Towards a Standard Feature Set of NIDS Datasets

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

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

Abstract. Network Intrusion Detection Systems (NIDSs) datasets are essential tools used by researchers for the training and evaluation of Machine Learning (ML)-based NIDS models. There are currently five datasets, known as NF-UNSW-NB15, NF-BoT-IoT, NF-ToN-IoT, NF-CSE-CIC-IDS2018 and NF-UQ-NIDS, which are made up of a common feature set. However, their performances in classifying network traffic, mainly using the multi-classification method, is often unreliable. Therefore, this paper proposes a standard NetFlow feature set, to be used in future NIDS datasets due to the tremendous benefits of having a common feature set. NetFlow has been widely utilised in the networking industry for its practical scaling properties. The evaluation is done by extracting and labelling the proposed features from four well-known datasets. The newly generated datasets are known as NF-UNSW-NB15-v2, NF-BoT-IoT-v2, NF-ToN-IoT-v2, NF-CSE-CIC-IDS2018-v2 and NF-UQ-NIDS-v2. Their performances have been compared to their respective original datasets using an Extra Trees classifier, showing a great improvement in the attack detection accuracy. They have been made publicly available to use for research purposes [1].

Index terms— Machine Learning, NetFlow, Network Intrusion Detection System

1 Introduction

Network Intrusion Detection Systems (NIDSs) aim to detect network attacks as they are passing through their targeted networks. Signature-based NIDSs match attack signatures to incoming threats, giving a high detection accuracy on precedented attacks, however, failing to detect zero-day or new variant of attacks. Therefore, researchers have investigated anomaly-based NIDSs that focus on matching attack behaviours and patterns [2]. Machine Learning (ML), an artificial intelligence tool, is capable of learning and extracting complex network attack patterns that may threaten computer networks if undetected [3]. All network intrusions have a unique set of events that would aid in their identification process. These patterns, also known as attack vectors, are extracted from network traffic transmitted through packets in the form of features. These features form network data flows which are ideally labelled as benign or one of the various attacks categories to follow a supervised ML methodology.

Real-world network data flows are challenging to obtain, mainly due to security and privacy concerns. Also, labelling production network flows is a rigorous process due to the uncontrolled environment. Therefore, researchers have designed network test-beds to generate synthetic datasets that consist of labelled network data flows [4]. The data flows are made of up several network features. The network features that make up the data flows are often pre-selected based on the authors' domain knowledge and extraction tools available. As a result, these datasets are very different in terms of their feature sets and therefore the information represented. Due to the great impact of data features on the performance of learning models [5], the evaluation of proposed ML-based NIDSs are often unreliable when tested on multiple datasets using their original feature sets. Finally, as certain features require a complex and advanced level of extraction, the feasibility of scalable deployments is questionable.

To address this gap, [6] published four new datasets, known as NF-UNSW-NB15, NF-BoT-IoT, NF-ToN-IoT and CIC-CSE-IDS2018 and an additional one, known as NF-UQ-NIDS, which is generated by merging the aforementioned datasets. These datasets are generated by converting four well-known modern NIDS datasets into a common NetFlow format. NetFlow is an industry-standard protocol of network traffic collection [7], and its practical and scalable properties will also enhance the deployments of the models. However, due to the insufficient information represented by the extracted features, these datasets lead to unreliable detection accuracy, in particular when performing multi-class experiments. Therefore, this paper proposes an extended NetFlow feature set to be used in future NIDS datasets. The importance of having a standard feature set

shared by all datasets is paramount. It will facilitate a reliable evaluation of proposed learning models across various network settings and attack scenarios. As part of the proposed feature set evaluation, four well-known datasets are utilised by extracting and labelling the proposed NetFlow features. The datasets have been named NF-UNSW-NB15-v2, NF-BoT-IoT-v2, NF-ToN-IoT-v2, NF-CSE-CIC-IDS2018-v2 and NF-UQ-NIDS-v2, and made publicly available for research purposes.

The paper is organised by stating the drawbacks of the existing datasets and how they can be overcome in Section 2. Section 3 explains the importance of having a standard and a common feature set to be used in future datasets. It illustrates the methodology of extracting the proposed feature set from network packets. It also explains and lists the samples' distribution of the published datasets. Finally, in Section 4, the generated datasets detection performances are compared to their respective original feature sets using an Extra Trees ensemble classifier. This paper provides the research community with a standard NetFlow feature set that should be used in future NIDS datasets. Five datasets have been generated, with the proposed feature set, using four existing datasets. A preliminary set of results have been collected while conducting binary- and multi-class classification experiments.

2 Limitations of Existing Datasets

Researchers have created engineered benchmark NIDS datasets because of the complexity in obtaining labelled realistic network traffic. A network testbed is designed to simulate network behaviour of multiple end nodes. The artificial network environment overcomes the security and privacy issues faced by real-world networks. Besides, labelling the network flows generated by such controlled environments is more reliable than the open-world nature of realistic networks. During the experiments, benign network traffic and various attack scenarios are generated and conducted over the network testbed. In the meanwhile, the network packets are captured in their native packet capture (pcap) format and dumped onto storage devices. A set of features are extracted from the pcap files using appropriate tools and methods, forming network data flows. The result is a data source of labelled network flows reflecting benign and malicious network behaviour. The generated datasets are published and made publicly accessible for use in the design and evaluation phases of ML-based NIDS models [8].

