information

Article

Toward Generating a New Cloud-Based Distributed Denial of Service (DDoS) Dataset and Cloud Intrusion Traffic Characterization

MohammadMoein Shafi [1,2], Arash Habibi Lashkari [1,2,*], Vicente Rodriguez [3] and Ron Nevo [3]

[1] Behaviour-Centric Cybersecurity Center (BCCC), School of Information Technology, York University, Toronto, ON M3J 1P3, Canada; moeinsh@yorku.ca
[2] Department of Electrical Engineering and Computer Science, York University, Toronto, ON M3J 1P3, Canada
[3] cPacket, Milpitas, CA 95035, USA; vrodriguez@cpacketnetworks.com (V.R.); ron.nevo@cpacketnetworks.com (R.N.)
* Correspondence: ahabibil@yorku.ca

Abstract: The distributed denial of service attack poses a significant threat to network security. Despite the availability of various methods for detecting DDoS attacks, the challenge remains in creating real-time detectors with minimal computational overhead. Additionally, the effectiveness of new detection methods depends heavily on well-constructed datasets. This paper addresses the critical DDoS dataset creation and evaluation domain, focusing on the cloud network. After conducting an in-depth analysis of 16 publicly available datasets, this research identifies 15 shortcomings across various dimensions, emphasizing the need for a new approach to dataset creation. Building upon this understanding, this paper introduces a new public DDoS dataset named BCCC-cPacket-Cloud-DDoS-2024. This dataset is meticulously crafted, addressing challenges identified in previous datasets through a cloud infrastructure featuring over eight benign user activities and 17 DDoS attack scenarios. Also, a Benign User Profiler (BUP) tool has been designed and developed to generate benign user network traffic based on a normal user behavior profile. We manually label the dataset and extract over 300 features from the network and transport layers of the traffic flows using NTLFlowLyzer. The experimental phase involves identifying an optimal feature set using three distinct algorithms: ANOVA, information gain, and extra tree. Finally, this paper proposes a multi-layered DDoS detection model and evaluates its performance using the generated dataset to cover the main issues of the traditional approaches.

Keywords: distributed denial of service (DDoS); cloud-based DDoS; network traffic dataset; cloud-based network traffic analysis; network layer traffic analysis; transport layer traffic analysis; network traffic characterization; DDoS dataset

Citation: Shafi, M.; Lashkari, A.H.; Rodriguez, V.; Nevo, R. Toward Generating a New Cloud-Based Distributed Denial of Service (DDoS) Dataset and Cloud Intrusion Traffic Characterization. *Information* **2024**, *15*, 195. https://doi.org/10.3390/info15040195

Academic Editor: Ge Yu

Received: 26 February 2024
Revised: 22 March 2024
Accepted: 28 March 2024
Published: 31 March 2024

1. Introduction

The DDoS attack poses a significant threat to network security, aiming to overwhelm target networks with malicious traffic [1]. Detecting and mitigating DDoS attacks pose considerable challenges due to their sophisticated and dynamic nature [2,3]. Traditional detection methods often struggle to accurately identify and stop DDoS attacks in real time, primarily because of their inability to distinguish between legitimate and malicious traffic among the flood of incoming packets. Moreover, as cyber attackers continually evolve tactics, traditional detection mechanisms become increasingly ineffective in identifying newly emerging DDoS attack variants [4,5]. Consequently, there is a growing recognition of the need to leverage artificial intelligence (AI) techniques to enhance DDoS detection capabilities [6,7].

AI-based methods, including machine learning (ML) and deep learning (DL) algorithms, rely heavily on data to learn and discern various network behaviors associated with

different network activities. However, the effectiveness of these detection methods hinges on the availability of high-quality and comprehensive datasets for training and evaluation purposes. Therefore, at the heart of advancing DDoS detection methods capabilities lies the availability of reliable and comprehensive evaluation datasets [8].

However, creating and preparing datasets for DDoS detection entails a multifaceted process that demands meticulous planning, execution, and expertise. Establishing the infrastructure for dataset creation requires setting up an isolated environment that accurately mimics real-world network conditions while ensuring scalability, reliability, and security, demanding deep knowledge of network architecture and system administration. Generating representative traffic patterns and attack scenarios necessitates a comprehensive understanding of cyber threats and network protocols, and orchestrating the generation of benign and malicious traffic in a controlled manner demands sophisticated tools. Rigorous analysis is imperative to extract meaningful insights and validate dataset quality, requiring advanced anomaly detection and classification algorithms. Without clear planning and execution, the resulting dataset may lack fidelity and fail to adequately represent real-world scenarios, undermining the effectiveness of DDoS detection algorithms. Therefore, ensuring the integrity and comprehensiveness of the dataset creation process is critical for advancing DDoS detection capabilities and enhancing network security.

This research addresses the critical domain of DDoS dataset generation and evaluation, emphasizing the pivotal role of well-constructed datasets in enhancing the efficacy of DDoS detection and characterization algorithms and systems. Through an exhaustive analysis of 16 publicly available datasets, we identify significant shortcomings across various dimensions, underscoring the imperative for a novel approach to dataset development. By shedding light on the challenges and deficiencies in existing datasets, this research aims to pave the way for creating more reliable and comprehensive DDoS datasets, thereby advancing state-of-the-art DDoS detection and mitigation strategies.

Drawing upon this foundational understanding, the pioneering contribution in this work is the introduction of a new cloud-based DDoS dataset named BCCC-cPacket-Cloud-DDoS-2024 [9]. The development of this dataset involves a comprehensive examination of diverse aspects, ranging from attack trends to the generation of benign user traffic and behavior profiling. The paper proposes a roadmap for creating a network traffic dataset. A cloud infrastructure featuring eight distinct benign activities and 17 varied DDoS attack scenarios is established to tackle challenges extracted from previous datasets. Also, a Benign User Profiler (BUP) [10] tool is designed and implemented for generating network traffic based on benign user behavior. The dataset is manually labeled, and over 300 features are extracted from the network and transportation layers of the network traffic flows using the Network and Transportation Layers Flow Analyzer (NTLFlowLyzer) [11].

Next, to dataset creation, this research proposes a network traffic characterization model to detect different network activities. In crafting a robust detection model, we advocate a multi-layered approach to maximize cost effectiveness. The first layer focuses on distinguishing benign traffic from non-benign activity, while subsequent layers identify specific network behaviors for detailed classification. This approach enhances efficiency and minimizes computational costs.

Finally, the experimental evaluations using the generated dataset validate the efficacy of the proposed detection model in accurately identifying various network activities. The proposed model demonstrates strong performance through extensive testing against diverse DDoS attack scenarios and benign activities, highlighting its potential to strengthen network security defenses.

The main contributions of this research are:

- Introducing BCCC-cPacket-Cloud-DDoS-2024 [9], a new cloud-based DDoS dataset.
- Design and development of a Benign User Profiler (BUP) [10] tool to generate benign background traffic.
- Design and development of a DDoS characterization model.
- Introducing the cloud-based network traffic dataset creation roadmap.

The paper is organized as follows: Section 2 provides background information and delves into the relevant literature. Section 3 explains the different steps involved in creating a network dataset. Section 4 details the construction process of the new dataset, detailing each step involved in its creation. Section 5 expounds upon the traffic characterization model proposed in this study. The findings from the experimental endeavors are outlined in Section 6. Section 7 offers a thorough analysis and interpretation of the experimental results. Lastly, Section 8 concludes the paper by summarizing key insights and delineating potential avenues for future research in this domain.

2. Literature Review

This section reviews existing research in this domain, including available cloud-based DDoS attack detection, traffic analysis models, and available DDoS network traffic datasets. Finally, the section highlights the shortcomings of the previous works and the areas where the current literature could be improved.

2.1. Cloud-Based DDoS Attack Detection and Traffic Analysis

In recent years, the proliferation of cloud computing has introduced new challenges in ensuring the security and availability of cloud network resources [12]. Among these challenges, DDoS attacks targeting cloud infrastructures have emerged as a significant concern due to their potential to disrupt network services and exhaust server resources [13]. This section focuses on cloud-based DDoS attack detection through traffic analysis, discussing various approaches proposed in the literature to mitigate this threat.

To begin with, ref. [14] proposes a DDoS detection system based on the C4.5 algorithm and signature detection techniques to enhance security in cloud environments. In a similar work, by [15], the authors introduce the Scattered Denial-of-Service Mitigation Tree Architecture (SDMTA), tailored for hybrid cloud environments. Their architecture integrates network monitoring for efficient detection and mitigation of DDoS attacks. In a more straightforward study, by [16], the authors propose a solution leveraging machine learning algorithms like support vector machine, naive Bayes, and random forest for DDoS attack classification in cloud environments.

In another study, by [17], they propose integrating HTTP GET flooding and MapReduce processing for rapid DDoS attack detection in cloud computing environments. In another study, by [13], the authors address the challenge of identifying the sources of DDoS attacks in cloud environments, proposing a third-party auditor (TPA)-based packet trace-back approach. Leveraging Weibull distribution for analysis, their solution aims to provide efficient attack alerts and mitigate the impact of attacks on cloud users.

In [18], the authors tackle the challenge of detecting low-rate DDoS attacks resembling normal traffic flow. Their proposed solution utilizes soft computing techniques, including hidden Markov models (HMMs) and random forest algorithms, demonstrating improved classification accuracy in cloud environments. In another study, by [3], they propose a feature selection–whale optimization algorithm–deep neural network (FS-WOA-DNN) method for DDoS attack detection, emphasizing effective feature selection and classification techniques.

Lastly, ref. [19] develop a DDoS detection system integrated with OpenStack, employing machine learning algorithms and raw socket programming for monitoring network traffic. Their system effectively identifies and notifies administrators about DDoS attacks in private cloud environments, showing promising results in experimental evaluations.

Despite the advancements in cloud-based DDoS attack detection and traffic analysis, several limitations persist across existing works. Notably, the lack of publicly available datasets restricts the proposed solutions' reproducibility and comparative analysis. Additionally, many studies rely on synthetic or limited datasets, which may not accurately reflect real-world cloud attack scenarios. Moreover, challenges such as the adaptability of detection systems to evolving attack strategies, scalability issues in large-scale cloud environments, and the potential for false positives/negatives pose ongoing challenges for

researchers and practitioners in this domain. Future research efforts should address these limitations to enhance the robustness and effectiveness of DDoS detection mechanisms in cloud environments.

2.2. Available DDoS Attack Datasets

In the dynamic realm of network security, the progress and assessment of DDoS attacks heavily rely on the availability and quality of datasets [20,21]. Recognizing the limitations and strengths of existing datasets is pivotal for researchers seeking to advance the field. For this, we have analyzed the top publicly available DDoS datasets in the literature, enumerated as follows. Evaluating these datasets is vital for shaping the conception of innovative datasets, exemplified by the DDoS dataset under consideration in this paper. The objective is to aid in recognizing areas for improvement and lay the groundwork for developing more robust and representative datasets in the DDoS domain.

1. KDD99 1998-99 [22];
2. CAIDA (2004) [23];
3. CAIDA (2007) [24];
4. CAIDA (2017) [25];
5. CAIDA (2021) [26];
6. CDX (2009) [27];
7. Kyoto (2009) [28];
8. ISCX2012 [29];
9. ADFA (2013) [30];
10. CTU-13 [31];
11. UNSW-NB15 [32];
12. CIC-IDS2017 [33];
13. CSE-CIC-IDS2018 [34];
14. CIC-DDoS2019 [35];
15. SR-BH 2020 [36];
16. CUPID (2022) [37].

