A Comparative Analysis of Datasets for Intrusion Detection in Software-Defined Networks

Francesco Di Gennaro[1,2,*,†], Alessandro Cucchiarelli[2,†], Christian Morbidoni[3,†] and Luca Spalazzi[2,†]

[1] *IMT School for Advanced Studies, Lucca, Italy*
[2] *Università Politecnica delle Marche, Ancona, Italy*
[3] *Università degli Studi G. d'Annunzio, Pescara, Italy*

Abstract

Software-Defined Networking (SDN) offers centralized management, programmability, flexibility and scalability but has significant security risks, especially DDoS attacks against the SDN controller, threatening network availability. Machine learning (ML) and deep learning (DL) show promise in mitigating these threats, but their success depends on available datasets quality. Existing SDN datasets often focus narrowly on specific DDoS scenarios or synthetic environments, limiting their real-world applicability. This paper analyzes SDN threats datasets, evaluating their methodologies, features and ML applications. It highlights strengths like realistic traffic emulation and accessibility, alongside limitations such as narrow attack coverage and synthetic biases. A roadmap is proposed to guide the generation of new datasets, emphasizing diverse attacks, richer features, realistic augmentation and public access to enable robust ML/DL-based SDN security solutions.

Keywords

Dataset, IDS, SDN, DDoS, Machine Learning

1. Introduction

Software-Defined Networking (SDN) is an emerging technology that offers centralized control, programmability, scalability and flexibility in the management of complex networks. SDN enables dynamic resource allocation and simplified network management through the decoupling of the control and data planes.

However, this centralized architecture also introduces significant security vulnerabilities[1], particularly in the form of Distributed Denial of Service (DDoS) attacks[2], which can overwhelm the SDN controller resources and compromise the availability of network access. For the mentioned reason, there is a need to mitigate these vulnerabilities to ensure the resilience and robustness of SDN deployments, especially because these networks are more and more complex and extended. Researchers have focused on developing intrusion detection systems and mitigation techniques to mitigate these threats, focusing on leveraging machine learning and deep learning models[3].

At the core of these solutions there are datasets that simulate real-world SDN traffic conditions, including both benign and malicious traffic[3]. These datasets offer a benchmark for ML/DL-based systems by offering insights into network behavior and enabling improved attack detection. However, they vary in design, scope, attack target and features, exhibiting both strengths and limitations. This stresses the need for enhanced and representative data to successfully ensure modern SDN security.

The aim of the paper is to examine the methodologies used to create these datasets, the network topologies and tools employed and the traffic captured, highlighting their contributions to the field and the gaps that remain to be filled. By critically evaluating these datasets, the aim is to provide a foundation for future research in developing resilient, adaptable and scalable security solutions for SDN

Joint National Conference on Cybersecurity (ITASEC & SERICS 2025), February 03-8, 2025, Bologna, IT

[*] Corresponding author.

[†] These authors contributed equally.

s1117033@pm.univpm.it (F. Di Gennaro); a.cucchiarelli@univpm.it (A. Cucchiarelli); christian.morbidoni@unich.it (C. Morbidoni); l.spalazzi@univpm.it (L. Spalazzi)

environments.
The roadmap of the paper is the following: Section 2 discusses works related to creation of dataset in SDN environment. The section 3 presents an analysis of the paper mentioned in the Section 2 based on network architecture. In Section 4 the discussion based on elements of Section 2 and 3 is conducted. Section 5 deals with the conclusion and perspectives of future works. The Appendix A provides the network architectures employed by each paper.

2. Literature Review

Software-Defined Networks (SDNs) present peculiar security challenges, particularly due to their centralized architecture, which increases vulnerability to attacks targeting the availability of network resources. To address these threats, datasets have been developed to support research in intrusion detection and machine learning (ML)-based security solutions, each with strengths and limitations.
In the sub-paragraph 2.8 the Table 1 provides an overview of the features that will be detailed in the following subsection for each dataset.

2.1. Niyaz et al. (NSJ) Dataset

The dataset[4] was published in 2017. It analyzes DDoS attack detection in SDN across different vectors with a combination of various flooding techniques on the following protocols: TCP, UDP and ICMP. Seven specific types of attacks were examined. These attacks were simulated with tools like hping3[1] in a testing environment, varying packet sizes and rates to overwhelm SDN infrastructure, even though specific rates were not quantified.
Although the study's primary focus is DDoS attacks, it also acknowledges the potential applicability of its methodology to other intrusion types, such as network scanning and malware propagation.
Traffic was collected over three days from a real-world wireless home network for benign traffic and a segregated lab environment for malicious traffic. The data was captured using tcpdump, manipulated with bit-twist and replayed via tcpreplay. The experimental setup for malign traffic generation utilized a SDN testbed consisting of a POX[2] controller, OpenFlow-enabled switches and virtual machines running on the hypervisor VMware ESXi, which ensures centralized traffic monitoring and feature extraction as shown in the Figure 1.
The POX controller hosted application employed for the detection and enabled communication based on Openflow protocol, through central management of all the traffic channels. A set of 68 features was extracted from TCP, UDP and ICMP traffic, including packet counts, entropy and flags like SYN and ACK.
Although the dataset combines real-world and simulated traffic, it is not explicitly available for public use.
As regards the detection, the system employed a Stacked Autoencoder (SAE) both for feature reduction and for classification, outperforming other models like softmax classifiers and neural networks.
The key contribution of the paper is the simulation of different attacks using TCP, UDP and ICMP protocols. The SDN testbed using the POX controller demonstrated its ability to centralize monitoring and feature extraction, with 68 features that offer input for machine learning based detection.

