A Detailed Analysis of the CICIDS2017 Data Set

Iman Sharafaldin, Arash Habibi Lashkari(✉), and Ali A. Ghorbani

Canadian Institute for Cybersecurity (CIC), University of New Brunswick (UNB), Fredericton, Canada
{Isharafa,A.Habibi.L,Ghorbani}@unb.ca

Abstract. The likelihood of suffering damage from an attack is obvious with the exponential growth in the size of computer networks and the internet. Meanwhile, intrusion detection systems (IDSs) and intrusion prevention systems (IPSs) are one of the most important defensive tools against the ever more sophisticated and ever-growing frequency of network attacks. Anomaly-based research in intrusion detection systems suffers from inaccurate deployment, analysis and evaluation due to the lack of an adequate dataset. A number of datasets such as DARPA98, KDD99, ISC2012, and ADFA13 have been used by the researchers to evaluate the performance of their proposed intrusion detection and intrusion prevention approaches. Based on our study of 16 datasets since 1998, many are out of date and unreliable. There are various shortcomings: lack of traffic diversity and volume, incomplete attack coverage, anonymized packet information and payload which does not reflect the current reality, or they lack some feature set and metadata. This paper focused on CICIDS2017 as the last updated IDS dataset that contains benign and seven common attack network flows, which meets real world criteria and is publicly available. It also evaluates the effectiveness of a set of network traffic features and machine learning algorithms to indicate the best set of features for detecting an attack category. Furthermore, we define the concept of superfeatures which are high quality derived features using a dimension reduction algorithm. We show that the random forest algorithm as one of our best performing algorithm can achieve better results with superfeatures versus top selected features.

Keywords: Intrusion detection · IDS dataset · DoS · Web attack · Infiltration · Brute force · Superfeature

1 Introduction

Intrusion detection plays a vital role in the network defense process by alerting security administrators about malicious behaviors such as intrusions, attacks, and malware. Having an IDS is a mandatory line of defense for protecting critical

The first two authors contributed equally to this work.

P. Mori et al. (Eds.): ICISSP 2018, CCIS 977, pp. 172–188, 2019.
https://doi.org/10.1007/978-3-030-25109-3_9

networks against ever increasing intrusive activities. Research on IDS has flourished However, researchers struggle to find comprehensive and valid datasets to test and evaluate their proposed techniques [14] and [developing—extracting—filtering out—isolating] a suitable dataset is a significant challenge itself [19].

Many datasets cannot be shared due to the privacy issues. Those that do become available are heavily anonymized and do not reflect current trends as evidenced by the lack of traffic variety and attack diversity A perfect dataset is yet to be realized [1,19]. It should also be mentioned that benchmark datasets need to be updated periodically [Nehinbe 2011]. Due to malware evolution and the continuous evolution of attack strategies. Since 1999, Scott *et al.* [26], Heideman and Papadopulus , Ghorbani *et al.* [10], Nehinbe [19], Shiravi *et al.* [1], and Sharfaldin *et al.* [9] have tried to propose an evaluation framework for IDS datasets. According to the latest research and proposed evaluation framework, 11 characteristics are critical for a comprehensive and valid IDS dataset: attack diversity, anonymity, available protocols, complete capture, complete interaction, complete network configuration, complete traffic, feature set, heterogeneity, labelling, and metadata [9].

Our Contributions: We make three contributions in this paper First, we generate a new IDS dataset, named CICIDS2017, that has all 11 characteristics above with updated common attacks such as DoS, DDoS, brute force, XSS, SQL injection, infiltration, port scan and botnet. The dataset is completely labelled and has over 80 network traffic features extracted and calculated for all benign and intrusive flows using CICFlowMeter software which is publicly available at the Canadian Institute for Cyber Security website [12]. Second, the paper analyzes the generated dataset to select the best feature sets to detect different attacks. Finally we execute seven common machine learning algorithms to evaluate our dataset.

The rest of the paper is organized as follows. Section 1.1 surveys 16 datasets generated between 1998 and 2017. Section 2 describes 11 features to look for in datasets. Section 3 describes in more details the characteristics of the new dataset. Section 4 defines and describes superfeatures and an analysis of them.

This paper is an extension of the one published in the ICISSP proceedings [35] with defining the superfeatures as the high quality features that are made by linear or non-linear combination of set of basic features along with analysis and visualization of the generated dataset.

1.1 Available Datasets

In this section, we survey 11 IDS datasets made available since 1998 discussing their shortcomings that point to the need for a new comprehensive and reliable dataset.

DARPA (Lincoln Laboratory 1998–99): This dataset was constructed for network security analysis and exposed the issues associated with the artificial injection attacks and benign traffic. This dataset includes e-mail, browsing, FTP, telnet, IRC, and SNMP activity. It contains attacks such as DoS, guess password, buffer

overflow, remote FTP, Synflood, Nmap, and Rootkit. Its shortcomings include: it does not represent real-world network traffic, it lacks false positives, and lacks actual attack data records. It is thus outdated for evaluating IDSs on modern networks both in terms of attack types and network infrastructure [2,16].

KDD'99 (University of California, Irvine 1998–99): This dataset is an updated version of DARPA98 and was made by processing the tcpdump portion. It contains different attacks such as Neptune-DoS, pod-DoS, Smurf-DoS, and buffer-overflow [6]. The benign and attack traffic are merged together in a simulated environment. It has a large number of redundant records and is studded with data corruptions that lead to skewed testing results [31]. NSL-KDD was created using KDD [31] to address some of KDD's shortcomings [16].

