Dataset: Large-scale Urban IoT Activity Data for DDoS Attack Emulation

Arvin Hekmati
hekmati@usc.edu
University of Southern California
Los Angeles, California, USA

Eugenio Grippo
egrippo@usc.edu
University of Southern California
Los Angeles, California, USA

Bhaskar Krishnamachari
bkrishna@usc.edu
University of Southern California
Los Angeles, California, USA

ABSTRACT

As IoT deployments grow in scale for applications such as smart cities, they face increasing cyber-security threats. In particular, as evidenced by the famous Mirai incident and other ongoing threats, large-scale IoT device networks are particularly susceptible to being hijacked and used as botnets to launch distributed denial of service (DDoS) attacks. Real large-scale datasets are needed to train and evaluate the use of machine learning algorithms such as deep neural networks to detect and defend against such DDoS attacks. We present a dataset from an urban IoT deployment of 4060 nodes describing their spatio-temporal activity under benign conditions. We also provide a synthetic DDoS attack generator that injects attack activity into the dataset based on tunable parameters such as number of nodes attacked and duration of attack. We discuss some of the features of the dataset. We also demonstrate the utility of the dataset as well as our synthetic DDoS attack generator by using them for the training and evaluation of a simple multi-label feed-forward neural network that aims to identify which nodes are under attack and when.

CCS CONCEPTS

• **Security and privacy** → *Malware and its mitigation.*

KEYWORDS

DDoS attacks, datasets, neural networks, machine learning, botnet

1 INTRODUCTION

Technological evolution has made possible the deployment of large internet of thing (IoT) systems that are able to connect multiple different sensors and actuators, allowing them to communicate and exchange enormous amounts of data. Such large-scale IoT systems consisting of thousands of sensor nodes are being proposed, for example, in the context of smart-city applications [5, 20, 23]. As these IoT systems grow in size and complexity, they are increasingly vulnerable from a Cybersecurity perspective [1, 3, 17].

Most significantly, they are liable to be hacked into and hijacked by malicious entities and then used as part of massive botnets as a launching ground for distributed denial of service (DDoS) attacks, potentially affecting millions of end-users [7, 11, 13, 14, 37, 38]. A famous example of an IoT-based DDoS attack was the Mirai botnet, first identified in August 2016 by MalwareMustDie, a whitehat security research group. Afterward, some of the biggest DDOS attacks in history were performed by Mirai botnet and its mutated variants. 400,000 nodes infected by this malware executed DDoS attacks on websites with a massive peak of 1.1 Tbps data transfer [4, 15, 18, 19, 28, 36].

One of the major steps to defend against such DDoS attacks is to detect them successfully, as close to real-time as possible. To this end, security researchers have turned to machine learning tools such as deep learning networks for DDoS attack detection [27]. The latest development in technology (parallel computing, high computational processing speeds, GPU's, etc.) have made possible the vertiginous development of deep neural networks (NN's) [2, 39], however their performance is very much dependent on the availability of rich datasets to train them. The training of NN models will be challenged in the near future as botnet attacks become increasingly complex and grow in volume infecting massive portions of IoT networks. Therefore, it is important to build and maintain suitable large-scale IoT dataset repositories, useful for training such models, to allow the research community to stay ahead of malicious attack developments.

There have been some prior efforts on collecting or creating synthetic DDoS attack datasets that could be used for training data-driven deep learning models [12, 24, 25, 29, 34, 35]. However many of these studies either are not specifically focused on IoT devices [6, 8–10, 26, 31–33], while others present data from a limited number of nodes (e.g., [16, 21, 30]). Further, while today's DDoS attacks such as Mirai are often relatively easy to detect because their traffic volumes dramatically exceed normal traffic (e.g., [22] reports nearly 100% true positives for an autoencoder-based detection scheme on the Mirai botnet), we believe that future attacks will be more cleverly camouflaged so that the traffic generated during the attack from a given node matches the traffic volume from a benign node. Thus, beyond datasets that show the behavior of today's attacks, there is still a need for large-scale datasets that characterize the benign activity of real IoT devices, which could be used as a basis to emulate more challenging future attacks.

We present in this work first a data set obtained from a real urban IoT system in a large city consisting of more than 4000 spatially distributed sensors. The data consists of the binary activity status of each node at a granularity of 30 seconds over a period of one month under a benign (non-attacked) setting.