2.2. Zerbini et al. (ZCA) Dataset

The authors of the paper[5] explore crucial aspects of security in network environments, with focus on the analysis, detection and mitigation of both DDoS and port scan attacks. The research provides a framework for network security research, with emphasis on dataset generation, traffic analysis and the adoption of machine learning models. It investigates various DDoS attack typologies such as TCP-SYN, UDP and ICMP floods, as well DDoS attacks on the application layer targeting web servers. The research also considers attack rates, ranging from high rate floods that saturate bandwidth to low

[1] Documentation available at https://www.kali.org/tools/hping3/.
[2] Documentation available at https://noxrepo.github.io/pox-doc/html/.

rate and stealthy attacks aimed at evading detection by targeting resources on the application layer.
In addition to DDoS attacks, the dataset contains other types of intrusion, such as port scan.

Traffic was collected using a virtualized testbed simulating realistic conditions, with tcpdump that is used to collect packet data and statistical features. The following simulation tools were employed: Hping3, LOIC[3], Metasploit, Hydra[4] and SQLMap[5] in order to generate traffic, while the network architecture was based on a tree topology with a root switch and four subnets consisting of 20 hosts each. Open vSwitch was employed for its compatibility with OpenFlow and robust capabilities in handling network traffic. The controller platform leveraged the POX controller, integrating discrete wavelet transforms (DWT), anomaly detection and mitigation modules.

The network architecture is represented in the Figure 1. Although the two-day simulation produced only six features and focused primarily on traditional DDoS attacks, the dataset supports research in network security and intrusion detection. It is noted to be available for further study. Various machine learning models, including Decision Trees, Random Forests, Naive Bayes, K-Nearest Neighbors, Support Vector Machines and Multi-Layer Perceptrons were applied to assess the dataset, demonstrating its utility in advancing machine learning applications for SDN security.

The study provides a systematic approach to dataset generation and traffic analysis for network security, emphasizing diverse DDoS attack typologies and application-layer attacks. It also explores additional threats like password guessing and web application exploits. Although in a virtualized testbed, the dataset deals with realistic network scenarios. The network architecture, based on a tree topology with Open vSwitch and a POX controller, integrates anomaly detection mechanisms. Even though if focuses on six features and traditional DDoS scenarios, the dataset demonstrates its applicability for research, with machine learning models like Random Forest and SVM that show high performances for treaths detection.

2.3. InSDN Dataset

The InSDN dataset[6] offers a representation of security challenges in SDN environments. It addresses different attack types and includes advanced simulation tools. It spans multiple DDoS attack modes, such as TCP-SYN, UDP, ICMP floods and attacks like HTTP Flood conducted using Slowloris[6] and Torshammer[7].

Beyond DDoS, the dataset addresses also other significant threats, including password-guessing attacks, web application vulnerabilities, botnet and exploitation of services like Samba.

Traffic collection was performed in a controlled virtualized environment using Tcpdump for packet capture and CICFlowMeter for feature extraction, encompassing both normal and malicious traffic against various SDN layers.

The testbed's network architecture consists of four virtual machines, including a Kali Linux attacker virtual machine, an ONOS[8]-based SDN controller, a Mininet[9] emulator with Open vSwitch[10] and a Metasploitable-2 machine for vulnerability testing, as it is possible to see in the scheme of network architecture in Figure 2. Over 80 features were initially extracted and later reduced to 48 SDN-specific metrics.

The dataset is publicly available, supporting different research needs. The dataset was tested using common machine learning algorithms that are: Decision Tree, Random Forests, AdaBoost, K-Nearest Neighbor, Naive Bayes, Support Vector Machines and Multi-Layer Perceptrons. These models have demonstrated their effectiveness in detecting common attacks like DDoS and probing, even though challenges remain in identifying complex patterns such as User-to-Root (U2R) attack.

[3] Documentation available at: https://github.com/NewEraCracker/LOIC/releases.
[4] Documentation available at: https://www.kali.org/tools/hydra/.
[5] Documentation available at: https://sqlmap.org/.
[6] Documentation available at://github.com/gkbrk/slowloris.
[7] Documentation available at: https://github.com/Karlheinzniebuhr/torshammer.
[8] Open Network Operating System, documentation available at https://opennetworking.org/onos/.
[9] Virtual network simulator available at https://mininet.org/.
[10] Multilayer virtual switch available at https://www.openvswitch.org/.

The InSDN dataset, in conclusion, is a robust and diverse resource for advancing SDN security research. It captures a wide range of attack types, including both volumetric and DDoS attacks on application layer, as well as other significant threats like password guessing, botnet activities and web attacks. The virtualized environment ensures realistic traffic conditions, with a rich set of features that were extracted to support detailed analysis. Its publicly available status enhances accessibility, making it a valuable resource for machine learning research. Evaluations with different machine learning models like Random Forest and AdaBoost, demonstrate the dataset's capability in detecting common threats, but some aspects needed to be deepened, such as the detection of complex patterns like U2R exploitation.