DEFCON (The Shmoo Group, 2000–2002): The DEFCON8 dataset created in 2000 contains port scanning and buffer oveflow attacks, whereas DEFCON10, created in 2002, contains port scan and sweeps, bad packets, administrative privilege, and FTP by telnet protocol attacks. In this dataset, the traffic produced during the capture the flag (CTF) competition is different from real world network traffic since it mainly consists of intrusive traffic as opposed to normal background traffic. This dataset was used to evaluate alert correlation techniques [11,18].

CAIDA (Center of Applied Internet Data Analysis 2002–2016): This organization has three datasets (a) CAIDA OC48includes different types of data observed on an OC48 link in San Jose (b) CAIDA DDOS which includes one-hour DDoS attack traffic split of 5-min pcap files, and (c) CAIDA Internet traces 2016 which is passive traffic traces from CAIDA's Equinix-Chicago monitor on the high-speed internet backbone. Most of CAIDAs datasets are very specific to particular events or attacks and are anonymized with their payload, protocol information, and destination. These are not useful benchmarking datasets due to a number of shortcomings, see [1,5,7,8,22] for details.

LBNL (Lawrence Berkeley National Laboratory and ICSI 2004–2005): The dataset is full header network traffic recorded at a medium-sized site. It does not have payload and suffers from heavy anonymization to remove any information which could identify an individual IP [17].

ISOT (Intrusion Dataset 2008): The dataset contains malicious and non-malicious datasets [36]. The benign part was created by combining (a) a dataset from the traffic lab at Ericsson Research which contain different benign traffic such as web browsing, gaming and torrent traffic; and (b) a dataset from the Lawrence Berkeley National Lab (LBNL) that contains different benign traffic such as traffic for web, email and streaming media applications. The malicious part contains Storm and Waledac botnet traffic. The three datasets were merged using their own method. It contains 23 subnets of normal traffic and one subnet for malicious traffic. Each flow has seven flow-based and four host-based features.

CDX (United States Military Academy 2009): This dataset captures network warfare competitions and can be utilized to generate modern, labelled

datasets. It includes web, email, DNS lookups, and other service traffic. The attackers used attack tools such as Nikto, Nessus, and WebScarab to carry out reconnaissance and attacks automatically. It can be used to test IDS alert rules, but suffers from the lack of traffic diversity and volume [25].

Kyoto (Kyoto University 2009): This dataset was gathered from honeypots, so there is no labelling or anonymization, but it has a limited view of network traffic because only attacks directed at the honeypots can be observed. It has ten extra features, such as IDS_detection, malware_detection, and Ashula_detection, than previous datasets which are useful in NIDS analysis and evaluation. Normal traffic was simulated by only DNS and mail traffic data, which is not reflective of real world normal traffic. So there are no false positives which are important for minimizing the number of alerts [15,23,28].

Twente (University of Twente 2009): This dataset includes three services, OpenSSH, Apache web server and Proftp, using authtident on port 113 and captured data from a honeypot network by Netflow. There is some simultaneous network traffic such as authident, ICMP, and IRC traffic, which are not completely benign or malicious. Moreover, this dataset contains some unknown and uncorrelated alerts traffic. It is labelled and is more realistic, but the lack of volume and diversity of attacks is obvious [29].

UMASS (University of Massachusetts 2011): The dataset includes trace files, which are network packets, and some traces on wireless applications [UMASS 2011] [Nehinbe 2011]. It was generated using a single TCP-based download request attack scenario. The dataset is not useful for testing IDS and IPS techniques due to the lack of variety of traffic and attacks [30].

ISCX2012 (University of New Brunswick 2012): This dataset has two [CICIDS2017 5] profiles, the alpha-profile which carried out various multi-stage attack scenarios, and the beta-profile, which is the benign traffic generator and generates realistic network traffic with background noise. It includes network traffic with HTTP, SMTP, SSH, IMAP, POP3, and FTP protocols with full packet payload. However, since it does not contain any HTTPS traces, and HTTPS represents nearly 70% of todays network traffic, the distribution of the simulated attacks is therefore not realistic. Moreover, the distribution of the simulated attacks is not based on real world statistics [1].

ADFA (University of New South Wales 2013): This dataset includes normal training and validating data and 10 attacks per vector [4]. It contains FTP and SSH password brute force, Java based Meterpreter, add new superuser, Linux Meterpreter payload and C100 Webshell attacks. In addition to the lack of attack diversity and variety of attacks, the behaviors of some attacks in this dataset are not well separated from the normal behavior [32,34].

CTU-13 (CTU University 2013): This dataset was created by CTU University, Czech Republic [37]. The dataset contains botnet and benign traffic and background communication traffic. This dataset uses bidirectional Netflow records. They defined 13 different scenarios and captured specific malware traffic

Table 1. Comparing available IDS datasets based on the dataset evaluation framework [35].

