NetFlow Datasets for Machine Learning-Based Network Intrusion Detection Systems

Mohanad Sarhan[1(✉)], Siamak Layeghy[1], Nour Moustafa[2], and Marius Portmann[1]

[1] University of Queensland, Brisbane, QLD 4072, Australia
{m.sarhan,siamak.layeghy}@uq.net.au, marius@ieee.org
[2] University of New South Wales, Canberra, ACT 2612, Australia
nour.moustafa@unsw.edu.au

Abstract. Machine Learning (ML)-based Network Intrusion Detection Systems (NIDSs) have become a promising tool to protect networks against cyberattacks. A wide range of datasets are publicly available and have been used for the development and evaluation of a large number of ML-based NIDS in the research community. However, since these NIDS datasets have very different feature sets, it is currently very difficult to reliably compare ML models across different datasets, and hence if they generalise to different network environments and attack scenarios. The limited ability to evaluate ML-based NIDSs has led to a gap between the extensive academic research conducted and the actual practical deployments in the real-world networks. This paper addresses this limitation, by providing five NIDS datasets with a common, practically relevant feature set, based on NetFlow. These datasets are generated from the following four existing benchmark NIDS datasets: UNSW-NB15, BoT-IoT, ToN-IoT, and CSE-CIC-IDS2018. We have used the raw packet capture files of these datasets, and converted them to the NetFlow format, with a common feature set. The benefits of using NetFlow as a common format include its practical relevance, its wide deployment in production networks, and its scaling properties. The generated NetFlow datasets presented in this paper have been labelled for both binary- and multi-class traffic and attack classification experiments, and we have made them available for to the research community [1]. As a use-case and application scenario, the paper presents an evaluation of an Extra Trees ensemble classifier across these datasets.

Keywords: Network intrusion detection system · NetFlow · Machine learning · Network datasets · Network features

1 Introduction

Anomaly-based Network Intrusion Detection Systems (NIDSs) aim to learn and extract complex network behaviours to classify incoming traffic into various

Z. Deze et al. (Eds.): BDTA 2020/WiCON 2020, LNICST 371, pp. 117–135, 2021.
https://doi.org/10.1007/978-3-030-72802-1_9

attacks and benign classes [2]. Network attack vectors can be obtained from various features transmitted through network traffic, such as packet counts/sizes, protocols, services and flags. Each network attack's type has a different identifying pattern, known as a set of events that may compromise the security principles of networks if undetected [3]. The fact that these patterns are learnt from network traffic data shows the importance of data collection for Machine Learning (ML) training and evaluation stages. Real network data is challenging to obtain due to security and privacy issues. Also, production networks do not generate labelled flows, which is necessary for following a supervised ML learning approach.

As such, researchers have used network testbeds to create synthetic datasets that are publicly available for research purposes [4]. These NIDS datasets contain labelled network flows that are made up of certain features extracted from network traffic. These features are pre-determined by the datasets' authors based on their domain knowledge and tools used during their extraction. Network data features have a great impact on the performance of ML-based NIDSs [5]. Over the past few years, researchers have evaluated their proposed models on datasets using their original sets of features. However, as these features are very different, evaluating ML models often is not reliable, as each ML-based NIDS is trained and validated using different data features. Moreover, due to their complex techniques of extraction, these network feature sets might not be feasible for collection or storage in real-production networks.

In order to address this gap, we have created five NIDS datasets, which all have the same sets of features that facilitate reliable NIDS evaluation over multiple datasets. These datasets are created by converting four well-known modern NIDS datasets into NetFlow format. NetFlow is a widely deployed protocol of network flow collection [6]. Obtaining NetFlow features from existing NIDS datasets will enable researchers to evaluate ML models across various datasets using the same set of features. Moreover, it will also determine the performance of NetFlow features in detecting various attack types present in the datasets.

The rest of this paper is organised as below. Section 2 illustrates the limitations faced by existing datasets and how they can be overcome. Section 3 explains the importance and methodology of creating NetFlow datasets as well as the distribution of various benign/attack flows in the newly created datasets. Finally, the new datasets are evaluated in Sect. 4, by comparing their binary-class and multi-class classification performance to the original features of their corresponding datasets. The main contribution of this paper is to provide the research community with five NetFlow datasets, with the same feature sets, using four existing benchmark datasets, along with an initial set of results collected while evaluating the new datasets using binary-class and multi-class classification experiments.

2 Limitations of Existing Datasets

Due to the complexity in obtaining labelled real-world network flows, researchers have generated synthetic benchmark NIDS datasets. These datasets are made publicly available for use in the training and testing stages of the ML-based NIDSs. Currently, there are more than 15 NIDS datasets available in the field [7] containing labelled network flows. These datasets reflect network benign behaviour combined with synthetic attack scenarios. Each dataset contains a few attack categories conducted over a testbed network. The packets are captured, during the experiments, in the packet capture (pcap) format, and then pre-determined network features are extracted from these pcap files. A key stage of designing an ML-based NIDS is the selection of these features. The selected features must be feasible in count and extraction's complexity for efficient storage and collection. The sets should also provide adequate information for the efficient classification by the ML model.