While previous datasets have significantly contributed to the advancement of DDoS research, it is crucial to acknowledge and address their inherent limitations. Recognizing these shortcomings underscores the importance of developing new datasets, such as the DDoS dataset generated in this research, to overcome existing challenges and propel the field forward. By learning from the successes and pitfalls of previous works, researchers can inform the design of more realistic and comprehensive datasets, ultimately fostering the development of more effective and adaptable DDoS detection methods. Below, we have compiled a list highlighting the shortcomings identified in the previous datasets based on the works of [8,35,38].

1. **Imbalanced class distribution:** The imbalanced class distribution in datasets often mirrors real-world scenarios, where certain DDoS attacks are more prevalent than others. Addressing this limitation is vital as it ensures that detection models are trained on data that accurately reflects the wild attacks' distribution.
2. **Limited diversity of attacks:** Datasets with limited diversity fail to capture the full spectrum of DDoS attacks encountered in real-world networks. This shortfall hampers the effectiveness of detection methods by neglecting to train models on a comprehensive range of attack types and techniques.
3. **Outdated threat scenarios:** The inclusion of outdated threat scenarios in datasets may lead to the development of ill-equipped detection models to handle emerging DDoS threats. This limitation highlights the need for datasets that continuously evolve to reflect the evolving landscape of DDoS attacks in real-world environments.

4. **Lack of Realistic Network Traffic:** Realistic network traffic patterns are essential for training accurate DDoS detection models. Datasets lacking such traffic fail to capture network behavior's intricacies, hindering detection methods' effectiveness in real-world deployment scenarios.
5. **Absence of encrypted traffic:** With an increasing prevalence of encryption in network communications, datasets lacking encrypted traffic fail to simulate real-world conditions accurately. Including encrypted traffic in datasets is crucial for effectively training detection models capable of handling encrypted DDoS attacks.
6. **Insufficient labeling accuracy:** Inaccurate labeling of data instances undermines the reliability of datasets and, consequently, the effectiveness of detection models trained on them. Ensuring high labeling accuracy is paramount to developing robust DDoS detection mechanisms.
7. **Limited incorporation of user behavior:** User behavior plays a significant role in DDoS attack detection, yet datasets often overlook this aspect. Incorporating user behavior data into datasets enhances the reality of training data, leading to more effective detection models in real-world scenarios.
8. **Incompatibility with modern protocols:** Datasets that do not support modern network protocols fail to reflect the current state of network communications. Ensuring compatibility with modern protocols is essential for developing detection models that address contemporary DDoS threats.
9. **Limited exploration of low-rate DDoS attacks:** Low-rate DDoS attacks pose unique challenges that are usually overlooked in datasets. By exploring these attack types, datasets can better prepare detection models to identify and mitigate low-rate DDoS attacks in real-world scenarios.
10. **Lack of realistic DDoS traffic variability:** Variability in DDoS traffic patterns is essential for training robust detection models capable of adapting to evolving attack strategies. Datasets lacking such variability fail to prepare detection mechanisms for real-world deployment adequately.
11. **Absence of hybrid DDoS scenarios:** Hybrid DDoS attacks combine multiple attack vectors, presenting complex challenges for detection and mitigation. Including hybrid attack scenarios in datasets is crucial for training detection models capable of identifying and mitigating these sophisticated threats.
12. **Insufficient exploration of DDoS amplification techniques:** Datasets often overlook the exploration of DDoS amplification techniques, which attackers commonly use to magnify the impact of their attacks. Understanding and mitigating these techniques requires datasets that adequately represent such attack scenarios.
13. **Inadequate representation of application-layer DDoS attacks:** Application-layer DDoS attacks target specific services or applications, posing unique challenges for detection and mitigation. Datasets must include instances of application-layer attacks to train detection models effectively.
14. **Non-Inclusion of insider threats:** Insider threats present a significant risk to network security, yet datasets often overlook this threat vector. Including instances of insider threats in datasets is essential for training detection models capable of identifying and mitigating such risks.
15. **Absence of multi-modal data:** Multi-modal data, incorporating various data types such as network traffic, system logs, and user behavior, provides a more comprehensive view of DDoS attacks. Datasets lacking multi-modal data fail to capture the complexity of real-world attack scenarios, limiting the effectiveness of detection models.

Recognizing the importance of overcoming these limitations, this research aims to address the first ten challenges outlined in this analysis.

3. Dataset Creation Roadmap

Network dataset creation is a fundamental aspect of network security research, enabling the evaluation and validation of defense mechanisms against cyber threats. The

process involves careful planning, infrastructure setup, data generation, and analysis to simulate real-world scenarios accurately. In this work, we present a comprehensive roadmap for creating network datasets, addressing the challenges and intricacies encountered at each stage.

Establishing a structured framework and roadmap for dataset creation is paramount due to the complexity of the task. A well-defined roadmap streamlines the process, ensuring systematic progression from initial conception to the final dataset. It provides researchers with a clear direction, facilitates collaboration, and enhances reproducibility, fostering advancements in network security research.

Creating a network dataset poses formidable challenges owing to network traffic's dynamic and heterogeneous nature. The general roadmap is illustrated in Figure 1. In the following, we explain each step highlighted in the roadmap:

1. **Scope Definition**
   Firstly, defining the scope of the target network demands a deep understanding of the specific environment under study, such as an e-commerce company network encompassing diverse user interactions and transactions. This necessitates comprehensive data collection and analysis, often complicated by the sheer volume and variety of network activities.
2. **Infrastructure Preparation**
   The preparation of infrastructure, whether cloud-based or otherwise, constitutes a critical initial step in dataset creation. A robust infrastructure ensures scalability, reliability, and performance, essential for generating and analyzing network traffic data. However, configuring and maintaining the infrastructure can be arduous, requiring expertise in network administration and resource optimization to mitigate potential bottlenecks and ensure seamless operation.
3. **Defining Users and Entities**
   Defining the corresponding users and entities within the network, along with their respective profiles, is essential for generating realistic traffic patterns. This involves categorizing users based on their roles, behaviors, and privileges and identifying network entities such as servers, clients, and applications. However, accurately characterizing user profiles and entity interactions poses challenges, particularly in large-scale networks with diverse user demographics and complex system architectures.
4. **Designing Benign Traffic Generator**
   A benign traffic generator design based on the defined user profiles is crucial for simulating legitimate, realistic, and real-world network activities. However, developing an effective traffic generator balances realism with efficiency and scalability. Generating diverse and realistic traffic patterns while avoiding bias or over-representing specific user behaviors requires careful consideration of traffic generation and profile definition, often necessitating iterative refinement and validation.
5. **Studying Attack Trends**
   Analyzing historical attack trends is essential for understanding prevalent threats and vulnerabilities in network environments. However, identifying relevant attack vectors and trends amidst evolving cyber threats can be challenging, requiring continuous monitoring and analysis of security incidents and threat intelligence sources. Moreover, extrapolating past attack trends to anticipate future threats necessitates robust analytical frameworks and predictive modeling techniques.
6. **Attack Selection and Implementation**
   Selecting suitable attack scenarios and implementing them within the network environment involve various complexities. Identifying realistic attack scenarios that align with the network's characteristics and threat landscape requires in-depth knowledge of common attack methodologies and their potential impact on network infrastructure. Furthermore, developing and deploying attack implementations necessitates expertise in security testing methodologies and adherence to ethical considerations to prevent unintended consequences or system compromise.

7. **Data Capturing and Analysis**
   Capturing raw network data in the form of PCAP files is essential for capturing the intricacies of network traffic and facilitating subsequent analysis. However, capturing and storing network traffic data at scale poses challenges regarding data volume, storage capacity, and processing overhead. Moreover, ensuring the integrity and confidentiality of captured data while adhering to privacy regulations requires robust data anonymization and encryption mechanisms.
8. **Development of Traffic Analyzer**
   Designing and developing a network traffic analyzer to convert raw PCAP files into analyzed data (e.g., CSV files) is crucial for extracting meaningful insights from captured network traffic. However, developing an efficient and accurate traffic analyzer addresses various technical challenges, such as packet parsing, protocol decoding, and traffic classification. Additionally, ensuring the scalability and reliability of the analyzer across diverse network environments and traffic patterns requires rigorous testing and optimization.
9. **Data Labeling and Testing**
   Labeling the resulting dataset and conducting comprehensive testing and analysis is essential for validating the quality and reliability. However, manually labeling network traffic data for attack and benign activities can be labor intensive and error-prone, necessitating automated labeling techniques and human validation processes. Moreover, thorough testing and analysis of the dataset against predefined metrics and ground truth scenarios are crucial for assessing its effectiveness in simulating real-world network conditions and evaluating defense mechanisms.

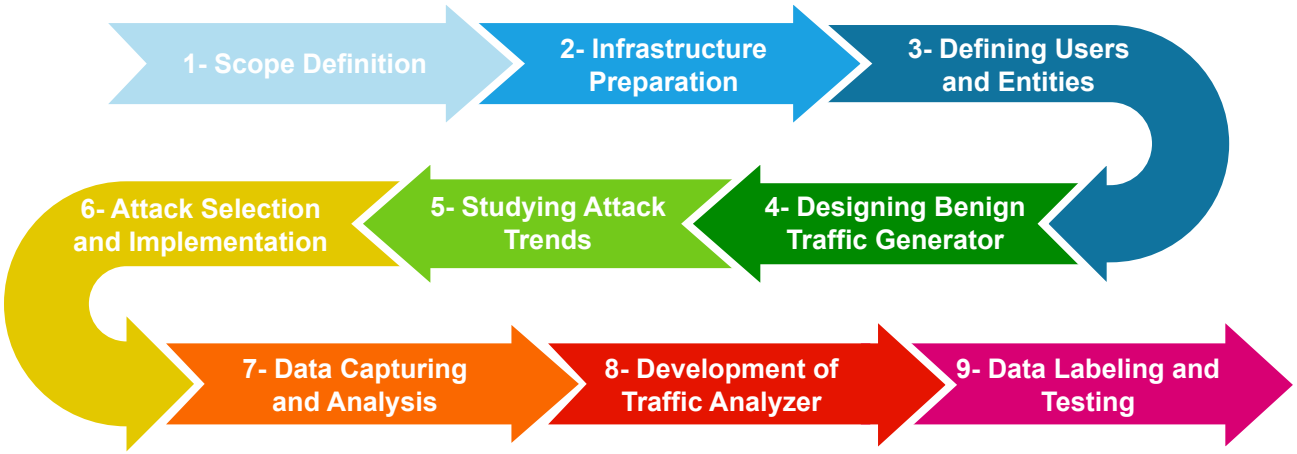

Figure 1. Network dataset creation roadmap.

In conclusion, creating network datasets requires a structured and systematic approach encompassing various stages, from infrastructure setup to data analysis and validation. Despite the challenges encountered at each step, a well-defined roadmap facilitates the generation of high-quality datasets, essential for advancing network security research and enhancing cybersecurity defenses. By addressing these challenges and adopting best practices, researchers can simulate real-world network environments effectively and develop robust defense strategies against emerging cyber threats.

4. The New Dataset

This section discusses the details of the generated dataset, named BCCC-cPacket-Cloud-DDoS-2024 [9]. The discussion starts with an explanation of the infrastructure and attack scenarios. Then, we discuss benign scenarios and the definitions of benign user behaviors in this work. Finally, this section provides the data analyzed from the Network and Transportation Layers Flow Analyzer (NTLFlowLyzer) [11] and the dataset details.