2.4. Novaes et al. (NCLP) Dataset

The dataset proposed by NCLP[7] deals with the security problems of Software-Defined Networking, with emphasis on both DDoS and Portscan attacks. The dataset provides a valuable base for SDN security research, where various attack scenarios and methodologies are addressed. It encompasses many types of DDoS attacks, including UDP, SYN, TFTP, DNS, NTP floods, WebDDoS and Apache remote memory exhaustion attacks. In addition to DDoS, Portscan attacks are conducted, simulating attempts to identify and exploit open ports.
The dataset simulates high and low intensity DDoS attacks, using the Scapy tool to generate traffic with different rates and durations, trying to reflect real-world conditions. Traffic was collected in a Mininet environment over two days: normal traffic was captured on the first day, while the second day included mixed traffic, analyzed in intervals of one second to grant sufficient granularity.
The Floodlight SDN controller, an OpenFlow compliant platform, managed the network, collecting flow and port statistics for analysis. The topology of the network consisted of a star with six switches and 120 hosts as shown in the architecture in Figure 2, trying to simulate real SDN.
Six features were captured, including bits per second, packets per second and entropy measures for source and destination IPs and ports, in order to train machine learning models.
The dataset is publicly available in order to increase its utility. The authors evaluated the dataset using common machine learning models, but also including a novel LSTM-Fuzzy Logic model, which outperformed traditional algorithms like k-Nearest Neighbors, Support Vector Machines and Multi-Layer Perceptrons. This last model outperformed the others in terms of accuracy as regards the prediction of normal traffic and detection of anomalies, making the dataset a good resource for developing security models for SDN environments.
This dataset offers a good basis for SDN security research, addressing various DDoS attack types, which include volumetric and application based attacks and Portscan scenarios. By simulating high and low intensity attacks in an environment simulated with Mininet, the authors provide traffic data that is suitable for machine learning applications. The star topology controlled by a single appliance, that is the Floodlight SDN controller, assures scalability and applicability to real world conditions.
Six features enable the traffic analysis, while the accessibility of the dataset extends its utility. The advanced models used, like the LSTM-Fuzzy Logic, demonstrate its potential as regards the detection of anomalies and normal traffic, making it a valuable metric for developing resilient SDN security mechanisms.

2.5. Yungaicela et al.'s (Y-NV-RP-DJM-C) SDN-SlowRate DDoS Dataset

The dataset[8] was published in 2022. It was designed for deep learning systems which deal with application layer attacks.
The research is based on a real data center topology managed by the ONOS SDN controller, with a focus on application based layer and slow-rate DDoS attack. The slow HTTP read attack is the primary DDoS variant considered, which is designed to deplete server resources by opening long duration connections without triggering timeouts.
Even though it is innovative in concentrating on slow rate attacks, the dataset does not take into consideration high rate and volumetric DDoS cases. Traffic capture was performed on a physical testbed

representing SDN-enabled physical structure and it consists of HP Aruba and NEC hardware. Traffic data was captured with CICFlowMeter, with detailed packet level and flow level statistics.

Legitimate traffic sources consisted of FTP and video streaming, allowing also the evaluation of false positives and system performance. The network architecture tries to simulate a realistic data center with spine leaf topology, real appliances for traffic simulation and generation. The ONOS controller grants centralized management, with IDS and IPS integration supported as shown in the Figure 3.

The dataset is available on Internet and presents 13 features for each instance. The authors applied LSTM model for the evaluation of the performances and demonstrated its quality for identifying slow rate DDoS attacks in SDN.

In conclusion, the SDN-SlowRate-DDoS dataset tries to fill the gap in SDN security analyzing slow rate DDoS attacks. Its realistic configuration managed by the ONOS controller ensures applicability. Moreover, the feature set of 13 flow metrics provides a good basis for testing IDS systems. However, its focus on slow rate attacks limits the applicability to different DDoS scenarios, providing chances for future enhancements.

2.6. Aladaileh et al. (AAHHBA) Dataset

This dataset[9] was published in 2022 and was obtained via a POX controller and 64 hosts in a llinear topology. SDN domain is targeted with low rate and high rate DDoS attacks. The dataset takes into consideration different DDoS attack types, that is to say situations with single or multiple attackers targeting one or more victims.

The traffic throughput are defined as low rate (5 packets per second) and high rate (33 packets per second) for simulating various attack intensities. Dataset generation is achieved in Mininet to cover the simulation of the network topology with a POX controller managing SDN activities and Kali Linux used for attack traffic deployment.

Traffic data is sampled during 60 minutes simulations at 5 seconds intervals across eight attack conditions. Even though the paper is innovative for the use of mixed traffic rates, the dataset contains only seven features, among them there are source and destination IP addresses to calculate entropy values for packet randomness over time. The topology of the network includes one single SDN controller, OpenFlow switches and 64 hosts, thus providing flexibility in simulating diverse attack configurations.

The study does not account for any other forms of attacks apart from DDoS and relies on statistical methods such as Renyi joint entropy for its analysis, with future possibility to include machine learning models for the generation of rules, even though the dataset is not open source. However, its limited feature set and focus on entropy leave room for further enhancement in the future.