Due to lack of a standard set of features for generating NIDS datasets, the authors of these datasets have applied their own domain knowledge to create network features, which they believe would aid in the classification process. As a result, each available dataset has been created with an almost exclusive set of network features. The variance of information represented in each dataset has caused limitations in the field that keeps aggravating with the new releases and production of NIDS datasets. The two main issues of having different feature sets in benchmark datasets are 1. dimensional overload due to collection and storage of various features, some of which are irrelevant and 2. inability to evaluate an ML model's generalisation across multiple NIDS datasets using a targeted or a proposed feature set. We believe the unreliable evaluation methods have caused a gap between the extensive academic research conducted and the actual deployments of ML-based NIDS models in the real-world.

Identifying the ideal set of network features to be used in NIDS datasets has been an ongoing research topic over the last decade. However, due to the subjection to the datasets used in the experiments, the identified feature sets have been custom to each dataset. These sets are also subjected to the feature selection techniques and ML models used to identify and evaluate them respectively. Moreover, due to the differences in datasets', the selected or identified features can not be evaluated using other datasets, simply due to their absence. The rest of this section discusses four of the most recent and common publicly available NIDS datasets. These datasets have been released within the last five years so they represent modern behavioural network attacks.

- **UNSW-NB15.** The Cyber Range Lab of the Australian Centre for Cyber Security (ACCS) released the widely used, UNSW-NB15, dataset in 2015. The IXIA PerfectStorm tool was utilised to generate a hybrid of testbed-based benign network activities as well as synthetic attack scenarios. Tcpdump tool was implemented to capture a total of 100 GB of pcap files. Argus and Bro-IDS now called Zeek, and twelve additional algorithms were used to extract the dataset's original 49 features [8]. The dataset contains 2,218,761 (87.35%) benign flows and 321,283 (12.65%) attack ones, that is, 2,540,044 flows in total.
- **BoT-IoT.** The Cyber Range Lab of the Australian Centre for Cyber Security (ACCS) designed a network environment in 2018 that consists of normal and

botnet traffic [9]. The Ostinato and Node-red tools were utilised to generate the non-IoT and IoT traffic respectively. A total of 69.3GB of pcap files were captured and Argus tool was used to extract the dataset's original 42 features. The dataset contains 477 (0.01%) benign flows and 3,668,045 (99.99%) attack ones, that is, 3,668,522 flows in total.

- **ToN-IoT.** A recent heterogeneous dataset released in 2020 [10] that includes telemetry data of Internet of Things (IoT) services, network traffic of IoT networks and operating system logs. In this paper, we utilise the portion containing network traffic flows. The dataset is made up of a large number of attack scenarios conducted in a representation of a medium-scale network at the Cyber Range Lab by ACCS. Bro-IDS, now called Zeek, was used to extract the dataset's original 44 features. The dataset is made up of 796,380 (3.56%) benign flows and 21,542,641 (96.44%) attack samples, that is, 22,339,021 flows in total.
- **CSE-CIC-IDS2018.** A dataset released by a collaborative project between the Communications Security Establishment (CSE) & Canadian Institute for Cybersecurity (CIC) in 2018 [11]. The victim network consisted of five different organisational departments and an additional server room. The benign packets were generated by network events using the abstract behaviour of human users. The attack scenarios were executed by one or more machines outside the target network. The CICFlowMeter-V3 tool was used to extract the original dataset's 75 features. The full dataset has 13,484,708 (83.07%) benign flows and 2,748,235 (16.93%) attack flows, that is, 16,232,943 flows in total.

Figure 1 illustrates all the shared and exclusive features of these datasets. As seen, the list of features shared by all four datasets includes only 3 features, and the pairwise shared features numbers vary from 1 to 5. Most of the features are exclusive to individual datasets. This has made it challenging for researchers to measure the performance of their proposed ML models using the same set of features across the four datasets. Apart from the small number of shared features, other differences make it even more difficult for using these datasets in the evaluation of ML-based NIDSs. The first issue is the vast differences in the ratio of the benign/attack flows. The UNSW-NB15 and CSE-CIC-IDS2018 datasets have very high benign to attack ratios (20 and 7.2 respectively) whereas for the ToN-IoT and BoT-IoT datasets this ratio is about 0.2 and 0.02 respectively.

The next issue is the number and type of features in each dataset. The UNSW-NB15 and ToN-IoT datasets have approximately the same number of original features. The CSE-CIC-IDS2018 dataset has almost double the number of their features and the BoT-IoT dataset has a slightly lower number. The original feature sets in UNSW-NB15, BoT-IoT and CSE-CIC-IDS2018 contain handcrafted features that are not present in network traffic but are statistically measured from other features, such as the average or sum of the number of bytes transferred over the last 100 s. All these differences, and the necessity of having multiple NIDS datasets with the common ground feature set, to generalise the evaluation of NIDSs, has led to the generation of the new datasets. This will