The network traffic features that form these data flows are critical as they need to represent an adequate amount of security events that would aid in the model's classification between the benign and attack classes. They also need to be feasible in count and extraction's complexity for scalable and practical deployments. A key task of designing an ML-based NIDS is the selection of the utilised data features. However, due to the lack of a standard feature set in generating NIDS datasets, the authors have applied their domain experience in the selection of these features. As a result, each available dataset is made up of its own unique set of features, that their authors believe would lead to the best possible results in the classification stages. Each of the current feature sets is almost exclusive and completely different from other sets, sharing only a small number of features. The current evaluation methods of ML models across multiple datasets requires the usage of the unique feature sets presented by each dataset.

The differences in the information represented by each dataset's feature set have caused limitations and concerns regarding the reliability of the evaluation methods followed. The three main issues of not having a benchmark feature set are 1. unfeasible deployments due to collection and storage of various features, some of which are irrelevant due to the lack of security events and 2. inability to evaluate an ML model's generalisation using a targeted feature set across multiple datasets and 3. lack of a universal merged dataset containing network data flows from multiple sources. It is believed that 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 [9]. The rest of this section discusses four of the most recent and widely-used 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 [10]. 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.

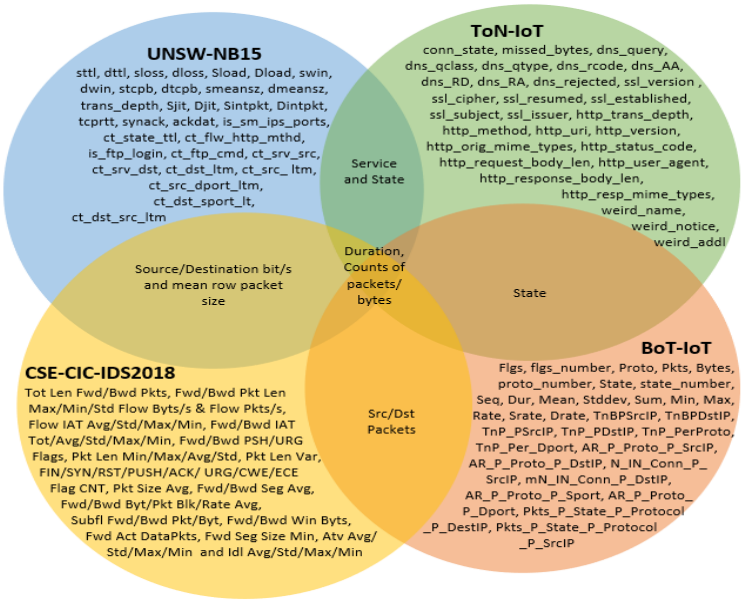

Fig. 1: Venn diagram of the shared and exclusive features of four NIDS datasets

- **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 [11]. 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 [12] that includes telemetry data of Internet of Things (IoT) services, network traffic of IoT networks and operating system logs. In this paper, the portion containing network traffic flows is utilised. 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 [13]. 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.

In Figure 1, the shared and unique features of the aforementioned datasets are displayed. The set of features available in all four datasets contains 3 features, and the pairwise shared features numbers vary from 1 to 5. As most of the features are exclusive to individual datasets, the evaluation of proposed ML models using a targeted feature set across the four datasets is challenging. Moreover, the ratio of the classes i.e. benign and attack flows is extremely varied in each dataset. Where the UNSW-NB15 and CSE-CIC-IDS2018 datasets have very high benign to attack ratios, whereas the ToN-IoT and BoT-IoT datasets are mainly made up of attack samples, which do not represent a realistic network behaviour. Also, some of the features in the UNSW-NB15, BoT-IoT and CSE-CIC-IDS2018 datasets are handcrafted features that are not originally found in network packets but are statistically calculated based on other features, such as the total number of bytes transferred over the last 100 seconds. All these differences in the information presented by datasets have led to the design of a standard feature set for NIDS datasets.

3 Benchmarking a Standard Feature Set

Due to the aforementioned limitations faced by the different feature sets, in this paper, a standard feature set is proposed to be used across all NIDS datasets. The feature set will be evaluated and benchmarked to be used in the releases of new NIDS datasets to efficiently design future NIDS. The design of ML-based NIDS requires a proposed feature set that will be implemented in the final design, the choice of these features significantly alters the performance of the learning models. By having a standard feature set, researchers can evaluate the model's classification ability based on their chosen features, across multiple datasets and hence different attack scenarios conducted over several network environments to make sure their measured model performance well generalises. Moreover, by having datasets sharing a common ground feature set, they can be merged creating a universal comprehensive source of data. This method can be used in federated learning techniques where a single model is trained across multiple data sources [14]. Finally, having a standard feature set will grant control over the information presented by datasets and therefore, ensuring they are the same and modifying the standard set as required. We believe that a standard feature set will narrow the gap between the research experiments and practical deployments of ML-based NIDS [9].

3.1 NetFlow

The collection and storage of network traffic are important for organisations to monitor, analyse and audit their network environments. However, network traffic tends to overload in volume and therefore are aggregated in terms of flows. A network flow is a sequence of packets, in either uni- or bi-direction, between two unique endpoints sharing some attributes such as source/destination IP address and L4 (transport layer) ports, and the L4 protocol, also known as the five-tuple [6]. A network flow can also be enhanced with additional features, each representing details on the respective traffic. The information provided by these features can contain security events that are essential in analysing network traffic in case of a threat [15]. 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 [16]. NetFlow evolved over the years, where version 9 is the most common due to its larger variety of features and bidirectional flow support [17]. Most of the network devices such as routers and switches are capable of extracting NetFlow records. This is a great motivation for standardising NetFlow features for NIDS datasets, as the level of complexity and resources required to collect and store them is lower.