4.1. Infrastructure

The proposed cloud architecture and its intended utilization are outlined in this section, providing an overview of the selected approach and justifying the design of such a cloud architecture. The proposed architecture in Figure 2 shows a comprehensive network configuration to present a real-world corporate environment. This architecture is designed to emulate the complexities and vulnerabilities inherent in such environments, including the potential for DDoS attacks. We have structured the architecture within this environment to reflect typical organizational dynamics. The victim's VPC represents the internal network of a typical company. It includes four Windows machines for daily normal user activities, such as web browsing and email checking. These users exemplify the end-users typical of the company's operational environment.

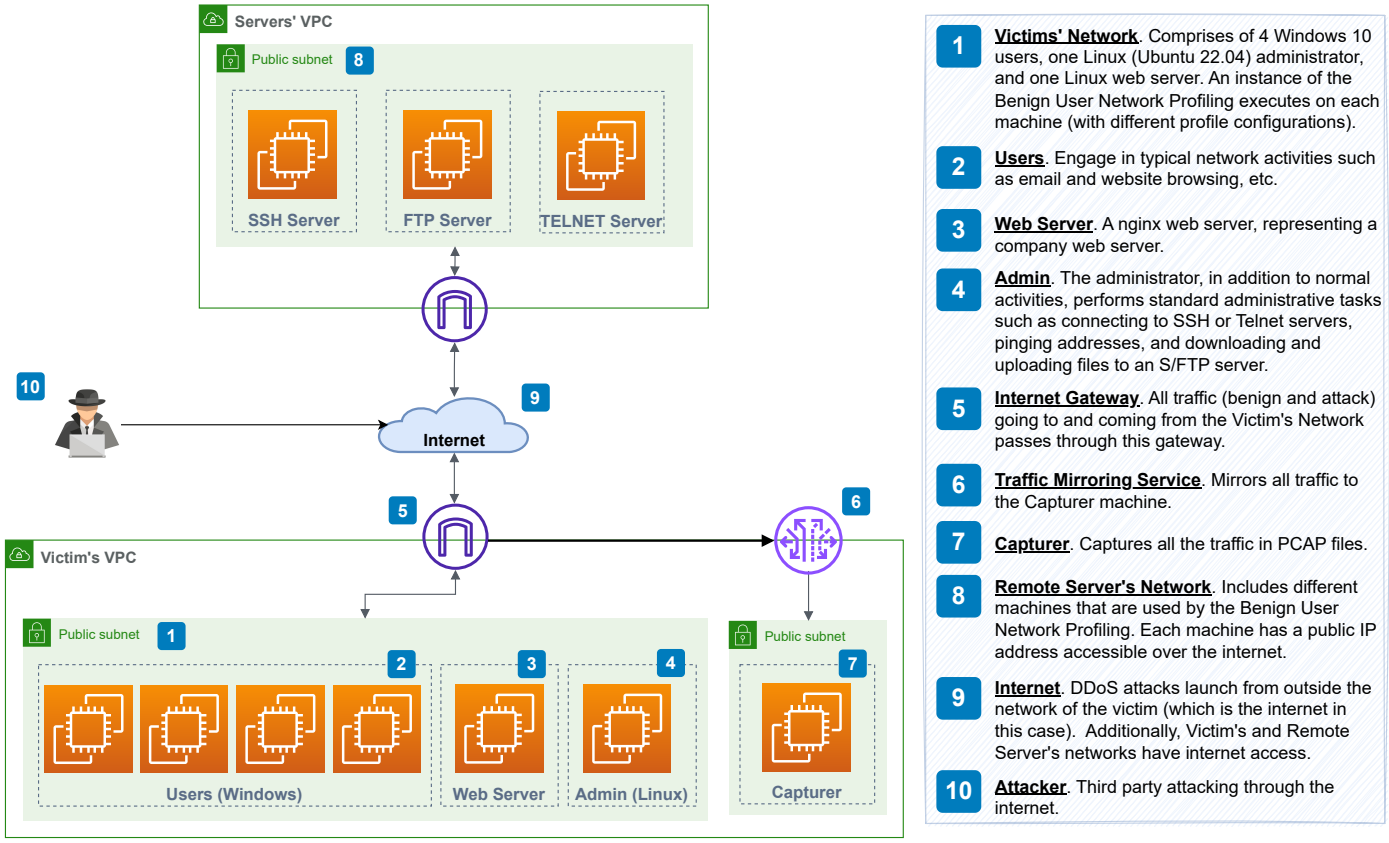

Figure 2. Cloud architecture.

Notably, the system introduces an administrator user with standard user activities and advanced administrative functions. The administrative tasks assigned to the administrator involve establishing secure connections to remote machines and employing protocols such as SSH or TELNET. Additionally, the administrator engages in FTP and SSH activities on remote servers. Specific servers, such as the SSH and TELNET servers, have been strategically placed in a dedicated virtual private cloud (VPC) named "Servers VPC" to facilitate these administrative functions.

Moreover, the critical infrastructure, such as the web server, resides alongside other machines in the victim's VPC used for hosting the company website using the Nginx web application. The architecture leverages the AWS Traffic Mirroring service to ensure comprehensive monitoring of network activities, reflecting our commitment to enhancing the visibility of potential security threats within the local network. Microsoft Windows

Server 2022 has been employed as the operating system for server machines, while the web server, administrator, and capturer machines are Ubuntu Server 22.04 LTS.

It is imperative to mention that the architecture incorporates an attacker component responsible for executing attacks through the internet. A third-party service has been engaged to carry out these attacks. Noteworthy is the absence of mitigation services or firewalls impeding network traffic, aligning with the simulated vulnerability of the corporate environment under study. Also, public service ports, including 80, 443, 22, etc., have been opened on each machine to mimic real-world conditions. Furthermore, all routing tables, ACL groups, and security groups permit the passage of all traffic, contributing to the realism of the simulation.

4.2. Attack Scenarios

While this study primarily focuses on DDoS attack scenarios targeting the victim's VPC, we acknowledge the possibility of broader attack vectors beyond the scope of the investigation in this work. Labeling the non-benign data outside the scheduled attack scenarios as "suspicious" underscores the awareness of potential threats from the internet. The attack scenarios section concentrates on the orchestrated execution of diverse attacks, including identifying attack types, the number of attacks, and their scheduling. This strategic approach involves launching 17 TCP-based DDoS attacks targeting the designated system, which are as follows:

- TCP-SYN-Valid;
- TCP-bypassV1;
- TCP-KILLALL-V2;
- TCP-IGMP;
- TCP-SYN;
- KILLER-TCP;
- TCP-CONTROL;
- TCP-Flag-MIX;
- TCP-Flag-SYN;
- TCP-Flag-SYNACK;
- TCP-Flag-ACK;
- TCP-Flag-ACKPSH;
- TCP-Flag-RSTACK;
- TCP-Flag-SYNTIME;
- TCP-Flag-SYNTFO;
- TCP-Flag-OSYN;
- TCP-Flag-OSYNP.

These attacks are diligently chosen to cover a broad spectrum of attack types, ensuring a thorough assessment of system resilience. The dataset web page [9] explains each attack scenario and provides further details. Specific timeframes are allocated for each attack to maintain a structured and manageable methodology. Each attack spans twenty minutes, followed by a ten-minute rest interval. The rest interval avoids the traffic overlap from two attacks, as sometimes the attacks will not be finished on time. Potential complications during the subsequent labeling procedure are mitigated by allowing for a resting period after each attack. This fact has been considered in the labeling process as well.

This deliberate scheduling optimizes data collection and streamlines the subsequent labeling process. The detailed schedule for the attack scenarios is provided in Table A1. It is important to emphasize that background traffic was still collected throughout the attack periods to simulate the real-world scenario, enabling us to differentiate and analyze benign and attack traffic simultaneously.

4.3. Benign User Profiling

This subsection outlines the typical user activities creation, investigating previous traffic generators aligned with leveraging the proposed Benign User Profiler (BUP) [10] tool designed to emulate genuine user behavior and produce realistic benign traffic data.

4.3.1. Available Benign User Traffic Generators

Ensuring clean and realistic benign and malicious data is crucial. However, prior research has predominantly emphasized the quality of malicious data, neglecting the significance of accurate ground truth for any detection system. The academic landscape for traffic generation is somewhat limited, with existing studies dating back over a decade. Based on investigation in this area, these are the available tools:

- PackETH [39];
- Iperf [40];
- D-ITG [41];
- Ostinato [42];
- SolarWinds [43];
- Packet Sender [44];
- Nping [45];
- NetScan [46];
- TRex [47].

A recurring theme emerges when reviewing the various network traffic generation tools. These tools are primarily geared towards benchmarking and stress testing rather than simulating benign, real-world user behavior. While tools like "SolarWinds" or "Packet Sender" offer valuable features for testing network performance, they lack a specific focus on emulating authentic user interactions based on predefined profiles. The importance of a benign traffic generator that accurately replicates user behavior cannot be overstated. Within the broader context of constructing an ideal DDoS dataset, the imperative of faithfully replicating genuine user interactions takes center stage. Consequently, a primary objective of this work is to design and implement a dedicated tool that addresses the imperative of generating authentic benign traffic.

4.3.2. Proposed Benign User Traffic Generator

Comprehensive research is conducted across different domains to facilitate the selection of appropriate user behaviors. Drawing insights from previous works [48–53], here is a comprehensive list of behaviors:

- ***Web Browsing (normal and admin user)*** Web browsing behavior encompasses various types of websites that reflect normal user interactions on an average weekly day. Drawing upon research insights from previous works [48,50,51], the following website categories are integrated into the web browsing behavior:

  - Shopping;
  - Social networks;
  - Downloading;
  - Music streaming;
  - Food;
  - News checking.
  - Video watching;
  - Taxi;

  To ensure authenticity, the Firefox browser is used for this study. This choice is driven by its widespread usage, open-source nature, and reputation for user privacy and security features [54]. Importantly, our approach involves using a real browser for interactions rather than relying on scripted requests, distinguishing this work from other benign user profiles and contributing to the reality of the generated traffic across various online activities. Additionally, factors like the number of open tabs, time of the day, and time spent on each website [48] are considered for a comprehensive and accurate generation of benign network traffic.
- ***Emailing (normal and admin user)***
  The Gmail web server is selected as the primary platform for simulating email-related behaviors, including sending and receiving. This decision is grounded in the widespread use of Gmail, ensuring that the benign network traffic generated accurately reflects typical email interactions. Configuring users to send emails to each other at regular intervals ensures a controlled and reliable simulation of both sending and receiving activities. Additionally, the selected approach allows the attachments for each email and provides a comprehensive representation of benign traffic associated with email communication.

