applied sciences

MDPI

Article

Generating Network Intrusion Detection Dataset Based on Real and Encrypted Synthetic Attack Traffic

Andrey Ferriyan [1,*,†], Achmad Husni Thamrin [1,†], Keiji Takeda [2,†] and Jun Murai [3,†]

[1] Graduate School of Media and Governance, Keio University, Kanagawa 252-0882, Japan; husni@sfc.keio.ac.jp
[2] Faculty of Environment and Information Studies, Keio University, Kanagawa 252-0882, Japan; keiji@sfc.keio.ac.jp
[3] Keio University, Tokyo 108-8345, Japan; jun@sfc.keio.ac.jp
* Correspondence: andrey@keio.jp
† These authors contributed equally to this work.

Abstract: The lack of publicly available up-to-date datasets contributes to the difficulty in evaluating intrusion detection systems. This paper introduces HIKARI-2021, a dataset that contains encrypted synthetic attacks and benign traffic. This dataset conforms to two requirements: the content requirements, which focus on the produced dataset, and the process requirements, which focus on how the dataset is built. We compile these requirements to enable future dataset developments and we make the HIKARI-2021 dataset, along with the procedures to build it, available for the public.

Keywords: network intrusion detection system; network intrusion datasets; encrypted network traffic; https; tls

Citation: Ferriyan, A.; Thamrin, A.H.; Takeda, K.; Murai, J. Generating Network Intrusion Detection Dataset Based on Real and Encrypted Synthetic Attack Traffic. *Appl. Sci.* **2021**, *11*, 7868. https://doi.org/10.3390/app11177868

Academic Editor: Rafael T. de Sousa, Jr.

Received: 2 August 2021
Accepted: 21 August 2021
Published: 26 August 2021

1. Introduction

It is challenging to estimate how much malicious detection methods have improved in the intrusion detection system (IDS) field. Training IDSs that employ machine learning depends on the available datasets, but obtaining a reliable dataset for comparison is difficult. Among the factors that make it difficult to compare datasets are a lack of proper documentation of the methods [1], a lack of comparison methodology [2], and a lack of important features, such as ground-truth labels, and publicly available and real-world environment traffic. Furthermore, network traffic nowadays is mainly being encrypted for communication security and privacy, and only very few datasets reflect this situation.

The dataset is an important part to build machine learning-based IDS models. The process starts with capturing traffic either as a packet or flow from the internet. Afterward, the captured traffic is compiled into a specific type of data containing network-related features, including labeling. A general machine learning-based IDS can be shown in Figure 1. Labeling is a crucial process for the dataset. Handling ground-truth is a real challenge, especially when experts cannot determine whether the traffic is an attack or benign. This is a reason why researchers use synthetic traffic. However, this implies the generated traffic is not representative of the real world environment. In a nutshell, the process of making a dataset starts with capturing traffic, and ends with the final preprocessing phase. The final result from the preprocessing phase is a labeled dataset. Each data point is classified into malicious or benign. The file contains tabular data in a human-readable format, such as a CSV file, or binary form, such as an IDX file. The number of detected malicious or false alarms can be used to benchmark the dataset.

The existing datasets lack reliably encrypted traces and practicality to produce as the basis to build the comprehensive model for the detection of new attacks. Most of the existing research that employs encrypted traffic are focused on different scopes, such as traffic classification and analysis [3]. Although such research exists [4], the dataset is not publicly available, due to the sensitivity of the data.

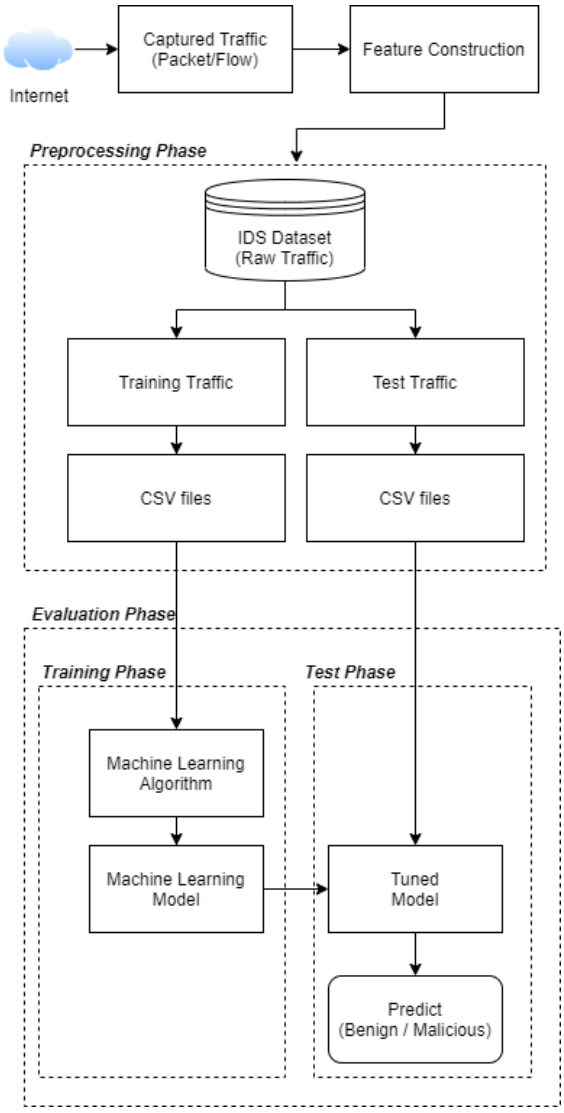

Figure 1. A general machine learning-based IDS flow.

Benchmark datasets are an important basis to evaluate and compare the quality among different IDS. Based on the detection methods, there are three types of IDS: signature-based, anomaly-based, and a combination of signature-based and anomaly-based. These three types of IDS benchmark their systems with the KDD99 dataset, which is obsolete. The signature-based one focuses on building automatic signature generation [5], while the anomaly-based focuses on observing an outlier from the legitimate profile [6]. The signature-based type relies on a pattern-matching method to identify and attempt to match with the signatures database. When an attack attempt matches with the signature, an alert is raised. The signature-based type has the highest accuracy and lowest false alarm rate but this type cannot detect unknown attacks. While the anomaly-based type might detect unknown attacks by comparing abnormal traffic with the normal traffic, the ratio of false alarm rates remains high.

In this paper, we present a tool and requirements for making a new dataset created by generating encrypted network traffic in a real-world environment. Our contributions are two-fold. First, we propose new requirements for creating new datasets. Second, we create a new IDS dataset that covers the network traffic with encrypted traces. The dataset is labeled with attacks, such as brute force login and probing. The packet traces with payload are provided along with the background traffic and ground-truth data. We extract and adopt more than 80 features from the CICIDS-2017 dataset for the ground-truth,

benign traffic, and malicious traffic by using Zeek [7], an open source network security monitoring tool.

The paper is organized as follows. In Section 2, we review the existing datasets and we provide the most important features from their dataset, such as the duration of capturing of the network traffic, what kind of attack they implemented, and what format of data they used. From the review, we summarize the requirements that need to be satisfied to build a practical, implemented dataset and compare it among the existing datasets in Section 3. In Section 4, we describe the dataset generation methodology along with the attack traffic generation and explain the characteristics of the attack traffic. Subsequently, we describe the network configuration for generating network traffic, the scenarios, the tools and code we used to generate, and the duration of capturing the network features. In Section 5, we analyze the dataset and provide information on how the labeling works. Finally, the last section concludes this paper.

2. Review of Existing Datasets

Many researchers have published papers based on generated IDS datasets, such as ISCX [8], UNSW-NB15 [9], and UGR'16 [10]. In this section, we introduce several IDS datasets with their characteristics. We highlight several important requirements from their perspective.

