message on behalf of a real client with high sequence numbers, causing the AP to not accept frames.
	Authentication request	Involves sending a large number of authentication request frames; AP overloads can cause it to shut down and drop the wireless network.
	Fake Power Saving	Takes advantage of the Power Saving mechanism by sending a null frame on behalf of the victim with the power saving bit set to 1.
	Clear-to-Send	Relies on the Request-to-Send/Clear-to-Send mechanism; causes STA to wait for a transmission that never occurs;
	Request-to-Send	Similar to a CTS flood, involves sending a large number of RTS frames, which prevents other clients from accessing the medium.
	Beacon	Involves sending multiple beacon frames with different SSIDs; causes confusion for end users attempting to connect to the correct network.
	Probe Request	Drain resources from the AP; sends large number of probe request frames.
	Probe Response	Involves flooding a victim with a large number of probe response frames.
Impersonation	Honeypot	Wireless network created by an attacker designed to attract unsuspecting victims.
	Evil Twin	Wireless network created by an attacker that is an exact replica of an existing network used by the victim.
	Caffe Latte	Attacking wireless networks where direct access to the access point is not necessary.
	Hirte	Extension of the Caffe Latte attack in which ARP packets are fragmented to collect more IVs from the connected device; easier to crack WEP key.
Injection	ARP Injection	Injecting a fake ARP Request into the wireless network
	Fragmentation	The attacker first performs a fake authentication with the Access Point and then receives at least one frame. Attacker can guess the first 8 bytes of the keystream. Then constructs a frame with a known payload, breaks it into fragments.
	Chop-Chop	Dropping the last byte of the encrypted frame and then guessing a valid Integrity Check Value (ICV).

IEEE 802.11i, also known as WPA2, was an improvement to the original IEEE 802.11 standard that aimed to improve protocol security. It significantly augments and expands the well-known AWID2 corpus by capturing and analyzing traces of wide range of IEEE 802.1X Wi-Fi network attacks. AWID3 is expected to be a great improvement in the design and evaluation of IDSs. Attacks from the wireless MAC layer to higher ones that are common to IEEE 802.3 networks.

CAIDA

Center of Applied Internet Data Analysis (2002-2016) created the CAIDA data set and made it widely available to the research community who provide data or network access. CAIDA contains three datasets: CAIDA OC48 (contains various types of data observed on an OC48 link in San Hose), CAIDA DDoS (contains one hour of DDoS attack traffic divided into 5-minute pcap files), and CAIDA Internet Traces 2016 (CAIDA Equinix-Chicago High-speed Internet backbone passive traffic traces) are three datasets contained in CAIDA. CAIDA datasets collect a variety of data types in geographically and topologically diverse regions. Because of numerous flows, these benchmarking are ineffective (Proebstel, 2008).

CAIDA OC48

The CAIDA OC48 Peering Point Traces Dataset (2002-2003) contains anonymized passive traffic traces collected from large ISP’s west coast OC48 peering point from 2002 to 2003. The payload is removed and IP addresses anonymized using CryptoPAn prefix-preserving anonymization tool with the same key for all traces in this dataset. The CAIDA OC48 data is useful for research on the internet traffic characteristics such as application breakdown, security events, topological distribution, and flow volume and duration. These traces can be read by any program that supports the pcap (tcpdump) format (CAIDA, 2020a).

CAIDA DDoS

This dataset contains the traffic traces of a flooding DDoS attack over a one-hour period. The attack’s goal was to consume the computing resource of the targeted server. IP addresses have been pseudonymized, and their payloads and non-attack traffic have been removed from the dataset for security reasons, limiting its usability. This dataset found its application in detecting low rate stealthy as well as high-rate flooding DDoS attacks (Behal & Kumar, 2016). This type of DoS attack attempts to prevent access to the targeted server by consuming computing resources on the server and by consuming all computing resources on the server as well as all network bandwidth connecting the server to the Internet. The one-hour trace is divided into 5-minute pcap files. Only attack traffic to the victim and responses to the attack from the victim are included in the traces. Traces in this dataset are anonymized using CryptoPAn prefix-preserving anonymization using a single key. The payload has been removed from all packets (CAIDA, 2020b).

CAIDA Internet Traces

The CAIDA Internet Traces dataset contains three subsets:

- 2008-2014: contains anonymized passive traffic traces from CAIDA's equinix-chicago and equinix-sanjose high-speed Internet backbone connections.
    - the first available traffic trace is an hourly traffic trace collected during the DITL 2008 measurement event;
    - contains anonymized packet headers in pcap format for a single direction of the bidirectional OC129 link at the equinix-chicago monitors;
    - a one-hour recording resulted in 83GB compressed pcap files;
    - a monthly one-hour trace is collected;
    - traffic traces are anonymized using CryptoPan prefix-preserving anonymization;
    - during recording, packets are truncated to a specified length (64-96 B) to avoid excessive packet loss due to disk I/O overload.
    - payload is removed from all packets; only header information at the transport layer is retained;
    - the Endace network cards used to record these traces provide timestamps with nanosecond precision;
- 2015-2016: contains anonymized passive traffic traces from CAIDA's equinix-chicago monitors on high-speed Internet backbone links.
- 2018-2019: contains anonymized passive traffic traces from CAIDA's equinix-nyc monitor.

Starting with the 2010 traces, the original nanosecond timestamps are provided as separate ascii files alongside the pcap files. The traces can be read with any software that can read pcap (tcpdump) files (CAIDA, 2019).

CIC-IDS-2017

In 2018, the Canadian Institute for Cybersecurity (CIC) created the CIC-IDS-2017 dataset. The dataset consists of ~2.8 million benign and malicious records with 77 features and ~128 thousands current common attack covering 11 criteria (see Table 2) with 14 types of attacks (Sharafaldin et al, 2018). For this dataset, the authors examined the abstract behavior of 25 users based on HTTP, HTTPS, FTP, SSH, and email protocols. The attacks implemented include brute-force FTP, brute-force SSH, DoS, Heartbleed, web attack, Infiltration, Botnet and DDoS (UNB University of New Brunswick: Canadian Institute for Cybersecurity, 2018).