	Network	Traffic	Label.	Interact.	Captu.	Protocols					Attack diversity							Ano.	Heter.	Features	Meta.
						HTTP	HTTPS	SSH	FTP	Email	Browser	Bforce	DoS	Scan	Bdoor	DNS	Other				
DARPA	YES	NO	YES	YES	YES	YES	NO	YES	YES	YES	NO	YES	YES	YES	NO	NO	YES	NO	NO	NO	YES
KDD'99	YES	NO	YES	YES	YES	YES	NO	YES	YES	YES	NO	YES	YES	YES	NO	NO	YES	NO	NO	YES	YES
DEFCON	NO	NO	NO	YES	YES	YES	NO	YES	NO	NO	NO	NO	NO	YES	YES	NO	YES	-	NO	NO	NO
CAIDAs	YES	YES	NO	NO	NO	-	-	-	-	-	NO	NO	YES	YES	NO	YES	YES	YES	NO	NO	YES
LBNL	YES	YES	NO	NO	NO	YES	NO	YES	NO	NO	-	-	-	YES	-	-	-	YES	NO	NO	NO
CDX	NO	NO	NO	YES	YES	YES	NO	YES	YES	YES	NO	NO	YES	YES	NO	YES	-	-	NO	NO	NO
KYOTO	YES	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	NO	NO	YES	YES
TWENTE	YES	YES	YES	YES	YES	YES	NO	YES	YES	NO	NO	YES	NO	YES	NO	NO	YES	-	-	NO	YES
UMASS	YES	NO	YES	NO	YES	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	YES	-	-	NO	NO
ISCX2012	YES	NO	YES	YES	YES	YES	NO	YES	YES	YES	YES	YES	YES	YES	YES	NO	YES	NO	YES	NO	YES
ADFA2013	YES	YES	YES	YES	YES	YES	NO	YES	YES	YES	YES	YES	NO	NO	YES	NO	YES	NO	-	NO	YES
ISOT	YES	YES	YES	YES	YES	YES	NO	NO	NO	YES	NO	NO	NO	NO	NO	NO	YES	YES	NO	NO	YES
SSHCure	YES	YES	YES	YES	YES	NO	NO	YES	NO	NO	NO	YES	NO	NO	NO	NO	NO	NO	NO	NO	YES
CTU-13	YES	YES	YES	YES	YES	YES	NO	NO	NO	NO	YES	NO	YES	YES	YES	NO	YES	NO	NO	YES	YES
UGR'16	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	NO	NO	YES	NO	YES	NO	YES	YES	NO	NO	YES
CICIDS2017	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	NO	YES	YES	YES

for each scenario. They used a Windows XP SP2 virtual machine as a guest and a Linux Debian as their host. All of them were connected to the university network. As for labeling, all traffic was initially labeled as background traffic. Traffic that originated from switches, proxies, and legitimate computers was labeled as benign. All traffic which came from infected machines was labeled as botnet.

SSHCure (University of Twente 2014): This dataset contains SSH attacks on a campus network [38]. SSHCure contains Netflow records that were exported from Cisco 6500 series routers. It has two parts which were collected over a month on the campus of the UT. Each part represents different scenarios. The first part contains SSH traffic targeting honeypots and the second part contains SSH traffic from normal servers. There are 11348 attack records.

UGR'16 (University of Granada 2016) This dataset was created by the University of Granada and is designed for the evaluation of cyclostationarity-based network IDSs [39]. The dataset was captured over four months in a tier-3 ISP. They anonymized Netflow records. The dataset offers little attack variety. Also, they mixed botnet captures in a controlled environment with background traffic that reduces the quality of the dataset.

CICIDS2017 (Canadian Institute for Cybersecurity 2017): The dataset was created by Canadian Institute for Cybersecurity. They proposed a novel systematic approach by defining two types of profiles to create a valid dataset. The dataset contains a variety of up to date multi stages attacks such as Heartbleed and different types of DoS and DDoS attacks. Furthermore, a variety of modern protocols are included. It has 80 features for each Netflow record and is in CSV format, making importing it into machine learning software easy [35].

2 Comparing Current Datasets

Finding a suitable IDS dataset is a significant challenge since many datasets cannot be shared due to privacy issues. Also, most of the available datasets are heavily anonymized and do not reflect the real-world trends. According to our last dataset evaluation framework published in 2016 [9], a dataset should meet 11 criteria.

Complete Network Configuration. Having a complete computer network is the foundation of an online dataset to represent the real world. Several attacks have only revealed themselves in a complete network with numerous PCs, servers, routers, and firewall. So it is necessary to have a realistic configuration in the testbed to capture the real effects of attacks.

Complete Traffic. Traffic is a sequence of packets from a source, which can be a host, router, or switch, to a destination, which may be another host, a multicast group, or a broadcast domain. Based on the traffic generation technique, it is possible to have real, pseudo-realistic, or synthetic traffic in a dataset. The pseudo-realistic has partially the real world traffic, such as having simulated human behavior traffics with real attack scenarios.

Labeled Dataset. While a dataset for evaluating different discovery mechanisms in this domain is important, tagging and labeling data are also important. If there are no correct [accurate — informative] labels, without a doubt, it is not possible to use a dataset and the results of any analysis is invalid and unreliable. For example, in network datasets, after converting pcaps to netflows, it is desireable to have [accurate — informative] labels which are informative, useful and understandable for users and not merely "benign" or "malicious".

Complete Interaction. For the correct interpretation of the results, one of the vital features is the amount of available information on anomalous behaviour. So, having all network interactions such as within or between internal LANs is one of the major requirements for a valuable dataset.

Complete Capture. It is essential to capture all traffic to calculate the false-positive percentage of an IDS system. It seems some of the datasets remove traffic which is non-functional or I s not labeled.