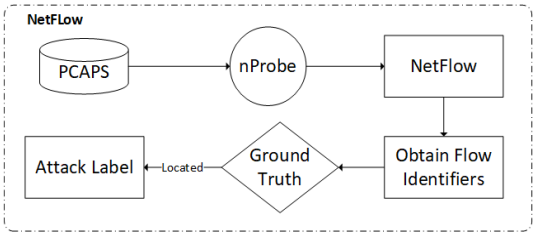

Fig. 2: NetFlow datasets' extraction and labelling procedure

NetFlow v9 features have been utilised to form the proposed feature set, listed and described in Table 1. There are 43 features in total with some providing information on general flow statistics and others on specific protocol applications. All the features are flow-based, meaning they are extracted from packet headers and does not depend on the payload information which is often encrypted due to privacy concerns. The features are numerical in type for efficient ML experiments. These features might contain useful security events to enhance the models' intrusions detection capabilities. Having a standard feature set 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.

3.2 Datasets

Figure 2 shows the procedure of creating the NetFlow datasets by extracting flows from the pcap files and labelling them based on the grand truths provided. The nProbe tool by Ntop [18] was utilised to convert the

Table 1: List of the proposed standard NetFlow features

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
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
TCP_FLAGS	Cumulative of all TCP flags
CLIENT_TCP_FLAGS	Cumulative of all client TCP flags
SERVER_TCP_FLAGS	Cumulative of all server TCP flags
DURATION_IN	Client to Server stream duration (msec)
DURATION_OUT	Client to Server stream duration (msec)
MIN_TTL	Min flow TTL
MAX_TTL	Max flow TTL
LONGEST_FLOW_PKT	Longest packet (bytes) of the flow
SHORTEST_FLOW_PKT	Shortest packet (bytes) of the flow
MIN_IP_PKT_LEN	Len of the smallest flow IP packet observed
MAX_IP_PKT_LEN	Len of the largest flow IP packet observed
SRC_TO_DST_SECOND_BYTES	Src to dst Bytes/sec
DST_TO_SRC_SECOND_BYTES	Dst to src Bytes/sec
RETRANSMITTED_IN_BYTES	Number of retransmitted TCP flow bytes (src->dst)
RETRANSMITTED_IN_PKTS	Number of retransmitted TCP flow packets (src->dst)
RETRANSMITTED_OUT_BYTES	Number of retransmitted TCP flow bytes (dst->src)
RETRANSMITTED_OUT_PKTS	Number of retransmitted TCP flow packets (dst->src)
SRC_TO_DST_AVG_THROUGHPUT	Src to dst average thpt (bps)
DST_TO_SRC_AVG_THROUGHPUT	Dst to src average thpt (bps)
NUM_PKTS_UP_TO_128_BYTES	Packets whose IP size <= 128
NUM_PKTS_128_TO_256_BYTES	Packets whose IP size > 128 and <= 256
NUM_PKTS_256_TO_512_BYTES	Packets whose IP size > 256 and <= 512
NUM_PKTS_512_TO_1024_BYTES	Packets whose IP size > 512 and <= 1024
NUM_PKTS_1024_TO_1514_BYTES	Packets whose IP size > 1024 and <= 1514
TCP_WIN_MAX_IN	Max TCP Window (src->dst)
TCP_WIN_MAX_OUT	Max TCP Window (dst->src)
ICMP_TYPE	ICMP Type * 256 + ICMP code
ICMP_IPV4_TYPE	ICMP Type
DNS_QUERY_ID	DNS query transaction Id
DNS_QUERY_TYPE	DNS query type (e.g. 1=A, 2=NS..)
DNS_TTL_ANSWER	TTL of the first A record (if any)
FTP_COMMAND_RET_CODE	FTP client command return code

pcaps into NetFlow version 9 format and selecting 43 features to be extracted. The dump format is chosen as text flows, in which each feature is separated by a comma (,) to be utilised as CSV files. In the last step, two label features are created 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 data 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 sample is labelled as a benign flow, class 0.

In this paper, the proposed feature set has been extracted using four well-known datasets; UNSW-NB15, BoT-IoT, ToN-IoT and CSE-CIC-IDS2018. Their publicly available pcap files and ground truth events have been utilised in the features extraction and labelling processes respectively. The generated datasets have been named NF-UNSW-NB15-v2, NF-BoT-IoT-v2, NF-ToN-IoT-v2, NF-CSE-CIC-IDS2018-v2 and NF-UQ-NIDS-v2. The later dataset is a merge of all the other datasets, which is a practical advantage of having a common feature set. Table 2 lists the NetFlow datasets and compares their properties to the original datasets in terms of the Feature Extraction (FE) tool utilised, the number of features, files size and the benign to attack samples ratio. As illustrated, two NetFlow datasets are corresponding to each original NIDS dataset, where v2 is the expanded version. The fifth NetFlow dataset is the comprehensive dataset that combines all the four.