Table 2 – CIC-IDS-2017 criteria and description
Таблица 2 – CIC-IDS-2017 критерии и описание
Табела 2 – CIC-IDS-2017 критеријуми и опис

No	Criteria	Description
1	Complete Network configuration	A complete network topology includes Modem, Firewall, Switches, Routers, and presence of a variety of operating systems such as Windows, Ubuntu and Mac OS X.
2	Complete Traffic	By having a user profiling agent and 12 different machines in Victim-Network and real attacks from the Attack-Network.
3	Labelled Dataset	Section 4 and Table 2 show the benign and attack labels for each day. Also, the details of the attack timing will be published on the dataset document.
4	Complete Interaction	As Figure 1 shows, we covered both within and between internal LAN by having two different networks and Internet communication as well.
5	Complete Capture	Because we used the mirror port, such as tapping system, all traffics have been captured and recorded on the storage server.
6	Available Protocols	Provided the presence of all common available protocols, such as HTTP, HTTPS, FTP, SSH and email protocols.
7	Attack Diversity	Included the most common attacks based on the 2016 McAfee report, such as Web based, Brute force, DoS, DDoS, Infiltration, Heart-bleed, Bot and Scan covered in this dataset.
8	Heterogeneity	Captured the network traffic from the main Switch and memory dump and system calls from all victim machines, during the attacks execution.
9	Feature Set	Extracted more than 80 network flow features from the generated network traffic using CICFlowMeter and delivered the network flow dataset as a CSV file. See our PCAP analyzer and CSV generator.
10	MetaData	Completely explained the dataset which includes the time, attacks, flows and labels in the published paper.
11	Day, Date, Description, Size	Days of normal network activity and attacks.

CSE-CIC-2018

A joint project between the Communication Security Establishment (CSE) and the CIC produced the CSE-CIC-2018 dataset, which included detailed descriptions of intrusions along with abstract distribution models for applications, protocols, or lower-level network entities.

The final data set included seven different attack scenarios: Brute-Force, Hearth Bleed, Botnet, DoS, DDoS, Web Attacks and Infiltration (see Table 3). The attack infrastructure consists of 50 machines and the victim organization consists of 5 departments and includes 420 machines and 30 servers (Kali Linux).

Table 3 – CSE-CIC-2018 attacks and tools
Таблица 3 – CSE-CIC-2018 атаки и инструменты
Табела 3 – CSE-CIC-2018 напади и алати

Attack	Tools	Victim
Bruteforce (1 day)	FTP – Patator; SSH – Patator	Ubuntu 16.4 (Web Server)
DoS (1 day)	Hulk, GoldenEye, Slowloris, Slowhttptest	Ubuntu 16.4 (Apache)
DoS (1 day)	Heartleech	Ubuntu 12.04 (Open SSL)
Web (2 days)	Damn Vulnerable Web App (DVWA); In-house selenium framework (XSS, Brute-force);	Ubuntu 16.4 (Web Server)
Infiltration (2 days)	First level: Dropbox download in a windows machine; Second Level: Nmap and portscan;	Windows Vista & Macintosh
Botnet (1 day)	Ares: remote shell, file upload/download, capturing screenshots and key logging	Windows Vista, 7, 8.1, 10 (32-bit) and 10 (64-bit)
DDoS & PortScan (2 days)	Low Orbit Ion Canon for UDP, TCP, HTTP requests	

The dataset includes the captured network traffic and the system logs of each machine, as well as 80 features extracted from the captured traffic using CICFlowMeter-V3. CICFlowMeter is a network traffic flow generator that produces bidirectional flows (Biflow), where the first packet determines the forward (source to destination) and reverse (destination to source) directions, hence the 83 statistical features such as Duration, Number of packets, Number of bytes, Length of packets, etc.

The application output is in the CSV file format with six columns labeled for each flow, namely FlowID, SourceIP, DestinationIP, SourcePort, DestinationPort, and Protocol with more than 80 network traffic features. Normally, TCP flows are terminated when the connection is broken (by the FIN packet), while UDP flows are terminated by a flow timeout. The flow timeout value can be set arbitrarily according to the particular scheme, e.g. 600s for TCP and UDP. A list of extracted features can be found at (UNB University of New Brunswick: Canadian Institute for Cybersecurity, 2017).

The dataset shows class imbalance as about 17% of the instances contain abnormal traffic. The data set is not used as a treasure trove for signature-based IDS, but to promote anomaly-based intrusion detection (Levy & Khoshgoftaar, 2020).

1998/1999 DARPA intrusion detection evaluation dataset

The DARPA dataset was produced by the Lincoln Laboratory of the Massachusetts Institute of Technology (MIT) in 1998 and 1999. The dataset consists of two parts: online and offline. All network traffic including the total payload of each packet, was recorded in tcp dump format and made available for analysis. In these evaluations, the data was in the form of sniffed network traffic, Solaris BSM audit data, Windows NT audit data (1999 DARPA), and file system snapshots, and an attempt was made to identify intruders that had penetrated a test network during the data collection period. The IDSs are tested in an offline evaluation using network traffic and audit logs collected from a simulated network (Lippmann et al, 2000). The test network consisted of a mixture of real and simulated machines; background traffic was artificially generated by the real and simulated machines while attacks were carried out against the real computers.

The DARPA dataset is used to measure the detection rate and false alarm rate for network traffic consisting of four types of attacks: Denial of Service (DoS), probing (Probe/Scan attacks), and two types of privilege escalation attacks – User to Root (U2R) and Remote to Local (R2L). The 1998 DARPA Intrusion Detection Evaluation Dataset (1998 DARPA) contains 41 features and a class. In total, there are 409021 records with classes labeled as either normal or one of the 22 attack types. However, only 409020 records can be used, primarily because of errors in the records within the dataset (see Table 4) (Khor et al, 2009).

Table 4 – 1998 DARPA record types
Таблица 4 – 1998 DARPA виды записи
Табела 4 – 1998 DARPA класе записа

Record type	Number of records
Normal	97277
Denial of Service	391458
Probe	4107
Remote to Local	1126
User to Root	52

The DARPA 1999 consists of weeks 1-3 of training data and weeks 4-5 of testing data. Weeks one and three contain normal traffic and week two contains labeled attacks (Thomas et al, 2008). The descriptions of the attacks are listed in Table 5.