The paper provides a datased focused on low rate and high rate DDoS attacks in SDN environments, based on a simulated network topology using Mininet and a POX controller. The schematic of the network architecture is depicted in Figure 3.

It demonstrates various attack scenarios, from a single attacker to multiple ones and categorizes traffic based on its intensity. Traffic data collection and the analysis based on entropy provide insights into packet randomness over time. Nevertheless, the dataset relies on only seven features, moreover the exclusion of some attacks, such as TCP and ICMP floods, limits its application. The accessibility would be required to assure its utility in the scientific research.

Despite this, the innovative aspects of this paper, that is to say the focus on statistical methodology and potential integration with machine learning, makes it useful for further research in SDN security.

2.7. Ahuja et al. (ASMK) Dataset

ASMK[10] article deals with DDoS attack detection in Software-Defined Networking (SDN) through automated machine learning-based attack detection for enhancing security. The research proposes a machine learning-based framework for DDoS detection in SDN architecture, keeping in view real-time responsiveness, efficiency and scalability for SDN controller security.

It primarily targets volumetric DDoS attacks that exhaust controller resources, including TCP SYN, UDP and ICMP flood attacks. The framework simulates both high-rate and moderate-rate attack scenarios, emulating realistic volumetric traffic patterns.

Traffic was emulated in a Mininet-based SDN testbed, both generating normal traffic and malicious traffic. Certain key tools like Mininet, hping3 and Wireshark supported traffic generation as well as traffic analysis.

The networking infrastructure consisted of a centralized tree topology with OpenFlow-switched nodes and a single SDN controller, in emulation of real-world exposure to controller-directed attacks. The network architecture is described in Figure 4. The study utilized the POX controller for traffic monitoring and data plane interaction, extracting 23 features such as packet sizes, flow durations and protocol-specific statistics to support machine learning model training.

Evaluations were done with Decision Tree, Random Forest and Neural Networks, with Random Forests being the best at classifying benign against malicious flows. Although the dataset includes synthetic and simulated traffic for complete analysis, it is not publicly released. This limitation, along with the focus of the framework only towards DDoS attacks, helps to highlight potential areas for future research, particularly in broadening applicability to more general network anomalies.

In conclusion, the study offers a structured approach to DDoS detection in SDN , emphasizing scalability and efficiency. The focus on SDN-specific attack scenarios and machine learning applications provides valuable insights, although the dataset's unavailability and limited attack diversity suggest opportunities for further research and dataset expansion.

2.8. High and Low-Rates Dataset-Based DDoS Attacks Against SDN (HLD-DDoSDN) Dataset

This study[10] proposes a benchmark dataset, HLD-DDoSDN, to mitigate the threats posed by Distributed Denial of Service (DDoS) attacks in SDN platforms based on machine learning and to evaluate detection performance.

This study introduces the HLD-DDoSDN dataset, designed to evaluate DDoS attack detection in SDN environments by simulating realistic traffic conditions. The dataset focuses on DDoS attack typologies, including TCP SYN, UDP and ICMP floods, with both high-rate (33.33 packets/second) and low-rate (5 packets/second) to emulate both aggressive and stealthy strategies. However, it exclusively targets DDoS flooding attacks, without addressing other attacks. Traffic was collected using a Mininet-based SDN testbed with a linear topology, incorporating a single POX SDN controller, an OpenFlow vSwitch and 64 hosts, as shown in Figure 4.

Attack traffic was generated with spoofed IPs and randomized source ports, while normal traffic was simulated concurrently. Tools such as Scapy crafted attack packets, Wireshark captured traffic for analysis and CICFlowMeter extracted 71 traffic features, including packet size distributions and inter-arrival times, to enable robust analysis.

The labeled dataset, adhering to benchmark standards and being made publicly available, provides a valuable source of data for SDN security research. Machine learning evaluations employed a Deep Multilayer Perceptron (D-MLP) model with high detection accuracy for binary and multiclass classification tasks. The performance metrics highlight the model's robustness, reinforcing the dataset's utility for developing and testing DDoS detection systems in SDN.

In summary, this paper gives a contribution to SDN security research by addressing existing dataset limitations and introducing a detection framework. Future research can expand the dataset with more attack vectors and test scalability on more complex SDN environments.

3. Network Architecture Complexity

In evaluating datasets for intrusion detection systems in Software-Defined Networking, it is important to highlight that network features and attack plans alone are insufficient for a comprehensive assessment. In order to build a solid evaluation framework, the network architecture complexity should also be taken into account[11].

Table 1
Comparison of datasets based on network features.

Dataset Name	Protocols for DDoS Attacks				DDoS Attacks Variation Rates		Controller	Number of Features	Dataset Availability
	TCP	UDP	ICMP	Appl. Layer	High	Low			
NSJ	✓	✓	✓		✓		POX	68	
ZCA	✓	✓	✓	✓	✓		POX	6	✓
InSDN	✓	✓	✓	✓	✓		ONOS	83	✓
NCLP	✓	✓	✓	✓	✓	✓	Floodlight	6	✓
Y-NV-RP-DJM-C	✓	✓		✓		✓	ONOS	13	✓
AAHHBA		✓			✓	✓	POX	7	
ASMK	✓	✓	✓		✓		Ryu	23	✓
HLD-DDoSDN	✓	✓	✓		✓	✓	POX	71	✓

While attack categories and network features give valuable insight into the nature of attacks and traffic behavior, they fail to fully incorporate the scalability and operational problems of actual-world SDN deployments. The architectural aspects also play a crucial role in accurately simulating and evaluating the network's reaction to attacks, particularly with regards to ML-based IDS models[12].