- **NF-UNSW-NB15-v2** The NetFlow-based format of the UNSW-NB15 dataset, named NF-UNSW-NB15, has been expanded with additional NetFlow features and labelled with its respective attack categories. The total number of data flows are 2,390,275 out of which 95,053 (3.98%) are attack samples and 2,295,222 (96.02%) are benign. The attack samples are further classified into nine subcategories, Table 3 represents the NF-UNSW-NB15-v2 dataset's distribution of all flows.

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
NF-UNSW-NB15-v2	**2021**	**nProbe**	**43**	**0.41**	**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
NF-BoT-IoT-v2	**2021**	**nProbe**	**43**	**5.60**	**0.0 to 10.0**
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
NF-ToN-IoT-v2	**2021**	**nProbe**	**43**	**2.47**	**3.6 to 6.4**
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-CSE-CIC-IDS2018-v2	**2021**	**nProbe**	**43**	**2.80**	**8.8 to 1.2**
NF-UQ-NIDS	2020	nProbe	12	1.0	7.7 to 2.3
NF-UQ-NIDS-v2	**2021**	**nProbe**	**43**	**12.5**	**3.3 to 6.7**

Table 3: NF-UNSW-NB15-v2 distribution

Class	Count	Description
Benign	2295222	Normal unmalicious flows
Fuzzers	22310	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	2299	A group that presents a variety of threats that target web applications through ports, emails and scripts.
Backdoor	2169	A technique that aims to bypass security mechanisms by replying to specific constructed client applications.
DoS	5794	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	31551	Are sequences of commands controlling the behaviour of a host through a known vulnerability.
Generic	16560	A method that targets cryptography and causes a collision with each block-cipher.
Reconnaissance	12779	A technique for gathering information about a network host and is also known as a probe.
Shellcode	1427	A malware that penetrates a code to control a victim's host.
Worms	164	Attacks that replicate themselves and spread to other computers.

- **NF-BoT-IoT-v2** An IoT NetFlow-based dataset generated by expanding the NF-BoT-IoT dataset. 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 37,763,497 out of which 37,628,460 (99.64%) are attack samples and 135,037 (0.36%) are benign. There are four attack categories in the dataset, Table 4 represents the NF-BoT-IoT-v2 distribution of all flows.

Table 4: NF-BoT-IoT-v2 distribution

Class	Count	Description
Benign	135037	Normal unmalicious flows
Reconnaissance	2620999	A technique for gathering information about a network host and is also known as a probe.
DDoS	18331847	Distributed Denial of Service is an attempt similar to DoS but has multiple different distributed sources.
DoS	16673183	An attempt to overload a computer system's resources with the aim of preventing access to or availability of its data.
Theft	2431	A group of attacks that aims to obtain sensitive data such as data theft and keylogging

- **NF-ToN-IoT-v2** The publicly available pcaps of the ToN-IoT dataset are utilised to generate its NetFlow records, leading to a NetFlow-based IoT network dataset called NF-ToN-IoT. The total number of data flows are 16,940,496 out of which 10,841,027 (63.99%) are attack samples and 6,099,469 (36.01%), Table 5 lists and defines the distribution of the NF-ToN-IoT dataset.

Table 5: NF-ToN-IoT-v2 distribution

Class	Count	Description
Benign	6099469	Normal unmalicious flows
Backdoor	16809	A technique that aims to attack remote-access computers by replying to specific constructed client applications
DoS	712609	An attempt to overload a computer system's resources with the aim of preventing access to or availability of its data.
DDoS	2026234	An attempt similar to DoS but has multiple different distributed sources.
Injection	684465	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	7723	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	1153323	covers a variety of attacks aimed at retrieving passwords by either brute force or sniffing.
Ransomware	3425	An attack that encrypts the files stored on a host and asks for compensation in exchange for the decryption technique/key.
Scanning	3781419	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	2455020	Cross-site Scripting is a type of injection in which an attacker uses web applications to send malicious scripts to end-users.

- **NF-CSE-CIC-IDS2018-v2** The original pcap files of the CSE-CIC-IDS2018 dataset are utilised to generate a NetFlow-based dataset called NF-CSE-CIC-IDS2018-v2. The total number of flows are 18,893,708 out of which 2,258,141 (11.95%) are attack samples and 16,635,567 (88.05%) are benign ones, Table 6 represents the dataset's distribution.

Table 6: NF-CSE-CIC-IDS2018-v2 distribution

Class	Count	Description
Benign	16635567	Normal unmalicious flows
BruteForce	120912	A technique that aims to obtain usernames and password credentials by accessing a list of predefined possibilities
Bot	143097	An attack that enables an attacker to remotely control several hijacked computers to perform malicious activities.
DoS	483999	An attempt to overload a computer system's resources with the aim of preventing access to or availability of its data.
DDoS	1390270	An attempt similar to DoS but has multiple different distributed sources.
Infiltration	116361	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	3502	A group that includes SQL injections, command injections and unrestricted file uploads

- **NF-UQ-NIDS-v2** 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 a 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 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 75,987,976 records, out of which 25,165,295 (33.12%) are benign flows and 50,822,681 (66.88%) are attacks. Table 7 lists the distribution of the final attack categories.