Table 5 – DARPA attack classes and descriptions
Таблица 5 – DARPA классы атак и описание
Табела 5 – DARPA класе напада и опис

Attack class	Attack type	Description
Probe	ipsweep, lsdomain, mscan	Scans a computer network or a DNS server to find valid IP addresses
	portsweep, mscan	Scans a computer network or a DNS server to find active ports
	queso, mscan	Scans a computer network or a DNS server to find hostoperating system types
	satan	Scans a computer network or a DNS server to find known vulnerabilities
DoS (Designed to disrupt a host or network service)	selfping	Solaris operating system crash
	tcpreset	Active termination of all TCP connections to a specific host
	arppoison	Corruption of ARP cache entries for a victim not in others' caches
	crashiis	Crashes the Microsoft Windows NT web server
	Dosnuke	Crashes Windows NT
R2L (Attacker who does not have an account on a victim machine)	guest, dict	Gains local access to the machine
	ppmacro	Exfiltrates files from the machine
	framespoof	Modifies data in transit to the machine
	ppmacro	NT power point macro attack
	framespoof	Man-in-middle web browser attack
	netbus	NT trojan-installed remote administration tool
	sshtrojan	Linux trojan SSH server
	ncftp	Linux FTP file access-utility with a bug that allows remote commands to run on a local machine
U2R (Local user on a machine is able to obtain privileges	ntfsdos, sqlattack	Secret attacks, where a user who is allowed to access the special files exfiltrates them

The DARPA 1999 test data consisted of 190 instances of Probe (37), DoS (63), R2L (53) and U2R (37) attacks. The following types of attacks were added to the training set (see Table 6).

Table 6 – DARPA attack types
Таблица 6 – DARPA классы атак
Табела 6 – DARPA класе напада

Attack class	Attack type
DoS	appache2, back, land, mailbomb, neptune, pod, processtable, teardrop, smurf, syslogid, udpstorm, warexzlient
Probe	ntinfoscan, iligal-sniffer
R2L	ftpwrite, httptunnel, imap, named, netcat, phf, sendmail, snmpget, xlock, xsnoop
U2R	casesen, eject, fdlformat, flbconfig, loadmodule, nukepw, perl, ppmacro, ps, secret, srchole, xterm, yaga

The main criticisms of the DARPA data relate to: (1) the software used to generate traffic on the testbed, which is not publicly available, (2) the evaluation criteria do not take into account the system resources used, (3) the ease of use, (4) the type of system it is on, (5) the procedures used in building the dataset and performing the evaluation, (6) the background data does not include background noise such as packet storms, (7) strange packets, (8) anomalous Internet traffic that is not caused by malicious behavior, etc.

ISCX 2012

The ISCX 2012 dataset has two profiles. Alpha performs various multi-stage attacks, and Betha generates actual network traffic with background noise. The dataset contains network traffic for HTTP, SMP, SSH, IMAP, POP3, and FTP protocols but no HTTPS traces. The distribution of simulated attacks is not based on real world statistics (Sharafaldin et al, 2018).

The dataset shows realistic network behavior and includes various intrusion scenarios. It is shared as a complete network capture with all internal traces to evaluate the payloads for a deep data packet analysis. In addition, the dataset includes seven days of both normal and malicious network traffic activity.

The dataset was created using profiles that contain abstract representations of network traffic actions and behaviors. ISCX-IDS 2012 contains two different profiles to create network traffic behaviors and scenarios (Table 7) (Khan et al, 2019).

Table 7 – ICSX 2012 criteria
Таблица 7 – ICSX 2012 критерии
Табела 7 – ICSX 2012 критеријуми

No	Criteria	Description
1	Realistic network traffic	Ideally, a dataset should not exhibit any unintended properties, both network and traffic wise. This is to provide a clearer picture of the real effects of attacks over the network and the corresponding responses of workstations. For this reason, it is necessary for the traffic to look and behave as realistically as possible. This includes both normal and anomalous traffic. Any artificial post-capture trace insertion will negatively affect the raw data and introduce possible inconsistencies in the final dataset. Consequently, all such adjustments are highly discouraged.
2	Labelled Dataset	A labeled dataset is of immense importance in the evaluation of various detection mechanisms. Hence, creating a dataset in a controlled and deterministic environment allows for the distinction of anomalous activity from normal traffic; therefore, eliminating the impractical process of manual labeling.
3	Total interaction capture:	The amount of information available to detection mechanisms are of vital importance as this provides the means to detect anomalous behaviour. In other words, this information is essential for post-evaluation and the correct interpretation of the results. Thus, it is deemed a major requirement for a dataset to include all network interactions, either within or between internal LANs.
4	Complete Capture	Privacy concerns related to sharing real network traces have been one of the major obstacles for network security researchers as data providers are often reluctant to share such information. Consequently, most such traces are either used internally, which limits other researchers from accurately evaluating and comparing their systems, or are heavily anonymized with the payload entirely removed resulting in decreased utility to researchers. In this work, the foremost objective is to generate network traces in a controlled testbed environment, thus completely removing the need for any sanitization and thereby preserving the naturalness of the resulting dataset.
5	Diverse intrusion scenarios	Attacks have increased in frequency, size, variety, and complexity in recent years. The scope of threats has also changed into more complex schemes, including service and application-targeted attacks. Such attacks can cause far more serious disruptions than traditional brute force attempts and also require a more in-depth insight into IP services and applications for their detection. Through executing attack scenarios and applying abnormal behaviour, the aim of this objective is to perform a diverse set of multistage attacks; each carefully crafted and aimed towards recent trends in security threats. This objective often labels many of the available datasets as ineffective and unfit for evaluating research results.
6	Day, Date, Description, Size	7 days of normal network activity and attacks

The ISCX IDS 2012 dataset is publicly available for researchers at (UNB University of New Brunswick: Canadian Institute for Cybersecurity, 2012).

KDD Cup '99

The KDD Cup ‘99 dataset is a collection of data transfer from a virtual environment and used for 5th Knowledge Discovery and Data Mining Tools competition. It is a subset of the 1998 DARPA dataset collected by simulating network traffic in a medium sized U.S. Air Force LAN (TCP dump data) over a nine-week period.

The dataset was collected and distributed at the Massachusetts Institute of Technology (MIT) Lincoln Laboratory. The KDD Cup ’99 consists of the full KDD Cup ’99 dataset, which includes simulation of normal connections and four attack classes (Probe, DoS, R2L, U2R), a 10% KDD dataset for training the classifiers, and a KDD test dataset intendend for testing (Gifty Jeya et al, 2012, pp.28-32).

The structure of the full dataset is given in Table 8.