The Appendix A provides an overview of the schematics of the network architectures employed in the detailed papers. Table 2 provides an overview for assessing the complexity of network architectures as described in the datasets. Facilitates comparative evaluations of architectural sophistication, focusing on structural and operational characteristics critical to scalability and representation of realism.

The SDN controller is fundamental to the architecture, enabling the separation of the control and data planes, centralizing management and providing essential data for machine learning tasks such as traffic optimization and anomaly detection. Without a controller, the lack of centralized data access significantly hinders ML applications[13] and is assimilated to traditional networking concept. For this reason, studies underscore the controller's pivotal role in enabling data-driven approaches for SDN management[14].

The number of Open vSwitch (OVS) instances reflects a network's scalability and complexity. While single-switch setups mimic traditional networks, they fail to represent real-world SDN scenarios. Multiple OVS instances allow more realistic evaluations of dynamic network conditions[13], essential for training ML models.

Tools like Mininet frequently adopt such setups for comprehensive SDN simulations[14].

The nature of the environment, whether virtualized or hybrid with physical devices, determines the realism of network simulations. Virtual systems are efficient and cost-saving for prototyping but lack the dynamics of real world setups, where hardware heterogeneity and latency have a vital role to play[15].

The number of hosts is another critical parameter, as hosts generate the traffic patterns essential for network simulations. More hosts ensure diverse and realistic traffic patterns, which are beneficial for ML training and testing[13], while fewer hosts can result in oversimplification. Studies highlight the importance of simulating diverse host densities to simulate real-world environments and train efficient ML models[16].

Table 2
Comparison of datasets based on network architecture characteristics.

Dataset Name	Controller Presence	Number of Open vSwitch Instances	Environment Type	Number of Hosts
NSJ	Yes (Only during Attack traffic collection)	Single	Virtualized	5 (normal traffic collection)/15 (mixed traffic)
ZCA	Yes	Multiple	Virtualized	80
InSDN	Yes	Multiple	Virtualized	5
NCLP	Yes	Multiple	Virtualized	120
Y-NV-RP-DJM-C	Yes	Multiple	Hybrid	20
AAHHBA	Yes	Single	Virtualized	64
ASMK	Yes	Multiple	Virtualized	13
HLD-DDoSDN	Yes	Single	Virtualized	63

4. Discussion

This section examines the strengths and weaknesses of current datasets for intrusion detection in SDN environments, highlighting their innovations and limitations. The subsection 4.1 focuses on the strengths of existing datasets, including their ability to emulate realistic traffic patterns, provide different attack types and integrate wide feature sets essential for training ML models. The subsection 4.2 addresses their weaknesses, that is to say, for instance, limited attack variability, use of synthetic traffic and lack of real-time features, which limit their usability in complex and dynamic SDN scenarios. In the end, the subsection 4.3 provides a roadmap for designing new generation datasets to overcome these challenges and support security research.

4.1. Strenght Points

The importance of realistic datasets in training effective intrusion detection systems (IDS) for Software-Defined Networking (SDN) cannot be overstated. Realistic datasets mimicking real-world SDN traffic scenarios, including high-rate and low-rate DDoS attacks, are important in the development of detection systems that can operate in dynamic and complex environments[3]. For instance, datasets like InSDN cover a wide range of attack types, which makes them versatile and suitable for testing a variety of detection methods.

Richness in features is another important trait. Databases such as those of NSJ and InSDN feature rich data that plays a significant role when training robust machine learning models. Rich features allow models to efficiently recognize and react to complex patterns of attacks[17].

Besides, public access is an important aspect for increasing reproducibility and community-based improvements. Datasets like HLD-DDoSDN and InSDN facilitate community-based improvements and the reproducibility of results across different research settings[3].

Additionally, many datasets integrate with modern SDN controllers like POX, ONOS and Floodlight, ensuring practical relevance to real-world SDN architectures[16].

By combining realistic traffic simulations, wide attack plan and rich feature sets, these datasets play a significant role in advancing the effectiveness of ML-driven IDS in SDN environments.

4.2. Weakness Points

Most existing datasets, for instance, NSJ and AAHHBA, have limited attack diversity, with most of them only accounting for a limited number of DDoS attack types, that is to say TCP SYN, UDP and ICMP floods. This does not account for more sophisticated hybrid attacks, where attackers combine volumetric, protocol-based and application-layer methods, along with more intelligent stealthy methods like low-and-slow attacks or adversarial evasion techniques.

Consequently, these datasets are less representative of the diverse threat landscape encountered in real-world networks. Furthermore, narrow feature sets are another issue, with some datasets, offering minimal statistical data. This lack of richness in features restricts the ability to explore correlations between traffic patterns and anomalies, limiting the effectiveness of ML models. As a result, these datasets are less adaptable and may not perform well in detecting complex or subtle attack patterns.

Many datasets are also predominantly focused on DDoS attacks, ignoring other critical security threats in SDN environments, such as attacks directed to the controller. This narrow focus leads to a blind spot in detection systems trained on such datasets, leaving them vulnerable to other types of intrusions.