4 Evaluation

In this section, the proposed NetFlow feature set is evaluated across five NIDS datasets; NF-UNSW-NB15-v2, NF-BoT-IoT-v2, NF-ToN-IoT-v2, NF-CSE-CIC-IDS2018-v2 and NF-UQ-NIDS-v2. An ensemble ML classi-

Table 7: NF-UQ-NIDS-v2 distribution

Class	Count	Class	Count
Benign	25165295	Scanning	3781419
DDoS	21748351	Fuzzers	22310
Reconnaissance	2633778	Backdoor	18978
Injection	684897	Bot	143097
DoS	17875585	Generic	16560
Brute Force	123982	Analysis	2299
Password	1153323	Shellcode	1427
XSS	2455020	MITM	7723
Infilteration	116361	Worms	164
Exploits	31551	Ransomware	3425
Theft	2431		

fier, known as Extra Trees, that belongs to the trees family has been utilised for this purpose. The evaluation is conducted by comparing the classifier performance to the corresponding metrics of the original datasets. As part of the data pre-processing, the flow identifiers such as IDs, source/destination IP and ports, timestamps and start/end time are dropped to avoid learning bias towards attacking and victim end nodes. For the UNSW-NB15 and NF-UNSW-NB15-v2 datasets, The Time To Live (TTL)-based features i.e., `sttl`, `dttl`, `ct_state_ttl`, `MIN_TTL` and `MAX_TTL` are dropped due to their extreme correlation with the labels.

Additionally, the min-max normalisation technique has been applied to scale all the datasets' values between 0 and 1. Finally, an Extra Trees classifier has been designed using 50 randomised decision trees estimators. A class weight parameter has been set using Equation 1, due to the extreme imbalance in both the binary- and multi-class labels. Carious classification metrics are collected 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). For a fair evaluation, five cross-validation splits are conducted and the mean is measured.

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

4.1 Binary-class Classification

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
NF-UNSW-NB15-v2	**99.73%**	**0.9845**	**0.97**	**97.07%**	**0.16%**	**5.92**
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
NF-BoT-IoT-v2	**100.00%**	**0.9987**	**1.00**	**100.00%**	**0.26%**	**3.90**
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
NF-ToN-IoT-v2	**99.64%**	**0.9959**	**1.00**	**99.76%**	**0.58%**	**8.47**
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-CSE-CIC-IDS2018-v2	**99.35%**	**0.9829**	**0.97**	**96.89%**	**0.31%**	**21.75**
NF-UQ-NIDS	97.25%	0.9669	0.94	95.66%	2.27%	14.35
NF-UQ-NIDS-v2	**97.90%**	**0.9830**	**0.98**	**97.12%**	**0.52%**	**14.18**

In Table 8, the binary-class attack detection performance of the datasets have been measured and compared to the original datasets using various metrics. The NF-UNSW-NB15-v2 dataset has significantly increased the performance of the NF-UNSW-NB15 and UNSW-NB15 datasets with an AUC of 0.9845 compared to 0.9485 and 0.9545 respectively. It also used the lowest amount of prediction time amongst the other datasets. The NF-BoT-IoT-v2 dataset has achieved a similar attack detection accuracy to the BoT-IoT dataset with the same DR but a lower FAR and prediction time, resulting in an AUC of 0.9987 and a prediction time of 3.90 µs compared to an AUC of 0.9948 and a prediction time of 7.62 µs of its parent

BoT-IoT dataset. NF-BoT-IoT-v2 has achieved significantly higher performance than NF-BoT-IoT with and accuracy of 100% compared to 93.82%.

The detection results of NF-ToN-IoT-v2 dataset are superior to its original ToN-IoT dataset, obtaining a 0.9965 AUC and 1.00 F1 Score. Compared to NF-ToN-IoT, it achieved a higher DR of 99.76% but a higher FAR of 0.58%. The accuracy achieved by NF-ToN-IoT-v2 is 99.64%, which is higher than ToN-IoT (97.86%) and similar to NF-ToN-IoT (99.66%). The NF-CSE-CIC-IDS2018-v2 dataset performance was much more efficient than the CSE-CIC-IDS2018 and NF-CSE-CIC-IDS2018-v2 dataset. It achieved a high DR of 96.89% and a low FAR of 0.31%, however it required 21.75 µs per sample prediction which is similar to its parent dataset. The overall accuracy achieved is 99.35%, which is higher than both the CSE-CIC-IDS2018 (98.31%) and NF-CSE-CIC-IDS2018 (95.33%) datasets. The merged NF-UQ-NIDS-v2 dataset achieved an accuracy of 97.90%, a DR of 97.12% and a FAR of 0.52%, outperforming the NF-UQ-NIDS dataset with a lower prediction time of 14.18 µs.

Figure 3 visually represents the F1 score obtained when applying an *Extra Trees* classifier on the three different versions of five NIDS datasets in three different formats; the originally published as well as basic and proposed NetFlow feature sets. This fair comparison between the NetFlow feature sets demonstrates the benefit of having a common feature set across multiple datasets. It enables efficient comparison of different attack types detection conducted over multiple network scenarios using a common feature set. Overall, version two of the NetFlow datasets containing the proposed feature set made up of 43 features, has outperformed the original and sta versions in terms of attacks detection performance. All of the datasets have significantly achieved a higher F1 score than their respective datasets except for NF-ToN-IoT-v2 achieving an equal F1 score of 1. It is clear that using version two of the NetFlow datasets made up of the proposed feature set, achieves a higher classification performance. However, further feature selection experiments are required to identify the key features in the new feature set to enhance the collection and storage tasks.