Available Protocols. There are many different types of traffic. Some are vital for testing an IDS system such as bursty traffic which is an uneven pattern of data transmission and can cover some protocols such as HTTP and FTP. Interactive traffic includes sessions that consist of short request and response pairs such as applications involving real-time interaction with users (e.g., web browsing, online purchasing). In latency-sensitive traffic, e.g. VOIP and video conferencing, the user has an expectation that data will be delivered on time. In non-Real-time traffic, such as news and mail traffic, timely delivery is not important. A complete dataset should have both normal and anomalous traffic.

Attack Diversity. In recent years, threats have expanded their scopes into intricate scenarios such as application and app attacks. The types of attacks are changing daily. So, having the ability to test and analyze IDS and IPS systems by these new attacks and threat scenarios is one of the most important requirements that an on-line dataset should support. We categorized attacks into seven major groups based on the 2016 McAfee report, browser-based, brute force, DoS, scan or enumeration, backdoors, DNS, and other attacks (e.g., Heartbleed, Shellshock, and Apple SSL library bug).

Anonymity. Most of the datasets removed their payload due to privacy issues which decreases [deminishes] the usefulness of the dataset for some detection mechanisms, especially deep packet inspection (DPI).

Heterogeneity. Ideally, for IDS research, network traffic logs from various sources, e.g., operating systems and network equipment, would be available as they could be used for a complete test covering all aspects of the detection process. A homogeneous dataset using a single source type can be useful for analyzing a specific type of detection systems.

Feature Set. The main goal of providing a dataset is to let other researchers test and analyze their systems. One of the main challenges is to calculate and analyze the features. It is possible to extract features from different type of data sources such as traffic or logs using feature extraction applications.

Metadata. Lack of proper documentation is one of the main issues with datasets. Most do not have documentation or, even if they do, it is incomplete. Insufficient information about the network configuration, operating systems for attacker and victim machines, attack scenarios, and other vital information detracts from the usefulness of a dataset.

As Table 1 shows, among the 16 publicly available IDS datasets since 1998, just CICIDS2017 [35] covers all 11 criteria.

3 Selecting a Dataset

We selected CICIDS2017 based on the evaluation table (Table 1). Only this dataset covered all 11 evaluation criteria [35]. It includes two networks, namely attack network and victim network. The victim network has a highly secure infrastructure with firewall, router, switches and most of the common operating systems along with an agent that provides the benign behaviors on each PC. The attack network is completely separated infrastructure designed by a router and switch and a set of PCs with public IPs and different operating systems for executing the attack scenarios. Table 2 shows the victim and attackers IPs and operating systems [35].

Table 2. Operating systems and IPs [35].

	Machine	OS	IPs
Victim-Network	Servers	Win Server 2016 (DC and DNS)	192.168.10.3
		Ubuntu 16 (Web Server)	192.168.10.50–205.174.165.68
		Ubuntu 12	192.168.10.51–205.174.165.66
	PCs	Ubuntu 14.4 (32, 64)	192.168.10.19–192168.10.17
		Ubuntu 16.4 (32-64)	192.168.10.16–192.168.10.12
		Win 7 Pro	192.168.10.9
		Win 8.1-64	192.168.10.5
		Win Vista	192.168.10.8
		Win 10 (Pro 32-64)	192.168.10.14–192.168.10.15
		Mac	192.168.10.25
	Firewall	Fortinet	
Attackers	PCs	Kali	205.174.165.73
		Win 8.1	205.174.165.69
		Win 8.1	205.174.165.70
		Win 8.1	205.174.165.71

Generating the realistic background traffic is one of the highest priorities on IDS/IPS datasets. This dataset, used a CIC-B-Profile system [35], which is responsible for profiling the abstract behavior of human interactions and generates natural benign background traffic. The B-Profile for this dataset extracts the abstract behavior of 25 users based on the HTTP, HTTPS, FTP, SSH, and email protocols.

Since CICIDS2017 is intended for network security and intrusion detection purposes, it simulates seven attack families, namely: brute force attack, heartbleed attack, botnet, DoS attack, DDoS attack, web attack and infiltration attack. Table 3 shows the attacks for one week [35]. (CICIDS2017 is publicly available at http://www.unb.ca/cic/datasets/IDS2017.html)

4 Superfeature

In this paper we define "superfeature" to mean a high quality feature that is made by a linear or non-linear combination of basic or derived features. One of the main methods to extract superfeatures is to use dimension reduction techniques. Although the ability of dimension reduction techniques (like t-SNE and PCA) to visualize anomalies has been proven, the problem is that it is not always possible to interpret the meaning of the different axes of the visualization. One can extract superfeatures by applying dimensional reduction techniques and mapping data-points from higher dimensions to lower dimensions. Not all dimensional reduction techniques can do this efficiently. For example neighbor embedding techniques are not suitable for our work because they ruin the structure of the space by optimizing their cost function. The most appropriate unsupervised dimensional reduction technique which we found was singular value decomposition (SVD). It was chosen because it can provide insights about the relation of superfeatures (reduced dimensions) and features and one can be aware of the most influential features in the selected superfeatures.

4.1 Singular Value Decomposition

The singular value decomposition (SVD) is a matrix factorization. If A is an $n \times m$ matrix, then we can decompose A as a product of three different factors [40]:

$$A = U \Sigma V^*, \tag{1}$$

Table 3. Daily label of dataset [35].