- ***Systemic (normal and admin user)***
  This category pertains to the traffic related to operating system (OS) services. The choice to focus on systemic activities stems from the need to capture network traffic associated with routine system-level operations. The rationale behind prioritizing systemic activities is capturing network traffic related to routine system-level operations. It is crucial to clarify that this approach does not involve generating such traffic; instead, the natural network activity of the OS on each machine is enabled for routine service updates and other essential functions. This intentional focus contributes to the dataset's authenticity, facilitating a more comprehensive evaluation of DDoS detection methods in routine system-level operations scenarios.
- ***Command Line (admin user)***
  This category explicitly addresses admin user behaviors and involves activities related to the Linux terminal. It encompasses tasks such as updating package lists, installing packages, creating, modifying, and deleting directories and files, and other administrative tasks. Simulating these command-line activities generates benign network traffic representative of the administrative functions carried out through the terminal interface.
- ***SSH or Remote Command Line (admin user)***
  Like the command-line activity, the SSH or remote command-line category involves executing commands through an SSH session to a remote machine. This distinct category acknowledges the unique nature of remote command-line operations. These interactions are simulated to generate benign network traffic representative of administrative tasks conducted remotely, thereby enhancing the authenticity of generating the benign traffic across diverse scenarios.
- ***File Transfer, FTP server (admin user)***
  This category is dedicated to activities related to FTP operations and focuses on benign network traffic associated with file transfers. Different file sizes and formats are simulated for both downloading from and uploading to an FTP server. This part ensures comprehensive coverage of everyday file transfer activities associated with administrative tasks.
- ***File Transfer, SCP (admin user)***
  This category addresses secure file transfers between different machines using SCP. While SCP is associated with SSH, we make it a separate category. Simulating sending and receiving files with various formats and sizes to and from another machine allows us to generate benign network traffic that accurately reflects the secure file transfer activities associated with machine-to-machine interactions. This approach enhances the authenticity of benign traffic generation and contributes to a comprehensive understanding of secure file transfer activities in admin user behaviors.

4.3.3. Benign Scenarios

This subsection delves into characterizing general benign scenarios considered during ordinary operations by regular and admin users within the victim network. Figure 3 shows a sample of the benign scenarios for a regular Windows user (see Appendix A). The new dataset includes two days of pure benign data. Notably, the benign scenarios remain consistent across all days, including attack days, ensuring uniformity in benign behavior, a practice mirroring real-world scenarios, where consistent user behavior is maintained during both benign and challenging periods.

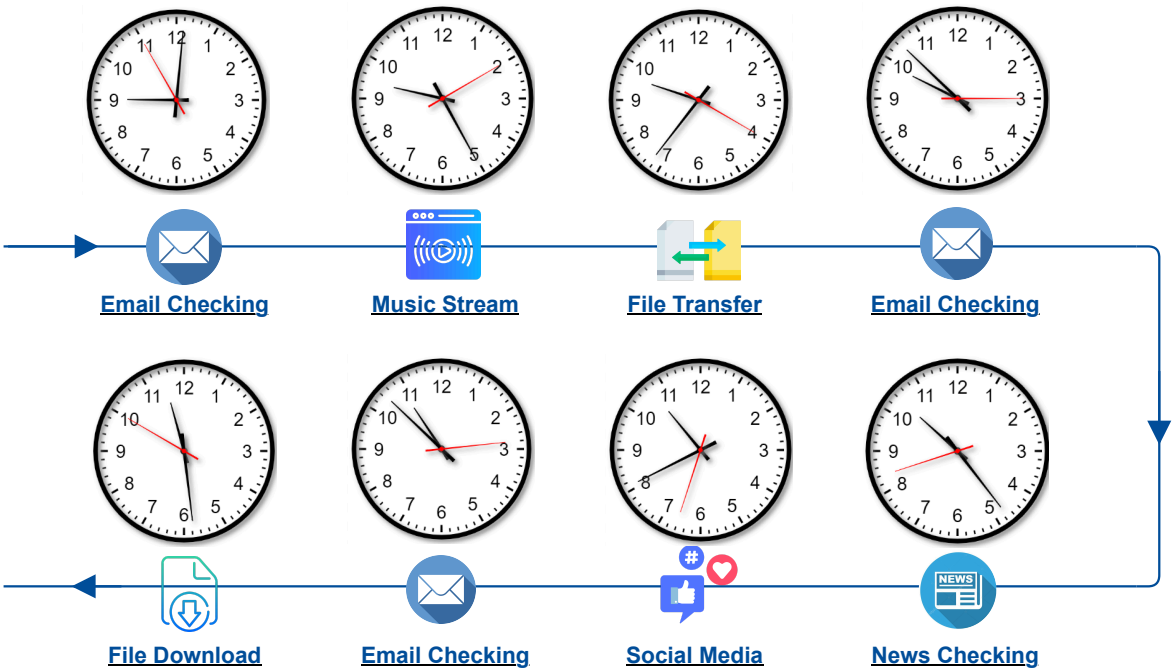

Figure 3. An example of a benign user behavior profile.

4.4. Data Capture

This subsection elucidates the data-capture process, which spans from 09:00 to 17:00 daily. Table 1 furnishes detailed information regarding the captured data, delineating the specific days allocated for benign and adversarial traffic capture. The data highlights a significant trend, indicating that over 90 percent of all user behavior on the network side, irrespective of the specific activity, is associated with the TCP protocol. This prominence underscores the importance of analyzing TCP behavior and implementing effective methods to mitigate potential malicious activities.

Table 1. Captured data information.

Date	# of IP Packets	# of TCP Packets	# of UDP Packets
Thursday 14 December	37,552,636 (99.75%)	35,213,784 (93.54%)	2,335,531 (6.20%)
Saturday 16 December	38,823,009 (99.81%)	36,151,292 (92.94%)	2,666,822 (6.86%)
Monday 18 December	28,302,904 (99.85%)	24,628,272 (86.89%)	3,671,949 (12.95%)
Tuesday 19 December	23,550,922 (99.89%)	21,653,847 (91.84%)	1,890,837 (8.02%)
Sum of All	**128,229,471**	**117,647,195**	**10,565,139**

4.5. Data Labeling (CSV File Generation)

Given the raw data's intricate nature, sophisticated packet analysis techniques are essential. This involves scrutinizing the packet-level information to extract relevant features that can offer a clear perspective on the data's underlying behavior and characteristics. The aim is to transform the raw packets into a well-organized data sheet that not only aids understanding but also facilitates the application of learning-based algorithms.

For this analysis, a thorough evaluation of publicly available network traffic analyzers was conducted, including NTLFlowLyzer [11], CICFlowMeter [33], NFStream [55], and others. After a diligent investigation and comparison, it was concluded that NTLFlowLyzer was the most suitable choice for this work. The decision to select NTLFlowLyzer was based on its comprehensive feature set, impressive performance metrics, and notable capability to handle and analyze cloud traffic effectively, as all the packets have at least one layer of the

VXLAN encapsulation because of traffic mirroring. This tool aligns well with the research objectives, providing a robust platform for examining network data.

The findings are presented by organizing the analyzed data into Tables 2 and 3. These tables contain information in the form of generated CSVs detailing the data per label and activity. A distinctive label named "suspicious" is introduced in the tables. This label is assigned to activities that do not align with predefined benign or attack scenarios. The labeling of attack data is determined based on the timeline of the attacks, and the labeling of benign data is based on purely benign days' data. Data falling outside of these categories are labeled as "suspicious".

Table 2. Analyzed data information; distribution of labels across different days.

Date	# of Benign Flows	# of Attack Flows	# of Suspicious Flows	Sum of All Flows
Thursday 14 December	105,087	0	0	105,087
Saturday 16 December	189,678	0	0	189,678
Monday 18 December	68,444 (∼21%)	220,276 (∼69%)	31,810 (∼10%)	320,530
Tuesday 19 December	49,990 (∼58%)	8193 (∼10%)	27,296 (∼32%)	85,479
Sum of All Flows	**413,199 (∼59%)**	**228,469 (∼33%)**	**59,106 (∼8%)**	**700,774**

Table 3. Analyzed data information based on flow count, distribution of activities across different days.

ID	Activity	Thursday	Saturday	Monday	Tuesday	Sum
1	**Benign**	85,853	159,007	28,746	28,678	**302,284**
2	**Benign-SSH**	1333	1410	120	122	**2985**
3	**Benign-FTP**	329	97	28	29	**483**
4	**Benign-Email-Receive**	480	458	245	212	**1395**
5	**Benign-Email-Send**	596	558	442	342	**1938**
6	**Benign-Systemic**	2814	14,333	17,590	13,337	**48,074**
7	**Benign-Web Browsing HTTP-S**	10,471	12,603	21,114	7084	**51,272**
8	**Benign-TELNET**	3211	1212	159	186	**4768**
9	**Suspicious**	-	-	31,810	27,296	**59,106**
10	**Attack-TCP-Valid-SYN**	-	-	**8043**	-	-
11	**Attack-TCP-BYPass-V1**	-	-	**138,368**	-	-
12	**Attack-Killall-v2**	-	-	**6033**	-	-
13	**Attack-TCP-IGMP**	-	-	**7251**	-	-
14	**Attack-TCP-SYN**	-	-	**6953**	-	-
15	**Attack-Killer-TCP**	-	-	**6254**	-	-
16	**Attack-TCP-Control**	-	-	**5744**	-	-
17	**Attack-TCP-Flag-MIX**	-	-	**7416**	-	-
18	**Attack-TCP-Flag-SYN**	-	-	**7845**	-	-
19	**Attack-TCP-Flag-ACK**	-	-	**10,683**	-	-
20	**Attack-TCP-Flag-SYN-ACK**	-	-	**8204**	-	-
21	**Attack-TCP-Flag-ACK-PSH**	-	-	**7482**	-	-

Table 3. *Cont.*

ID	Activity	Thursday	Saturday	Monday	Tuesday	Sum
22	**Attack-TCP-Flag-RST-ACK**	-	-	-	**1445**	-
23	**Attack-TCP-Flag-SYN-TFO**	-	-	-	**3631**	-
24	**Attack-TCP-Flag-SYN-TIME**	-	-	-	**1360**	-
25	**Attack-TCP-Flag-OSYN**	-	-	-	**867**	-
26	**Attack-TCP-Flag-OSYNP**	-	-	-	**890**	-

5. Proposed Traffic Characterization Model

The section presents a new traffic characterization model designed to enhance the efficiency of identifying diverse benign and malicious activities within a network. The model's architecture, depicted in Figure 4, presents a multi-layered approach to improve the accuracy of activity classification while addressing various practical considerations, including complexity and interoperability. Adopting a multi-layered structure comprising classification and identification layers is rooted in balancing computational efficiency, resource utilization, and model performance. This segmentation allows for streamlined processing and optimized resource allocation throughout the classification pipeline.

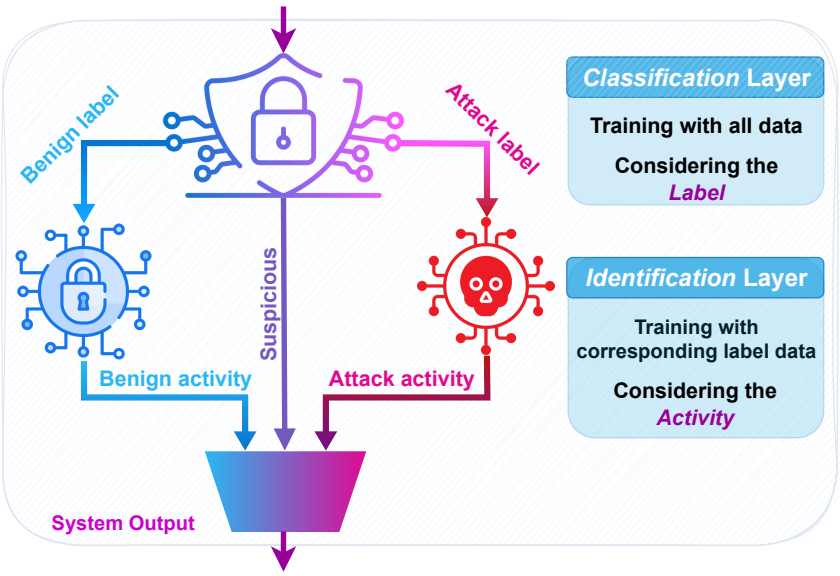

Figure 4. Proposed model.