2.1. KDD99

The KDD99 dataset was created in 1999, using tcpdump, and was built based on the data captured by the DARPA 98 IDS evaluation program [11]. The training data are around four gigabytes of compressed TCP data from seven weeks of network traffic. The network traffic contains attack traffic and normal traffic. The capture of the network traffic was done in a simulated environment. The dataset contains a total of 24 attack types, which fall into four main categories: Denial of Service (DOS), Remote to Local (R2L), User to Root (U2R), and probing. KDD99 has been used extensively in IDS research. The report [12] showed that during 2010–2015, 125 published papers performed IDS evaluation using KDD99. While this dataset is considered inadequate for evaluation, such as a lot of redundant instances recorded, the main problem is that the dataset is not up to date with the recent situation and attack vectors. Although many researchers were already convinced with this information, studies from another group of researchers argued that this dataset is the most widely used for benchmarking [13] or to limit their study only for KDD99 [14].

2.2. MAWILab

MAWI was built in 2001 and consists of a set of labels locating traffic anomalies in the MAWI archive [15]. This dataset contains tcpdump packet traces captured from an operational testbed network in a link between Japan and the United States. The archive contains 15 min of daily traces. This dataset is huge with a very long period. The labeled MAWI archive is known as MAWILab, obtained from a graph-based methodology that combines different and independent anomaly detectors [16]. MAWI archives labeling is based on inferences that results in no ground-truth traffic that can be used for evaluation. The label has three classes: anomalous for a true anomaly, suspicious indicates that the traffic is likely to be anomalous, and notice is assigned as an anomaly but it does not reach a consensus from all anomaly detectors. Several researchers used MAWILab for anomaly detection [17] and generating labeled flow [18]. The limitation of this dataset is that the packet traces are captured for 15 min each day. The header information is available in the packet traces but the payload is removed.

2.3. CAIDA (Center of Applied Internet Data Analysis)

CAIDA has several different types of datasets, categorized as ongoing, one-time snapshot, and complete [19]. CAIDA collects the data from different locations, and each of the datasets has different characteristics, such as Distributed Denial of Services (DDoS)

attack, UDP probing, BGP monitoring, IPv4 census with passive traffic traces captured from a darknet, an academic ISP, and a residential BGP with active measurements of ICMP ping, HTTP GET and traceroutes. Most of the datasets are anonymized with IP addresses and the payload, which severely reduces their usefulness. Based on their catalog, during 2017–2020, most of the papers related to IDS and security focused on Denial of Service (DoS) [20,21], Distributed Denial of Service (DDoS) [22], DNS security [23], Network Telescope Daily Randomly, and Uniformly Spoofed Denial-of-Service (RSDoS) Attack Metadata. Each record contains the IP address of the attack victim, the number of distinct attacker IPs in the attack, the number of distinct attacker ports and target ports, the cumulative number of packets observed in the attack, the cumulative number of bytes seen for the attack, the maximum packet rate seen in the attack as the average per minute, the timestamp of the first and the last observed packet of the attack, the autonomous system number of target_IP at the time of the attack, and the country and continent geolocation of target_IP at the time of the attack. This dataset is updated every day.

2.4. SimpleWeb

SimpleWeb is a dataset collected from the network of the University of Twente [24]. This dataset contains packet headers of all packets that are transmitted over the uplink of access to the internet. The packets are captured with tcpdump and Netflow version 5. The payload from the packets is removed because it contains sensitive information, such as HTTP requests or conversations of instant messaging applications. The labeled dataset for suspicious traffic is collected by using a honeypot server. Despite no ground-truth data being available, researchers still use it to compare with the different real-world environment (e.g., campus network, backbone link) [25], while others employ it for background traffic for botnet detection [26], and to evaluate publicly available dataset for similarity searches to detect network threats [27].

2.5. NSL-KDD

NSL-KDD is an updated dataset that tries to solve some of the inherent problems in the KDD99 dataset [28]. The NSL-KDD dataset contains features and labels indicating either normal or an attack, with specific types of attacks. Every instance in the training set contains a single connection session, which is divided into four groups, such as basic features from the network connection, content-related features, time-related features, and host-based traffic features. Each instance is labeled either as normal or attack. These attacks are categorized into four groups: Denial of Service (DoS), User to Root (U2R), Remote to Local (R2L), and Probing. Many researchers use it as a benchmark to help them to compare their intrusion detection systems performance. Several groups of researchers used different scopes, such as IoT-based networks [29] and Vehicular Ad Hoc Network (VANET) [30]. The former is for SYN flood, UDP flood, and Ping of Death (PoD) detection, while the latter is mostly for DDoS detection. Other researchers used different methods and switched from conventional machine learning to deep learning based methods [31–33].

2.6. IMPACT

IMPACT is a marketplace of cyber-risk data. The data distribution and tool repository are provided by multiple providers and stored and accessed from multiple hosting sites [34]. The datasets related to cyber-attacks, such as the daily feed of network flow data produced by Georgia Tech Information Security Center's malware analysis system, updates once a year. These datasets are only open for specific countries based on approval by the Department of Homeland Security (DHS).

2.7. UMass

UMass is a trace repository provided by the University of Massachusetts Amherst [35]. The network-attack-relevant data are provided with various type of data, such as traffic flow from the TOR network [36], a trace of attack simulation on peer-to-peer data sharing

network [37], passive localization attack simulation with reality mining dataset [38] containing sensor data (proximity, location, location labels, etc.), and survey data (personal attributes, research group, position, neighborhood of apartment, and lifestyle).

2.8. Kyoto

This dataset was created between 2006 and 2015 by Kyoto University through honeypot servers. This dataset was created using Bro 2.4 (the former name of Zeek) with 24 statistical features consisting of 14 features extracted based on the KDD99 dataset and an additional 10 features, such as IDS_detection, Malware_detection, Ashula_detection, Label, Source_IP_Address, Source_Port_Number, Destination_IP_Address, Destination_Port_Number, Start_Time, and Protocol [39]. The information is limited to the attack information targeting the honeypot server. There are no packet traces or information about the payload. Furthermore, the information on how to label the dataset is not found [40]. Several published papers using the Kyoto dataset focused on anomaly detection, especially on the feature analysis [41], feature dimensionality reduction and ensemble classifier [42].

2.9. IRSC

This dataset was created by Indian River State College and consists of network flows and full packet capture [43]. The dataset represents a real-world environment in which the attack traffic has two different types: the controlled version, which is synthetically created by the team, and the uncontrolled version, which are the real attacks. The flow based traffic created with the Silk [44] and the full packet capture created with the Snort IDS [45]. The additional source of flow data was produced from the Cisco firewall using NetFlow version 9. While the authors stated that the dataset is a complete capture with payload and flow data, unfortunately, it is not publicly available.

2.10. UNSW-NB15

UNSW-NB15 was created using a commercial penetration tool in the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS). This tool can generate hybrid synthetically modern normal activities and contemporary attack behaviors from network traffic [9]. They collected tcpdump traces for a total duration of 31 h. From these network traces, they extracted 49 features categorized into five groups: flow features, basic features, content features, time features, and additional generated features. Feature and statistical analyses are the most common methods used in several published papers employing UNSW-NB15 [46–48]. While [46] could obtain 97% accuracy by using 23 features, [47] incorporated the XGBoost algorithm for feature reduction, using several traditional machine learning algorithms for evaluation, such as Artificial Neural Network (ANN), Logistic Regression (LR), k-Nearest Neighbor (kNN), Support Vector Machine (SVM) and Decision Tree (DT).