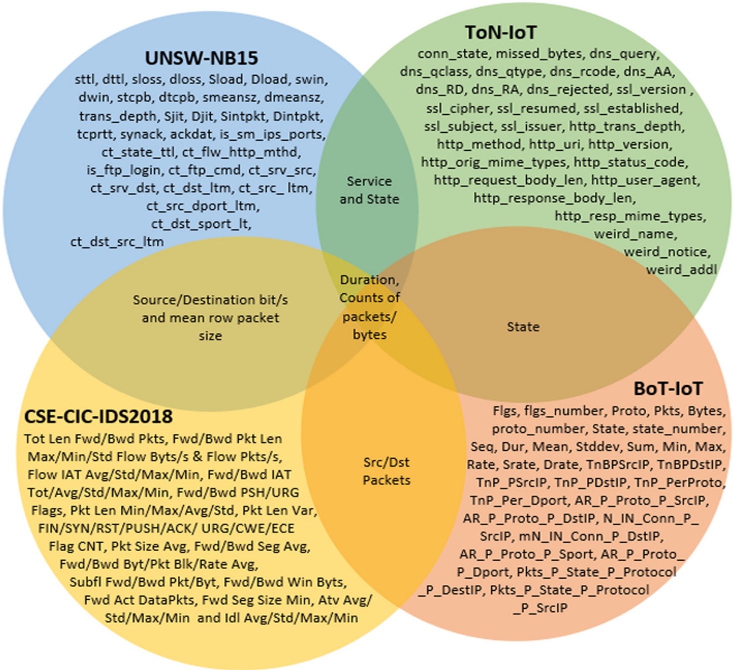

Fig. 1. List of the shared and exclusive features of four NIDS datasets

enable researchers to evaluate their proposed ML-based NIDS model's performance across various network designs and attack scenarios to make sure their measured model performance well generalises.

3 NetFlow Datasets

3.1 NetFlow

Collecting and recording network traffic is necessary to monitor and analyse network environments. There are two main trends for this process, capturing the complete network packets, and extracting a summary in the form of flows. While packet capturing provides full access to the network information, it is not scalable as it might necessitate large-capacity data storage to record a short period of traffic. The large volume of such data not only makes it difficult for the analysis, but it also faces privacy and security concerns. The alternative method is extracting network traffic summary as flows, which is very common in the networking industry due to its practical relevance and scaling properties. A network flow identifies a sequence of packets between two endpoints sharing a number of attributes. The packets flow can be unidirectional or bidirectional. These common attributes include; source/destination IP address and L4 (transport layer) ports, and the L4 protocol. These shared attributes are often referred to as the five-tuple.

The information provided by network flows are essential to analyse network traffic for security events [12]. The network flows can be represented in various formats where the NetFlow is the de-facto industry standard developed and proposed by Darren and Barry Bruins from Cisco in 1996 [13]. Other network hardware manufacturers have also implemented and adopted their protocols such as *NetStream* by Huawei, *Jflow* by Juniper, *Cflow* by Alcatel-Lucent, *Rflow* by Ericsson and *s-flow* that is supported by 3Com/HP, Dell, and Netgear. In response to the need for a universal standard of flow information, the Internet Engineering Task Force (IETF) has developed a new protocol, named Internet Protocol Flow Information Export (IPFIX) which is based on Cisco NetFlow. Similar to NetFlow, IPFIX considers a flow to be any number of packets sharing certain characteristics observed in a specific time-frame. NetFlow evolved over the years, where version 9 is the most common due to its larger variety of features and bidirectional flow support [14].

NetFlow makes it possible to convert any available dataset into a common ground feature set. Accomplishing that, researchers would be able to compare datasets efficiently and most importantly evaluate their proposed ML-based NIDS models using the same set of features across various datasets and attack types. Most of the production network devices such as routers and switches are capable of extracting NetFlow records. This is a great motivation for evaluating the performance of NetFlow features in terms of attack detection, as the level of complexity and resources required to collect and store them is lower. Moreover, the generated datasets sharing the same set of features can be merged together to generate a universal NIDS dataset containing data flows from various network environments consisting of various attack scenarios. Finally, the same set of features can be extracted from any future generated NIDS dataset and be merged into the current ones, increasing the value of the datasets.

3.2 Conversion

Figure 2 shows the procedure of creating the NetFlow datasets by extracting flows (in NetFlow format) from the pcap files of the original datasets, and labelling extracted flow records based on the grand truths provided by dataset authors. We utilised the publicly available pcap files of each dataset to generate

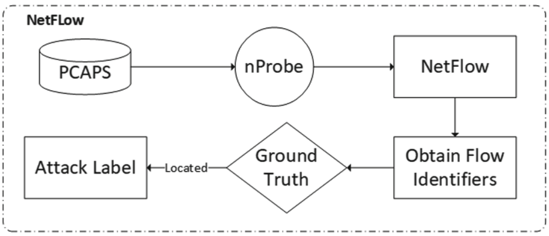

Fig. 2. NetFlow datasets' extraction and labelling procedure

Table 1. List of NetFlow fields included in the datasets proposed in this paper

Feature	Description
IPV4_SRC_ADDR	IPv4 source address
IPV4_DST_ADDR	IPv4 destination address
L4_SRC_PORT	IPv4 source port number
L4_DST_PORT	IPv4 destination port number
PROTOCOL	IP protocol identifier byte
TCP_FLAGS	Cumulative of all TCP flags
L7_PROTO	Layer 7 protocol (numeric)
IN_BYTES	Incoming number of bytes
OUT_BYTES	Outgoing number of bytes
IN_PKTS	Incoming number of packets
OUT_PKTS	Outgoing number of packets
FLOW_DURATION_MILLISECONDS	Flow duration in milliseconds

the NetFlow datasets. The nProbe tool by Ntop [15] was utilised to convert the pcaps into NetFlow version 9 format and selecting 12 features to be extracted. Table 1 lists the extracted NetFlow features along with their brief description. Using nProbe we create a text file listing the pcaps path of the original datasets. We specify NetFlow version 9 due to its popularity.