Machine learning (ML) algorithms have primarily been used to learn benign and malicious behaviors. In contrast to employing deep learning (DL)-based architectures, the chosen model architecture emphasizes interoperability and transparency, which are crucial considerations in industry collaborations. While DL-based approaches offer compelling capabilities, their integration within industry settings can present challenges related to compatibility and interpretability. Furthermore, DL-based algorithms may require extensive labeled data for training, making them less feasible in scenarios with limited annotated datasets. We aim to provide industry stakeholders with clear insights into the decision-making processes underlying activity classification by opting for a more interpretable model architecture.

In the proposed model, the initial layer focuses on training with labeled data to establish the classification of each input. Subsequently, a dual-path architecture is implemented within the second layer, featuring two distinct models. One model is exclusively trained with benign data, while the other is trained solely with attack data. The training process

in the second layer is centered around considering the specific activity associated with the input. If the first layer classifies the input label as benign, the subsequent path directs the input to the benign model. Conversely, if the first layer identifies the input label as an attack, the input is routed to the corresponding attack model. In cases where the first layer outputs a suspicious label, the activity is similarly classified as suspicious. Ultimately, the final output of the system corresponds to the identified activity based on the cascaded processing in the multi-layered structure.

The proposed dual-path architecture within the second layer further enhances the model's flexibility and interpretability. The system can effectively distinguish between normal and malicious network activities by training separate models exclusively with benign and attack data. This approach mitigates the risk of overfitting and addresses concerns regarding the black-box nature of DL-based algorithms. In conclusion, the proposed traffic characterization model offers a sophisticated approach to distinguishing activities within network traffic. Its multi-layered structure balances computational efficiency with classification accuracy, while considerations of interoperability and interpretability ensure its suitability for real-world applications.

6. Experimental Results

This section delves into the comprehensive exploration of the experimental results. It commences with an analysis of the results of feature selection. Subsequently, it explains the implementation of the proposed model. Following this, the experiment scenarios are elucidated. Finally, the selected features are fed into the learning algorithms for classification and identification across varied experiment scenarios.

6.1. Feature Selection

This section comprehensively explores three key algorithms: analysis of variance (ANOVA), information gain, and ensemble learning with the extra tree classifier. Table 4 presents the results of the selected features for each algorithm based on the label classification.

Table 4. Selected features across different algorithms. F: forward, B: backward, D: delta, T: time, L: length.

	1st 10 Features	2nd 10 Features	3rd 10 Features	4th 10 Features
Analysis of Variance (ANOVA)	max hdr byte, min hdr byte, mean hdr byte, med hdr byte, mode hdr byte, F max hdr byte, F min hdr byte, F mean hdr byte, F std hdr byte, F med hdr byte	F cov hdr byte, F mode hdrbyte, F var hdr byte, B std hdr byte, B cov hdr byte, B var hdr byte, F init win byte, B init win byte, rst flag counts, B rst flag counts	psh flag % in total, rst flag % in total, F psh flag % in total, F syn flag % in total, B psh flag % in total, F psh flag % in F pkts, B psh flag % in B pkts, B rst flag % in B pkts, B pkts IAT mean, B pkts IAT max	B pkts IAT min, B pkts IAT total, B pkts IAT med, B pkts IAT mode, handshake duration, handshake state, mean B pkts DT, med B pkts DT, skew pkts DL, mode F pkts DL
Information Gain	duration, total hdr byte, max hdr byte, min hdr byte, mean hdr byte, med hdr byte, mode hdr byte, F total hdr byte, F max hdr byte, F min hdr byte	F mean hdr byte, F med hdr byte, F mode hdr byte, F init win byte, pkts rate, B pkts rate, F pkts rate, syn flag % in total, ack flag % in total, F syn flag % in total	pkts IAT mean, packet IAT max, packet IAT min, packet IAT total, pkts IAT med, pkts IAT mode, F pkts IAT mean, F pkts IAT max, F pkts IAT min, F pkts IAT total	F pkts IAT med, F pkts IAT mode, B pkts IAT total, handshake duration, mean pkts DT, var pkts DT, std pkts DT, med pkts DT, med B pkts DT, med F pkts DT

Table 4. *Cont.*

	1st 10 Features	2nd 10 Features	3rd 10 Features	4th 10 Features
Extra Tree	total hdr byte, max hdr byte, min hdr byte, mean hdr byte, med hdr byte, mode hdr byte, F total hdr byte, F max hdr byte, F min hdr byte, F mean hdr byte	F med hdr byte, F mode hdr byte, F init win byte, B init win byte, pkts rate, B pkts rate, F pkts rate, rst flag counts, B rst flag counts, syn flag % in total	rst flag % in total, F syn flag % in total, B rst flag % in total, F psh flag % in F pkts, F syn flag % in F pkts, B psh flag % in B pkts, pkts IAT mean, packet IAT max, packet IAT min, packet IAT total	pkts IAT med, pkts IAT mode, F pkts IAT mean, F pkts IAT max, F pkts IAT min, F pkts IAT total, F pkts IAT med, F pkts IAT mode, B pkts IAT total, B pkts IAT mode

6.2. Selecting and Implementing the Learning Algorithms

As the random forest (RF) demonstrates adeptness in handling multiple classes, accommodating diverse feature sets, and robustly managing imbalanced class distributions, this algorithm was selected for each layer in the proposed model. So, the model will be suitable for addressing the complexity inherent in distinguishing between 8 benign, 17 attack, and 1 suspicious activity categories. The ensemble nature of RF, which aggregates predictions from various decision trees, enhances the model's ability to discern intricate patterns within each activity class. Given the hierarchical architecture characterized by cascaded processing across multiple layers, the ensemble learning paradigm of RF proves instrumental in accurately classifying the 26 diverse activity classes. This amalgamation of hierarchical modeling and ensemble learning contributes significantly to the overall efficiency of activity identification within the multi-layered framework.

6.3. Experiment Scenarios and Performance Results

Seven distinct experiment scenarios, referred to as tasks, have been defined for a thorough assessment of the proposed model, as detailed below:

- ***Task 1:*** It involves classifying data into three categories: benign, suspicious, and attack.
- ***Task 2:*** It focuses on identifying specific attack activities within the dataset.
- ***Task 3:*** It concentrates on the identification of different benign activities.
- ***Task 4:*** It involves identifying both suspicious and benign activities.
- ***Task 5:*** It extends the identification challenge to both suspicious and attack activities.
- ***Task 6:*** It entails identifying attack activities and the benign label.
- ***Task 7:*** The final task encompasses the broadest identification challenge, requiring the model to classify all activities in the dataset.

Success in these seven tasks is imperative for any model working with this dataset. The model must effectively learn each label and activity to understand the underlying network characteristics comprehensively. Additional experiments were conducted to gauge the model's performance compared to alternative approaches, employing four well-established machine learning algorithms: naive Bayes (NB), support vector machine (SVM), random forest (RF), and XGBoost. Table 5 summarizes the results obtained from each task, providing a comparative overview of the performance of the proposed model against the four alternative machine learning algorithms. This comparative analysis simplifies the assessment of the proposed model's effectiveness in handling the complexity and diversity of the activity identification task.

Table 5. Performance results with top 40 features.

Task	*Model*	*Precision*	*Recall*	*F1-Score*	*Model*	*Precision*	*Recall*	*F1-Score*
Task 1	*NB*	0.71	0.46	0.48	*Logistic Reg.*	0.62	0.62	0.62
	SVM	0.56	0.60	0.58	*KNN*	0.91	0.91	0.91
	RF	0.94	0.94	0.94	*Decision Tree*	0.84	0.85	0.84
	XGBoost	0.94	0.94	0.94	*Extra Tree*	0.94	0.94	0.94
	Proposed	0.94	0.94	0.94	*Bagging*	0.93	0.94	0.93
Task 2	*NB*	0.55	0.58	0.56	*Logistic Reg.*	0.58	0.59	0.58
	SVM	0.55	0.58	0.56	*KNN*	0.75	0.70	0.72
	RF	0.79	0.72	0.75	*Decision Tree*	0.71	0.69	0.70
	XGBoost	0.85	0.71	0.76	*Extra Tree*	0.75	0.75	0.75
	Proposed	0.79	0.77	0.78	*Bagging*	0.80	0.72	0.75
Task 3	*NB*	0.76	0.13	0.10	*Logistic Reg.*	0.58	0.49	0.53
	SVM	0.50	0.45	0.48	*KNN*	0.89	0.88	0.89
	RF	0.96	0.94	0.95	*Decision Tree*	0.85	0.82	0.83
	XGBoost	0.96	0.95	0.95	*Extra Tree*	0.93	0.93	0.93
	Proposed	0.96	0.96	0.96	*Bagging*	0.94	0.92	0.93
Task 4	*NB*	0.64	0.11	0.08	*Logistic Reg.*	0.45	0.49	0.47
	SVM	0.37	0.40	0.38	*KNN*	0.91	0.90	0.91
	RF	0.95	0.92	0.93	*Decision Tree*	0.86	0.84	0.85
	XGBoost	0.95	0.94	0.94	*Extra Tree*	0.95	0.92	0.93
	Proposed	0.96	0.92	0.93	*Bagging*	0.93	0.93	0.93
Task 5	*NB*	0.41	0.46	0.43	*Logistic Reg.*	0.51	0.56	0.53
	SVM	0.41	0.46	0.43	*KNN*	0.69	0.69	0.69
	RF	0.75	0.73	0.73	*Decision Tree*	0.61	0.67	0.64
	XGBoost	0.78	0.74	0.74	*Extra Tree*	0.74	0.74	0.74
	Proposed	0.88	0.84	0.86	*Bagging*	0.72	0.71	0.71
Task 6	*NB*	0.58	0.29	0.19	*Logistic Reg.*	0.37	0.48	0.40
	SVM	0.35	0.50	0.40	*KNN*	0.85	0.85	0.85
	RF	0.88	0.86	0.87	*Decision Tree*	0.81	0.82	0.81
	XGBoost	0.91	0.86	0.87	*Extra Tree*	0.84	0.86	0.85
	Proposed	0.97	0.96	0.97	*Bagging*	0.82	0.80	0.81
Task 7	*NB*	0.51	0.27	0.17	*Logistic Reg.*	0.48	0.63	0.53
	SVM	0.29	0.46	0.35	*KNN*	0.84	0.84	0.84
	RF	0.85	0.85	0.85	*Decision Tree*	0.74	0.78	0.75
	XGBoost	0.86	0.86	0.85	*Extra Tree*	0.85	0.86	0.85
	Proposed	0.91	0.91	0.91	*Bagging*	0.84	0.84	0.84

7. Analysis and Discussion

This section analyzes the experimental results of the previous section.

7.1. Feature Selection Analysis

This subsection carefully analyzes feature selection, examining each algorithm and exploring selected and non-selected feature categories in detail.

7.1.1. Selected Features Analysis

This subsection explores each category of selected features, which are then amalgamated for a more comprehensive analysis.