Table 8 – KDD Cup ’99 file content
Таблица 8 – KDD Cup ’99 содержание файлов
Табела 8 – KDD Cup ‘99 садржај фајлова

File	File content
kddcup.names	List of features
kdd.data.gz	Full data set (uncompressed)
kdd.cup.data_10_percent.gz	10% subset (compressed)
kddcup.newtestdata_10_percent_unlabeled.gz	1.4M, 45M uncompressed
kddcup.testdata.unlabeled.gz	11.2M, 430M uncompressed
kddcup.testdata. unlabeled_10_percent.gz	1.4M, 45M uncompressed
corrected.gz	Test data with corrected labels
training_attack_types	List of attack types
typo-correction.txt	Short description of corrections to the data set

The full KDD Cup ‘99 dataset contains 4,898,431 single connection records, each of which consists of 41 features labeled as normal or attacks (Tavallaee et al, 2009).

The number of instances is given in Table 9. The attack classes are described in Table 10.

Table 9 – KDD Cup '99 instance number
Таблица 9 – Номер случая в базе KDD Cup '99
Табела 9 – Број инстанци у KDD Cup '99 бази

Attack class	**Training set**	**10% Training set**	**Test set**
Normal	492,708	97,278	60,593
Probe	41,102	4,107	4,166
DoS	3,883,370	391,458	229,853
R2L	1,126	1,126	16,347
U2R	52	52	70

Table 10 – KDD Cup '99 attack classes
Таблица 10 – Классы атак в базе KDD Cup '99
Табела 10 – Класе напада у бази KDD Cup '99

Attack class	**Attack type**
Probe	ipsweep, nmap, portsweep, satan
DoS	back, land, neptune, pod, smurf, teardrop
R2L	fpwrite, spy, phf, guesspasswd, imap, warezclient, warezmaster, multihop
U2R	rootkit, perl, loadmodule, bufferoverflow

The features describing the connections can be classified into four categories:

- *Basic features* – determined from the packet header without examining the contents of the packet.

- *Content features* – determined by analyzing the content of the TCP packet (number of unsuccessful attempts to login to the system).

- *Time features* – determine the duration of the connection from a source IP address to a destination IP address. The connection is a sequence of data packets that begin and end at predefined times.

- *Traffic features* – are based on a window that has an interval of a certain number of connections (suitable for describing attacks that last longer than the interval of the specific time features).

The features are listed in Table 11.

Table 11 – KDD Cup '99 features
Таблица 11 – Атрибуты базы KDD Cup '99
Табела 11 – Атрибути у бази KDD Cup '99

No	Feature	Description
1.	duration	length of connection
2.	protocol type	type of protocol (TCP, UDP...)
3.	service	destination service (ftp, telnet...)
4.	flag	status of connection
5.	source bytes	No. of B from source to destination
6.	destination bytes	No. of B from destination to source
7.	land	If the source=destination address are the same land=1/if not, 0
8.	wrong fragments	No. of wrong fragments
9.	urgent	No. of *urgent* packets
10.	hot	No. of *hot* indicators
11.	failed logins	No. of unsuccessful attempts at login
12.	logged in	If logged in=1/if login failed 0
13.	# compromised	No. of *compromised* states
14.	root shell	If a command interpreter with a root account is running root shell=1/if not, then 0
15.	su attempted	If *su* command is attempted=1, otherwise=0
16.	# root	No. of *root* accesses
17.	# file creations	No. of operations that create new files
18.	# shells	No. of active command interpreters
19.	# access files	No. of file creation operations
20.	# outbound cmds	No. of outbound commands in an ftp session
21.	is host login	is host login=1 if the login is on the *host login* list/if not 0
22.	is guest login	If a guest is logged into the system = 1 otherwise 0
23.	count	No. of connections to the same host as the current connection at a given interval
24.	srv count	No. of connections to the same service as the current connection at a given interval
25.	serror rate	% of connections with SYN errors
26.	srv error rate	% of connections with SYN errors
27.	rerror rate	% of connections with REJ errors
28.	srv rerror rate	% of connections with REJ errors
29.	same srv rate	% of connections to the same service
30.	diff srv rate	% of connections to different services
31.	srv diff host rate	% of connections to different hosts
32.	dst host count	No. of connections to the same destination
33.	dst host srv count	No. of connections to the same destination that use the same service
34.	dst host same src rate	% of connections to the same destination that use the same service
35.	dst host srv rate	% of connections to different hosts on the same system
36.	dst host same srv port rate	% of connections to a system with the same source port
37.	dst host srv diff host rate	% of connections to the same service coming from different hosts
38.	dst host serror rate	% of connections to a host with an S0 error
39.	dst host srv serror rate	% of connections to a host and specified service with an S0 error
40.	dst host serror rate	% of connections to a host with an RST error
41.	dst host srv serror rate	% of connections to a host and specified service with an RST error

The main criticism is that the KDD Cup ’99 dataset is not an authentic simulation of real network traffic. Other problems include complexity of the training and test sets, the impact of duplicates to machine learning algorithms, the number of attack instances of attack is too high relative to the number of instances of normal traffic, the relationship between each attack category is not realistic, the instances of individual attacks are similar to the instances of normal traffic for the R2L attack types, etc.

Kyoto 2006+

The Kyoto 2006+ dataset contains records of real network traffic data collected from November 2006 to December 2015 on five different computer networks inside and outside Kyoto University (Protić, 2018, pp.587-589). The first part of the dataset contains records collected from ~350 honeypots, including two darknet sensors with ~300 unused IP addresses and other IDSs (Song et al, 2011; Singh et al, 2015; Najafabadi et al, 2016). To generate traffic, the authors developed a server that was deployed on the same network as the honeypots. The first part of the Kyoto 2006+ dataset recorded from 2006 to 2009 consists of 24 features containing ~90 million instances. Fourteen statistical features were derived from the KDD-Cup ’99 dataset (KDD, 1999; Ashok Kumar & Venugopalan, 2018). The authors also added 10 additional features that were used exclusively to detect anomalies. In the observation period, more than 50 million sessions with normal traffic, 43 million sessions with known attacks and 425,000 sessions with unknown attacks were recorded. As a part of the Kyoto 2006+ dataset, a total of 20GB of data was collected from November 2009 to December 2015 (Park et al, 2018). The IDS Bro was used to convert packet-based traffic into a session format. (Demertzis, 2018; McCarthy, 2014). IDS Bro is a behavioral and signature-based analysis framework that provides detailed information about the hypertext transfer protocol (HTTP), the domain name system (DNS), the secure shell (SSH) communication protocol, and irregular network behavior (Song et al, 2011). It is suitable for high-performance network monitoring, protocol analysis, and real-time application layer status reporting. The Bro event engine is responsible for receiving and converting the internet protocol (IP) packets into events that are passed to the policy script interpreter that generates the output. DoS, exploits, malware, port scans and shell code attacks were recorded with no additional information about a specific attack. The Kyoto 2006+ dataset does not provide detailed information on attacks parameters.