The dump format is chosen as text flows, in which each feature is separated by a comma (,) to be utilised as CSV files. The maximum number of flows in a file is 100 m dumped in a maximum of 100 m seconds, and nProbe is set not to modify the original pcaps timestamps. In the last step, we create two label features by matching the five flow identifiers; source/destination IPs and ports and protocol to the ground truth attack events published with the original datasets. If a flow is located in the attack events it would be labelled as an attack, class 1, in the binary label and its respective attack's type would be recorded in the attack label, otherwise, the record is labelled as a benign flow, class 0.

3.3 Datasets

Table 2 lists the NetFlow datasets, and compares their properties to the original datasets, in terms of Feature Extraction (FE) tool, number of features, files size and the benign to attack samples ratio. As illustrated, there is one NetFlow dataset corresponding to each original NIDS dataset, and the fifth NetFlow dataset is the comprehensive dataset that combines all the four.

- **NF-UNSW-NB15.** The NetFlow-based format of the UNSW-NB15 dataset, named NF-UNSW-NB15, has been developed and labelled with its respective attack categories. The total number of data flows are 1,623,118 out of which 72,406 (4.46%) are attack samples and 1,550,712 (95.54%) are benign. The

Table 2. Specifications of the datasets proposed in this paper, compared to the original datasets that have been used to generate them

Dataset	Release year	Feature extraction tool	Number of features	CSV size (GB)	Benign to attack samples ratio
UNSW-NB15	2015	Argus, Bro-IDS and MS SQL	49	0.55	8.7 to 1.3
NF-UNSW-NB15	**2020**	**nProbe**	**12**	**0.11**	**9.6 to 0.4**
BoT-IoT	2018	Argus	42	0.95	0.0 to 10
NF-BoT-IoT	**2020**	**nProbe**	**12**	**0.05**	**0.2 to 9.8**
ToN-IoT	2020	Bro-IDS	44	3.02	0.4 to 9.6
NF-ToN-IoT	**2020**	**nProbe**	**12**	**0.09**	**2.0 to 8.0**
CSE-CIC-IDS2018	2018	CICFlowMeter-V3	75	6.41	8.3 to 1.7
NF-CSE-CIC-IDS2018	**2020**	**nProbe**	**12**	**0.58**	**8.8 to 1.2**
NF-UQ-NIDS	**2020**	**nProbe**	**12**	**1.0**	**7.7 to 2.3**

attack samples are further classified into nine subcategories, Table 3 represents the NF-UNSW-NB15 dataset's distribution of all flows.

- **NF-BoT-IoT.** An IoT NetFlow-based dataset generated using the BoT-IoT dataset, named NF-BoT-IoT. The features were extracted from the publicly available pcap files and the flows were labelled with their respective attack categories. The total number of data flows are 600,100 out of which 586,241 (97.69%) are attack samples and 13,859 (2.31%) are benign. There are four attack categories in the dataset, Table 4 represents the NF-BoT-IoT distribution of all flows.
- **NF-ToN-IoT.** We utilised the publicly available pcaps of the ToN-IoT dataset to generate its NetFlow records, leading to a NetFlow-based IoT network dataset called NF-ToN-IoT. The total number of data flows are 1,379,274 out of which 1,108,995 (80.4%) are attack samples and 270,279 (19.6%) are benign ones, Table 5 lists and defines the distribution of the NF-ToN-IoT dataset.
- **NF-CSE-CIC-IDS2018.** We utilised the original pcap files of the CSE-CIC-IDS2018 dataset to generate a NetFlow-based dataset called NF-CSE-CIC-IDS2018. The total number of flows are 8,392,401 out of which 1,019,203 (12.14%) are attack samples and 7,373,198 (87.86%) are benign ones, Table 6 represents the dataset's distribution.
- **NF-UQ-NIDS.** A comprehensive dataset, merging all the aforementioned datasets. The newly published dataset represents the benefits of the shared dataset feature sets, where the merging of multiple smaller datasets is possible. This will eventually lead to a bigger and more universal NIDS dataset containing flows from multiple network setups and different attack settings. It includes an additional label feature, identifying the original dataset of each flow. This can be used to compare the same attack scenarios conducted over two or more different testbed networks. The attack categories have been modified to combine all parent categories. Attacks named DoS attacks-Hulk, DoS attacks-SlowHTTPTest, DoS attacks-GoldenEye and DoS attacks-Slowloris

Table 3. NF-UNSW-NB15 distribution

Class	Count	Description
Benign	1550712	Normal unmalicious flows
Fuzzers	19463	An attack in which the attacker sends large amounts of random data which cause a system to crash and also aim to discover security vulnerabilities in a system
Analysis	1995	A group that presents a variety of threats that target web applications through ports, emails and scripts
Backdoor	1782	A technique that aims to bypass security mechanisms by replying to specific constructed client applications
DoS	5051	Denial of Service is an attempt to overload a computer system's resources with the aim of preventing access to or availability of its data
Exploits	24736	Are sequences of commands controlling the behaviour of a host through a known vulnerability
Generic	5570	A method that targets cryptography and causes a collision with each block-cipher
Reconnaissance	12291	A technique for gathering information about a network host and is also known as a probe
Shellcode	1365	A malware that penetrates a code to control a victim's host
Worms	153	Attacks that replicate themselves and spread to other computers