2.11. UGR'16

This dataset was created from several NetFlow v9 collectors located in the network of a Spanish ISP [10]. It is composed of two different types of datasets that are split in weeks. First, the calibration set contains real background traffic data, and second, the test data contain real background traffic and synthetically generated traffic data with well-known types of attacks. Due to the nature of the NetFlow data, payloads from the network traffic were not included. The types of attacks implemented in this dataset are Low-rate DoS, Port scanning, and Botnet traffic. Between 2017 and 2021, we found mixed methods from several published papers, such as [49,50], Rajagopal et al. [49], who argued that conventional machine learning methods were ineffective and instead used stacking ensembles to improve performance and reliable predictions, while [50] proposed hybridized multi-model system to improve the accuracy of detecting the intrusion. Ref. [51] addressed imbalanced data problems by producing synthetic data with the Generative Adversarial Network (GAN).

2.12. CICIDS-2017

This dataset was created by the Canadian Institute for Cybersecurity at University of Brunswick in 2017. The purpose of CICIDS-2017 was intrusion detection, and it consisted of several attack scenarios. In this dataset, the attack profiles were produced using publicly available tools and codes. Six attack profiles were implemented, such as brute force, heartbleed, botnet, DoS, DDoS, web attack, and infiltration attack. The realistic background traffic was generated, using a B-Profile system [52]. The B-Profile system extracted 25 behaviors of users based on several protocols, such as HTTP, HTTPS, FTP, SSH, and SMTP. The network traffic features were captured with CICFlowMeter [53], which extracted 80 features from the pcap file. The flow label included SourceIP, SourcePort, DestinationIP, DestinationPort, and Protocol. Mixed methods are used, incorporating CICIDS-2017 to detect specific attacks such as DoS attack [54] by using feature reduction, web-attack detection [55], and anomaly web traffic [56] with ensemble architecture and feature reduction. Others are improving the AdaBoost-based method [57] to counter the imbalance of the training data [58], and combining feature selection and information gain to find relevant and significant features and to improve accuracy and execution time.

3. Dataset Requirements

While the authors of ISCX [8], UGR'16 [10], and CICIDS-2017 [53] introduce a new dataset and provide extensive requirements about the dataset, their works have different research objectives and scope. In contrast to their earlier dataset, our work is a complement to fill the gap, missing from the previous requirement.

3.1. Requirements for IDS Evaluation Datasets

Generally, different datasets have different assets and requirements. Shiravi et al. [8] focused on accurate labeling in the dataset by building a systematic profile to generate the dataset. They argued that the network traffic should be as realistic as possible, so a complete capture in a realistic network must be satisfied. It will impact anonymity and lead to potential privacy issues. Fernandez et al. [10] provided only flow information and focused on the duration of the capturing. Furthermore, a flow format with only 5-tuple is not enough and needs additional features if the malicious traffic is delivered via an encrypted protocol, such as HTTPS. We found that the requirements to build an IDS dataset from Sharafaldin et al. [52] is extensive. Unfortunately, their generated traffic comes from an emulated network, which is missing a realistic environment. In addition, the information about ground-truth and how the labeling works was not found in their paper and, thus, has the potential to be inaccurate and unreliable for analysis. Cordero et al. [59] created a tool called ID2T and we found that their requirements are better in practical terms. They categorized the requirements into functional and non-functional ones. Functional requirements focus on what is needed to construct datasets, while the non-functional requirements specify several criteria that need to be satisfied to be of practical use.

All of the requirements have high similarity. However, none of the works highlighted the importance of encrypted traffic in the dataset, and this is one of the emphases in our requirements. We derived our requirements for datasets based on the above works as well as by reviewing the existing datasets which described that the quality of the dataset mostly affects the outcome of the NIDS system. We classified the requirements into content requirements and process requirement. The content requirements are similar to [59], such as functional requirement, which focuses on what is needed to construct a dataset, and [8] on complete network traces and realistic network traffic capture. The process requirement is similar to that of [10] in the documentation point. While this is not enough, the information on how to produce a new dataset and practical to implement does not exist.

The proposed requirements try to fill the gap of information from previous datasets. Based on our content requirements, we found at least four missing points:

(1) Most of the datasets are not anonymized, such as KDD99, SimpleWeb, NSL-KDD, Kyoto, IRSC, and UNSW-NB15. We chose to preserve privacy by anonymizing only a specific part of the background traffic based on the Crypto-Pan algorithm.
(2) The majority of the datasets are impractical to generate, such as KDD99, CAIDA, NSL-KDD, IMPACT, UMass, IRSC, UNSW-NB15, and CICIDS-2017.
(3) They do not have ground-truth data, such as MAWILab, CAIDA, SimpleWeb, IMPACT, UMass, Kyoto, and CICIDS-2017.
(4) As for encryption information, most of the datasets contain non-encrypted traffic, except for MAWILab, UGR'16, and CICIDS-2017. These datasets neither focused on nor classified encrypted traffic. However, HIKARI-2021 is focused on encrypted traffic.

The content requirements focus on the assets of the dataset to achieve a practical way to produce a dataset, while the process requirement specifies the information on how the dataset is built, so a new dataset can be built in the future using the same process. We list these requirements below along with some descriptions of each item.

3.1.1. Content Requirements

(1) Complete capture: complete capture of the network traffic, such as communication between host, broadcast message, domain lookup query, the protocol being used. The most important thing from complete capture is that both flow data and pcap should be available.
(2) Payload: payload is not needed for a flow-based approach. However, having comprehensive information and extracting the most out of the data is important. HIKARI-2021 is the dataset that provides labeled encrypted traffic, while the well-known datasets do not focus on encrypted traffic. There is a possibility that a full payload captured might be useful in the future.
(3) Anonymity: synthetic traffic should provide full packet capture, while real traffic must anonymize certain packets to preserve privacy.
(4) Ground-truth: the datasets should provide realistic traffic from a real production network, compared with the synthetic traffic, and ensure no unlabeled attack in the ground-truth.
(5) Up to date: both packet traces from flow data and pcap should be always accessible by repeating the capturing process of the network traffic. Because the data are subject to change over time, repeating the procedures guarantees that the dataset always obtains the latest information.
(6) Labeled dataset: correctly labeling data as malicious or benign is important for accurate and reliable analysis. The labeling process is a manual task and determined by the flow with a combination of the source IP address, source port, destination IP address, destination port, and protocol.
(7) Encryption Information: information on how to establish benign or malicious traffic must be stated. We are focused on application layer attacks, such as brute force and probing that employ HTTPS with TLS version 1.2 to deliver the attacks.

3.1.2. Process Requirement

Methods: producing a new dataset with specific requirements and practical implementation is important. Therefore, the methods should cover information on how to generate the dataset, how to generate the benign and attack traffic, how the background traffic is being captured, how the labeling process works, and how to implement it in the network. Furthermore, we need to determine what scenarios and how to deliver the synthetic attack in the network. In addition, the information of what features and how many can be extracted from the packet traces should be declared. Information on how to make a new dataset should be available in detail and practical to generate.

3.2. Comparison of the Existing Datasets against the above Requirements

Comparisons between IDS datasets are shown in Table 1, where we assess the datasets in Section 2, based on the requirements that we set in Section 3.1.

Table 1. Comparison of IDS datasets based on the requirements in Section 3.1.