The reliance on synthetic traffic generated in emulated environments like Mininet presents another challenge. While these environments offer control and repeatability, they often fail to replicate the unpredictable dynamics of real-world networks, such as heterogeneous device interactions, varying user behaviors and external factors like latency and hardware failures. As a result, machine learning models trained on these datasets may not generalize well to actual production environments.

Dataset variety and volume are also a concern because some datasets contain either too minimal traffic volatility or volume. Datasets collected over the short term or from individual testbeds cannot mimic the long-term patterns and varied scenarios on which realistic model training should have to be established. Additionally, the absence of device diversity further reduces the relevance of these datasets compared to dynamic and global SDN settings.

Moreover, the majority of datasets have no informative attributes to train machine learning time attack detection systems, without features like timestamps, inter-arrival time and flow state transitions that are essential in time-sensitive anomaly detection. The absence of such features in datasets precludes the development of IDS systems that can provide real-time responses, which are critical in high-risk environments where low latency and quick responses are essential. Table 3 qualitatively maps some characteristics related to the considerations made in the previous paragraph to the different datasets analyzed in this paper. The colors used in the table have the following meanings:

- green: represents "yes" for binary classification and "high" for quantitative features.
- yellow: indicates an intermediate value between high and low intervals.
- red: denotes "no" for binary classification and "low" for quantitative features.

Table 3
Graphical view of Dataset comparison.

	NSJ	ZCA	InSDN	NCLP	Y-NV-RP-DJ-M-C	AAHHBA	ASMK	HLD-DDoSDN
Variety of attack typologies	🟥	🟥	🟨	🟥	🟥	🟥	🟥	🟥
Feature richness	🟩	🟥	🟩	🟥	🟨	🟥	🟨	🟩
Public availability	🟥	🟩	🟩	🟩	🟩	🟥	🟩	🟩
Presence of supported controllers	🟩	🟩	🟩	🟥	🟩	🟩	🟥	🟩
Likelihood of topology compared to real-world SDN scenarios	🟥	🟥	🟨	🟥	🟩	🟥	🟥	🟥
Device heterogeneity	🟥	🟥	🟨	🟥	🟩	🟥	🟥	🟥
Dataset size	🟥	🟥	🟨	🟥	🟩	🟨	🟥	🟩
Real-time statistics	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥

Among the various proposed datasets, dataset Y-NV-RP-DJ-M-C (corresponding to the fourth column) demonstrates the highest performance, as indicated by the predominance of green boxes. This dataset contains only two red boxes, indicating the absence of attack types beyond DDoS attacks and the lack of real-time statistics due to insufficient information on resource usage.

4.3. Roadmap for the Generation of a New Dataset

To address the limitations of existing datasets, a comprehensive dataset should integrate advanced attack diversity, including hybrid DDoS, ransomware and zero-day exploits, to reflect evolving threats. It should also enrich feature sets with temporal trends, entropy measurements and real-time signals to support dynamic detection methods.

Traffic capture across different SDN topologies and controller platforms, such as POX, ONOS and RYU, can record diverse network configurations, while the combination of synthetic and real-world traffic ensures practical relevance along with scalability.

Detailed temporal annotations, such as timestamps and inter-arrival times, enable real-time anomaly detection. Additionally, datasets should reflect large-scale networks for scalability testing and be made publicly accessible with clear documentation and preprocessing tools.

Mitigating misalignment with real SDN architectures through novel network designs and traffic generation methodologies will continue to enhance dataset applicability.

By including these enhancements, a new dataset may fill in gaps and be a valuable resource for advancing SDN security research.

5. Conclusion

The study highlights the central role played by datasets in driving ML and DL-based remedies for SDN network security. A critical assessment of existing datasets revealed some major strengths including traffic simulation, extended sets of features and integration with modern SDN controllers.

Nevertheless, the paper shows primary weaknesses of current datasets in table 3. The majority of them have poor attack diversity and are only able to identify volumetric DDoS attacks without addressing increasing threats like application-layer attacks. Also, reliance on simulated environments reduces the overall generalizability of models that learn from these datasets to actual networks, which are more varied and complex. The lack of temporal data and real-time annotations further restricts the development of time-sensitive intrusion detection systems. To address these limitations, this paper proposes a future direction for the creation of next-generation datasets. It emphasizes hybrid attacks, more sophisticated feature extraction, realistic data augmentation and multi-scenario traffic collection across various topologies and controllers.

Public accessibility and detailed documentation are also highlighted as essential for fostering widespread adoption and collaborative improvement. By implementing these strategies, future datasets can better reflect real-world network conditions, enabling the development of robust, scalable and adaptive security mechanisms for SDN environments.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-3.5/4.0 in order to paraphrase and reword. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Y. Maleh, Y. Qasmaoui, K. E. Gholami, Y. Sadqi, S. Mounir, A comprehensive survey on SDN Security: Threats, mitigations, and future directions, Journal of Reliable Intelligent Environments 8 (2022) 29–48. URL: https://link.springer.com/article/10.1007/s40860-022-00171-8#citeas. doi:10.1007/s40860-022-00171-8.