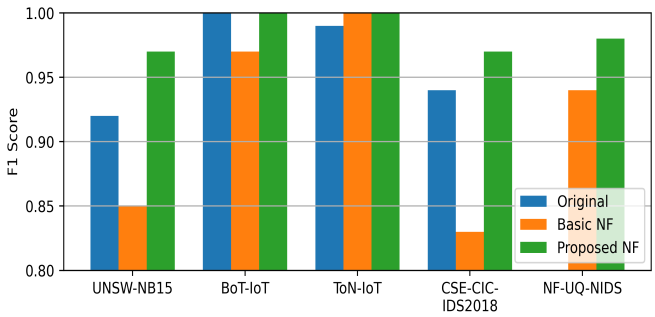

Fig. 3: Binary-class classification F1 score

4.2 Multi-class Classification

To further evaluate the proposed NetFlow feature set, multi-class classification experiments are conducted to measure the DR and F1 score of each class category present in the datasets. The weighed average of the DR, F1 Score and prediction time are also calculated. Tables 9, 10 11, 12 and 13 represent the performances of the NF-UNSW-NB15-v2, NF-BoT-IoT-v2, NF-ToN-IoT-v2, NF-CSE-CIC-IDS201-v2 and NF-UQ-NIDS-v2 datasets respectively. The original datasets performances are also provided for comparison purposes.

In Table 9, the benefits of using the NF-UNSW-NB15-v2 over the former datasets are realised by increasing the weighted F1 score to 0.99 from 0.98 and decreasing the prediction time to 8.81 µs. The DR of certain attack types such as fuzzers, generic and worms have significantly improved while the others have remained at slightly the same rate. The detection of the analysis, backdoor and DoS attacks are still unreliable even when using the expanded feature set, further analysis is required to identify the missing key features. However, due to their small number of samples, the overall accuracy of the NF-UNSW-NB15-v2 is higher (98.90%) than UNSW-NB15 (98.19%) and NF-UNSW-NB15 (97.62%).

Table 10 shows that the NF-BoT-IoT-v2 dataset is achieving the almost perfect multi-classification performance similar to the BoT-IoT dataset of a 99.99% accuracy and 1.00 F1 Score. However, the four attack

Table 9: NF-UNSW-NB15-v2 multi-class classification results

	UNSW-NB15		NF-UNSW-NB15		NF-UNSW-NB15-v2	
Class Name	**DR**	**F1 Score**	**DR**	**F1 Score**	**DR**	**F1 Score**
Benign	99.72%	1.00	99.02%	0.99	99.85%	1.00
Analysis	4.39%	0.03	28.28%	0.15	30.89%	0.17
Backdoor	13.96%	0.08	39.17%	0.17	40.30%	0.18
DoS	13.63%	0.18	31.84%	0.41	29.57%	0.36
Exploits	83.25%	0.80	81.04%	0.82	80.41%	0.84
Fuzzers	50.50%	0.57	62.63%	0.55	80.57%	0.85
Generic	86.08%	0.91	57.13%	0.66	85.15%	0.90
Reconnaissance	75.90%	0.80	76.89%	0.82	80.02%	0.83
Shellcode	53.61%	0.59	87.91%	0.75	87.67%	0.69
Worms	5.26%	0.09	52.91%	0.55	85.98%	0.69
Weighted Average	**98.19%**	**0.98**	**97.62%**	**0.98**	**98.90%**	**0.99**
Prediction Time (µs)	**9.94**		**9.35**		**8.81**	

categories are almost fully detected except for the theft attacks, where only 83.01% were successfully detected. The expanded NetFlow feature set has increased the accuracy from 73.58% to 99.99% and the F1 Score from 0.77 to 1.00. Which is a significant improvement that overcomes the drawbacks of the original NetFlow datasets, despite the slight increase in prediction time.

Table 10: NF-BoT-IoT-v2 multi-class classification results

	BoT-IoT		NF-BoT-IoT		NF-BoT-IoT-v2	
Class Name	**DR**	**F1 Score**	**DR**	**F1 Score**	**DR**	**F1 Score**
Benign	99.58%	0.99	98.65%	0.43	99.76%	1.00
DDoS	100.00%	1.00	30.37%	0.28	99.99%	1.00
DoS	100.00%	1.00	36.33%	0.31	99.99%	1.00
Reconnaissance	100.00%	1.00	89.95%	0.90	99.93%	1.00
Theft	91.16%	0.95	88.06%	0.18	83.01%	0.85
Weighted Average	**100.00%**	**1.00**	**73.58%**	**0.77**	**99.99%**	**1.00**
Prediction Time (µs)	**12.63**		**9.19**		**11.86**	

In Table 11, the NF-ToN-IoT-v2 dataset has shown outstanding results when conducting multi-classification experiments. It notably outperformed the ToN-IoT and NF-ToN-IoT datasets by increasing the weighted F1 score to 0.98 from 0.87 and 0.60 respectively. It also required a much lower prediction time than the basic NetFlow dataset. The expanded feature set has increased the DR of all attack types except for DoS, MITM and XSS attacks. Overall the NF-ToN-IoT-v2 dataset has aided in the enhanced detection of the attacks present in this dataset, with an accuracy of 98.05% confirming the reliability of the expanded NetFlow feature set.