Dataset	Comp. Capture	Payload	Anonymity	Ground-Truth	Up to Date Traffic	Labeled	Encryption	Practical to Generate
KDD99 [11]	Yes	Yes	No	Yes	No	Yes	No	No
MAWILab [15]	Yes	No	Yes	No	Yes	Yes	Yes	Yes
CAIDA [19]	Yes	No	Yes	No	Yes	No	No	No
SimpleWeb [24]	Yes	No	No	No	No	No	No	Yes
NSL-KDD [28]	Yes	Yes	No	Yes	No	Yes	No	No
IMPACT [34]	Yes	No	Yes [1]	No	Yes	No	No	No
UMass [35]	Yes	Yes	-	No	No	No	No	No
Kyoto [39]	Yes	Yes	No	No	No	Yes	No	Yes
IRSC [43]	Yes	Yes	No	Yes	No	Yes	No	No
UNSW-NB15 [9]	Yes	Yes	No	Yes	No	Yes	No	No
UGR'16 [10]	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
CICIDS-2017 [53]	Yes	Yes	Yes	No	No	Yes	Yes [2]	No

[1] Mix datasets with partial anonymization; [2] Mix data between un-encrypted data, such as HTTP, and encrypted data, such as SSH.

We were unable to find the information regarding the anonymity of the UMass dataset; therefore, no indicator was given. As for the IMPACT dataset, this platform has many datasets, some parts of which are anonymized, while others are not. In the CICIDS-2017 dataset, one part of the traffic has samples for encrypted traffic with benign and attack profiles.

We have four observations from the above comparisons. First, there is a need for encrypted samples of benign and attack traffic. We found that [15] in their dataset have information on whether the traffic is anomalous or suspicious but it depends on the anomaly detectors. The payload from the packet traces was not included. This limited the capability of IDS because many attacks cannot be detected only by network flow with only 5-tuple attributes. In addition, [53] in their datasets included the traffic from benign and attack profiles from SSH. While this is beneficial, the diversity of the attack needs to be expanded to applications, such as browser attacks, or with different protocols, such as HTTPS, and we did not find that this protocol exists in their dataset. Second, we found that most of the datasets are not anonymized. The reason is probably that their testing beds are in a controlled environment or they have consent with their activity. The former is the best option with the consequences that the traffic will have more synthetic traffic while reducing the real traffic. The latter is preferred if they can preserve privacy. Furthermore, privacy can be maintained by anonymizing the traffic, but being highly anonymized may decrease the results of the analysis [8,60,61]. Third, we found that most of the datasets do not have ground-truth data and background traffic, which make the analysis limited only to their model. Fourth, there is a need for a methodology on how to create a new dataset. This is due to the nature of the network environment that is subject to change over time. How to create new datasets with the practical implementation is important, so researchers may make their dataset and evaluate it with their environment. This methodology can be a guideline for IDS researchers to follow for making a practical dataset.

4. HIKARI-2021 Generation Methodology

In this section, we explain our methodology for producing our dataset, which we call HIKARI-2021. The process starts with creating a victim network, where background traffic is captured, and attackers generate synthetic benign traffic, using a benign profile (details in Section 4.3), and malicious traffic, using an attacker profile (details in Section 4.4). The attacker traffic is captured in the attacker network. We do this to differentiate between synthetic benign and malicious traffic. Distinguishing between benign and malicious traffic is based on several criteria (details in Section 4.4). We then process the packet traces to anonymize the background traffic and extract the features. The packet traces and extracted features, as well as the documentation, constitute the produced dataset.

We are focused on application layer attacks that employ HTTPS. Based on the report from the 2021 Data Breach Investigation, 80% of the attack vectors come from application-layer attacks. There are many attacks on the internet but we are not focused on how many attacks we can generate. Based on the survey from netcraft.com and websitesetup.org, WordPress, Joomla, and Drupal are among the ten most popular open-source CMSs, with the combined market share of almost 50%. Based on the information from CVE, more than 300 vulnerabilities existed for WordPress from 2006 to 2021, 92 vulnerabilities for Joomla from 2004 to 2021, and 202 vulnerabilities for Drupal from 2002 to 2021. More than half of the vulnerabilities from these three CMSs are part of Brute Force and Probing. Furthermore, the goal of this research is not in the attack diversity but in what kind of attack we can deliver in the encrypted network. We decided to focus on common application-layer attacks, such as brute force and probing. In addition, the IDS researcher may build their script based on our tool to enrich the attack, such as SQL Injection, Denial of Service, etc.

4.1. Network Configuration for Generating Dataset

Figure 2 shows our network configuration, where attackers are on a separate network from the victims. The format of the data we captured is pcap. The important point in this configuration is as follows:

(1) The attacker network with two machines is deployed with CentOS 7 and CentOS 8. There are no specific criteria of the attackers' machines as long as they can run Bash and Python scripts. The Python version is 3.8.8 from Miniconda 3.

(2) In the victim network, three machines are deployed with one Debian 8 machine running Joomla 3.4.3, and two Debian 9 machines running Drupal 8.0 and WordPress 5.0. There are no specific criteria for the OS version for the victim network, and the three different Content Management Systems (CMS) such as Drupal, WordPress, and Joomla use default themes and plugins. These three open-source CMSs were chosen based on their popularity. These machines are used for collecting the background traffic.

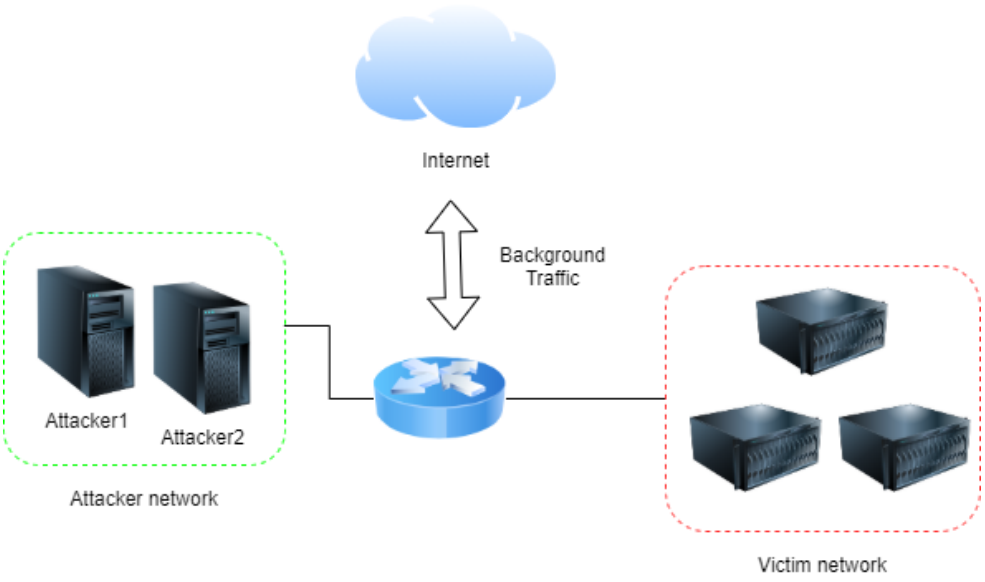

Figure 2. Network configuration to generate dataset.

4.2. Background Profile

Generating realistic data is important. For the background traffic, we captured all the data without any filter or firewall in the victim network. Therefore, there is a possibility that the background traffic may contain malicious traffic or attacks. To preserve privacy without degrading the result of the analysis, we anonymized several pieces of information, such as IP address and the payload.

4.3. Benign Profile

To generate the benign profile, we considered using a profile similar to human behavior. To achieve it, we used Selenium [62], which runs two headless browsers: Google Chrome and Mozilla Firefox. These two browsers act like humans by clicking random links from multiple websites, sign up as a user, sign in, post an article to the target victim's server, and sign out. To behave like a human and to avoid being detected as a bot or web spider, we used several configurations, such as user-agent and random delay, for every sequence of action. The addresses of the websites are from Alexa's top 1 million visitors [63]. The benign profile was developed with Python script; this activity simulates benign traffic. All benign traffic is captured without anonymizing the payload. The type of traffic generated is HTTPS only.