Days	Labels
Monday	Benign
Tuesday	BForce, SFTP and SSH
Wednes.	DoS and Hearbleed Attacks, slowloris, Slowhttptest, Hulk and GoldenEye
Thurs.	Web and Infiltration Attacks, Web BForce, XSS and Sql Inject, Infiltration Dropbox Download and Cool disk
Friday	DDoS LOIT, Botnet ARES, PortScans (sS, sT, sF, sX, sN, sP, sV, sU, sO, sA, sW, sR, sL and B)

where U represents an orthogonal $n \times n$ matrix, also, V is an orthogonal $m \times m$ matrix, V^* is the transpose of V, and Σ is an $n \times m$ sparse matrix with all zero values except for its diagonal entries, which are nonnegative real numbers. If σ_{ij} is the i, j entry of Σ, then $\sigma_{ij} = 0$ unless $i = j$ and $\sigma_{ii} = \sigma_i \geq 0$. The σ_i are the "singular values" and the columns of u and v are respectively the right and left singular vectors. Singular values are considered to be ordered so that

$$\sigma_1 \geq \sigma_2 \geq \cdots .$$

5 Analysis and Result

First we select two eigenvectors of SVD as our superfeatures. In order to show the efficiency of our selected superfeatures, we calculated them for each attack and then we compared them with the top two individual features from our feature selection in Table 4 on previous research [35]. In order to build the decomposition matrix, we used a set of 8000 randomly selected benign flows and 2000 attack flows for each attack (from our training dataset), which is considerably small in comparison to the whole dataset. Then we used a random forest classifier and 5 fold cross validation for the top two individually selected features for each attack and the top two selected superfeatures. We choose a random forest classifier because it is among the best classifiers in Table 5 [35]. As Table 8 shows, superfeatures outperform the top individual features in all selected attacks.

We can consider our dataset as a matrix that every row corresponds a Netflow and that column corresponds a feature. Now given this matrix we can decompose it by SVD and then interpret the result [33]. Matrix U is Netflow to superfeature similarity matrix. Also, matrix Σ represents strength of each superfeature. Finally, matrix V^* represents features to superfeatures similarity matrix.

In order to interpret "features to superfeatures" relation, we use matrix V^* in the SVD formula. Also, we defined a threshold of 0.1 to remove unimportant relations. Tables 6 and 7 represent relationships between superfeatures and features for all attacks. CICIDS2017 contains 80 features. In these tables we consider only meaningful relations: that means we removed all super-feature to feature relations with value zero for both top superfeatures. As is evident from Tables 6 and 7, the most influential features for generating superfeatures are flow duration, inter-arrival time related features (for flow, forward and backward categories) and idle time related features. One of the main reasons might be that all of these attacks contain same characteristics and they are all anomalies. Also, all of these attacks contains some bursty behaviors in comparison with benign traffic and because of this kind of behavior the flow duration, idle time and IAT related features are so pronounced as to indicate superfeatures.

Table 4. Feature selection [35].

Label	Feature	Weight
Benign	B.Packet Len Min	0.0479
	Subflow F.Bytes	0.0007
	Total Len F.Packets	0.0004
	F.Packet Len Mean	0.0002
DoS GoldenEye	B.Packet Len Std	0.1585
	Flow IAT Min	0.0317
	Fwd IAT Min	0.0257
	Flow IAT Mean	0.0214
Heartbleed	B.Packet Len Std	0.2028
	Subflow F.Bytes	0.1367
	Flow Duration	0.0991
	Total Len F.Packets	0.0903
DoS Hulk	B.Packet Len Std	0.2028
	B.Packet Len Std	0.1277
	Flow Duration	0.0437
	Flow IAT Std	0.0227
DoS Slowhttp	Flow Duration	0.0443
	Active Min	0.0228
	Active Mean	0.0219
	Flow IAT Std	0.0200
DoS slowloris	Flow Duration	0.0431
	F.IAT Min	0.0378
	B.IAT Mean	0.0300
	F.IAT Mean	0.0265
SSH-Patator	Init Win F.Bytes	0.0079
	Subflow F.Bytes	0.0052
	Total Len F.Packets	0.0034
	ACK Flag Count	0.0007
FTP-Patator	Init Win F.Bytes	0.0077
	F.PSH Flags	0.0062
	SYN Flag Count	0.0061
	F.Packets/s	0.0014
Web Attack	Init Win F.Bytes	0.0200
	Subflow F.Bytes	0.0145
	Init Win B.Bytes	0.0129
	Total Len F.Packets	0.0096

Table 4. (*continued*)

Label	Feature	Weight
Infiltration	Subflow F.Bytes	4.3012
	Total Len F.Packets	2.8427
	Flow Duration	0.0657
	Active Mean	0.0227
Bot	Subflow F.Bytes	0.0239
	Total Len F.Packets	0.0158
	F.Packet Len Mean	0.0025
	B.Packets/s	0.0021
PortScan	Init Win F.Bytes	0.0083
	B.Packets/s	0.0032
	PSH Flag Count	0.0009
DDoS	B.Packet Len Std	0.1728
	Avg Packet Size	0.0162
	Flow Duration	0.0137
	Flow IAT Std	0.0086

Table 5. The performance examination results [35].