Table 11: NF-ToN-IoT-v2 multi-class classification results

	ToN-IoT		NF-ToN-IoT		NF-ToN-IoT-v2	
Class Name	**DR**	**F1 Score**	**DR**	**F1 Score**	**DR**	**F1 Score**
Benign	89.97%	0.94	98.97%	0.99	99.44%	0.99
Backdoor	98.05%	0.31	99.22%	0.98	99.79%	1.00
DDoS	96.90%	0.98	63.22%	0.72	98.76%	0.99
DoS	53.89%	0.57	95.91%	0.48	89.41%	0.91
Injection	96.67%	0.96	41.47%	0.51	90.14%	0.91
MITM	66.25%	0.16	52.81%	0.38	37.45%	0.45
Password	86.99%	0.92	27.36%	0.24	97.16%	0.97
Ransomware	89.87%	0.11	87.33%	0.83	97.29%	0.98
Scanning	75.05%	0.85	31.30%	0.08	99.67%	1.00
XSS	98.83%	0.99	24.49%	0.19	96.83%	0.96
Weighted Average	**84.61%**	**0.87**	**56.34%**	**0.60**	**98.05%**	**0.98**
Prediction Time (µs)	**12.02**		**21.21**		**12.15**	

Table 12 presents the detection results of the NF-CSE-CIC-IDS2018-v2 dataset that has improved the DR of most of the attacks present in the dataset, achieving an accuracy of 96.90 and F1 Score of 0.98. Most attacks were fully detected with a DR ranging between 99% to 100%. However, the detection of certain attack types such as Brute Force, DDoS attack-HOIC, infiltration and SQL injection is still unreliable when using the expanded feature set. Their respective F1 score is lower due to a high number of false positives. Further

feature selection analysis is required to enhance the dataset feature set. Overall, the performance of the NF-CSE-CIC-IDS2018-v2 dataset is superior to the CSE-CIC-IDS2018 and NF-CSE-CIC-IDS2018 datasets, despite having an increased prediction time of 27.28 µs compared to 24.17 µs and 17.29 µs respectively.

Table 12: NF-CSE-CIC-IDS2018-v2 multi-class classification results

	CSE-CIC-IDS2018		**NF-CSE-CIC-IDS2018**		**NF-CSE-CIC-IDS2018-v2**	
Class Name	**DR**	**F1 Score**	**DR**	**F1 Score**	**DR**	**F1 Score**
Benign	89.50%	0.94	69.83%	0.82	99.69%	1.00
Bot	99.92%	0.99	100.00%	1.00	100.00%	1.00
Brute Force -Web	71.36%	0.01	50.21%	0.52	28.05%	0.01
Brute Force -XSS	72.17%	0.72	49.16%	0.39	29.34%	0.00
DDoS attack-HOIC	100.00%	1.00	45.66%	0.39	57.33%	0.73
DDoS attack-LOIC-UDP	83.59%	0.82	80.98%	0.82	99.29%	1.00
DDoS attacks-LOIC-HTTP	99.93%	1.00	99.93%	0.71	100.00%	1.00
DoS attacks-GoldenEye	99.97%	1.00	99.32%	0.98	100.00%	1.00
DoS attacks-Hulk	100.00%	1.00	99.65%	0.99	100.00%	1.00
DoS attacks-SlowHTTPTest	69.80%	0.60	0.00%	0.00	100.00%	1.00
DoS attacks-Slowloris	99.44%	0.62	99.95%	1.00	99.99%	1.00
FTP-BruteForce	68.76%	0.75	100.00%	0.79	100.00%	1.00
Infilteration	36.15%	0.08	62.66%	0.04	39.58%	0.43
SQL Injection	49.34%	0.30	25.00%	0.22	41.44%	0.00
SSH-Bruteforce	99.99%	1.00	99.93%	1.00	100.00%	1.00
Weighted Average	**90.28%**	**0.94**	**71.92%**	**0.80**	**96.90%**	**0.98**
Prediction Time (µs)	**24.17**		**17.29**		**27.28**	

Table 13 compares the attacks detection results of the merged NIDS dataset; NF-UQ-NIDS-v2 compared to its former (NF-UQ-NIDS) dataset. Most of the attacks DR has increased by using NF-UQ-NIDS-v2 dataset. The detection of attacks such as DoS, Generic, Worms, DDoS, Injection, password, scanning and XSS has significantly improved. However, attacks such as infiltration and MITM have been detected less accurately. Also, the time consumed to predict a single test sample has increased from 14.74 µs to 25.67 µs. An increased accuracy from 70.81% to 96.96% and F1 score from 0.79 to 0.97 confirms the enhanced detection capabilities of the expanded feature set across 20 attack types.

Table 13: NF-UQ-NIDS-v2 multi-class classification results

	NF-UQ-NIDS		**NF-UQ-NIDS-v2**	
Class Name	**DR**	**F1 Score**	**DR**	**F1 Score**
Analysis	69.63%	0.21	78.43%	0.24
Backdoor	90.95%	0.92	89.61%	0.93
Benign	71.70%	0.83	93.45%	0.96
Bot	100.00%	1.00	100.00%	1.00
Brute Force	99.94%	0.85	98.16%	0.74
DoS	55.54%	0.62	99.46%	1.00
Exploits	80.65%	0.81	85.16%	0.84
Fuzzers	63.24%	0.54	80.58%	0.84
Generic	58.90%	0.61	85.41%	0.88
Infilteration	60.57%	0.03	21.62%	0.19
Reconnaissance	88.96%	0.88	98.24%	0.76
Shellcode	83.89%	0.15	89.35%	0.34
Theft	87.22%	0.15	81.66%	0.22
Worms	52.97%	0.46	87.20%	0.71
DDoS	77.08%	0.69	99.43%	1.00
Injection	40.58%	0.50	90.03%	0.90
MITM	57.99%	0.10	35.97%	0.43
Password	30.79%	0.27	97.09%	0.97
Ransomware	90.85%	0.85	96.82%	0.87
Scanning	39.67%	0.08	97.36%	0.98
XSS	30.80%	0.21	95.72%	0.95
Weighted Average	**70.81%**	**0.79**	**96.93%**	**0.97**
Prediction Time (µs)	**14.74**		**25.67**	