4.4. Attacker Profile

The attack traffic is generated synthetically, first by targeting a specific page for user login of the CMSs, and second by scanning their vulnerability. Both of the attacks are delivered via the HTTPS protocol. The attacks are delivered on different days with different scenarios (details in Section 4.5). The types of attacks are as follows:

(1) Brute force attack: this attack is the most famous for cracking passwords. The attacker usually repeatedly tries to gain the target over and over using all possible combinations using a dictionary of possible common passwords [64]. We developed a script that mimics a brute force attack, using a browser to deliver the attack. We intentionally added a user to the three different CMSs with the role as an admin and password, which we took randomly from [64]. The purpose is to make sure that the brute force attack is delivered successfully.
(2) Brute force attack with different attack vectors: while the first type of attack uses the browser as the attack vector, the second attack uses a different attack vector, XMLRPC. We developed a script that accesses XMLRPC for gaining valid credential access.
(3) Probing: this is also called vulnerability probing. This script scans the web applications, such as Joomla, WordPress, and Drupal to find their vulnerability. The tools for vulnerability scanning are publicly available. For this dataset, the scripts used these probing scripts: droopescan [65] for WordPress and Drupal, and joomscan [66] for Joomla.

We provide the template script to customize the attack profile so researchers may use it for making custom attacks using different vectors. Distinguishing between an attack profile and benign profile is based on the source IP address, source port, destination IP address, destination port, protocol, and the day both of the profiles being generated. In addition, to determine benign traffic, any destination addresses in the Alexa list are considered benign.

4.5. Scenarios

We captured the traffic non-consecutively between 28 March and 4 May 2021, with each capture session lasted for 3 to 5 hours. In the first scenario, no attack traffic was generated, and only background traffic was being captured. In the second scenario, brute force attack traffic was generated for 2 days. Furthermore, a brute force with different attack vectors was generated in the third scenario. In the last scenario, scanning vulnerabilities of WordPress, Joomla, and Drupal were generated.

4.6. Dataset Preprocessing

The traces were captured using tcpdump with full packet capture. As for the background traffic, we fully captured the traffic but then we anonymized it to maintain privacy. To preserve privacy, we used a Crypto-PAn based algorithm [67]. The complete dataset contains several files: pcap files from background traffic, and synthetic attacks. The flowmeter files with pkl and CSV are available for downloads [68]. The preprocessing flow from pcap files into CSV files is presented in Figure 3.

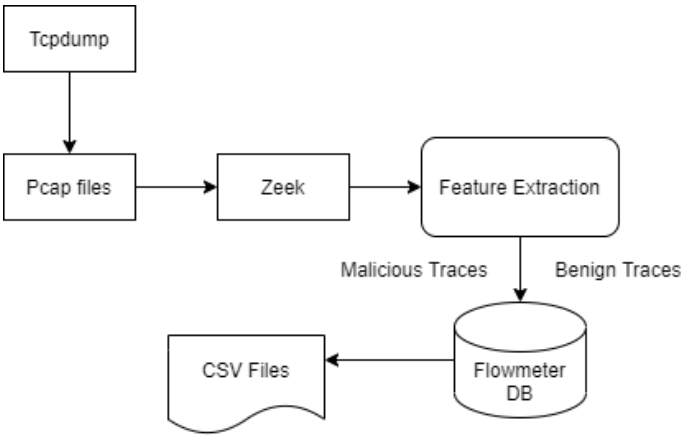

Figure 3. The preprocessing flow of HIKARI-2021 dataset.

4.7. Labeling Process

During background traffic validation, we found malicious cryptomining traffic. The result comes from the Zeek rules, which shows that some traffic is that of malicious cryptomining, such as XMRIGCC. We then separated and added it as a new attack, which we categorized as XMRIGCC CryptoMiner. Labels were applied on a per-flow basis. In the background traffic, we did not find any attack besides the cryptomining. Other than background, our labeling was based on the generated synthetic rules, such as source IP address, source port, destination IP address, destination port, and protocol. The dataset consists of two labels: traffic_category and label. The former represents the name of the traffic category, while the latter is only a single value with 0 representing Benign, and 1 representing Attack as shown in Table 2.

Table 2. Labeled features information.

Traffic Category	Label	Total Flows (Flowmeter)	No. Encrypted Session
Background	Benign	170,151	36,782
Benign	Benign	347,431	116,309
Bruteforce	Attack	5884	5884
Bruteforce-XML	Attack	5145	5145
Probing	Attack	23,388	23,388
XMRIGCC CryptoMiner	Attack	3279	0

4.8. Feature Description

HIKARI-2021 features were extracted using Zeek. Table 3 shows the features while Figure 4 displays a statistical description of the features. Most of the features were adopted from CICIDS-2017, while uid, originh, originp, responh, responp, traffic_category, and Label were derived from Zeek.

Table 3. List of features in HIKARI-2021.

No	Feature	No	Feature	No	Feature
1	uid	30	flow_ECE_flag_count	59	flow_iat.avg
2	originh	31	fwd_pkts_payload.min	60	flow_iat.std
3	originp	32	fwd_pkts_payload.max	61	payload_bytes_per_second
4	responh	33	fwd_pkts_payload.tot	62	fwd_subflow_pkts
5	responp	34	fwd_pkts_payload.avg	63	bwd_subflow_pkts
6	flow_duration	35	fwd_pkts_payload.std	64	fwd_subflow_bytes
7	fwd_pkts_tot	36	bwd_pkts_payload.min	65	bwd_subflow_bytes
8	bwd_pkts_tot	37	bwd_pkts_payload.max	66	fwd_bulk_bytes
9	fwd_data_pkts_tot	38	bwd_pkts_payload.tot	67	bwd_bulk_bytes
10	bwd_data_pkts_tot	39	bwd_pkts_payload.avg	68	fwd_bulk_packets
11	fwd_pkts_per_sec	40	bwd_pkts_payload.std	69	bwd_bulk_packets
12	bwd_pkts_per_sec	41	flow_pkts_payload.min	70	fwd_bulk_rate
13	flow_pkts_per_sec	42	flow_pkts_payload.max	71	bwd_bulk_rate
14	down_up_ratio	43	flow_pkts_payload.tot	72	active.min
15	fwd_header_size_tot	44	flow_pkts_payload.avg	73	active.max
16	fwd_header_size_min	45	flow_pkts_payload.std	74	active.tot
17	fwd_header_size_max	46	fwd_iat.min	75	active.avg
18	bwd_header_size_tot	47	fwd_iat.max	76	active.std
19	bwd_header_size_min	48	fwd_iat.tot	77	idle.min
20	bwd_header_size_max	49	fwd_iat.avg	78	idle.max
21	flow_FIN_flag_count	50	fwd_iat.std	79	idle.tot
22	flow_SYN_flag_count	51	bwd_iat.min	80	idle.avg
23	flow_RST_flag_count	52	bwd_iat.max	81	idle.std
24	fwd_PSH_flag_count	53	bwd_iat.tot	82	fwd_init_window_size
25	bwd_PSH_flag_count	54	bwd_iat.avg	83	bwd_init_window_size
26	flow_ACK_flag_count	55	bwd_iat.std	84	fwd_last_window_size
27	fwd_URG_flag_count	56	flow_iat.min	85	traffic_category
28	bwd_URG_flag_count	57	flow_iat.max	86	Label
29	flow_CWR_flag_count	58	flow_iat.tot		