- ***Header-Related Features***
  Examining header-related features across three feature selection algorithms unveils a consistent trend where the top 10 selected features consistently pertain to header bytes. This initiates a detailed analysis focusing on header bytes in this context, examining specific scenarios and characteristics.
  The TCP header values exhibit limited patterns for each network flow in benign scenarios. For instance, in a benign context, a standardized handshake procedure occurs at the beginning of each flow, resulting in uniform header values. Any deviations or anomalies in these header-related patterns, particularly those at the onset of a flow, become easily detectable. Such anomalies include DDoS TCP handshake, DDoS TCP SYN (various TCP SYN scenarios), and DDoS TCP SYN-ACK, where attackers aim to exhaust system resources by initiating and keeping open connections to prevent benign connections from establishing. Notably, features like the handshake state and flow duration offer valuable insights into the underlying behavior and nature of the flow.
  The selected features underscore the significance of both the header and handshake categories in effectively distinguishing between DDoS attacks and benign data. However, more than using these features alone may be required in complex attack scenarios employing a customary handshake. Additional features become necessary in such cases. The prominence of header-related features arises from the fact that, in general DDoS scenarios, attackers often employ predefined packets or requests, increasing the likelihood of similar header-related options. This similarity simplifies the differentiation between DDoS and benign data. For instance, a SYN flood might generate numerous connection requests to a destination port while changing only the source port value in the TCP header.
  Furthermore, as detailed in Section 4.2, a significant proportion of DDoS attacks in this dataset manipulate header values. Consequently, features related to the TCP header emerge as the most informative. Notably, among the top 40 selected features, approximately 75% are directly derived from header values. These feature categories include header bytes, init win bytes, flag percentage, and handshake-related metrics. In contrast, categories such as delta time, IAT, and packet rate, which are not directly linked to the TCP header, contribute to a comprehensive approach to detecting TCP-based DDoS attacks.
  This underscores the importance of considering all aspects of the TCP header for effective detection. For example, a TCP-ACK flood attack inundates the target with a high volume of ACK requests, disrupting the flag distribution within header bytes compared to regular traffic.
- ***Flag-related features***
  DDoS attacks often manipulate TCP flags, such as SYN, RST, and PSH, to disrupt standard communication patterns. Within this category, anomalies in the distribution of flags can signal specific attack types. For instance, an abnormal SYN–ACK ratio may indicate a TCP-SYN flood attack, while a dynamic flag distribution strategy could mimic standard traffic patterns, challenging detection mechanisms. This highlights the importance of analyzing flag percentages for a nuanced understanding of attack tactics.
- ***Delta-time-related features***
  Attackers disrupt regular communication by introducing variations in the time intervals between successive packets. Unusual values in "delta time" features can indicate irregular packet transmission patterns. For example, pulsing DDoS attacks involve rhythmic variations in inter-packet delta times, making it challenging for defenders to predict attack patterns. Conversely, specific DDoS attacks use consistently low delta time between packets to maximize traffic volume and increase the likelihood of

successful disruption. Identifying and analyzing these delta time patterns are crucial to robust detection mechanisms.

- ***Inter-arrival time (IAT)-related features***
  The inter-arrival time reflects variations in packet transmission intervals. Bursty DDoS traffic exhibits irregular spikes in packet transmission, causing variations in inter-arrival times. Recognizing and distinguishing these bursts is critical for identifying potential attacks amidst benign traffic. Moreover, sophisticated DDoS attacks may involve coordinated sequences with specific IAT patterns, such as rapid bursts followed by brief periods of inactivity. Understanding and detecting these coordinated sequences enhances the accuracy of identifying intricate attack strategies.
- ***Rate-related features***
  DDoS attacks typically exhibit abnormally high packet rates to cause network congestion and service disruption. Elevated values in the "packet rate" feature can signal the presence of a DDoS attack, especially when compared to the baseline packet rates observed during regular network activity. This can indicate an adaptive strategy, where attackers dynamically adjust the packet rate during an attack to adapt to changing network conditions or evade static detection thresholds. Conversely, some DDoS attacks intentionally maintain a low and stealthy packet rate to avoid detection. Identifying and analyzing deviations from the expected baseline requires a nuanced analysis of packet rate dynamics.
- ***All together***
  The consistent presence of anomalies across multiple feature categories collectively enhances attack detection accuracy. Recognizing patterns across these diverse features contributes to a more robust detection mechanism. Examining the correlations between different feature categories reveals more comprehensive attack signatures. For instance, a high packet rate combined with abnormal flag percentages may indicate a sophisticated DDoS strategy. Understanding these interdependencies allows for a deeper understanding of the evolving nature of attacks and improves detection accuracy.
  Moreover, attackers may dynamically adjust their strategies throughout an attack. Continuous monitoring of features enables the identification of evolving attack patterns. A nuanced understanding of the dynamic nature of attacks is becoming imperative for developing adaptive detection mechanisms capable of responding to emerging threats in real-time. Incorporating diverse, informative feature categories into consideration provides a holistic and detailed perspective on network traffic. This inclusive approach empowers the development of robust and adaptive detection models adept at identifying various DDoS attack tactics. By comprehensively examining correlations and evolving patterns, these models prove effective in staying ahead of attackers and responding dynamically to the intricate landscape of cyber threats.
- ***Online detection strategy***
  In an online detection system, where computing a single feature value for a potentially high number of open flows can be resource-intensive and time consuming, it is crucial to adopt an approach that optimizes both time and resources. The strategy involves a structured, multi-layered framework, where features are computed in an order that facilitates early detection with minimal computational cost.
  The first layer of this framework prioritizes features that are easy to calculate and contribute to early detection. Notably, features associated with the handshake scenario emerge as critical components of this initial layer. The rationale behind this prioritization is twofold. Firstly, in the initial stages of a network flow, decisions about its malicious nature cannot be accurately made by calculating features such as header bytes mean or packet rate. Secondly, by focusing on features related to the handshake process, the system can efficiently identify and halt potentially malicious incoming traffic lacking a valid handshake process. This early intervention conserves system resources by avoiding calculating additional feature values for such connec-

tions. Furthermore, this proactive approach ensures that more resources are available for benign users.

Another set of informative features for early detection includes "init win bytes". Attackers may manipulate the initial window size in TCP packets during DDoS attacks to impact the target's resource utilization. Anomalies or irregularities in the values of the "init win bytes" feature serve as indicators of potential attempts to exploit vulnerabilities, overwhelm network resources, or establish malicious connections. Beyond the early detection phase, attention shifts to flows characterized by a normal handshake process, usual flags, and regular features. A normal handshake process denotes that all TCP steps have been executed within a reasonable timeframe. The system monitors and calculates the other most informative features (other selected features) for such flows.

7.1.2. Not-Selected Features Analysis

In the context of network traffic analysis, the absence of certain features, such as packet delta length, header delta length, and payload delta length, can provide valuable insights into the nature of the data being examined. While the previous analysis has covered selected features, addressing why these specific features may be absent or are deemed non-informative for the given task is essential.

Firstly, the absence of "packet delta length" as a significant feature could be attributed to its limited relevance in DDoS detection and identification. Packet delta length typically measures the difference in length between consecutive packets. The packet delta length may fluctuate considerably when a network experiences diverse benign activities, protocols, and usage patterns. This variability can render it challenging to establish a reliable baseline for distinguishing between normal and malicious traffic. Additionally, for TCP-based DDoS attacks, attackers may strategically manipulate packet lengths to mimic legitimate traffic, further diminishing the discriminatory power of this feature.

Similarly, the exclusion of "header delta length" could be justified by its potential lack of discriminatory value in identifying DDoS attacks. "Header delta length" typically refers to changes in the lengths of headers between successive packets. However, in the case of TCP-based DDoS attacks, attackers often craft their packets to maintain valid header structures, making it challenging to differentiate attack traffic from benign traffic based solely on header length variations. As a result, this feature might not provide meaningful insights into distinguishing malicious activities.

Furthermore, the decision not to consider "payload delta length" could be reasoned by the inherent challenges associated with payload analysis in DDoS detection. Payloads in network traffic can exhibit significant diversity due to the vast array of benign activities and protocols. Attempting to identify DDoS attacks based on payload characteristics becomes impractical in such scenarios, as attackers often employ valid payload data to evade detection at the application layer. The similarity between benign and attack payloads under the TCP protocol further diminishes the utility of payload delta length as a discriminative feature.

In summary, the analysis's absence of "packet delta length", "header delta length", and "payload delta length" indicates a thoughtful consideration of their limited suitability for DDoS detection in the given context. The intricate nature of network traffic, the presence of diverse benign activities, and the tactics employed by attackers in crafting their traffic collectively lead us to conclude that these features may not offer meaningful insights or reliable discrimination in identifying DDoS attacks based on the TCP protocol.

7.2. Performance Analysis

In light of the obtained results, it can be deduced that the observed suspicious activities may be closely associated with the attack category. The analysis indicates a notable discrepancy in performance across various tasks, with particular emphasis on the task involving the identification between attack and suspicious activities. Across all models con-

sidered, this specific task demonstrated suboptimal performance, suggesting a challenge in accurately distinguishing between these two categories.

Concurrently, it is noteworthy that identifying benign labels and attack activities exhibited a commendable performance across the models under scrutiny. This disparity in task performance underscores the complexity involved in identifying and classifying potentially malicious activities, especially when distinguishing between attacks and activities categorized as suspicious. The findings imply a potential improvement in model robustness and training strategies, explicitly targeting the nuanced differentiation between attack and suspicious patterns.

The superior performance of the proposed model, as contrasted with traditional solutions, can be attributed to a deliberate design choice involving the utilization of models that concentrate on a reduced number of classes. In direct contrast to the conventional approach of incorporating all classes within a singular model, the proposed methodology involves the creation of individual models, each dedicated to a subset of classes. This strategic segmentation allows each model to specialize in and precisely learn the distinct behaviors associated with a limited number of classes. Consequently, the models achieve a more nuanced understanding of the targeted classes, avoiding the challenge of moderately learning many classes, which often leads to under-fitting.

The approach's efficacy is underscored by the premise that the models are better equipped to capture and comprehend the intricacies of class behaviors by grouping different classes and employing a dedicated classifier for each group. This preventive measure against under-fitting contributes significantly to the enhanced overall performance observed in the proposed model. The findings advocate a paradigm shift towards more focused and specialized models, demonstrating the potential benefits of class-specific learning in developing robust, high-performing solutions. Further exploration into the dynamics of class grouping and its impact on model generalization may offer valuable insights for refining and optimizing future iterations of the proposed methodology.

7.3. Addressing Previous Shortcomings

This subsection discusses how this work has effectively addressed the identified shortcomings outlined in Section 2.2:

1. **Imbalanced class distribution:** The dataset achieves a more balanced distribution, with a ratio of 60% benign to 40% non-benign data, including 8% labeled as suspicious. This balance ensures a more representative dataset for training and evaluation.
2. **Limited diversity of attacks:** This work incorporates a wide range of DDoS attacks, totaling 17 different attack types, surpassing the diversity found in existing datasets. This ensures comprehensive coverage of various attack scenarios and enhances the fidelity of the new dataset.
3. **Outdated threat scenarios:** A thorough analysis was conducted, leveraging reports from Microsoft [56–58] and Cloudflare [59] to identify and prioritize recent attack trends. Additionally, utilizing a third-party service for attack execution ensures the incorporation of up-to-date attack methodologies, addressing concerns regarding outdated threat scenarios.
4. **Lack of realistic network traffic:** The proposed approach includes the development of a Benign User Traffic generator capable of producing realistic benign user data. This contrasts with previous approaches that relied on simulated or less accurate benign data, thereby enhancing the realism of the new dataset.
5. **Absence of encrypted traffic:** The benign traffic generator is configured to generate diverse traffic, including encrypted traffic, mirroring real-world network scenarios more accurately. This ensures that the dataset encompasses the complexities of encrypted communication, which are often overlooked in previous datasets.
6. **Insufficient labeling accuracy:** Benign and attack scenarios were meticulously scheduled to ensure precise labeling, supported by experimental results validating the

accuracy of the labeling process across different algorithms. This meticulous approach enhances the reliability and trustworthiness of the new dataset.