To make the dataset useful for training machine learning tools for DDoS detection, we need to augment it through synthetic attack emulation. Therefore, in addition to the raw (benign) activity data, we also provide a dedicated script that generates attacks in the proposed dataset synthetically. Our script allows the setting of multiple parameters: number of nodes to be processed, total attack duration, attack ratio, starting time of the attack; and each particular node is equipped with a time-stamp and an output that varies binary between 0 (no attack) and 1 (attacked node). To illustrate the utility of our dataset and attack emulator, we design, train, and implement a simple supervised feed-forward NN model to detect malicious attacks utilizing the provided dataset. We make the dataset and the attack emulation script along with our illustrative NN model available as an open-source repository online at https://github.com/ANRGUSC/Urban_IoT_DDoS_Data.

The paper is structured as follows: section 2 presents the original dataset and some statistics about it. The attack and defense mechanism are presented in section 3. Finally, we conclude the paper in section 4.

2 ORIGINAL AND BENIGN ACTIVITY DATASETS

The original data has been collected from the activity status of real event-driven IoT nodes deployed in an urban area[1]. The original dataset contains three main features, the node ID, the location of the node in Latitude and Longitude, and a timestamp of the activity status of the IoT node. A record has been added to the original dataset whenever the activity status of a node changes. The raw dataset has 4060 nodes with one month worth of data. Table 1 presents two sample data points in the original dataset.

Table 1: Sample Data Points in Original Dataset

NODE	LAT	LNG	TIME	ACTIVE
33989	26.078	19.235	2021-01-01 23:00:00	1
5276	33.962	40.087	2021-01-01 23:00:01	0

Having a record of each node whenever the status of that node changes provides a bias towards the information of nodes that have more activity changes during the day. In order to overcome this issue, we also provide a script that takes the original dataset and generates a new benign activity dataset showing the activity status for each node every t_s seconds. In this way, all nodes in the benign activity dataset will have the same number of records. The script can generate a customized benign dataset by providing the beginning and ending date, the number of IoT nodes, and the time step, i.e., t_s.

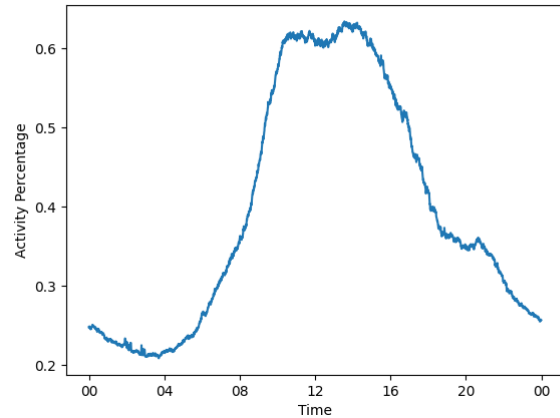[1]The source of this data has been anonymized for privacy and security reasons.

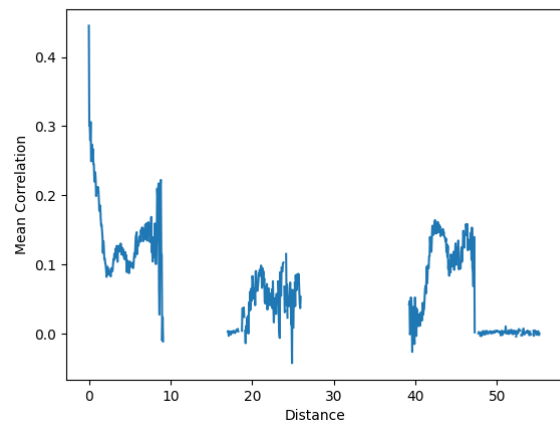**Figure 1: Active Nodes Percentage vs Time**

**Figure 2: Nodes Activity Mean Correlation vs Distances**

2.1 Dataset Statistics

In this subsection, some statistics of the datset are presented to illustrate the dataset's properties.

Figure 1 presents the mean number of active nodes versus the time of the day on one particular day of the dataset. As we can see, up to 65% of the nodes get activated around the middle of the day, but by midnight only about 20% of the nodes are active.