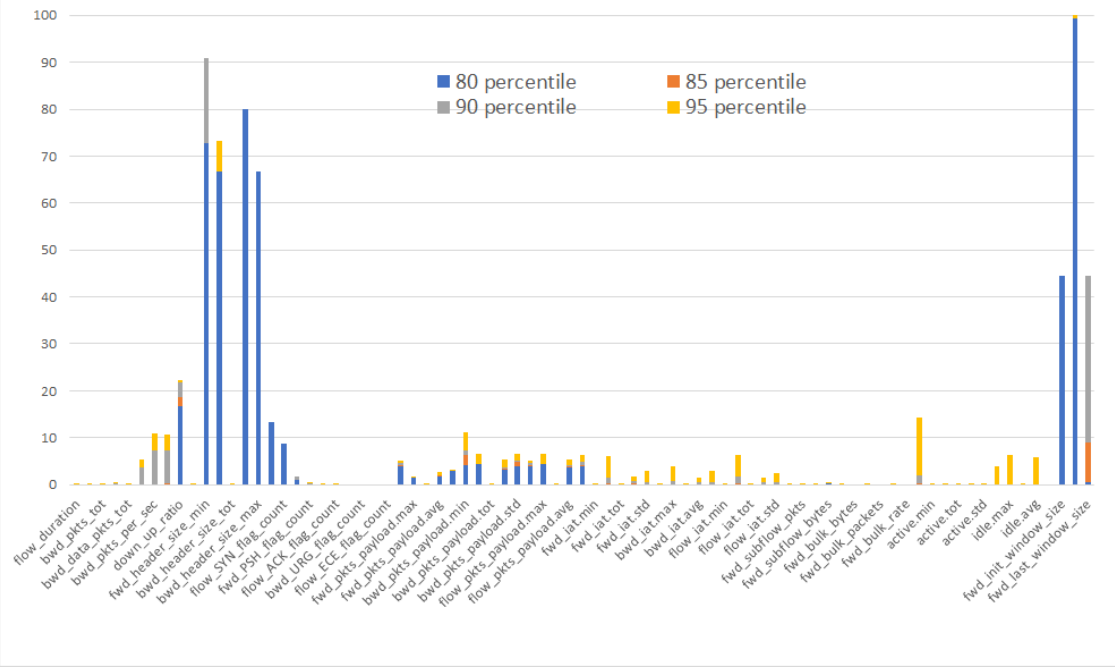

Figure 4. Most of the features are skewed, where the value of the 95th percentile is less than ten percent of the maximum value.

4.9. Performance Analysis

We conducted an examination using a basic performance analysis using four machine learning algorithms. Table 4 displays the performance of the examination results in Accuracy, Balanced Accuracy, Precision, Recall, and F1.

Table 4. Basic Performance Analysis.

Algorithm	Accuracy	Balanced Accuracy	Precision	Recall	F1
KNN	0.98	0.94	0.86	0.90	0.88
MLP	0.99	0.99	0.99	0.99	0.99
SVM	0.99	0.99	0.99	0.98	0.99
RF	0.99	0.99	0.99	0.99	0.99

5. Comparison of KDD99, UNSW-NB15, CICIDS-2017, and HIKARI-2021

Table 5 shows an analysis comparison among KDD99, UNSW-NB15, CICIDS-2017, and HIKARI-2021. The table consists of seven parameters: the number of unique IP addresses, simulation, duration of the data being captured, format data being collected, attack category, feature extraction tools, and the number of features extracted from each dataset. The number of unique IP addresses of CICIDS-2017 and HIKARI-2021 were from the unique destination IP addresses from the dataset. Partial means that the dataset is mixed between a simulation or synthetic and real-network environment.

Table 5. The dataset comparison of KDD99, UNSW-NB15, CICIDS-2017, and HIKARI-2021 [68].

Parameters	KDD99	UNSW-NB15	CICIDS-2017	HIKARI-2021
Number of unique IP address	11	45	16,960	7991
Simulation	Yes	Yes	Partial	Partial
Duration of the data being captured	5 weeks	16 h	65 h	39 h
Format of the data collected	3 types (tcpdump, BSM, dumpfile)	pcap files	pcap files	pcap files
Number of Attack categories	4	9	7	4
Feature extraction tools	Bro-IDS tool	Argus, Bro-IDS	CICFlowmeter	Zeek-IDS, python tools
Number of features	42	49	80	86

6. Conclusions and Future Work

Publicly available up-to-date datasets to benchmark and compare among IDS are important, especially as the network traffic is changing over time. There are two main contributions of this paper. First, we made a new requirement for building new datasets which are lacking in the existing datasets, such as anonymization, payload, ground-truth, encryption, and a practical method to implement it. Anonymizing certain data will prevent privacy issues, while capturing with the payload will enrich the information that we can collect for detecting malicious traffic within encrypted traffic. Providing the ground-truth data is crucial, so no unlabeled attack is recorded in the dataset. The lack of existing datasets with encrypted traffic, even though most present-day traffic use it for delivering attacks, has become our concern. Second, we generated a new IDS dataset called HIKARI-2021, which covers the network traffic with encrypted traces. The datasets were produced with a mix of ground-truth data, which are missing in the existing IDS datasets. The datasets are available publicly [68]. We adopted more than 80 features from CICIDS-2017 and added more features as a reference, such as a source IP address (originh), source port (originp), destination IP address, and destination port. We labeled each flow as benign or attack, where benign has two categories (Benign or Background), while attack has four (Bruteforce, Bruteforce-XML, Probing, and XMRIGCC CryptoMiner).

We want to highlight what makes our dataset different from the existing IDS datasets. This is based on our proposed ideal requirements. The first is from the content requirements,

such as complete capture, for which we provide all traces with pcap files (e.g., background traffic, benign, and attack); the payload is provided with the exception that we anonymize the background traffic, while anonymity is part of a requirement to preserve privacy.

The ground-truth and labeled are manually evaluated based on the source IP address, source port, destination IP address, destination port, and protocol. This process is to make sure that no unlabeled attack is in the ground-truth. The last requirement is encryption. This one of the most important requirements, as we know that unknown malicious traffic uses these attack vectors to deliver attacks.

The second is process requirement. It is to ensure that researchers can follow the guidelines to create their dataset. The information on how to generate the synthetic attacks and the network configuration should be available. We provided the scripts on how to capture and generate the synthetic attacks from the attack profile. The tools for mimicking human interaction, such as browsing and clicking random links, are available. These two profiles, the attack profile and benign profile, are important for producing new data if researchers want to add more attack vectors and update the traffic with their own needs. The labeling process script to produce ground-truth data is provided. The process requirement can be implemented in the controlled environment so that researchers can make new datasets based on their network configuration. For a basic evaluation, we examined the performance of the HIKARI-2021 dataset in terms of Accuracy, Balanced Accuracy, Precision, Recall, and F1, using four machine learning algorithms.

In the future, we would like to extend our observation with the background traffic and add an evaluation. Because background traffic is uncertain and not labeled in the data, the possible approach for evaluation is using machine learning with unsupervised learning. Furthermore, we would like to make performance comparisons with the existing datasets and proceed with the analysis of application identification, as this is important because malicious traffic may be disguised using reserved ports to bypass firewalls or IDS and blend with normal network activity.

Author Contributions: Conceptualization, A.F.; data curation, A.F.; funding acquisition, K.T., J.M.; investigation, A.F.; methodology, A.F.; analysis, A.F.; resources, K.T., J.M.; supervision, K.T., J.M.; validation, A.H.T.; writing—original draft, A.F.; writing—review and editing, A.H.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.4782195 (accessed on 10 May 2021).