Instead, the feature Label determines whether the session is normal or not (Ting, 2011). Table 12 presents the Kyoto 2006+ dataset.

Table 12 – Kyoto 2006+ dataset
Таблица 12 – Kyoto 2006+ база данных
Табела 12 – Kyoto 2006+ база података

No	Feature	Description
1	Duration – basic	Length of the connection (in seconds)
2	Service – basic	Connection's server type (dns, ssh, other)
3	Source bytes – basic	No of data bytes sent by the source IP address
4	Destination bytes – basic	No of data bytes sent by the destination IP address
5	Count	No of connections whose source IP address and destination IP address are the same to those of the current connection in the past two seconds
6	Same_srv_rate	% of connections to the same service in the Count feature
7	Serror_rate	% of connections that have 'SYN' errors in the Count feature
8	Srv_serror_rate	% of connections that have 'SYN' errors in Srv_count
9	Dst_host_count	No of connections whose source IP address is also the same to that of the current connection
10	Dst_host_srv_count	No of connections whose service type is also the same to that of the current connection
11	Dst_host_same_src_port_rate	% of connections whose source port is the same to that of the current connection in the Dst_host_count feature
12	Dst_host_serror_rate	% of connections that have 'SYN' errors in the Dst_host_count feature
13	Dst_host_srv_serror_rate	% of connections that have 'SYN' errors in the Dst_host_srv_count feature
14	Flag	The state of the connection at the time of connection was written (tcp, udp)
15	IDS_detection	Reflects if IDS triggered an alert for the connection
16	Malware_detection	Indicates if malware was observed at the connection
17	Ashula_detection.	Means if shellcodes and exploit codes were used in the connection
18	Label	Indicates whether the session was attack or not
19	Source_IP_Address	Source IP address used in the session
20	Source_Port_Number	Indicates the source port number used in the session
21	Destination_IP_Address	It was also sanitized
22	Destination_Port_Number	Indicates the destination port number used in the session
23	Start_Time	Indicates when the session was started
24	Duration	Indicates how long the session was being established

NSL-KDD

The NSL-KDD dataset is created from the KDD Cup '99 dataset. It corrects flaws in the KDD Cup '99 dataset caused by redundant records in the training set and duplicate records in the test set. Furthermore, the number of records in both the training set and the test sets is appropriate (Protic, 2018, pp.587-589). The training set contains 21 different attack types, while the test set contain 37 different attack types. The known attacks are those presented in the training set, while the additional 16 attacks are only available in the test set (see Table 13) (Nkiama et al, 2016). Normal traffic in the training set contains 67,343 instances, while normal traffic in the test set contains 9,711 instances.

Table 13 – Kyoto 2006+ attack classes
Таблица 13 – Классф атак в базе Kyoto 2006+
Табела 13 – Класе напада у бази Kyoto 2006+

Attack class	**Attack type – Training set**	**Attack type – Test set**
Probe	ipsweep, nmap, portsweep, satan	ipsweep, nmap, portsweep, satan
DoS	back, land, neptune, pod, smurf, teardrop	apache2, back, land, mailbomb, neptune, pod, processtable, smurf, teardrop, udpstorm, worm
R2L	guess_passwd, ftp_write, imap, multihop, phf, spy, warezmaster, warezclient	guess_passwd, ftp_write, imap, httptunnel, phf, multihop, named, snmpguess, snmpgetattack, sendmail, warezmaster, xlock, xsnoop
U2R	buffer_overflow, loadmodule, perl, rootkit	buffer_overflow, loadmodule, rootkit, perl, ps, sqlattack, xterm

UNSW-NB-15

In 2015, Moustafa et al introduced a hybrid academic intrusion detection UNSW-NB-15 dataset derived from real normal data, and the synthesized contemporary attack activities of network traffic. The dataset consists of raw network packets containing nine different attacks (Moustafa & Slay, 2015). Raw network packets from the UNSW-NB 15 dataset are generated by the IXIA PerfectStorm tool at the Cyber Range Lab of UNSW Canberra. The tcpdump tool was used to capture 100 GB of raw traffic (Pcap files).

This dataset contains nine types of attacks, namely Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms. The Argus and Bro-IDS tools are used and twelve algorithms are developed to generate a total of 49 features with the specified class label (UNSV Sydney, 2021). The total number of records is two million and 540,044 stored in the four CSV files:

- The ground truth table and the list of event files.
- One partition from this dataset was configured as a training and test set.
- The number of records in the training set is 175,341 records and the test set consists of 82,332 records of different attack and normal types.

This dataset is a collection of network packets exchanged between hosts (see Table 14) (Ahmad et al, 2022).

Table 14 – UNSW-NB-15 data type
Таблица 14 – Виды данных в базе UNSW-NB-15
Табела 14 – Врсте података у бази UNSW-NB-15

No	Data type	Description
1	Normal	Natural transaction data
2	Analysis	An attack targets web applications through emails, ports, or web scripts
3	Backdoor	Using backdoor to secure remote access
4	DoS	Attacks computer memory
5	Exploits	An instruction that takes advantage of bugs/errors caused by unintentional behavior on the network
6	Fuzzers	An attack to crash the system by inputting a lot of random data
7	Generic	A technique to clash the block-cipher configuration by using hash functions
8	Reconnaissance	A probe to evade network security controls by collecting relevant information
9	Shellcode	Code is used to exploit software vulnerabilities
10	Worms	A set of virus codes can be added to a computer system or other programs

Conclusion

The researchers worldwide investigate various cybersecurity issues, such as malicious attacks on computer networks. The main challenges in evaluating intrusion detection and intrusion prevention are the massive amounts of data in well-known and publicly available datasets. The majority of the datasets presented in this paper are simulations of real network traffic. Several are hybrid, and one is based on real network traffic. The size, number of features, purpose and type of attacks of each dataset vary.