Figure 2 shows the mean correlation between pairs of 500 randomly selected nodes in the dataset, in terms of their activity versus their distance. For this purpose, the pair distances between the nodes are calculated using Euclidean distance. The Pearson correlation has been used to calculate the correlation between nodes' activity in the one month worth of data. Then, the distances between nodes have been split into 1000 bins, and the average correlation for the nodes whose pair distances fall in the bin boundaries has been calculated for each bin. As we can see, generally, the farther the nodes are, their correlation will be lower (dropping from a mean
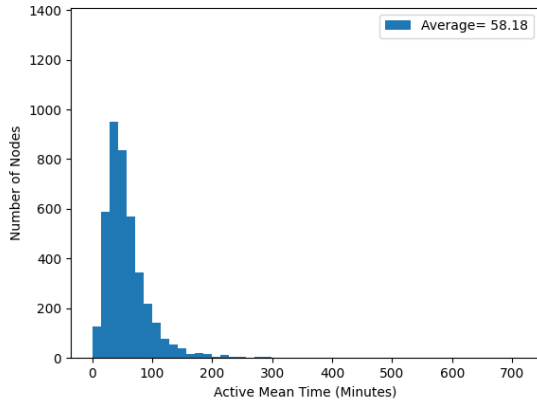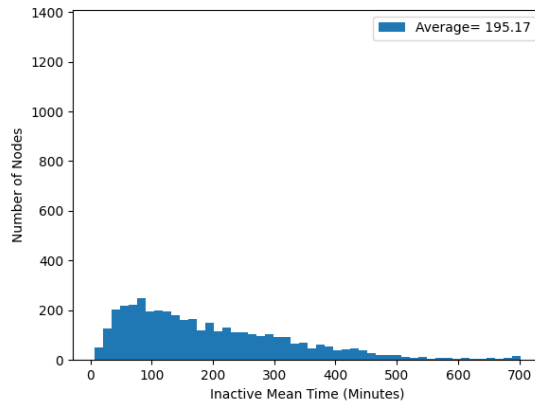
(a) Histogram of mean activity time per node - From 8 AM to 8 PM
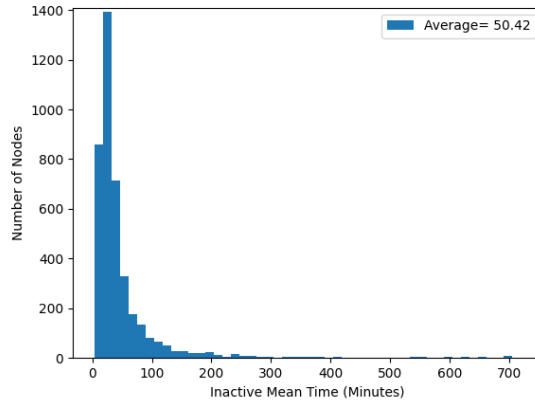
(b) Histogram of mean inactivity time per node - From 8 AM to 8 PM
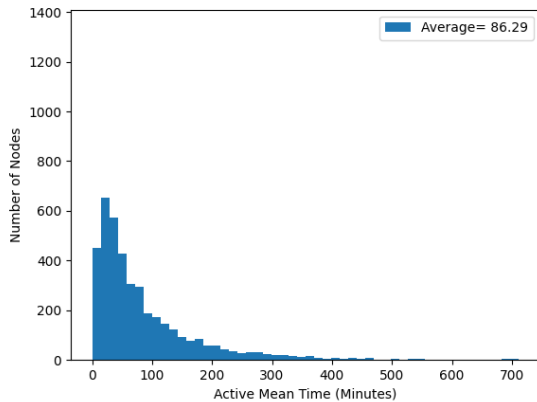
(c) Histogram of mean activity time per node - From 8 PM to 8 AM

(d) Histogram of mean inactivity time per node - From 8 PM to 8 AM

Figure 3: Histograms of mean activity and inactivity time per node

correlation of about 0.4 for nodes that are very close to each other to values around 0.1 or below for distances greater than 10 units).

Next, we present some statistics that show how long nodes tend to remain active or inactive. Figure 3 shows a histogram of the mean active time for all nodes as well as the mean inactive time, for both the day time (8 AM to 8 PM) and nighttime (8 PM to 8 AM). Figures 3a and 3c show the active mean time for the day and night, respectively. Figures 3b and 3d show the inactive mean time for the day and night, respectively. As we can see in these figures, both mean active times and mean inactive times tend to be higher in the nights compared to the days.