Acknowledgments: Our sincere appreciation for the Indonesia government, particularly LPDP (Lembaga Pengelola Dana Pendidikan Indonesia = Indonesia Endowment Fund for Education) that provides the scholarship to study at Keio University.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tavallaee, M.; Stakhanova, N.; Ghorbani, A.A. Toward credible evaluation of anomaly-based intrusion-detection methods. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2010**, *40*, 516–524. [CrossRef]
2. Aviv, A.J.; Haeberlen, A. Challenges in Experimenting with Botnet Detection Systems. In Proceedings of the 4th Workshop on Cyber Security Experimentation and Test (CSET 11), San Francisco, CA, USA, 8 August 2011.
3. Velan, P.; Čermák, M.; Čeleda, P.; Drašar, M. A survey of methods for encrypted traffic classification and analysis. *Int. J. Netw. Manag.* **2015**, *25*, 355–374. [CrossRef]
4. De Lucia, M.J.; Cotton, C. Identifying and detecting applications within TLS traffic. In Proceedings of the Cyber Sensing 2018, Orlando, FL, USA, 15–19 April 2018; Volume 10630. [CrossRef]
5. Kaur, S.; Singh, M. Automatic attack signature generation systems: A review. *IEEE Secur. Priv.* **2013**, *11*, 54–61. [CrossRef]
6. Ahmed, M.; Naser Mahmood, A.; Hu, J. A survey of network anomaly detection techniques. *J. Netw. Comput. Appl.* **2016**, *60*, 19–31. [CrossRef]
7. Zeek IDS. 2021. Available online: https://zeek.org (accessed on 10 May 2021).

8. Shiravi, A.; Shiravi, H.; Tavallaee, M.; Ghorbani, A.A. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. *Comput. Secur.* **2012**, *31*, 357–374. [CrossRef]
9. Moustafa, N.; Slay, J. The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set. *Inf. Secur. J. Glob. Perspect.* **2016**, *25*, 18–31. [CrossRef]
10. 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.* **2018**, *73*, 411–424. [CrossRef]
11. Lippmann, R.P.; Fried, D.J.; Graf, I.; Haines, J.W.; Kendall, K.R.; McClung, D.; Weber, D.; Webster, S.E.; Wyschogrod, D.; Cunningham, R.K.; et al. Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation. In Proceedings of the DARPA Information Survivability Conference and Exposition (DISCEX'00), Hilton Head, SC, USA, 25–27 January 2000; Volume 2; pp. 12–26. [CrossRef]
12. Siddique, K.; Akhtar, Z.; Khan, F.A.; Kim, Y. KDD Cup 99 data sets: A perspective on the role of data sets in network intrusion detection research. *Computer* **2019**, *52*, 41–51. [CrossRef]
13. Özgür, A.; Erdem, H. A review of KDD99 dataset usage in intrusion detection and machine learning between 2010 and 2015. *PeerJ* **2016**, *4*, e1954v1. [CrossRef]
14. Luo, C.; Wang, L.; Lu, H. Analysis of LSTM-RNN based on attack type of kdd-99 dataset. In Proceedings of the International Conference on Cloud Computing and Security, Haikou, China, 8–10 June 2018; Springer: Cham, Switzerland, 2018; pp. 326–333. [CrossRef]
15. Fukuda Lab Mawi Archive. 2021. Available online: https://fukuda-lab.org/mawilab (accessed on 10 May 2021).
16. Fontugne, R.; Borgnat, P.; Abry, P.; Fukuda, K. Mawilab: Combining diverse anomaly detectors for automated anomaly labeling and performance benchmarking. In Proceedings of the Co-NEXT '10: Conference on Emerging Networking EXperiments and Technologies, Philadelphia, PA, USA, 30 November–3 December 2010; pp. 1–12. [CrossRef]
17. Hafsa, M.; Jemili, F. Comparative study between big data analysis techniques in intrusion detection. *Big Data Cogn. Comput.* **2019**, *3*, 1. [CrossRef]
18. Kim, J.; Sim, C.; Choi, J. Generating labeled flow data from MAWILab traces for network intrusion detection. In Proceedings of the ACM Workshop on Systems and Network Telemetry and Analytics, Phoenix, AZ, USA, 25 June 2019; pp. 45–48. [CrossRef]
19. CAIDA Datasets. 2021. Available online: https://www.caida.org/catalog/datasets/completed-datasets/ (accessed on 10 May 2021).
20. Jonker, M.; King, A.; Krupp, J.; Rossow, C.; Sperotto, A.; Dainotti, A. Millions of targets under attack: A macroscopic characterization of the DoS ecosystem. In Proceedings of the 2017 Internet Measurement Conference, London, UK, 1–3 November 2017; pp. 100–113. [CrossRef]
21. Lutscher, P.M.; Weidmann, N.B.; Roberts, M.E.; Jonker, M.; King, A.; Dainotti, A. At home and abroad: The use of denial-of-service attacks during elections in nondemocratic regimes. *J. Confl. Resolut.* **2020**, *64*, 373–401. [CrossRef]
22. Hinze, N.; Nawrocki, M.; Jonker, M.; Dainotti, A.; Schmidt, T.C.; Wählisch, M. On the potential of BGP flowspec for DDoS mitigation at two sources: ISP and IXP. In Proceedings of the ACM SIGCOMM 2018 Conference on Posters and Demos, Budapest, Hungary, 20–25 August 2018; pp. 57–59. [CrossRef]
23. Hesselman, C.; Kaeo, M.; Chapin, L.; Claffy, K.; Seiden, M.; McPherson, D.; Piscitello, D.; McConachie, A.; April, T.; Latour, J.; et al. The DNS in IoT: Opportunities, Risks, and Challenges. *IEEE Internet Comput.* **2020**, *24*, 23–32. [CrossRef]
24. Barbosa, R.R.R.; Sadre, R.; Pras, A.; van de Meent, R. *Simpleweb/University of Twente Traffic Traces Data Repository*; Centre for Telematics and Information Technology, University of Twente: Enschede, The Netherlands, 2010.
25. Haas, S. Security Monitoring and Alert Correlation for Network Intrusion Detection. Ph.D. Thesis, Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky, Hamburg, Germany, 2020.
26. Wang, J.; Paschalidis, I.C. Botnet detection based on anomaly and community detection. *IEEE Trans. Control Netw. Syst.* **2016**, *4*, 392–404. [CrossRef]
27. Čermák, M.; Čeleda, P. Detecting Advanced Network Threats Using a Similarity Search. In Proceedings of the IFIP International Conference on Autonomous Infrastructure, Management and Security, Munich, Germany, 20–23 June 2016; Springer: Cham, Switzerland, 2016; pp. 137–141. [CrossRef]
28. Tavallaee, M.; Bagheri, E.; Lu, W.; Ghorbani, A.A. A detailed analysis of the KDD CUP 99 data set. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, Ottawa, ON, Canada, 8–10 July 2009; pp. 1–6. [CrossRef]
29. Liu, J.; Kantarci, B.; Adams, C. Machine learning-driven intrusion detection for contiki-NG-based IoT networks exposed to NSL-KDD dataset. In Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning, Linz, Austria, 13 July 2020; pp. 25–30. [CrossRef]
30. Gao, Y.; Wu, H.; Song, B.; Jin, Y.; Luo, X.; Zeng, X. A distributed network intrusion detection system for distributed denial of service attacks in vehicular ad hoc network. *IEEE Access* **2019**, *7*, 154560–154571. [CrossRef]
31. Su, T.; Sun, H.; Zhu, J.; Wang, S.; Li, Y. BAT: Deep learning methods on network intrusion detection using NSL-KDD dataset. *IEEE Access* **2020**, *8*, 29575–29585. [CrossRef]
32. Ding, Y.; Zhai, Y. Intrusion detection system for NSL-KDD dataset using convolutional neural networks. In Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence, Shenzhen, China, 8–10 December 2018; pp. 81–85. [CrossRef]