Algorithm	Pr	Rc	F1	Execution (sec.)
KNN	0.96	0.96	0.96	1908.23
RF	0.98	0.97	0.97	74.39
ID3	0.98	0.98	0.98	235.02
Adaboost	0.77	0.84	0.77	1126.24
MLP	0.77	0.83	0.76	575.73
Naive-Bayes	0.88	0.04	0.04	14.77
QDA	0.97	0.88	0.92	18.79

Also, we visualized different attacks in Fig. 1 by using the top two selected superfeatures in two-dimensional planes. Red 'A' characters represent attack flows and blue 'B' characters represent benign flows. We can observe that brute force and web attacks tend to aggregate in the far left of figures. On the other hand, DoS attack is spread in space which might be due to being a low volume attack making it hard to distinguish from benign traffic. As well, the DDoS attack is the most difficult to distinguish because it is similar to benign flows. Moreover, except the DoS attack, there are no malicious flows in the upper right of visualizations.

Table 6. Web attack, FTP and SSH bruteforce attack superfeatures and features relations.

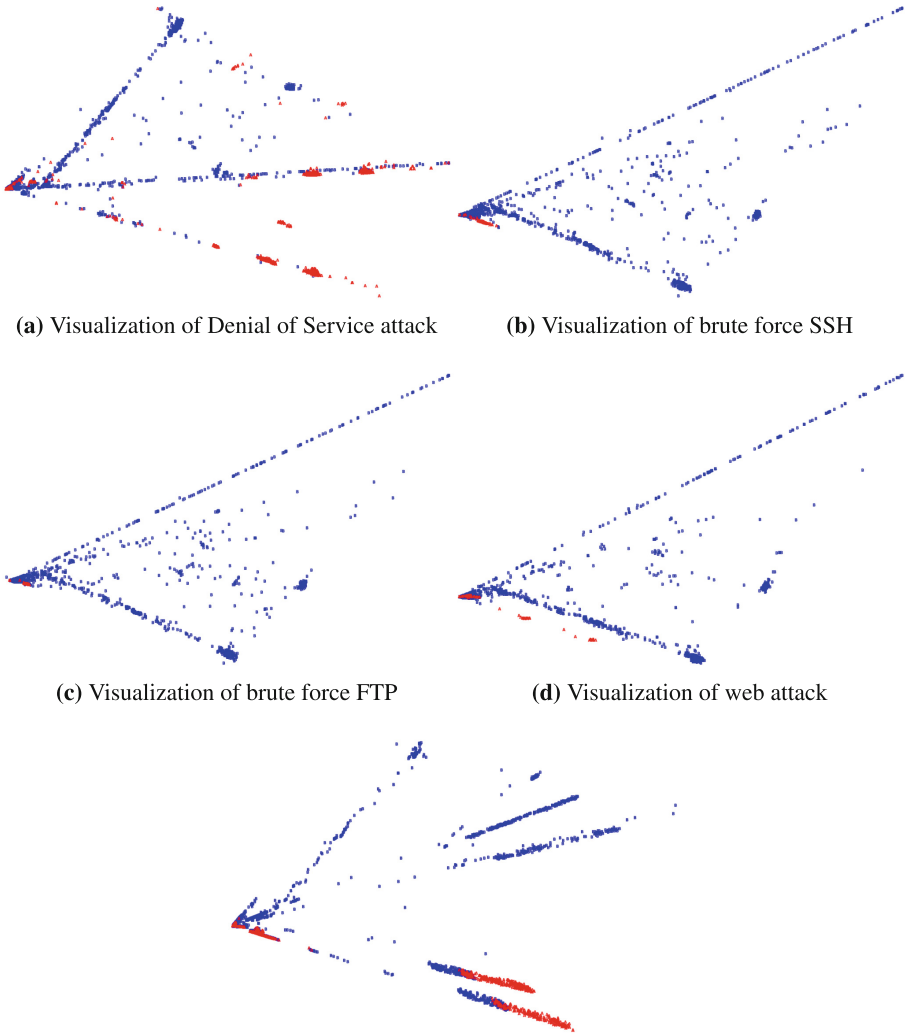
	Web attack		FTP bruteforce		SSH bruteforce	
	First SF	Second SF	First SF	Second SF	First SP	Second SP
Flow duration	0.5229	0.2536	0.5030	0.2882	0.5072	0.2800
Flow IAT Std	0	0.1487	0	0.1476	0	0.1462
Flow IAT Max	0.1787	0.2998	0.2037	0.2809	0.2015	0.2904
Fwd IAT Total	0.5190	0.2527	0.4983	0.2875	0.5025	0.2795
Fwd IAT Mean	0	0.2514	0.1126	0.2597	0.1072	0.2567
Fwd IAT Max	0.1780	0.2982	0.2025	0.2786	0.2003	0.2880
Fwd IAT Min	0	0.2661	0	0.2756	0	0.2709
Bwd IAT Total	0.4961	0.2602	0.4744	0.2979	0.4771	0.2966
Bwd IAT Mean	0	0.2481	0.1104	0.2495	0.1042	0.2437
Bwd IAT Max	0.1669	0.2725	0.1873	0.2510	0.1838	0.2537
Bwd IAT Min	0	0.2692	0	0.2707	0	0.2637
Idle Mean	0.1686	0.2883	0.1871	0.2666	0.1844	0.2718
Idle Max	0.1725	0.2912	0.1943	0.2682	0.1915	0.2743
Idle Min	0.1624	0.2908	0.1783	0.2697	0.1755	0.2743

SF: First Superfeature

Table 7. DDoS and DoS attack superfeatures and features relations.