3 ATTACK AND DEFENSE MECHANISM

This section presents how synthetic DDoS attacks are generated on the IoT nodes. Furthermore, here we define the training dataset features and also the detection mechanism.

3.1 Generating Attack Dataset

In this paper, we synthetically generate a DDoS attack on the IoT nodes by setting all attacked nodes to an active status for the duration of the attack. Note that this approach to attack generation is more coarse-grained than providing the volume of packets or packets generated with specific content and destinations during an attack. We plan to incorporate the generation of such additional fine-grained information during the attacks in the near-future. But even by focusing on activity status only, we are effectively emulating more challenging futuristic DDoS attacks that may be hard to detect from a single node's traffic.

A script is provided that can be used for generating attacks over the dataset. Three parameters can be set in generating the attacks: the start time of the attack, the duration of the attack, and the percentage of the nodes that go under attack. In our experiments, we used one week's worth of dataset for generating attacks. The

Table 2: Mean accuracy and recall for the NN detection model

	Mean Accuracy	Mean Recall
Training Dataset	0.94	0.93
Testing Dataset	0.88	0.84

attacks are started at 2 AM on each day of the week over all of the nodes with durations of 1, 2, 4, 8, 16 hours.

3.2 Generating Training Dataset

Given the attacked dataset, a labeled training dataset will be generated by calculating the mean activity time of the IoT nodes in the specified time windows. Note that, in the script provided for this paper, one could set the desired time windows for generating the training features. In this paper, we considered a list of 12 different time windows, namely as 1, 10, 30, minutes and 1, 2, 4, 8, 16, 24, 30, 36, 42, 48 hours, to calculate the mean activity time of the nodes.

3.3 Defense Mechanism

We train a feed-forward neural network to detect the DDoS attack on the IoT nodes based on their collective activity status over time. As noted before, detecting attacks based on activity status alone is more challenging than approaches based on measuring fine-grained traffic volumes or flow-level information.

The model we have trained is a simple binary classification to show a sample usage of the presented dataset. In this neural network, we have an input layer with 12 neurons. The input layer is followed by one hidden layer with 8 neurons and ReLU activation. A dropout of 20% and batch normalization is also used at the end of the hidden layer. The output is a single neuron with the Sigmoid activation function. In this experiment, we randomly selected 20 IoT devices nodes to generate attacks. We are training 20 different models for each IoT node using its data alone, each with different weights but all having the same architecture. The neural network model is trained for 500 epochs for each node to detect the attacked time slots in the dataset. We used one week's worth of data as the training dataset and another week as the testing dataset. Note that this is a simple approach that will not take into account any correlations in the data across different nodes, so there is scope for further improvement by developing more complex models that integrate the inputs from multiple IoT devices.

Table 2 present the mean recall and accuracy of the 20 models trained for detecting DDoS attacks. Figures 4 and 5 show the true attack attack (T), attack predictions true positive (TP) and false positives (FP) mean over all nodes vs time, for both training and testing dataset. In these figures we used the attack duration of 16 hours. As we can see, the attacked nodes are being detected very well in the training dataset with a few FP. On the other hand, we are getting around 84% recall on the testing dataset with a little bit more FP.