[2] A. K. Jain, H. Shukla, D. Goel, A comprehensive survey on DDoS detection, mitigation, and defense strategies in software-defined networks, Cluster Computing 27 (2024) 13129–13164. URL: https://doi.org/10.1007/s10586-024-04596-z. doi:10.1007/s10586-024-04596-z.

[3] A. A. Bahashwan, M. Anbar, S. Manickam, T. A. Al-Amiedy, M. A. Aladaileh, I. H. Hasbullah, A Systematic Literature Review on Machine Learning and Deep Learning Approaches for Detecting DDoS Attacks in Software-Defined Networking, Sensors 23 (2023). URL: https://www.mdpi.com/1424-8220/23/9/4441. doi:10.3390/s23094441.

[4] Q. Niyaz, W. Sun, A. Y. Javaid, A Deep Learning Based DDoS Detection System in Software-Defined Networking (SDN), EAI Endorsed Transactions on Security and Safety 4 (2017). doi:10.4108/eai.28-12-2017.153515.

[5] C. B. Zerbini, L. F. Carvalho, T. Abrão, M. L. Proença, Wavelet against random forest for anomaly mitigation in software-defined networking, Applied Soft Computing 80 (2019) 138–153. URL: https://www.sciencedirect.com/science/article/pii/S1568494619301115. doi:https://doi.org/10.1016/j.asoc.2019.02.046.

[6] M. S. Elsayed, N.-A. Le-Khac, A. D. Jurcut, InSDN: A Novel SDN Intrusion Dataset, IEEE Access 8 (2020) 165263–165284. doi:10.1109/ACCESS.2020.3022633.

[7] M. P. Novaes, L. F. Carvalho, J. Lloret, M. L. Proença, Long Short-Term Memory and Fuzzy Logic for Anomaly Detection and Mitigation in Software-Defined Network Environment, IEEE Access 8 (2020) 83765–83781. doi:10.1109/ACCESS.2020.2992044.

[8] N. M. Yungaicela-Naula, C. Vargas-Rosales, J. A. Perez-Diaz, E. Jacob, C. Martinez-Cagnazzo, Physical Assessment of an SDN-Based Security Framework for DDoS Attack Mitigation: Introducing the SDN-SlowRate-DDoS Dataset, IEEE Access 11 (2023) 46820–46831. doi:10.1109/ACCESS.2023.3274577.

[9] M. A. Aladaileh, M. Anbar, A. J. Hintaw, I. H. Hasbullah, A. A. Bahashwan, S. Al-Sarawi, Renyi Joint Entropy-Based Dynamic Threshold Approach to Detect DDoS Attacks against SDN Controller with Various Traffic Rates, Applied Sciences 12 (2022). URL: https://www.mdpi.com/2076-3417/12/12/6127. doi:10.3390/app12126127.

[10] A. A. Bahashwan, M. Anbar, S. Manickam, G. Issa, M. A. Aladaileh, B. A. Alabsi, S. D. A. Rihan, HLD-DDoSDN: High and low-rates dataset-based DDoS attacks against SDN, PLOS ONE 19 (2024) 1–29. URL: https://doi.org/10.1371/journal.pone.0297548. doi:10.1371/journal.pone.0297548.

[11] A. Khraisat, I. Gondal, P. Vamplew, J. Kamruzzaman, Survey of intrusion detection systems: techniques, datasets and challenges, Cybersecurity 2 (2019) 20. URL: https://doi.org/10.1186/s42400-019-0038-7. doi:10.1186/s42400-019-0038-7.

[12] A. Ahmad, E. Harjula, M. Ylianttila, I. Ahmad, Evaluation of Machine Learning Techniques for Security in SDN (2020). doi:10.1109/GCWkshps50303.2020.9367477.

[13] Y. Yoo, G. Yang, C. Shin, J. Lee, C. Yoo, Machine Learning-Based Prediction Models for Control Traffic in SDN Systems, IEEE Transactions on Services Computing 16 (2023) 4389–4403. doi:10.1109/TSC.2023.3324007.

[14] M. Paliwal, D. Shrimankar, O. Tembhurne, Controllers in SDN: A Review Report, IEEE Access 6 (2018) 36256–36270. doi:10.1109/ACCESS.2018.2846236.

[15] M. Hussain, N. Shah, R. Amin, S. S. Alshamrani, A. Alotaibi, S. M. Raza, Software-Defined Networking: Categories, Analysis, and Future Directions, Sensors 22 (2022). URL: https://www.mdpi.com/1424-8220/22/15/5551. doi:10.3390/s22155551.

[16] J. Xie, F. R. Yu, T. Huang, R. Xie, J. Liu, C. Wang, Y. Liu, A Survey of Machine Learning Techniques Applied to Software Defined Networking (SDN): Research Issues and Challenges, IEEE Communications Surveys & Tutorials 21 (2019) 393–430. doi:10.1109/COMST.2018.2866942.

[17] M. Wang, N. Yang, Y. Guo, N. Weng, Learn-IDS: Bridging Gaps between Datasets and Learning-Based Network Intrusion Detection, Electronics 13 (2024). URL: https://www.mdpi.com/2079-9292/13/6/1072. doi:10.3390/electronics13061072.