We presented datasets primarily used for intrusion detection, namely ADFA-LF, ADFA-WD, AWID, CAIDA, CIC-IDS-2017, CSE-CIC-2018, DARPA 98, SCX 2012, KDD Cup ’99, Kyoto 2006+, NSL-KDD and UNSW-NB15. The main characteristics and the comparative analysis are provided. The authors’ main goal is to assist researchers in selecting datasets that best meet their needs.

References

Ahmad, I., Haq, Q.E.U., Imran, M., Alassafi, M.O. & AlGhamdi, R.A. 2022. An efficient network intrusion detection and classification system. *Mathematics*, 10(3), art.number:530. Available at: https://doi.org/10.3390/math10030530.

Ashok Kumar, D. & Venugopalan, S.R. 2018. A Novel algorithm for Network Anomaly Detection using Adaptive Machine Learning. In: Saeed, K., Chaki, N., Pati, B., Bakshi, S. & Mohapatra, D. (Eds.) *Progress in Advanced Computing and Intelligent Engineering. Advances in Intelligent Systems and Computing*, 564. Singapore: Springer. Available at: https://doi.org/10.1007/978-981-10-6875-1_7.

Behal, S. & Kumar, K. 2016. Trends in validation of DDoS research. *International Conference on Computational Modeling and Security. Procedia Computer Science*, 85, pp.7-15. Available at: https://doi.org/10.1016/j.procs.2016.05.170.

Bohara, B., Bhuyan, J., Wu, F. & Ding, J. 2020. A Survey on the Use of Data Clustering for Intrusion Detection System in Cybersecurity. *International Journal of Network Security & Its Applications (IJNSA)*, 12(1), pp.1-18. Available at: https://doi.org/10.5121/ijnsa.2020.12101.

Borisniya, B. & Patel, D.R. 2015. Evaluation of Modified Vec tor Space Representation Using ADFA-LD and ADFA-WD Datasets. *Journal of Information Security*, 6(3), 250-264. Available at: https://doi.org/10.4236/jis.2015.63025.

-CAIDA. 2019. The CAIDA Anonymized Internet Traces Dataset (April 2008 - January 2019). *Caida.org,* December 3 [online]. Available at: https://www.caida.org/catalog/datasets/passive_dataset/ [Accessed: 10 June 2023].

-CAIDA. 2020a. The CAIDA “DDoS Attack 2007” Dataset. 2020. *Caida.org,* June 24 [online]. Available at: https://www.caida.org/catalog/datasets/ddos-20070804_dataset/ [Accessed: 10 June 2023].

-CAIDA. 2020b. The CAIDA OC48 Peering Point Traces. 2020. *Caida.org,* June 24 [online] Available at: https://www.caida.org/catalog/datasets/passive_oc48_dataset/ [Accessed: 10 June 2023].

Chen, J., Yang, T., He, B. & He, L. 2021. An analysis and research on wireless network security dataset. In: *2021 International Conference on Big Data Analysis and Computer Science (BDACS),* Kunming, China, pp.80-83, June 25-27. Available at: https://doi.org/10.1109/BDACS53596.2021.00025.

Demertzis, K. 2018. The Bro Intrusion Detection System. *Research Gate*. Available at: https://doi.org/10.13140/RG.2.2.35333.40168.

Ferriyan, A., Thamrin, A.H., Takeda, K. & Murai, J. 2021. Generating Network Intrusion Detection Dataset Based on Real and Encrypted Synthetic Attack Traffic. *Applied Sciences,* 11(17), art.number:7868. Available at: https://doi.org/10.3390/app11177868.

Jie, C., Jiawei, L., Shulin, W. & Sheng, Y. 2018. Feature selection in machine learning: A new perspective. *Neurocomputing*, 300, pp.70-79. Available at: https://doi.org/10.1016/j.neucom.2017.11.077.

Khan, M.A., Karim, Md.R. & Kim, Y. 2019. A Scalable and Hybrid Intrusion Detection System Based on the Convolutional-LSTM Network. *Symmetry*, 11(4), art.number:583. Available at: https://doi.org/10.3390/sym11040583.

Khraisat, A. Gondal, I., Vamplew, P. & Kamruzzaman, J. 2019. Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecurity*, 2(art.number:20). Available at: https://doi.org/10.1186/s42400-019-0038-7.

Khor, K.-C., Ting, C.-Y. & Amnuaisuk, S.-P. 2009. A Feature Selection Approach for Network Intrusion Detection. In: *2009 International Conference on Information Management and Engineering*, Kuala Lumpur, Malaysia, pp.133-137, April 3-5. Available at: https://doi.org/10.1109/ICIME.2009.68.

-KDD. 1999. SIGKDD-KDD Cup: KDD Cup 1999: Computer network intrusion detection. *Kdd.org* [online] Available at: https://kdd.org/kdd-cup/view/kdd-cup-1999 [Accessed: 10 June 2023].

Levy, J.L. & Khoshgoftaar, T.M. 2020. A survey and analysis of intrusion detection models based on CSE-CIC IDS 2018 Big Data. *Journal of Big Data* 7(art.number:104). Available at: https://doi.org/10.1186/s40537-020-00382-x.

Lippmann, R.P., Cunningham, R.K., Fried, D.J., Graf, I., Kendal, K.R., Webster, S.E. & Zissman, M.A. 2000. Results of DARPA 1998 Offline Intrusion Detection Evaluation. In: *Recent Advances in Intrusion Detection, RAID 99 Conference*, West Lafayette, Indiana, USA. September 7-9. [online] Available at: https://archive.ll.mit.edu/ideval/files/RAID_1999a.pdf [Accessed: 10 June 2023].

McCarthy, R. 2014. Network analysis with the Bro Network Security Monitor. *ADMIN Network & Security*, 24 [online] Available at: https://www.admin-magazine.com/Archive/2014/24/Network-analysis-with-the-Bro-Network-Security-Monitor/(language)/eng-US [Accessed: 10 June 2023].

Mighan, S.N. & Kahani, M.A. 2021. A novel scalable intrusion detection system based on deep learning. *International Journal of Information Security*, 20, pp.387-403. Available at: https://doi.org/10.1007/s10207-020-00508-5.

Moustafa, N. & Slay, J. 2015. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In: *2015 Military Communications and Information Systems Conference (MilCIS)*, Canberra, ACT, Australia, pp.1-6, November 10-12. Available at: https://doi.org/10.1109/MilCIS.2015.7348942.