4 CONCLUSION

We have presented a new spatio-temporal dataset describing the activity of a 4060-node event-based urban IoT deployment. We have
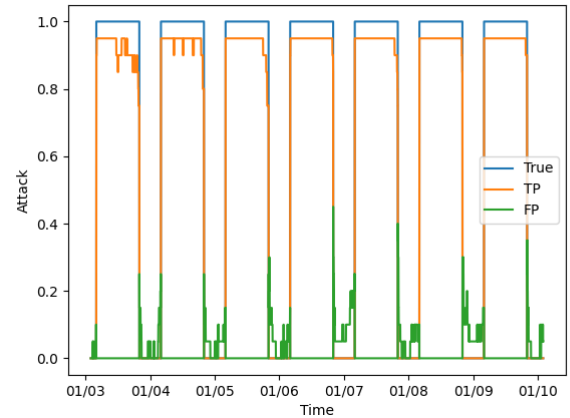
Figure 4: Training Dataset Attack Prediction vs Time
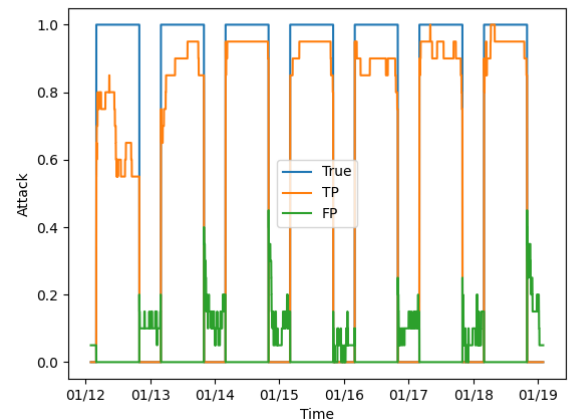
Figure 5: Testing Dataset Attack Prediction vs Time

also provided a script to create a benign dataset out of the original dataset to reduce bias toward nodes with more activity. We have shown some statistical analyses on the dataset to illustrate some its key properties. We have also presented a synthetic DDoS attack generator to generate attacks using the dataset, and illustrate the training and evaluation of a feed-forward neural network using the dataset as a way to detect such attacks. We hope that the dataset and tools we have provided are helpful to the community to undertake various types of research related to large-scale event-driven IoTs, including DDoS attack detection.

5 ACKNOWLEDGMENTS

and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

REFERENCES

[1] Dhuha Khalid Alferidah and NZ Jhanjhi. 2020. Cybersecurity Impact over Bigdata and IoT Growth. In *2020 International Conference on Computational Intelligence (ICCI)*. 103–108. https://doi.org/10.1109/ICCI51257.2020.9247722

[2] Ethem Alpaydin. 2020. *Introduction to machine learning*. MIT press.

[3] Ibrahim Alrashdi, Ali Alqazzaz, Esam Aloufi, Raed Alharthi, Mohamed Zohdy, and Hua Ming. 2019. AD-IoT: Anomaly Detection of IoT Cyberattacks in Smart City Using Machine Learning. In *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*. 0305–0310. https://doi.org/10.1109/CCWC.2019.8666450

[4] Manos Antonakakis, Tim April, Michael Bailey, Matthew Bernhard, Elie Bursztein, Jaime Cochran, Zakir Durumeric, J. Alex Halderman, Luca Invernizzi, Michalis Kallitsis, Deepak Kumar, Chaz Lever, Zane Ma, Joshua Mason, Damian Menscher, Chad Seaman, Nick Sullivan, Kurt Thomas, and Yi Zhou. 2017. Understanding the Mirai Botnet. In *Proceedings of the 26th USENIX Conference on Security Symposium* (Vancouver, BC, Canada) *(SEC'17)*. USENIX Association, USA, 1093–1110. https://doi.org/10.5555/3241189.3241275

[5] H. Arasteh, V. Hosseinnezhad, V. Loia, A. Tommasetti, O. Troisi, M. Shafie-khah, and P. Siano. 2016. Iot-based smart cities: A survey. In *2016 IEEE 16th International Conference on Environment and Electrical Engineering (EEEIC)*. 1–6. https://doi.org/10.1109/EEEIC.2016.7555867

[6] Frank Beer, Tim Hofer, David Karimi, and Ulrich Bühler. 2017. A new Attack Composition for Network Security. In *10. DFN-Forum Kommunikationstechnologien*, Paul Müller, Bernhard Neumair, Helmut Raiser, and Gabi Dreo Rodosek (Eds.). Gesellschaft für Informatik e.V., Bonn, 11–20.

[7] Elisa Bertino and Nayeem Islam. 2017. Botnets and internet of things security. *Computer* 50, 2 (2017), 76–79. https://doi.org/10.1109/MC.2017.62

[8] CAIDA. 2007. The CAIDA UCSD "DDoS Attack 2007" Datasett. https://www.caida.org/catalog/datasets/ddos-20070804_dataset. Accessed: 09-12-2021.

[9] DARPA. 2000. DARPA intrusion detection evaluation dataset, Lincoln Laboratory, Massachusetts Institute of Technology. https://www.ll.mit.edu/r-d/datasets/2000-darpa-intrusion-detection-scenario-specific-datasets. Accessed: 09-12-2021.