7. **Limited incorporation of user behavior:** The analysis of previous works in user network behavior informs the configuration of diverse user profiles within the benign traffic generator. This includes various behaviors such as web browsing, file transfer, and email checking, ensuring a more comprehensive representation of user activity within the dataset.
8. **Incompatibility with modern protocols:** The generated dataset mirrors realistic network traffic, encompassing a wide array of protocols, including modern ones like QUIC, DNS, and HTTPS. This ensures compatibility with modern network environments, addressing concerns about protocol compatibility present in previous datasets.
9. **Limited exploration of low-rate DDoS attacks:** The new dataset includes diverse attack strategies, ranging from low-rate to high-rate attacks, as evidenced by Table 3. This comprehensive coverage ensures that the dataset adequately represents the variability of DDoS attack intensities encountered in real-world scenarios.
10. **Lack of realistic DDoS traffic variability:** Utilizing a third-party service for DDoS attack execution ensures the incorporation of realistic attack data with diverse strategies. This contrasts with previous approaches that may have utilized packet generator tools without specific attack strategies, enhancing the variability and realism of the new dataset.

In conclusion, this research addresses various shortcomings present in existing datasets related to capturing malicious events in real-world scenarios. The fidelity and relevance of the newly generated dataset for studying network security and DDoS attack detection have been significantly improved through meticulous analysis and methodological enhancements. By achieving a more balanced class distribution, incorporating diverse attack types, prioritizing up-to-date threat scenarios, and ensuring the inclusion of realistic network traffic, diverse user behaviors, and compatibility with modern protocols, the dataset provides a comprehensive representation of real-world network environments. The rigorous approach to labeling accuracy and inclusion of low-rate DDoS attacks further contribute to the richness and variability of the dataset, enabling more robust evaluations and benchmarking of detection algorithms. These enhancements underscore the commitment to advancing state-of-the-art network security research and providing valuable resources for developing and evaluating effective DDoS mitigation strategies.

7.4. Comparison with Previous Datasets

Using the established methodology outlined in the work of [34], we employed a comprehensive framework for evaluating key metrics to compare the datasets. The results of this comparative analysis are presented in Table 6, revealing compelling insights into the performance of the new dataset relative to its predecessors in the last decade. The findings demonstrate that the new dataset surpasses all other datasets examined across various criteria, underscoring its superiority in effectively capturing and representing cloud network traffic dynamics.

Table 6. Datasets comparison.

Dataset	*Date*	*# Labels*	*# Features*	*Realistic Traffic*	*Data Distribution Benign–Malicious*	*Analyzer*	*User Profile*	*Cloud Env.*
ISCX2012	2012	6	18	✗	97-3	ISCXFlowMeter	✗	✗
CTU-13	2013	14	84	✓	-	Argus-NetFlow	✗	✗
UNSW-NB15	2015	10	157	✗	87-13	Argus-Bro-IDS	✗	✗
CICIDS2017	2017	14	80	✓	78-22	CICFlowMeter	✓	✗

Table 6. *Cont.*

Dataset	*Date*	*# Labels*	*# Features*	*Realistic Traffic*	*Data Distribution Benign–Malicious*	*Analyzer*	*User Profile*	*Cloud Env.*
CSE-CIC-IDS2018	2018	15	80	✓	83-17	CICFlowMeter	✓	✗
CIC-DDoS2019	2019	15	80	✓	10-90	CICFlowMeter	✓	✗
SR-BH2020	2020	13	32	✓	58-42	Not Public	✗	✗
CUPID	2022	2	80	✗	88-12	CICFlowMeter	✗	✗
BCCC-cPacket-Cloud-DDoS-2024	2024	26	322	✓	60-40	NTLFlowLyzer	✓	✓

8. Conclusions and Future Works

Establishing reliable and publicly accessible DDoS evaluation datasets is paramount for researchers and industry stakeholders. This paper explores the state-of-the-art generation of DDoS datasets, evaluates 16 publicly available datasets, and highlights 15 shortcomings across various perspectives. Following this, a comprehensive network dataset creation roadmap is introduced along with generating a new cloud-based DDoS dataset named BCCC-cPacket-Cloud-DDoS-2024 [9]. In summary, this paper aims to cover the shortcomings in generating DDoS datasets and establish a robust foundation for advanced research and development in the field by analyzing attack trends, benign traffic generation, user behavioral profiling, and cloud infrastructure.

The experimental and analytical phases identify an optimal feature set to distinguish between benign, suspicious, and DDoS attack traffic. This involves employing three distinct feature selection algorithms, ANOVA, information gain, and extra tree, each approaching the task uniquely. A new multi-layered detection model is designed and implemented to classify various activities proficiently. The proposed model thoroughly evaluates the performance and accuracy of the chosen features. Four well-known classifiers are implemented and tested with the selected feature set to ensure a comprehensive assessment. Comparing the performance of the proposed model with traditional approaches highlights its superior effectiveness.

The approach will be enhanced by considering diverse features and data sources, exploring different protocols and layers, including a broader range of attack types, and factoring in benign user behavior. The analysis will be extended to other network types, such as IoT and IIoT networks, and online detection mechanisms will be integrated for real-time threat response.

Author Contributions: M.S.: Designed and implemented the infrastructure, benign user network traffic generator, and detection model; scheduled and executed attacks and benign traffic; wrote and prepared the main manuscript text and conceptualization. A.H.L.: Project founder, Contributed to supervision, conceptualization, writing—review, editing, and securing the funding. V.R. and R.N.: Provided advice and suggestions as the industry collaborator and reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: The authors acknowledge the grant from the Natural Sciences and Engineering Research Council of Canada—NSERC (#RGPIN-2020-04701)—to Arash Habibi Lashkari and research fund from cPacket for Cloud Infrastructure and graduate student funding package.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: After publishing this paper, the new DDoS Dataset, "BCCC-cPacket-Cloud-DDoS-2024", will be publicly available on the website [9]. Additionally, the implementation code for the Benign User Profiler (BUP) will be accessible on the GitHub repository [10].

Acknowledgments: We thank cPacket for their invaluable financial and technical support in this research project.

Conflicts of Interest: The authors declare no conflicts of interest. Also, the funders had no role in the design of the study, in the collection, analysis, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

Appendix A

Table A1. Attack schedule.

Date	Time	Attack	Target	Target IP	Capturer
Monday, 18 December 2023	9:00–9:20	(1) TCP-SYN (Valid SYN)	Windows-machine-1	3.96.128.96 (10.0.9.208)	35.183.206.0 (10.0.17.180)
	9:30–9:50	(2) TCP-BYPass-V1	Linux-admin-1	99.79.45.168 (10.0.6.142)	
	10:00–10:20	(3) Killall-v2	Linux-webserver	15.222.45.224 (10.0.4.57)	
	10:32–10:52	(4) TCP-IGMP	Windows-machine-2	35.182.194.19 (10.0.4.132)	
	11:00–11:20	(5) TCP-SYN	Linux-admin-1	99.79.45.168 (10.0.6.142)	
	11:30–11:50	(6) Killer-TCP	Windows-machine-4	35.183.15.52 (10.0.3.52)	
	13:00–13:20	(7) TCP-Control	Linux-webserver	15.222.45.224 (10.0.4.57)	
	13:30–13:50	(8) TCP-MIX	Linux-admin-1	99.79.45.168 (10.0.6.142)	
	14:00–14:20	(9) TCP-SYN (syn flags only)	Windows-machine-2	35.182.194.19 (10.0.4.132)	
	14:30–14:50	(10) TCP-ACK	Windows-machine-3	3.99.186.200 (10.0.11.84)	
	15:00–15:20	(11) TCP-SYN-ACK	Linux-webserver	15.222.45.224 (10.0.4.57)	
	15:30–15:50	(12) TCP-ACK-PSH	Windows-machine-4	35.183.15.52 (10.0.3.52)	
Tuesday, 19 December 2023	9:00–9:20	(13) TCP-RST-ACK	Linux-webserver	15.222.45.224 (10.0.4.57)	3.99.150.239 (10.0.17.180)
	9:30–9:50	(14) TCP-SYN-TFO	Windows-machine-3	3.99.186.200 (10.0.11.84)	
	10:00–10:20	(15) TCP-SYN-TIME	Linux-admin-1	99.79.45.168 (10.0.6.142)	
	10:50–11:10	(16) TCP-OSYN	Windows-machine-1	3.96.128.96 (10.0.9.208)	
	11:20–11:40	(17) TCP-OSYNP	Linux-webserver	15.222.45.224 (10.0.4.57)	

Table A2. Sample Benign schedule for a normal user (Windows).

Start Times	Behavior	Detail
09:00, 09:25, 09:50, 10:15, 10:40, 11:05, 11:30, 11:55, 12:20, 12:45, 13:10, 13:35, 14:00, 14:25, 14:50, 15:15, 15:40, 16:05, 16:30, 16:55	**Email (Sending)**	All emails contain attachments as well.
09:17, 09:42, 10:07, 10:32, 10:57, 11:22, 11:47, 12:12, 12:37, 13:02, 13:27, 13:52, 14:17, 14:42, 15:07, 15:32, 15:57, 16:22, 16:47, 17:12	**Email (Reading)**	All emails contain attachments as well.
09:40	**Web Browsing (Music Streaming)**	It is a live stream.
12:30	**Web Browsing (Video Watching)**	YouTube
09:05	**Web Browsing (Video Watching)**	Continued with new video after finishing each one.
09:40, 13:40	**Web Browsing (News Checking)**	CBC.ca
10:00, 10:40, 14:20, 14:50, 15:20, 11:30, 12:03, 12: 36, 15:25, 15:35, 15:45	**Web Browsing (Downloading)**	File sizes: 5 GB, 228 MB, 4 MB, 4 MB, 4 MB, 1.7 KB, 1.7 KB, 1.7 KB, 100 MB, 100 MB, 100 MB
11:43	**Web Browsing (Food)**	UberEats
10:02, 10:22, 10:25, 10:45	**Web Browsing (Shopping)**	Amazon
13:30, 13:40	**Web Browsing (Shopping)**	Bestbuy
15:13	**Web Browsing (Social Media)**	LinkedIn
16:00	**Web Browsing (Taxi)**	Uber