33. Zhang, C.; Ruan, F.; Yin, L.; Chen, X.; Zhai, L.; Liu, F. A deep learning approach for network intrusion detection based on NSL-KDD dataset. In Proceedings of the 2019 IEEE 13th International Conference on Anti-Counterfeiting, Security, and Identification (ASID), Xiamen, China, 25–27 October 2019; pp. 41–45. [CrossRef]
34. IMPACT Cyber Trust. 2021. Available online: https://www.impactcybertrust.org/ (accessed on 10 May 2021).
35. UMass Trace Repository. 2021. Available online: http://traces.cs.umass.edu/index.php/Network/Network (accessed on 10 May 2021).
36. Nasr, M.; Bahramali, A.; Houmansadr, A. Deepcorr: Strong flow correlation attacks on tor using deep learning. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, Toronto, ON, Canada, 15–19 October 2018; pp. 1962–1976. [CrossRef]
37. Bissias, G.; Levine, B.N.; Liberatore, M.; Prusty, S. Forensic identification of anonymous sources in oneswarm. *IEEE Trans. Dependable Secur. Comput.* **2015**, *14*, 620–632. [CrossRef]
38. Eagle, N.; Pentland, A.S. Reality mining: Sensing complex social systems. *Pers. Ubiquitous Comput.* **2006**, *10*, 255–268. [CrossRef]
39. Kyoto Dataset. 2021. Available online: http://www.takakura.com/Kyoto_data (accessed on 10 May 2021).
40. 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, Salzburg, Austria, 10 April 2011; pp. 29–36. [CrossRef]
41. Singh, A.P.; Kaur, A. Flower pollination algorithm for feature analysis of kyoto 2006+ data set. *J. Inf. Optim. Sci.* **2019**, *40*, 467–478. [CrossRef]
42. Salo, F.; Nassif, A.B.; Essex, A. Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection. *Comput. Netw.* **2019**, *148*, 164–175. [CrossRef]
43. Zuech, R.; Khoshgoftaar, T.; Seliya, N.; Najafabadi, M.; Kemp, C. A New Intrusion Detection Benchmarking System. 2015. Available online: https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS15/paper/view/10368 (accessed on 30 April 2021).
44. Krystosek, P.; Ott, N.M.; Sanders, G.; Shimeall, T. *Network Traffic Analysis with SiLK*; Technical Report; Carnegie-Mellon Univeristy: Pittsburgh, PA, USA, 2019.
45. Snort IDS. 2021. Available online: https://snort.org/ (accessed on 10 May 2021).
46. Rajagopal, S.; Hareesha, K.S.; Kundapur, P.P. Feature Relevance Analysis and Feature Reduction of UNSW NB-15 Using Neural Networks on MAMLS. In *Advanced Computing and Intelligent Engineering*; Springer: Singapore, 2020; pp. 321–332. [CrossRef]
47. Kasongo, S.M.; Sun, Y. Performance Analysis of Intrusion Detection Systems Using a Feature Selection Method on the UNSW-NB15 Dataset. *J. Big Data* **2020**, *7*, 1–20. [CrossRef]
48. Kumar, V.; Das, A.K.; Sinha, D. Statistical analysis of the UNSW-NB15 dataset for intrusion detection. In *Computational Intelligence in Pattern Recognition*; Springer: Singapore, 2020; pp. 279–294. [CrossRef]
49. Rajagopal, S.; Kundapur, P.P.; Hareesha, K.S. A stacking ensemble for network intrusion detection using heterogeneous datasets. *Secur. Commun. Netw.* **2020**, *2020*. [CrossRef]
50. Radhakrishnan, C.; Karthick, K.; Asokan, R. Ensemble Learning based Network Anomaly Detection using Clustered Generalization of the Features. In Proceedings of the 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 18–19 December 2020; pp. 157–162. [CrossRef]
51. Yilmaz, I.; Masum, R.; Siraj, A. Addressing imbalanced data problem with generative adversarial network for intrusion detection. In Proceedings of the 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI), Las Vegas, NV, USA, 11–13 August 2020; pp. 25–30. [CrossRef]
52. Sharafaldin, I.; Gharib, A.; Lashkari, A.H.; Ghorbani, A.A. Towards a reliable intrusion detection benchmark dataset. *Softw. Netw.* **2018**, *2018*, 177–200. [CrossRef]
53. Lashkari, A.H.; 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, Porto, Portugal, 19–21 February 2017; pp. 253–262. [CrossRef]
54. Kshirsagar, D.; Kumar, S. An efficient feature reduction method for the detection of DoS attack. *ICT Express* **2021**. [CrossRef]
55. Kshirsagar, D.; Kumar, S. An ensemble feature reduction method for web-attack detection. *J. Discret. Math. Sci. Cryptogr.* **2020**, *23*, 283–291. [CrossRef]
56. Tama, B.A.; Nkenyereye, L.; Islam, S.R.; Kwak, K.S. An enhanced anomaly detection in web traffic using a stack of classifier ensemble. *IEEE Access* **2020**, *8*, 24120–24134. [CrossRef]
57. Yulianto, A.; Sukarno, P.; Suwastika, N.A. Improving AdaBoost-based Intrusion Detection System (IDS) Performance on CIC IDS 2017 Dataset. *J. Phys. Conf. Ser.* **2019**, *1192*, 012018. [CrossRef]
58. Stiawan, D.; Idris, M.Y.B.; Bamhdi, A.M.; Budiarto, R. CICIDS-2017 dataset feature analysis with information gain for anomaly detection. *IEEE Access* **2020**, *8*, 132911–132921. [CrossRef]
59. Cordero, C.G.; Vasilomanolakis, E.; Wainakh, A.; Mühlhäuser, M.; Nadjm-Tehrani, S. On generating network traffic datasets with synthetic attacks for intrusion detection. *ACM Trans. Priv. Secur.* **2021**, *24*, 1–39. [CrossRef]
60. Kenyon, A.; Deka, L.; Elizondo, D. Are public intrusion datasets fit for purpose characterising the state of the art in intrusion event datasets. *Comput. Secur.* **2020**, *99*, 102022. [CrossRef]
61. Varet, A.; Larrieu, N. Realistic Network Traffic Profile Generation: Theory and Practice. *Comput. Inf. Sci.* **2014**, *7*, 1–16. [CrossRef]
62. Selenium Python. 2021. Available online: https://selenium-python.readthedocs.io (accessed on 14 May 2021).

63. Alexa. 2021. Available online: https://www.alexa.com (accessed on 14 May 2021).
64. Daniel Miessler 10k Most Common Credentials. 2021. Available online: https://github.com/danielmiessler (accessed on 14 May 2021).
65. Droopescan. Available online: https://github.com/droope/droopescan (accessed on 30 April 2021).
66. Joomscan. Available online: https://github.com/OWASP/joomscan/releases (accessed on 30 April 2021).
67. Fan, J.; Xu, J.; Ammar, M.H.; Moon, S.B. Prefix-preserving IP address anonymization: Measurement-based security evaluation and a new cryptography-based scheme. *Comput. Netw.* **2004**, *46*, 253–272. [CrossRef]
68. Ferriyan, A.; Thamrin, A.H.; Takeda, K.; Murai, J. *HIKARI-2021: Generating Network Intrusion Detection Dataset Based on Real and Encrypted Synthetic Attack Traffic*; Zenodo: Geneva, Switzerland, 2021. [CrossRef]