[10] Derya Erhan and Emin Anarım. 2020. Boğaziçi University distributed denial of service dataset. *Data in brief* 32 (2020), 106187. https://doi.org/10.1016/j.dib.2020.106187

[11] Sebastián García, Alejandro Zunino, and Marcelo Campo. 2014. Survey on network-based botnet detection methods. *Security and Communication Networks* 7, 5 (2014), 878–903. https://doi.org/10.1002/sec.800

[12] Dilara Gümüşbaş, Tulay Yıldırım, Angelo Genovese, and Fabio Scotti. 2021. A Comprehensive Survey of Databases and Deep Learning Methods for Cybersecurity and Intrusion Detection Systems. *IEEE Systems Journal* 15, 2 (2021), 1717–1731. https://doi.org/10.1109/JSYST.2020.2992966

[13] Roger Hallman, Josiah Bryan, Geancarlo Palavicini, Joseph Divita, and Jose Romero-Mariona. 2017. IoDDoS-the internet of distributed denial of sevice attacks. In *2nd international conference on internet of things, big data and security. SCITEPRESS*. 47–58. https://doi.org/10.5220/0006246600470058

[14] Constantinos Kolias, Georgios Kambourakis, Angelos Stavrou, and Jeffrey Voas. 2017. DDoS in the IoT: Mirai and other botnets. *Computer* 50, 7 (2017), 80–84. https://doi.org/10.1109/MC.2017.201

[15] Constantinos Kolias, Georgios Kambourakis, Angelos Stavrou, and Jeffrey Voas. 2017. DDoS in the IoT: Mirai and Other Botnets. *Computer* 50, 7 (2017), 80–84. https://doi.org/10.1109/MC.2017.201

[16] Nickolaos Koroniotis and Nour Moustafa. 2019. University of New South Wales, The Bot-IoT Dataset. https://research.unsw.edu.au/projects/bot-iot-dataset. Accessed: 09-12-2021.

[17] Yang Lu and Li Da Xu. 2018. Internet of Things (IoT) cybersecurity research: A review of current research topics. *IEEE Internet of Things Journal* 6, 2 (2018), 2103–2115. https://doi.org/10.1109/JIOT.2018.2869847

[18] Artur Marzano, David Alexander, Osvaldo Fonseca, Elverton Fazzion, Cristine Hoepers, Klaus Steding-Jessen, Marcelo H. P. C. Chaves, Ítalo Cunha, Dorgival Guedes, and Wagner Meira. 2018. The Evolution of Bashlite and Mirai IoT Botnets. In *2018 IEEE Symposium on Computers and Communications (ISCC)*. 00813–00818. https://doi.org/10.1109/ISCC.2018.8538636

[19] Christopher D. McDermott, Farzan Majdani, and Andrei V. Petrovski. 2018. Botnet Detection in the Internet of Things using Deep Learning Approaches. In *2018 International Joint Conference on Neural Networks (IJCNN)*. 1–8. https://doi.org/10.1109/IJCNN.2018.8489489

[20] Camilo Alejandro Medina, Manuel Ricardo Pérez, and Luis Carlos Trujillo. 2017. IoT Paradigm into the Smart City Vision: A Survey. In *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*. 695–704. https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData.2017.109

[21] Yair Meidan, Michael Bohadana, Yael Mathov, Yisroel Mirsky, Asaf Shabtai, Dominik Breitenbacher, and Yuval Elovici. [n. d.]. University of California Irvine, Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/detection_of_IoT_botnet_attacks_N_BaIoT. Accessed: 09-12-2021.

[22] Yair Meidan, Michael Bohadana, Yael Mathov, Yisroel Mirsky, Asaf Shabtai, Dominik Breitenbacher, and Yuval Elovici. 2018. N-BaIoT—Network-Based Detection of IoT Botnet Attacks Using Deep Autoencoders. *IEEE Pervasive Computing* 17, 3 (2018), 12–22. https://doi.org/10.1109/MPRV.2018.03367731

[23] Ahmed Samy Nassar, Ahmed Hossam Montasser, and Nashwa Abdelbaki. 2017. A survey on smart cities' IoT. In *International Conference on Advanced Intelligent Systems and Informatics*. Springer, 855–864. https://doi.org/10.1007/978-3-319-64861-3_80

[24] Markus Ring, Sarah Wunderlich, Dominik Grüdl, Dieter Landes, and Andreas Hotho. 2017. Flow-based benchmark data sets for intrusion detection. In *Proceedings of the 16th European Conference on Cyber Warfare and Security. ACPI*. 361–369.