References

1. Aljuhani, A. Machine learning approaches for combating distributed denial of service attacks in modern networking environments. *IEEE Access* **2021**, *9*, 42236–42264. [CrossRef]
2. Bawany, N.Z.; Shamsi, J.A.; Salah, K. DDoS attack detection and mitigation using SDN: Methods, practices, and solutions. *Arab. J. Sci. Eng.* **2017**, *42*, 425–441. [CrossRef]
3. Agarwal, A.; Khari, M.; Singh, R. Detection of DDOS attack using deep learning model in cloud storage application. In *Wireless Personal Communications*; Springer: Berlin, Germany, 2021; Volume 127, pp. 1–21.
4. Aamir, M.; Zaidi, M.A. A survey on DDoS attack and defense strategies: From traditional schemes to current techniques. *Interdiscip. Inf. Sci.* **2013**, *19*, 173–200. [CrossRef]
5. Singh, J.; Behal, S. Detection and mitigation of DDoS attacks in SDN: A comprehensive review, research challenges and future directions. *Comput. Sci. Rev.* **2020**, *37*, 100279. [CrossRef]
6. Zeadally, S.; Adi, E.; Baig, Z.; Khan, I.A. Harnessing artificial intelligence capabilities to improve cybersecurity. *IEEE Access* **2020**, *8*, 23817–23837. [CrossRef]
7. Wu, H.; Han, H.; Wang, X.; Sun, S. Research on artificial intelligence enhancing internet of things security: A survey. *IEEE Access* **2020**, *8*, 153826–153848. [CrossRef]
8. Thakkar, A.; Lohiya, R. A review of the advancement in intrusion detection datasets. *Procedia Comput. Sci.* **2020**, *167*, 636–645. [CrossRef]
9. BCCC-Dataset. BCCC CPacket Cloud-based DDoS 2024. Behaviour-Centric Cybersecurity Center (BCCC). Available online: https://www.yorku.ca/research/bccc/ucs-technical/cybersecurity-datasets-cds (accessed on 8 March 2024).
10. BCCC-BUP. Benign User Profiler (BUP). Behaviour-Centric Cybersecurity Center (BCCC). Available online: https://github.com/ahlashkari/Benign-User-Profiler-BUP (accessed on 8 March 2024).
11. BCCC-NTLFlowLyzer. Network and Transport Layer Flow Analyzer (NTLFlowLyzer), Retrieved 10 February 2024. Behaviour-Centric Cybersecurity Center (BCCC). Available online: https://github.com/ahlashkari/NTLFlowLyzer (accessed on 8 September 2023).
12. Tabrizchi, H.; Kuchaki Rafsanjani, M. A survey on security challenges in cloud computing: Issues, threats, and solutions. *J. Supercomput.* **2020**, *76*, 9493–9532. [CrossRef]

13. Saxena, R.; Dey, S. DDoS attack prevention using collaborative approach for cloud computing. *Clust. Comput.* **2020**, *23*, 1329–1344. [CrossRef]
14. Zekri, M.; El Kafhali, S.; Aboutabit, N.; Saadi, Y. DDoS attack detection using machine learning techniques in cloud computing environments. In Proceedings of the 2017 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech), Rabat, Morocco, 24–26 October 2017
15. Kautish, S.; Reyana, A.; Vidyarthi, A. SDMTA: Attack detection and mitigation mechanism for DDoS vulnerabilities in hybrid cloud environment. *IEEE Trans. Ind. Inform.* **2022**, *18*, 6455–6463. [CrossRef]
16. Wani, A.R.; Rana, Q.; Saxena, U.; Pandey, N. Analysis and detection of DDoS attacks on cloud computing environment using machine learning techniques. In Proceedings of the 2019 Amity International Conference on artificial intelligence (AICAI), Dubai, United Arab Emirates, 4–6 February 2019; pp. 870–875.
17. Choi, J.; Choi, C.; Ko, B.; Kim, P. A method of DDoS attack detection using HTTP packet pattern and rule engine in the cloud computing environment. *Soft Comput.* **2014**, *18*, 1697–1703. [CrossRef]
18. Mugunthan, S. Soft computing based autonomous low rate DDOS attack detection and security for cloud computing. *J. Soft Comput. Paradig.* **2019**, *1*, 80–90.
19. Virupakshar, K.B.; Asundi, M.; Channal, K.; Shettar, P.; Patil, S.; Narayan, D. Distributed denial of service (DDoS) attacks detection system for OpenStack-based private cloud. *Procedia Comput. Sci.* **2020**, *167*, 2297–2307. [CrossRef]
20. Jindal, R.; Anwar, A. Emerging Trends of Recently Published Datasets for Intrusion Detection Systems (IDS): A Survey. *arXiv* **2021**, arXiv:2110.00773.
21. Chang, V.; Golightly, L.; Modesti, P.; Xu, Q.A.; Doan, L.M.T.; Hall, K.; Boddu, S.; Kobusińska, A. A survey on intrusion detection systems for fog and cloud computing. *Future Internet* **2022**, *14*, 89. [CrossRef]
22. 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.
23. Koga, R. Spoofer Data. Available online: https://catalog.caida.org/dataset/spoofer_data (accessed on 8 September 2023).
24. DDoS 2007 Attack. Available online: https://catalog.caida.org/dataset/ddos_attack_2007 (accessed on 8 September 2023).
25. CAIDA Randomly and Uniformly Spoofed Denial-of-Service Attack Metadata. Available online: https://catalog.caida.org/dataset/2017_imc_rsdos_targets (accessed on 8 September 2023).
26. Aggregated Daily RSDoS Attack Metadata (Corsaro 2). Available online: https://catalog.caida.org/dataset/telescope_corsaro2_daily_rsdos (accessed on 8 September 2023).
27. Sangster, B.; O'Connor, T.; Cook, T.; Fanelli, R.; Dean, E.; Morrell, C.; Conti, G.J. Toward Instrumenting Network Warfare Competitions to Generate Labeled Datasets. In Proceedings of the 2nd conference on Cyber Security Experimentation and Test (CSET), Montreal, QC, Canada, 10 August 2009.
28. Song, J.; Takakura, H.; Okabe, Y.; Eto, M.; Inoue, D.; Nakao, K. Statistical analysis of honeypot data and building of Kyoto 2006+ dataset for NIDS evaluation. In Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security, Salzburg, Austria, 10 April 2011; pp. 29–36.
29. 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]
30. Creech, G.; Hu, J. Generation of a new IDS test dataset: Time to retire the KDD collection. In Proceedings of the 2013 IEEE Wireless Communications and Networking Conference (WCNC), Shanghai, China, 7–10 April 2013; pp. 4487–4492.
31. Garcia, S.; Grill, M.; Stiborek, J.; Zunino, A. An empirical comparison of botnet detection methods. *Comput. Secur.* **2014**, *45*, 100–123. [CrossRef]
32. Moustafa, N.; Slay, J. UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In Proceedings of the 2015 Military Communications and Information Systems Conference (MilCIS), Canberra, Australia, 2015, pp. 1–6.
33. 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 (ICISSP), Porto, Portugal, 19–21 February 2017; pp. 253–262.
34. Sharafaldin, I.; Lashkari, A.H.; Ghorbani, A.A. Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp* **2018**, *1*, 108–116.
35. Sharafaldin, I.; Lashkari, A.H.; Hakak, S.; Ghorbani, A.A. Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy. In Proceedings of the 2019 International Carnahan Conference on Security Technology (ICCST), Chennai, India, 1–3 October 2019; pp. 1–8.
36. Riera, T.S.; Higuera, J.R.B.; Higuera, J.B.; Herraiz, J.J.M.; Montalvo, J.A.S. A new multi-label dataset for Web attacks CAPEC classification using machine learning techniques. *Comput. Secur.* **2022**, *120*, 102788. [CrossRef]
37. Lawrence, H.; Ezeobi, U.; Tauil, O.; Nosal, J.; Redwood, O.; Zhuang, Y.; Bloom, G. CUPID: A labeled dataset with Pentesting for evaluation of network intrusion detection. *J. Syst. Archit.* **2022**, *129*, 102621. [CrossRef]
38. Alhijawi, B.; Almajali, S.; Elgala, H.; Salameh, H.B.; Ayyash, M. A survey on DoS/DDoS mitigation techniques in SDNs: Classification, comparison, solutions, testing tools and datasets. *Comput. Electr. Eng.* **2022**, *99*, 107706. [CrossRef]
39. Packeth Sourceforge. Available online: http://packeth.sourceforge.net (accessed on 8 September 2023).

40. Iperf GitHub Page. Available online: https://github.com/esnet/iperf (accessed on 8 September 2023).
41. Distributed Internet Traffic Generator. Available online: http://traffic.comics.unina.it/software/ITG/ (accessed on 8 September 2023).
42. Ostinato. Available online: https://ostinato.org/ (accessed on 8 September 2023).
43. Solarwinds Traffic Generator Wan Killer. Available online: https://www.solarwinds.com/engineers-toolset/use-cases/traffic-generator-wan-killer (accessed on 8 September 2023).
44. Packet Sender. Available online: https://packetsender.com/ (accessed on 8 September 2023).
45. NMap. Available online: https://nmap.org/nping (accessed on 8 September 2023).
46. Net Scan Tools. Available online: https://www.netscantools.com/ (accessed on 8 September 2023).
47. Trex-tgn CISCO. Available online: https://trex-tgn.cisco.com (accessed on 8 September 2023).
48. Duarte Torres, S.; Weber, I.; Hiemstra, D. Analysis of search and browsing behavior of young users on the web. *Acm Trans. Web (Tweb)* **2014**, *8*, 1–54. [CrossRef]
49. Kumar, R.; Tomkins, A. A characterization of online browsing behavior. In Proceedings of the 19th International Conference on World Wide Web, Raleigh, NC, USA, 26–30 April 2010; pp. 561–570.
50. Wu, I.C.; Yu, H.K. Sequential analysis and clustering to investigate users' online shopping behaviors based on need-states. *Inf. Process. Manag.* **2020**, *57*, 102323. [CrossRef]
51. Möller, J.; van de Velde, R.N.; Merten, L.; Puschmann, C. Explaining online news engagement based on browsing behavior: Creatures of habit? *Soc. Sci. Comput. Rev.* **2020**, *38*, 616–632. [CrossRef]
52. Bakhshi, T.; Ghita, B. User traffic profiling. In Proceedings of the 2015 Internet Technologies and Applications (ITA), Wrexham, UK, 8–11 September 2015; pp. 91–97.
53. Varet, A.; Larrieu, N. Realistic network traffic profile generation: Theory and practice. *Comput. Inf. Sci.* **2014**, *7*, 1. [CrossRef]
54. Nelson, R.; Shukla, A.; Smith, C. Web Browser Forensics in Google Chrome, Mozilla Firefox, and the Tor Browser Bundle. In *Digital Forensic Education: An Experiential Learning Approach*; Springer Book: Berlin, Germany, 2020; pp. 219–241.
55. Aouini, Z.; Pekar, A. NFStream: A flexible network data analysis framework. *Comput. Netw.* **2022**, *204*, 108719. [CrossRef]
56. Azure DDoS Protection—2021 Q1 and Q2 DDoS Attack Trends. Available online: https://azure.microsoft.com/en-us/blog/azure-ddos-protection-2021-q1-and-q2-ddos-attack-trends/ (accessed on 8 September 2023).
57. Azure DDoS Protection—2021 Q3 and Q4 DDoS Attack Trends. Available online: https://azure.microsoft.com/en-us/blog/azure-ddos-protection-2021-q3-and-q4-ddos-attack-trends/ (accessed on 8 September 2023).
58. 2022 in Review: DDoS Attack Trends and Insights. Available online: https://www.microsoft.com/en-us/security/blog/2023/02/21/2022-in-review-ddos-attack-trends-and-insights/ (accessed on 8 September 2023).
59. Cloudflare DDoS Reports. Available online: https://radar.cloudflare.com/reports?q=DDoS (accessed on 8 September 2023).