Table 4. NF-BoT-IoT distribution

Class	Count	Description
Benign	13859	Normal unmalicious flows
Reconnaissance	470655	A technique for gathering information about a network host and is also known as a probe
DDoS	56844	Distributed Denial of Service is an attempt similar to DoS but has multiple different distributed sources
DoS	56833	An attempt to overload a computer system's resources with the aim of preventing access to or availability of its data
Theft	1909	A group of attacks that aims to obtain sensitive data such as data theft and keylogging

have been renamed to the parent DoS category. Attacks named DDOS attack-LOIC-UDP, DDOS attack-HOIC and DDoS attacks-LOIC-HTTP have been renamed to DDoS. Attacks named FTP-BruteForce, SSH-Bruteforce, Brute Force -Web and Brute Force -XSS have been combined as a brute-force category. Finally, SQL Injection attacks have been included in the injection attacks category. The NF-UQ-NIDS dataset has a total of 11,994,893 records, out of which 9,208,048 (76.77%) are benign flows and 2,786,845 (23.23%) are attacks. Table 7 lists the distribution of the final attack categories.

Table 5. NF-ToN-IoT distribution

Class	Count	Description
Benign	270279	Normal unmalicious flows
Backdoor	17247	A technique that aims to attack remote-access computers by replying to specific constructed client applications
DoS	17717	An attempt to overload a computer system's resources with the aim of preventing access to or availability of its data
DDoS	326345	An attempt similar to DoS but has multiple different distributed sources
Injection	468539	A variety of attacks that supply untrusted inputs that aim to alter the course of execution, with SQL and Code injections two of the main ones
MITM	1295	Man In The Middle is a method that places an attacker between a victim and host with which the victim is trying to communicate, with the aim of intercepting traffic and communications
Password	156299	covers a variety of attacks aimed at retrieving passwords by either brute force or sniffing
Ransomware	142	An attack that encrypts the files stored on a host and asks for compensation in exchange for the decryption technique/key
Scanning	21467	A group that consists of a variety of techniques that aim to discover information about networks and hosts, and is also known as probing
XSS	99944	Cross-site Scripting is a type of injection in which an attacker uses web applications to send malicious scripts to end-users

Table 6. NF-CSE-CIC-IDS2018 distribution

Class	Count	Description
Benign	7373198	Normal unmalicious flows
BruteForce	287597	A technique that aims to obtain usernames and password credentials by accessing a list of predefined possibilities
Bot	15683	An attack that enables an attacker to remotely control several hijacked computers to perform malicious activities
DoS	269361	An attempt to overload a computer system's resources with the aim of preventing access to or availability of its data
DDoS	380096	An attempt similar to DoS but has multiple different distributed sources
Infiltration	62072	An inside attack that sends a malicious file via an email to exploit an application and is followed by a backdoor that scans the network for other vulnerabilities
Web Attacks	4394	A group that includes SQL injections, command injections and unrestricted file uploads

4 Evaluation

For the evaluation of the newly published NetFlow datasets, we use an ML classifier and compared the classifier performances with the corresponding measures on the original datasets. We drop the flow identifiers such as IDs, source/destination IP and ports, timestamps and start/end time to avoid bias towards attacking or victim nodes. For UNSW-NB15, we additionally drop Time To Live (TTL) based features i.e., `sttl`, `dttl` and `ct_state_ttl`, due to their extreme correlation with the labels. Furthermore, we utilise the min-max normalisation technique to scale all datasets' values between 0 to 1. Finally, we apply an *Extra Trees ensemble* classifier, made up of 50 randomised decision trees estimators. The chosen classifier belongs to the 'trees' family and has proven to achieve reliable performances on NIDS datasets. Due to the extreme imbalance in all datasets' binary-class and multi-class labels, we set a custom class weight parameter, using Eq. 1.

$$Class\ Weight = \frac{Total\ Samples\ Count}{Number\ Of\ Classes \times Class\ Samples\ Count} \tag{1}$$

To reliably evaluate the datasets, we conduct five cross-validation splits and collect the average metrics such as accuracy, *Area Under the Curve (AUC)*, *F1 Score*, *Detection Rate (DR)*, *False Alarm Rate (FAR)* and time required to predict a single test sample in microseconds (µs).

Table 7. NF-UQ-NIDS distrubution

Class	Count	Class	Count
Benign	9208048	Scanning	21467
DDoS	763285	Fuzzers	19463
Reconnaissance	482946	Backdoor	19029
Injection	468575	Bot	15683
DoS	348962	Generic	5570
Brute Force	291955	Analysis	1995
Password	156299	Shellcode	1365
XSS	99944	MITM	1295
Infilteration	62072	Worms	153
Exploits	24736	Ransomware	142

4.1 Binary-Class Classification

In this experiment, we evaluate the attack detection performance of the NetFlow datasets compared to the original datasets. Table 8 lists the accuracy, AUC, F1 score, DR, FAR and prediction time results for both, the original and the NetFlow versions. The NF-UNSW-NB15 dataset achieved slightly lower performance than the UNSW-NB15 dataset, with almost the same DR but higher FAR, however, it used less time to predict the samples. The overall accuracy achieved by the NF-UNSW-NB15 dataset is 98.62% compared to 99.25% when using the UNSW-NB15 dataset. The NF-BoT-IoT dataset has achieved slightly lower classification performance, i.e. 93.70% DR and 0.97 F1 Score, compared to its parent BoT-IoT dataset which achieved a 100% DR and 1.00 F1 Score. The almost perfect results achieved by BoT-IoT has been deemed unreliable in a recent study [16], due to its extreme class imbalance of attack and benign samples which is unrealistic in a real-world network.