[25] Markus Ring, Sarah Wunderlich, Dominik Grüdl, Dieter Landes, and Andreas Hotho. 2017. Creation of Flow-Based Data Sets for Intrusion Detection. *Journal of Information Warfare* 16 (2017), 40–53. Issue 4. https://www.jstor.org/stable/26504117

[26] Iman Sharafaldin, Arash Habibi Lashkari, Saqib Hakak, and Ali A Ghorbani. 2019. Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy. In *2019 International Carnahan Conference on Security Technology (ICCST)*. IEEE, 1–8. https://doi.org/10.1109/CCST.2019.8888419

[27] Niharika Sharma, Amit Mahajan, and V Malhotra. 2016. Machine learning techniques used in detection of DoS attacks: a literature review. *International Journal of Advance Research in Computer Science and Software Engineering* 6, 3 (2016), 100–105.

[28] Hamdija Sinanović and Sasa Mrdovic. 2017. Analysis of Mirai malicious software. In *2017 25th International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*. IEEE, 1–5. https://doi.org/10.23919/SOFTCOM.2017.8115504

[29] UCI. 2021. University of California Irvine, KDD Archive. https://www.kdd.org/kdd-cup/view/kdd-cup-1999/Data. Accessed: 09-11-2021.

[30] Imtiaz Ullah and Qusay H. Mahmoud. 2020. A Scheme for Generating a Dataset for Anomalous Activity Detection in IoT Networks. In *Advances in Artificial Intelligence*, Cyril Goutte and Xiaodan Zhu (Eds.). Springer International Publishing, Cham, 508–520. https://doi.org/10.1007/978-3-030-47358-7_52

[31] UNB. 2017. Canadian Institute for Cybersecurity, "CICIDS2017," unb.ca, 2017. https://www.unb.ca/cic/datasets/ids-2017.html. Accessed: 09-13-2021.

[32] UNB. 2018. University of New Brunswick, "CSE-CIC-IDS2018 on AWS," 2018. https://www.unb.ca/cic/datasets/ids-2018.html. Accessed: 09-13-2021.

[33] UNB. 2019. University of New Brunswick, "DDoS Evaluation Dataset (CICD-DoS2019)",unb.ca, 2019. https://www.unb.ca/cic/datasets/ddos-2019.html. Accessed: 09-13-2021.

[34] UNB. 2021. University of New Brunswick, Canadian Institute for Cybersecurity. https://www.unb.ca/cic/datasets/. Accessed: 09-11-2021.

[35] UNSW. 2015. University of New South Wales, The UNSW-NB15 Dataset. https://research.unsw.edu.au/projects/unsw-nb15-dataset. Accessed: 09-11-2021.

[36] K Vengatesan, Abhishek Kumar, M Parthiban, Achintya Singhal, and R Rajesh. 2018. Analysis of Mirai Botnet Malware Issues and Its Prediction Methods in Internet of Things. In *International conference on Computer Networks, Big data and IoT*. Springer, 120–126. https://doi.org/10.1007/978-3-030-24643-3_13

[37] Saman Taghavi Zargar, James Joshi, and David Tipper. 2013. A survey of defense mechanisms against distributed denial of service (DDoS) flooding attacks. *IEEE communications surveys & tutorials* 15, 4 (2013), 2046–2069. https://doi.org/10.1109/SURV.2013.031413.00127

[38] Bruno Bogaz Zarpelo, Rodrigo Sanches Miani, Cludio Toshio Kawakani, and Sean Carlisto de Alvarenga. 2017. A Survey of Intrusion Detection in Internet of Things. *J. Netw. Comput. Appl.* 84, C (April 2017), 25–37. https://doi.org/10.1016/j.jnca.2017.02.009

[39] Amer Zayegh and N Bassam. 2018. *Neural network principles and applications*. Vol. 10. London: Pearson.