	DDoS		DoS	
	First SF	Second SF	First SF	Second SF
Flow duration	0.4312	0.1139	0.3912	0.1705
Flow IAT Std	0.1045	0.1039	0	0
Flow IAT Max	0.3652	0.1913	0.3487	0.1833
Fwd IAT Total	0.4269	0.1070	0.3902	0.1663
Fwd IAT Std	0.1440	0	0.1391	0
Fwd IAT Max	0.3709	0.1738	0.3486	0.1843
Fwd IAT Min	0	0	0	0
Bwd IAT Total	0.1777	0.7727	0.1854	0.7413
BWD IAT Mean	0	0		0.121337
Bwd IAT Std	0	0.1451	0	0.1530
Bwd IAT Max	0.1322	0.4359	0.1479	0.4036
Bwd IAT Min	0	0.2692	0	0
Idle Mean	0.2872	0.1656	0.3426	0.1840
Idle Std	0.1086	0	0	0
Idle Max	0.3640	0.1918	0.3477	0.1966
Idle Min	0.2101	0.1419	0.3380	0.2075

SF: First Superfeature



(a) Visualization of Denial of Service attack

(b) Visualization of brute force SSH

(c) Visualization of brute force FTP

(d) Visualization of web attack

(e) Visualization of Distributed Denial of Service Attack

Fig. 1. Visualization of different attacks by using superfeatures. (Color figure online)

Table 8. Accuracy of random forest using different feature selection and extraction scenarios.

	Considering all features	Considering top two superfeatures	Considering top two features
FTP brute force	0.9999	**0.9969**	0.9944
SSH brute force	0.9999	**0.9996**	0.9976
Web attacks	0.9998	**0.9982**	0.9911
DoS	0.9995	**0.9730**	0.9012
DDoS	0.9420	**0.8444**	0.6509

6 Conclusions

One of the fundamental concerns of researchers in the intrusion detection systems domain is the availability of representative datasets. We have analyzed 16 post 1998 publicly available IDS datasets and identified the following deficiencies: limited traffic diversity, insufficient traffic volume, anonymized packet information payload, constraints on variety of attacks, and lack of feature set and metadata. Our focus was on CICIDS2017, a publicly available IDS dataset including most current attacks in common use. Also, we defined the concept of superfeatures which are derived features extracted by using singular value decomposition. Then, we used random forest algorithm to compare the Accuracy of superfeatures with the best short feature set that were selected by using a random forest regressor algorithm. Finally, we proved that superfeatures demonstrate better accuracy than individual features.

Acknowledgements. The authors acknowledge the generous funding from the Atlantic Canada Opportunity Agency (ACOA) through the Atlantic Innovation Fund (AIF) and through grants from the National Science and Engineering Research Council of Canada (NSERC) to Dr. Ghorbani.

References

1. Shiravi, A., Shiravi, H., Tavallaee, M., Ghorbani, A.A.: Toward developing a systematic approach to generate benchmark datasets for intrusion detection. Comput. Secur. **31**(3), 357–374 (2012)
2. Brown, C., Cowperthwaite, A., Hijazi, A., Somayaji, A.: Analysis of the 1999 DARPA/Lincoln laboratory IDS evaluation data with NetaDHICT. In: 2009 IEEE SCISDA, pp. 1–7 (2009)
3. The Canadian Institute for Cybersecurity (CIC), CICFlowMeter: The network traffic flow generator and alanlyzer (2017). https://github.com/ISCX/CICFlowMeter
4. Creech, G., Hu, J.L.: Generation of a new IDS test dataset: time to retire the KDD collection. In: 2013 IEEE Wireless Communications and Networking Conference (WCNC), pp. 4487–4492 (2013)