The NF-ToN-IoT dataset's performance was superior to its original ToN-IoT dataset, achieving a 99.67% DR and 0.37% FAR, it also consumed less prediction time. The accuracy achieved is 99.66% proving its significance compared to the ToN-IoT dataset, 97.86%. The NF-CSE-CIC-IDS2018 dataset performance was less efficient than the CSE-CIC-IDS2018 dataset achieving a similar DR of 94.71% but a higher FAR of 4.59%, however significantly less time was consumed in prediction. The overall accuracy achieved is 95.33%, significantly lowering the 98.31% accuracy of the CSE-CIC-IDS2018 dataset. The merged NF-UQ-NIDS dataset achieved an accuracy of 97.25%, a DR of 95.66% and a FAR of 2.27%, achieving a reliable classification performance of 20 different attack categories. Figure 3 shows the AUC achieved using the *Extra Trees* classifier on the four newly published NetFlow-based datasets. This comparison is conducted by using the same set of features across all datasets.

This fair comparison demonstrates the benefit of the newly published datasets, which was not possible to achieve due to each dataset's unique set of features. Overall, the NetFlow datasets containing only eight features used in the classification experiments achieved a very similar attack detection performance compared to the original 36 features of the BoT-IoT, 38 features of both the UNSW-NB15 and ToN-IoT datasets and the 77 features of the CSE-CIC-IDS2018 dataset. We noticed a consistent prediction time decrease in using all the NetFlow datasets. Therefore, in terms of feasibility and practicality in real-world networks, using NetFlow features might lead to an overall superior performance if additional metrics are measured such as storage and computation power required to extract and store the utilised features.

Table 8. Binary-class classification results

Dataset	Accuracy	AUC	F1 score	DR	FAR	Prediction time (µs)
UNSW-NB15	99.25%	0.9545	0.92	91.25%	0.35%	10.05
NF-UNSW-NB15	**98.62%**	**0.9485**	**0.85**	**90.70%**	**1.01%**	**7.79**
BoT-IoT	100.00%	0.9948	1.00	100.00%	1.05%	7.62
NF-BoT-IoT	**93.82%**	**0.9628**	**0.97**	**93.70%**	**1.13%**	**5.37**
ToN-IoT	97.86%	0.9788	0.99	97.86%	2.10%	8.93
NF-ToN-IoT	**99.66%**	**0.9965**	**1.00**	**99.67%**	**0.37%**	**6.05**
CSE-CIC-IDS2018	98.31%	0.9684	0.94	94.75%	1.07%	23.01
NF-CSE-CIC-IDS2018	**95.33%**	**0.9506**	**0.83**	**94.71%**	**4.59%**	**17.04**
NF-UQ-NIDS	**97.25%**	**0.9669**	**0.94**	**95.66%**	**2.27%**	**14.35**

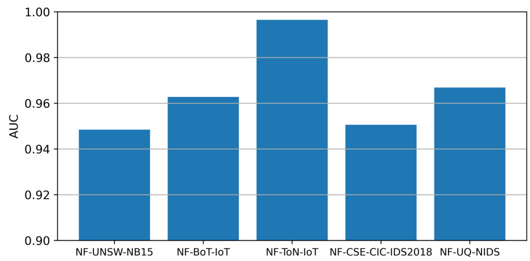

Fig. 3. Binary-class classification's AUC

Table 9. NF-UNSW-NB15 multi-class classification results

Class name	UNSW-NB15		NF-UNSW-NB15	
	DR	F1 score	DR	F1 score
Benign	99.72%	1.00	99.02%	0.99
Analysis	4.39%	0.03	28.28%	0.15
Backdoor	13.96%	0.08	39.17%	0.17
DoS	13.63%	0.18	31.84%	0.41
Exploits	83.25%	0.80	81.04%	0.82
Fuzzers	50.50%	0.57	62.63%	0.55
Generic	86.08%	0.91	57.13%	0.66
Reconnaissance	75.90%	0.80	76.89%	0.82
Shellcode	53.61%	0.59	87.91%	0.75
Worms	5.26%	0.09	52.91%	0.55
Weighted average	**98.19%**	**0.98**	**97.62%**	**0.98**
Prediction time (µs)	**9.94**		**9.35**	

4.2 Multi-class Classification

In this experiment, we measure the DR and F1 score of each attack's type present in each dataset. Tables 9, 10, 11, 12 and 13 list the DR and F1 score of each attack type for the NF-UNSW-NB15, NF-BoT-IoT, NF-ToN-IoT, NF-CSE-CIC-IDS2018 and NF-UQ-NIDS datasets respectively. The average accuracy and prediction time are calculated and the results are compared to their respective original datasets. In Table 9, we can conclude that by using the NF-UNSW-NB15 dataset, we can increase the DR of analysis, backdoor, DoS, fuzzers, shellcode and worms attacks, however, it was inefficient against generic attacks. The overall accuracy achieved which is 97.62% is slightly lower than the UNSW-NB15 dataset, 98.19%, due to the number of miss-correctly classified samples, however, the prediction time consumed was slightly lower.