A. Paper Network Architectures

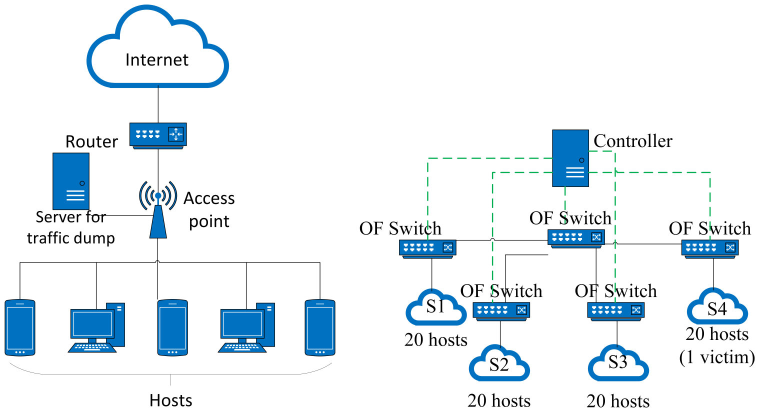

Figure 1: Respectively: NSJ and ZCA network architectures.

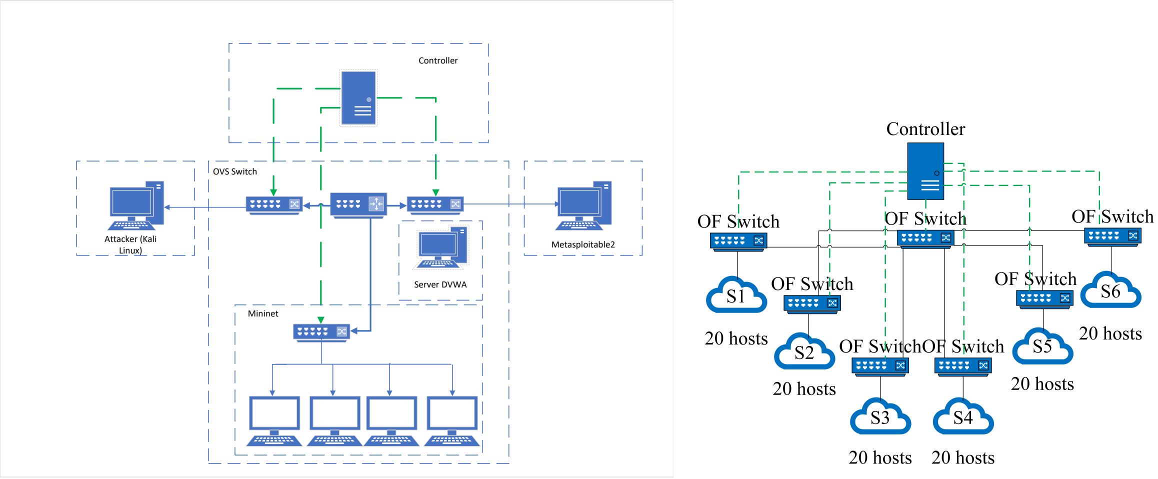

Figure 2: Respectively: InSDN and NCLP network architectures.

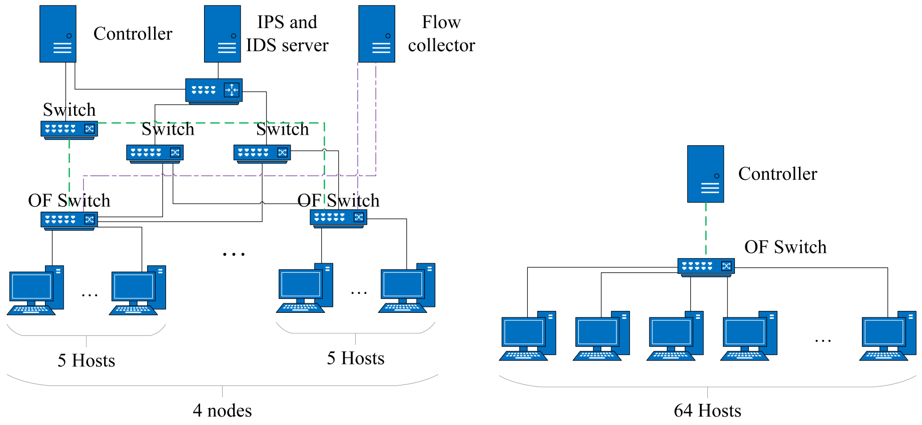

Figure 3: Respectively: Y-NV-RP-DJM-C and AAHHBA network architectures.

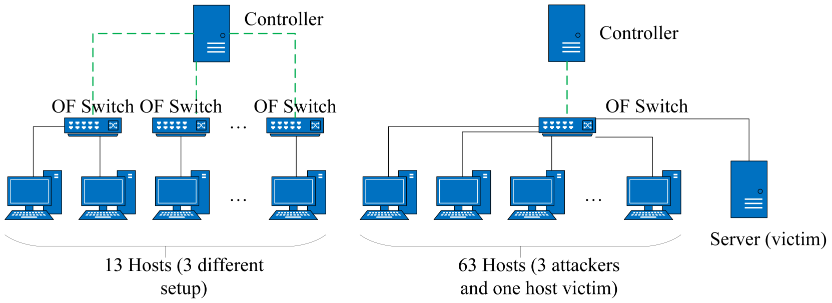

Figure 4: Respectively: ASMK and HLD-DDoSDN network architectures.