Najafabadi, M.N., Khoshgoftaar, T.M. & Selyia, N. 2016. Evaluating Feature Selection Methods for Network Intrusion Detection with Kyoto Data. *International Journal of Reliability, Quality and Safety Engineering*, 23(1), art.number:1650001. Available at: https://doi.org/10.1142/S0218539316500017.

Natkaniec, M. & Bednarz, M. 2023. Wireless Local Area Networks Threat Detection Using 1D-CNN. *Sensors,* 23(12), art.number:5507. Available at: https://doi.org/10.3390/s23125507.

Nkiama, H., Mohd Said, S.Z. & Saidu, M. 2016. A Subset Feature Elimination Mechanisms for Intrusion Detection System. *International Journal of Advanced Computer Science and Application*, 7(4), pp.148-157. Available at: https://doi.org/10.14569/IJACSA.2016.070419.

Omar, S., Ngadi, A. & Jebur, H.H. 2013. Machine Learning Techniques for Anomaly Detection: An Overview. *International Journal of Computer Applications*, 79(2), pp.33-41 [online] Available at:

https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=0278bbaf1db5df036f02393679d485260b1daeb7 [Accessed: 10 June 2023].

Park, K., Song, Y. & Cheong, Y. 2018. Classification of Attack Types for Intrusion Detection Systems Using a Machine Learning Algorithm. In: *2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService)*, Bamberg, Germany, pp.282-286, March 26-29. Available at: https://doi.org/10.1109/BigDataService.2018.00050.

Proebstel, E.P. 2008. *Characterizing and Improving Distributed Network-based Intrusion Detection Systems (NIDS): Timestamp Synchronization and Sampled Traffic*. Master thesis. Davis: University of California [online]. Available at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=ee123bb36e6d16ac9b70507e7ac614791dd8f759 [Accessed: 10 June 2023].

Protić, D. 2018. Review of KDD Cup '99, NSL-KDD and Kyoto 2006+ datasets. *Vojnotehnički glasnik/Military Technical Courier*, 66(3), pp.580-596. Available at: https://doi.org/10.5937/vojtehg66-16670.

Protić, D. & Stanković, M. 2020. Anomaly-Based Intrusion Detection: Feature Selection and Normalization Influence to the Machine Learning Models Accuracy. *European Journal of Formal Sciences and Engineering*, 3(1), pp.1-9. Available at: https://doi.org/10.26417/ejef.v2i3.p101-106.

Serkani, E., Gharaee, H. & Mohammadzadeh, N. 2019. Anomaly Detection Using SVM as Classifier and Decision Tree for Optimizing Feature Vectors. *The ISC International Journal of Information Security (ISeCure),* 11(2), pp.159-171 [online]. Available at: https://www.isecure-journal.com/article_91592_e825e0139e75d44a6b543ad437c18379.pdf [Accessed: 10 June 2023].

Sharafaldin, I., Lashkari, A.H. & Ghorbani, A.A. 2018. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. In: *Proceedings of the 4th International Conference on Information Systems Security and Privacy ICISSP,* Funchal, Madeira, Portugal, 1, pp.108-116, January 22-24. Available at: https://doi.org/10.5220/0006639801080116.

Singh, R., Kumar, H. & Singla, R.K. 2015. An intrusion detection system using network traffic profiling and online sequential extreme learning machine. *Expert Systems with Applications*, 42(22), pp.8609-8624. Available at: https://doi.org/10.1016/j.eswa.2015.07.015.

Soltani, M., Siavoshani, M.J. & Jahangir, A.H. 2021. A content based deep intrusion detection system. *International Journal of Information Security,* 21, pp.547-562. Available at: https://doi.org/10.1007/s10207-021-00567-2.

Song, J., Takakura, H., Okabe, Y., Eto, M., Inoue, D. & Nakao, K. 2011. Statistical analysis of honeypot data and building of Kyoto 2006+ dataset for NIDS evaluation. In: *EuroSys '11: Sixth EuroSys Conference: BADGERS '11 - Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*, Salzburg, Austria, pp.29-36, April 10-13. Available at: https://doi.org/10.1145/1978672.1978676.

Sudaroli Vijayakumar, D. & Ganapathy, S. 2018. Machine Learning Approach to Combat False Alarms in Wireless Intrusion Detection System. *Computer and Information Science* 11(3), pp.67-81. Available at: https://doi.org/10.5539/cis.v11n3p67.

Tavallaee, M., Bagheri, E., Lu, W. & Ghorbani, A. 2009. A Detailed Analysis of the KDD Cup '99 dataset. In: *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, Ottawa, ON, Canada, pp.1-6, July 8-10. Available at: Available at: https://doi.org/10.1109/CISDA.2009.5356528.

Thakkar, A. & Lohiya, R. 2020. A Review of the Advancement in Intrusion Detection Datasets. *Procedia Computer Science,* 167, pp.636-645. Available at: https://doi.org/10.1016/j.procs.2020.03.330.

Thomas, C., Sharma, V. & Balakrishnan, N. 2008. Usefulness of DARPA dataset for intrusion detection system evaluation. In: *Proceedings: Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security*, 6973, pp.1-8, March 16. Available at: https://doi.org/10.1117/12.777341.

Ting, K.M. 2011. Confusion Matrix. In: Sammut, C. & Webb, G.I. (Eds.) *Encyclopedia of Machine Learning*. Boston, MA: Springer. Available at: https://doi.org/10.1007/978-0-387-30164-8_157.

-UNB University of New Brunswick: Canadian Institute for Cybersecurity. 2018. *CSE-CIC/IDS2018 on AWS* [online] Available at: https://www.unb.ca/cic/datasets/ids-2018.html [Accessed: 10 June 2023].

-UNB University of New Brunswick: Canadian Institute for Cybersecurity. 2017. *Intrusion Detection Evaluation Dataset (CIC-IDS2017)* [online]. Available at: https://www.unb.ca/cic/datasets/ids-2017.html [Accessed: 10 June 2023].

-UNB University of New Brunswick: Canadian Institute for Cybersecurity. 2012. *Intrusion Detection Evaluation Dataset (ISCXIDS2012)* [online] Available at: https://www.unb.ca/cic/datasets/ids.html [Accessed: 10 June 2023].