5. T.C. Center for Applied Internet Data Analysis (CAIDA): The CAIDA OC48 Peering Point Traces Dataset, San Jose, California (2002)
6. I.U. University of California: KDD cup 1999 dataset (1999). http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
7. T.C. Center for Applied Internet Data Analysis (CAIDA): CAIDA DDoS attack dataset (2007)
8. T.C. Center for Applied Internet Data Analysis (CAIDA): CAIDA anonymized internet traces 2016 dataset (2016)
9. Gharib, A., Sharafaldin, I., Habibi Lashkari, A., Ghorbani, A.A.: An evaluation framework for intrusion detection dataset. In: 2016 International Conference on Information Science and Security (ICISS), Thailand, pp. 1–6 (2016)
10. Ghorbani, A.A., Lu, W., Tavallaee, M.: Network Intrusion Detection and Prevention: Concepts and Techniques. Springer, Boston (2010). https://doi.org/10.1007/978-0-387-88771-5
11. T.S. Group: Defcon 8, 10 and 11 (2000). https://www.defcon.org/
12. Habibi Lashkari, A., Draper Gil, G., Mamun, M.S.I., Ghorbani, A.A.: Characterization of tor traffic using time based features. In: Proceedings of the 3rd International Conference on Information Systems Security and Privacy (ICISSP), Portugal, pp. 253–262 (2017)
13. Heidemann, J., Papdopoulos, C.: Uses and challenges for network datasets. In: Cybersecurity Applications Technology Conference For Homeland Security, CATCH 2009, pp. 73–82 (2009)
14. Koch, R., Golling, M.G., Rodosek, G.D.: Towards comparability of intrusion detection systems: new data sets. In: Proceedings of the TERENA Networking Conference, p. 7 (2017)
15. Sato M., Yamaki H., Takakura H.: Unknown attacks detection using feature extraction from anomaly-based IDS alerts. In: 2012 IEEE/IPSJ 12th International Symposium on Applications and the Internet (SAINT), pp. 273–277 (2012)
16. McHugh, J.: Testing intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln laboratory. ACM Trans. Inf. Syst. Secur. **3**(4), 262–294 (2000)
17. Nechaev, B., Allman, M., Paxson, V., Gurtov, A.: Lawrence Berkeley National Laboratory (LBNL)/ICSI enterprise tracing project (2004)
18. Nehinbe, J.O.: A simple method for improving intrusion detections in corporate networks. In: Weerasinghe, D. (ed.) ISDF 2009. LNICST, vol. 41, pp. 111–122. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-11530-1_13
19. Nehinbe, J.O.: A critical evaluation of datasets for investigating IDSS and IPSS researches. In: IEEE 10th International Conference on CIS, pp. 92–97 (2011)
20. University of Massachusetts Amherst: Optimistic TCP hacking (2011). http://traces.cs.umass.edu
21. Pedregosa, F., et al.: Scikit-learn: machine learning in Python (2011)
22. Proebstel, E.P.: Characterizing and improving distributed network-based intrusion detection systems (NIDS): timestamp synchronization and sampled traffic. Master's thesis, University of California DAVIS, CA, USA (2008)
23. Chitrakar, R., Huang, C.: Anomaly based intrusion detection using hybrid learning approach of combining k-medoids clustering and Naive Bayes classification (2012)
24. Umer, M.F., Sher, M., Bi, Y.: Flow-based intrusion detection: techniques and challenges. Comput. Secur. **70**, 238–254 (2017). In: 8th WiCOM, pp. 1–5
25. Sangster, B., et al.: Toward instrumenting network warfare competitions to generate labeled datasets. In: 2009 USENIX. USENIX: The Advanced Computing System Association (2009)

26. Scott, P., Wilkins, E.: Evaluating data mining procedures: techniques for generating artificial data sets. Inf. Softw. Technol. **41**(9), 579–587 (1999)
27. Sharafaldin, I., Gharib, A., Habibi Lashkari, A., Ghorbani, A.A.: Towards a reliable intrusion detection benchmark dataset. Softw. Netw. **2017**, 177–200 (2017)
28. Song, J., Takakura, H., Okabe, Y., Eto, M., Inoue, D., Nakao, K.: Statistical analysis of honeypot data and building of Kyoto 2006+ dataset for NIDS evaluation. In: Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security, pp. 29–36. ACM (2011)
29. Sperotto, A., Sadre, R., Vliet, F., Pras, A.: A labeled data set for flow-based intrusion detection. In: Proceedings of the 9th IEEE International Workshop on IP Operations and Management, IPOM 2009, pp. 39–50 (2009)
30. Prusty, S., Levine, B.N., Liberatore, M.: Forensic Investigation of the OneSwarm Anonymous Filesharing System. In: ACM Conference on CCS (2011)
31. Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A.A.: A detailed analysis of the KDD cup 99 data set. In: 2009 IEEE SCISDA, pp. 1–6 (2009)
32. Xie, M., Hu, J.: Evaluating host-based anomaly detection systems: a preliminary analysis of ADFA-LD. In: Proceedings of the 6th IEEE International Congress on Image and Signal Processing (CISP 2013), pp. 1711–1716 (2013)
33. Skillicorn, D.: Understanding Complex Datasets: Data Mining with Matrix Decompositions. CRC Press, Boca Rato (2007). Evaluating host-based anomaly detection systems: a preliminary analysis of ADFA-LD. In: 2013 6th International Congress on Image and Signal Processing (CISP), vol. 03, pp. 1711–1716
34. Xie, M., Hu, J., Slay, J.: Evaluating host-based anomaly detection systems: application of the one-class SVM algorithm to ADFA-LD. In: 2014 11th FSKD, pp. 978–982 (2014)
35. Sharafaldin, I., Habibi Lashkari, A., Ghorbani, A.A.: Toward generating a new intrusion detection dataset and intrusion traffic characterization. In: 4th International Conference on Information Systems Security and Privacy (ICISSP), Portugal, January 2018 (2017)
36. Szabó, G., Orincsay, D., Malomsoky, S., Szabó, I.: On the validation of traffic classification algorithms. In: Claypool, M., Uhlig, S. (eds.) PAM 2008. LNCS, vol. 4979, pp. 72–81. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-79232-1_8
37. Garcia, S., Grill, M., Stiborek, J., Zunino, A.: An empirical comparison of botnet detection methods. Comput. Securi. **45**, 100–123 (2014)
38. Hofstede, R., Hendriks, L., Sperotto, A., Pras, A.: SSH compromise detection using NetFlow/IPFIX. ACM SIGCOMM Comput. Commun. Rev. **44**(5), 20–26 (2014)
39. Maciá-Fernández, G., Camacho, J., Magán-Carrión, R., García-Teodoro, P., Therón, R.: UGR '16: a new dataset for the evaluation of cyclostationarity-based network IDSs. Comput. Secur. **73**, 411–424 (2018)
40. De Lathauwer, L., De Moor, B., Vandewalle, J., B.S.S. by Higher-Order: Blind source separation by higher-order singular value decomposition. In: Proceeding of the 7th European Signal Processing Conference (EUSIPCO 1994), Edinburgh, UK, pp. 175–178 (1994)