Overall, the proposed NetFlow feature set has significantly improved the multi-class classification performance of the datasets as displayed in Figure 4. The performance is comparable to the original feature set of the datasets but remarkably superior to the basic NetFlow feature set. This demonstrates the benefits of adopting the proposed feature set across NIDS datasets and encourages researchers to generate their datasets in the proposed format for efficient and reliable ML experiments.

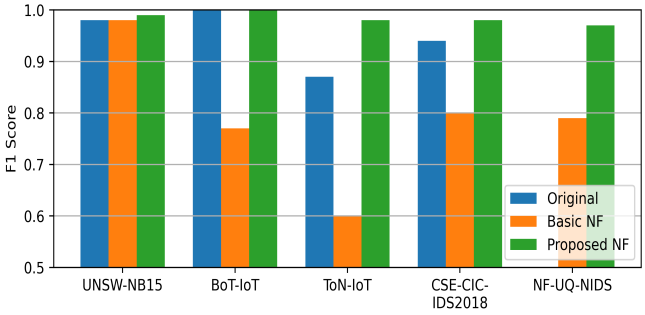

Fig. 4: Multi-class classification F1 score

5 Conclusion

A new network traffic feature set, listed in Table 2, has been proposed to be used in future NIDS datasets. The importance of having a standard benchmark feature set is that ML-based NIDS models can be reliably evaluated using a targeted feature set across multiple datasets and various attack scenarios. Also, multiple datasets can be merged, leading to a larger data source of labelled network flows collected over several network environments. As part of the proposed feature set evaluation, five new NIDS datasets have been generated and made publicly available to be used by researchers for a reliable evaluation of their proposed learning models. The feature set is displaying promising detection results in both binary- and multi-class classification experiments. There is a significant improvement over the original and basic NetFlow datasets' results, demonstrating the promising benefits of the proposed feature set. Almost all attack types were fully detected, however, certain ones require further feature selection analysis to identify key features that aid in their reliable detection. Overall, the proposed feature set attracts the tremendous benefits having a standard feature set across datasets. It achieves superior detection accuracy compared to other feature sets. It is made up of NetFlow features which are more practical in their extraction leading to scalable deployments. Therefore, it requires increased attention by researchers to be utilised in their proposed ML-based NIDS models.

References

1. "Netflow datasets." `http://staff.itee.uq.edu.au/marius/NIDS_datasets/`, 2020.
2. P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," *computers & security*, vol. 28, no. 1-2, pp. 18–28, 2009.
3. S. K. Sahu, S. Sarangi, and S. K. Jena, "A detail analysis on intrusion detection datasets," in *2014 IEEE International Advance Computing Conference (IACC)*, pp. 1348–1353, 2014.
4. A. Shiravi, H. Shiravi, M. Tavallaee, and A. A. Ghorbani, "Toward developing a systematic approach to generate benchmark datasets for intrusion detection," *Computers & Security*, vol. 31, no. 3, pp. 357 – 374, 2012.
5. A. Binbusayyis and T. Vaiyapuri, "Identifying and benchmarking key features for cyber intrusion detection: An ensemble approach," *IEEE Access*, vol. 7, pp. 106495–106513, 2019.
6. M. Sarhan, S. Layeghy, N. Moustafa, and M. Portmann, "Netflow datasets for machine learning-based network intrusion detection systems," 2020.
7. B. Claise, G. Sadasivan, V. Valluri, and M. Djernaes, "Cisco systems netflow services export version 9," 2004.
8. M. Ring, S. Wunderlich, D. Scheuring, D. Landes, and A. Hotho, "A survey of network-based intrusion detection data sets," *Computers & Security*, vol. 86, pp. 147 – 167, 2019.
9. R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," *2010 IEEE Symposium on Security and Privacy*, 2010.
10. N. Moustafa and J. Slay, "Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set)," *2015 Military Communications and Information Systems Conference (MilCIS)*, 2015.
11. N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "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.
12. N. Moustafa, "Ton-iot datasets," 2019.
13. I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," *Proceedings of the 4th International Conference on Information Systems Security and Privacy*, 2018.

14. Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, Jan. 2019.
15. B. Li, J. Springer, G. Bebis, and M. Hadi Gunes, "A survey of network flow applications," *Journal of Network and Computer Applications*, vol. 36, no. 2, pp. 567–581, 2013.
16. D. R. Kerr and B. L. Bruins, "Network Flow Switching and Flow Data Export," 2001.
17. 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.
18. Ntop, "nProbe, An Extensible NetFlow v5/v9/IPFIX Probe for IPv4/v6." `https://www.ntop.org/guides/nprobe/cli_options.html`, 2017.