-UNSV Sydney. 2021. The UNSW-NB15 Dataset. 2021. *UNSV Sydney*, June 02 [online] Available at: https://research.unsw.edu.au/projects/unsw-nb15-dataset [Accessed: 10 June 2023].

Xie, M., Hu, J., Yu, X. & Chang, E. 2014. Evaluating Host-Based Anomaly Detection Systems: Application of the Frequency-Based Algorithms to ADFA-LD. In: Au, M.H., Carminati, B. & Kuo, CC.J. (Eds.) *Network and System Security. NSS 2015. Lecture Notes in Computer Science*, 8792. Cham: Springer. Available at: https://doi.org/10.1007/978-3-319-11698-3_44.

Zhang, S., Xie, X. & Xu, Y. 2020. A Brute-Force Black-Box Method to Attack Machine Learning-Based Systems in Cybersecurity. *IEEE Access*, 8, pp.128250-128263. Available at: https://doi.org/10.1109/ACCESS.2020.3008433.

Угрозы кибербезопасности: Какой набор данных следует использовать для оценки системы обнаружения атак?

Даниела Д. Протич[а], *Миомир* М. Станкович[б]

[а] Вооруженные силы Республики Сербия, Генеральный штаб,
Управление информатики и телекоммуникаций (J-6),
Центр прикладной математики и электроники,
г. Белград, Республика Сербия, **корреспондент**

[б] Математический институт Сербской академии наук и искусств,
г. Белград, Республика Сербия

РУБРИКА ГРНТИ: 20.23.25 Информационные системы с базами знаний
ВИД СТАТЬИ: оригинальная научная статья

Резюме:

Введение/цель: Анализ многомерных наборов данных, используемых для обнаружения вторжений, становится настоящим вызовом для исследователей. В данной статье представлены самые используемые наборы данных. ADFA включает два набора данных, содержащих записи из Linux/Unix. AVID основан на фактических нормальных действиях и следах вторжений в Wi-Fi сеть стандарта IEEE 802.11. CAIDA собирает географические и топологические данные различных регионов. CIC-IDS-2017 основана на протоколах: HTTP, HTTPS, FTP, SSH и электронной почте. CSE-CIC-2018 включает абстрактные модели распространения для приложений, протоколы или сетевые модели нижнего уровня. DARPA содержит данные о сетевом трафике. Набор данных ISCX 2012 содержит различные виды многоэтапных атак и фактический сетевой трафик с фоновым шумом. KDD Cup ‘99 представляет собой смоделиорованную базу данных виртуальной сетевой среды. Kyoto 2006+ содержит записи о реальном сетевом трафике. Он используется исключительно для обнаружения аномалий. NSL-KDD корректирует недостатки в KDD Cup ’99, вызванные избыточными и дублирующимися записями. UNSW-NB-15 создан путем объединения реального и синтезированного трафика, который описывает атаки на сетевой трафик.

Методы: В данном исследовании использованы количественные и качественные методы. Рассматриваются научные референсы и общедоступная информация о вышеперечисленных базах данных.

Результаты: Наборы данных часто моделируются для достижения целей конкретной организации. Количество реальных наборов данных намного меньше количества моделируемых наборов данных. Обнаружение аномалий редко используется в современном мире.

Выводы: Представлены основные характеристики и сравнительный анализ наборов данных с точки зрения даты их создания, размера, количества атрибутов, видов трафика и назначения.

Ключевые слова: ADFA, AWID, CAIDA, CIC-IDS-2017, CSE-CIC-2018, DARPA, ISCX 2012, KDD Cup ’99, Kyoto 2006+, NSL-KDD, UNSW-NB15.

Напади на сајбер безбедност: који скуп података треба користити за евалуацију система за детекцију упада?

Данијела Д. Протић[а], *Миомир* М. Станковић[б]

[а] Војска Србије, Генералштаб, Управа за телекомуникације и информатику (Ј-6), Центар за примењену математику и електронику, Београд, Република Србија, **аутор за преписку**

[б] Математички институт Српске академије наука и уметности, Београд, Република Србија

ОБЛАСТ: рачунарске науке, електроника, телекомуникације
КАТЕГОРИЈА (ТИП) ЧЛАНКА: оригинални научни рад

Сажетак:

Увод: Анализа великих скупова података који се користе за детекцију упада постаје истраживачки изазов. У раду су представљени најчешће коришћени скупови података. ADFA садржи два скупа података са записима из Linux-а и Unix-а. AWID је заснован на реалној нормалној активности и активности упада у IEEE 802.11 Wi-Fi мрежу. CAIDA садржи податке са географских и тополошки различитих региона. CIC-IDS-2017 је базирана на протоколима: HTTP, HTTPS, FTP, SSH и email. CSE-CIC-2018 укључује апстрактне моделе дистрибуције за апликације, протоколе и мрежне ентитете нижег нивоа. DARPA садржи податке о мрежном саобраћају. ISCX 2012 је скуп података различитих вишестепених напада и стварног мрежног саобраћаја са позадинским шумом. KDD Cup ‘99 је симулирана база података виртуалног мрежног окружења. Kyoto 2006+ садржи записе реалног мрежног саобраћајаи користи се искључиво за детекцију аномалија. NSL-KDD коригује проблеме из KDD Cup ’99 изазване редундантним записима и дупликатима. UNSW-NB-15 настаје комбинацијом реалног и синтетизованог саобраћаја који описује активности типа напада на мрежни саобраћај.

Методе: Овај рад користи квалитативну и квантитативну технологију. Разматране су научне референце и јавно доступне информације о датим базама података.

Резултати: Базе података се често симулирају да би били испуњени циљеви које захтева одређена организација. Број реалних база података је веома мали у поређењу са симулираним базама података. Детекција аномалија данас се ретко користи.

Закључак: Приказане су главне карактеристике и компаративна анализа скупова података у погледу датума настанка, величине, броја атрибута, врсте саобраћаја и намене.

Кључне речи: ADFA, AWID, CAIDA, CIC-IDS-2017, CSE-CIC-2018, DARPA, ISCX 2012, KDD Cup ’99, Kyoto 2006+, NSL-KDD, UNSW-NB15.

Paper received on / Дата получения работы / Датум пријема чланка: 13.06.2023.
Manuscript corrections submitted on / Дата получения исправленной версии работы / Датум достављања исправки рукописа: 30.11.2023.
Paper accepted for publishing on / Дата окончательного согласования работы / Датум коначног прихватања чланка за објављивање: 01.12.2023.