Table 10 shows that the BoT-IoT dataset is achieving almost perfect multi-classification performances of a 100% accuracy and 1 F1 Score. Again, these results might be unreliable due to the extreme imbalance mentioned in [16]. In addition, there might be certain 'hidden label' features, such as the TTL-based features in the UNSW-NB15 dataset, that are extremely correlated to the attack types present in the dataset. The NF-BoT-IoT dataset was unreliable in the detection of the DDoS and DoS attacks. However, it achieved a 90% DR against reconnaissance and theft attacks. Although it achieved a lower DR of 73.58% and F1 Score of 0.77, the NetFlow dataset maintained the lower prediction time compared to the BoT-IoT dataset.

In Table 11, the NF-ToN-IoT dataset increased the DR of DoS attacks but lowered the DDoS, injection, MITM, password, scanning and XSS attacks compared to the ToN-IoT dataset. Further analysis is required to identify which fea-

Table 10. NF-BoT-IoT multi-class classification results

Class name	BoT-IoT		NF-BoT-IoT	
	DR	F1 score	DR	F1 score
Benign	99.58%	0.99	98.65%	0.43
DDoS	100.00%	1.00	30.37%	0.28
DoS	100.00%	1.00	36.33%	0.31
Reconnaissance	100.00%	1.00	89.95%	0.90
Theft	91.16%	95.37	88.06%	0.18
Weighted average	**100.00%**	**1.00**	**73.58%**	**0.77**
Prediction time (µs)	**12.63**		**9.19**	

tures of the original dataset were critical in the detection of the missed attacks and to be added to the NetFlow dataset. Overall, in multi-class classification, the NF-ToN-IoT dataset was not as effective in terms of overall accuracy and prediction time compared to the ToN-IoT dataset. It achieved a low prediction accuracy of 56.34% and a high prediction time of 21.21 µs. However, a binary-class classification deemed it was very efficient, therefore, it seems like the ML classifier is detecting the overall pattern of attacks present in the dataset, but not the pattern of individual attacks. We suspect that specific features present in the original dataset contain payload information that was enabling the ML classifier to detect certain attack types. Further analysis is required to investigate which features from the ToN-IoT dataset are necessary to identify each attack's type.

In Table 12, the performance of the NF-CSE-CIC-IDS2018 dataset can prove that attacks such as FTP-bruteforce and infiltration were better detected using the NetFlow features compared to the CSE-CIC-IDS2018 features. However, Brute Force -Web, Brute Force -XSS, DDOS attack-HOIC and SQL injection attack samples were mostly undetected by using the NetFlow features. The DoS attacks-SlowHTTPTest attack samples were fully undetected by the ML classifier. Similar to the NF-ToN-IoT dataset, the ML classifier was unable to efficiently detect the pattern of certain attack types. Overall, the accuracy and prediction time achieved while using the NF-CSE-CIC-IDS2018 dataset being 71.92% and 17.29 µs respectively were lower compared to the CSE-CIC-IDS2018 dataset.

Table 13 displays the full attack identification results of the merged dataset named NF-UQ-NIDS. The chosen ML classifier was efficient in the detection of certain attack's types such as backdoor, bot, bruteforce, exploits, shellcode, DDoS and ransomware. However, attacks such as analysis, DoS, fuzzers, generic, infiltration, worms, injection, MITM, password, scanning and XSS were not reliably detected. Further analysis is required to identify the features that are critical in identifying these attacks and to add them to the NetFlow features. The overall accuracy of 70.81% and prediction time 14.74 (µs) were achieved.

Table 11. NF-ToN-IoT multi-class classification results

Class name	ToN-IoT		NF-ToN-IoT	
	DR	F1 score	DR	F1 score
Benign	89.97%	0.94	98.97%	0.99
Backdoor	98.05%	0.31	99.22%	0.98
DDoS	96.90%	0.98	63.22%	0.72
DoS	53.89%	0.57	95.91%	0.48
Injection	96.67%	0.96	41.47%	0.51
MITM	66.25%	0.16	52.81%	0.38
Password	86.99%	0.92	27.36%	0.24
Ransomware	89.87%	0.11	87.33%	0.83
Scanning	75.05%	0.85	31.30%	0.08
XSS	98.83%	0.99	24.49%	0.19
Weighted average	**84.61%**	**0.87**	**56.34%**	**0.60**
Prediction time (µs)	**12.02**		**21.21**	

Table 12. NF-CSE-CIC-IDS2018 multi-class classification results

Class Name	CSE-CIC-IDS2018		NF-CSE-CIC-IDS2018	
	DR	F1 score	DR	F1 score
Benign	89.50%	0.94	69.83%	0.82
Bot	99.92%	0.99	100.00%	1.00
Brute Force -Web	71.36%	0.01	50.21%	0.52
Brute Force -XSS	72.17%	0.72	49.16%	0.39
DDOS attack-HOIC	100.00%	1.00	45.66%	0.39
DDOS attack-LOIC-UDP	83.59%	0.82	80.98%	0.82
DDoS attacks-LOIC-HTTP	99.93%	1.00	99.93%	0.71
DoS attacks-GoldenEye	99.97%	1.00	99.32%	0.98
DoS attacks-Hulk	100.00%	1.00	99.65%	0.99
DoS attacks-SlowHTTPTest	69.80%	0.60	0.00%	0.00
DoS attacks-Slowloris	99.44%	0.62	99.95%	1.00
FTP-BruteForce	68.76%	0.75	100.00%	0.79
Infilteration	36.15%	0.08	62.66%	0.04
SQL Injection	49.34%	0.30	25.00%	0.22
SSH-Bruteforce	99.99%	1.00	99.93%	1.00
Weighted average	**90.28%**	**0.94**	**71.92%**	**0.80**
Prediction time (µs)	**24.17**		**17.29**	

Table 13. NF-UQ-NIDS multi-class classification results

Class name	NF-UQ-NIDS	
	Detection rate	F1 score
Analysis	69.63%	0.21
Backdoor	90.95%	0.92
Benign	71.70%	0.83
Bot	100.00%	1.00
Brute Force	99.94%	0.85
DoS	55.54%	0.62
Exploits	80.65%	0.81
Fuzzers	63.24%	0.54
Generic	58.90%	0.61
Infilteration	60.57%	0.03
Reconnaissance	88.96%	0.88
Shellcode	83.89%	0.15
Theft	87.22%	0.15
Worms	52.97%	0.46
DDoS	77.08%	0.69
Injection	40.58%	0.50
MITM	57.99%	0.10
Password	30.79%	0.27
Ransomware	90.85%	0.85
Scanning	39.67%	0.08
XSS	30.80%	0.21
Weighted average	**70.81%**	**0.79**
Prediction time (µs)	**14.74**	

5 Conclusion

This paper provides the research community with five new NIDS datasets based on NetFlow features as shown in Table 2. These datasets can be used in ML-based NIDS training and evaluation stages. The datasets are showing positive results by achieving similar binary-class detection performance compared to the complete set of their corresponding original datasets. Though, in the case of multi-class detection experiments, the NF-BoT-IoT, NF-ToN-IoT and NF-CSE-CIC-IDS2018 datasets were not similarly efficient. Further feature analysis is required to identify the strength of each NetFlow feature, and how these datasets can be improved by adding key features from the original datasets to aid in the detection of missed attack types.

These published NetFlow datasets offer a promising performance, and serve three advantages; 1. the level of complexity and resources required to collect and store NetFlow features are lower, 2. proposed ML models can be evaluated using the same set of features across various attack types, and 3. datasets can be merged to generate a more comprehensive data source including collected over various network environments. Overall, the practicality and initial performance of NetFlow features' collection and attack detection, requires increased attention and interest by researchers in applying them into the real-world models for ML-based NIDS. Future works include enhancing the current datasets with additional NetFlow features which can potentially improve both the binary and multi-class classification performances. Finally, key features from the original datasets required to detect certain attack types must be identified to be included in NetFlow features.

References

1. "Netflow datasets." (2020). http://staff.itee.uq.edu.au/marius/NIDS_datasets/
2. Garcia-Teodoro, P., Diaz-Verdejo, J., Maciá-Fernández, G., Vázquez, E.: Anomaly-based network intrusion detection: techniques, systems and challenges. Comput. Secur. **28**(1–2), 18–28 (2009)
3. Sahu, S.K., Sarangi, S., Jena, S.K.: A detail analysis on intrusion detection datasets. In: 2014 IEEE International Advance Computing Conference (IACC), pp. 1348–1353 (2014)
4. 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)
5. Binbusayyis, A., Vaiyapuri, T.: Identifying and benchmarking key features for cyber intrusion detection: an ensemble approach. IEEE Access **7**, 106495–106513 (2019)
6. Claise, B., Sadasivan, G., Valluri, V., Djernaes, M.: Cisco systems netflow services export version 9 (2004)
7. Ring, M., Wunderlich, S., Scheuring, D., Landes, D., Hotho, A.: A survey of network-based intrusion detection data sets. Comput. Secur. **86**, 147–167 (2019)
8. Moustafa, N., Slay, J.: 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) (2015)
9. Koroniotis, N., Moustafa, N., Sitnikova, E.,Turnbull, B.: Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: BoT-IoT dataset, CoRR, vol. abs/1811.00701 (2018)
10. Moustafa, N.: ToN-IoT datasets (2019)
11. Sharafaldin, I., Habibi Lashkari, A., Ghorbani, A.A.: Toward generating a new intrusion detection dataset and intrusion traffic characterization. In: Proceedings of the 4th International Conference on Information Systems Security and Privacy (2018)
12. Li, B., Springer, J., Bebis, G., Hadi Gunes, M.: A survey of network flow applications. J. Netw. Computer Appl. **36**(2), 567–581 (2013)
13. Kerr, D.R., Bruins, B.L.: Network Flow Switching and Flow Data Export (2001)

14. Cisco Systems: Cisco IOS NetFlow Version 9 Flow-Record Format - White Paper. https://www.cisco.com/en/US/technologies/tk648/tk362/technologies_white_paper09186a00800a3db9.pdf (2011)
15. Ntop: nProbe, An Extensible NetFlow v5/v9/IPFIX Probe for IPv4/v6. https://www.ntop.org/guides/nprobe/cli_options.html (2017)
16. Al-Othman, Z., Alkasassbeh, M., Baddar, S. A.-H.: A state-of-the-art review on IoT botnet attack detection (2020)