Information Diversity based Detection for ON-OFF Low Strength DDoS Attacks in Smart Home IoT

Joseph Okonofua
Western Michigan University
Kalamazoo, MI, USA
joseph.okonofua@wmich.edu

Richard T. Meyer
Western Michigan University
Kalamazoo, MI, USA
richard.meyer@wmich.edu

Shameek Bhattacharjee
Western Michigan University
Kalamazoo, MI, USA
shameek.bhattacharjee@wmich.edu

ABSTRACT

In this paper, we propose a lightweight explainable machine learning approach that is device and attack-type agnostic and can detect IoT devices that are victims of low-intensity direct and reflective volumetric DDoS attacks launched in an ON-OFF manner. Specifically, our approach is based on a parameterized bio-inspired information-theoretic model that can capture small and subtle volumetric differences between attack versus benign byte volumes exchanged between IoT devices and the rest of the internet. Our approach has four main phases: (1) Feature Engineering involving a simple compression to achieve a universally reduced feature space for volumetric attacks; (2) Model Parameterization: identify appropriate parameters of a bio-inspired information-theoretic model and their appropriate pruned search spaces. (3) Parameter Learning: take a supervised approach for learning the optimal parameters of the explainable model using a local search. (4) Testing: We apply the learned parameters in the test set. To validate our approach, we use real datasets from 4 different types of IoT devices containing seven different kinds of attacks and varying DDoS attack volumes. Furthermore, we employ strategies to counter the inherent biases in attacked datasets to ensure unbiased evaluation.

CCS CONCEPTS

• **Security and privacy → Intrusion detection systems**.

ACM Reference Format:
Joseph Okonofua, Richard T. Meyer, and Shameek Bhattacharjee. 2024. Information Diversity based Detection for ON-OFF Low Strength DDoS Attacks in Smart Home IoT. In *25th International Conference on Distributed Computing and Networking (ICDCN '24), January 4–7, 2024, Chennai, India.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3631461.3631954

1 INTRODUCTION

Smart home IoT devices are vulnerable to direct and reflective DDoS attacks [4]. In direct attacks, the IoT device itself is the target, while in reflective DDoS, the IoT device acts as an intermediate (e.g., a Botnet Member) used by an attacker to reflect DDoS traffic onto target servers on the internet. In [1], malware reportedly found vulnerabilities in millions of IoT devices that were used to source DDOS traffic to bring down servers of various popular web servers.

ICDCN '24, January 4–7, 2024, Chennai, India

ACM ISBN 979-8-4007-1673-7/24/01…$15.00
https://doi.org/10.1145/3631461.3631954

1.1 Related Work

Prior approaches to IoT device attack detection are broadly classified into network-level data flow [10], attack-specific features [2], network segmentation [3], IoT network or device components classification [5], or adopt known machine learning techniques for anomaly detection [4], rate limiting at the victim side networks. Note that IoT devices for reflective DDoS can be used as a moving target to take down any server. Hence, rate limiting does not address the root cause; i.e., detecting the SH IoT devices reflecting DDoS traffic. However, IoT device-level detectors try to detect attacks at the source SH IoT network. If successful, it can prevent the spread of the attacks at the source IoT network's edge, like taking the IoT devices offline.

1.2 Key Challenges and Motivation

At the IoT device level, different variations in the attack strengths are possible. Some DDoS attack volumes could be very low. For example, [4] observes that low volumes of DDoS attacks from 10 pps to 30 pps are sufficient to make many (if not all) IoT devices unresponsive. Thus, reflective DDOS attacks using compromised IoT devices to reflect traffic will not be able to reflect traffic to the intended target if attack volumes are higher than 30pps. Hence, even though the amount of traffic per compromised IoT device is small, the sheer scale of the compromised IoT devices is enough to cripple target servers on the internet [1]. Therefore, we need to focus on low-strength DDoS attacks.

Second, every IoT device has widely different legitimate traffic patterns and drastically different ranges of traffic volume even under benign conditions. Since the amount of increase in the traffic per IoT device can be small, subtle variations between the legitimate and attack traffic make it difficult for detection systems to detect and isolate DDoS attacks.

Third, DDoS attack is an umbrella term that exploits vulnerabilities in several protocols. Current approaches look into the protocol-specific flows on each port per attack type, which creates the problem of too many features (e.g., 115 features in [11]) and increases the anomaly detection model's complexity. This in turn requires bulkier detection models such as autoencoders [11], and deep neural networks [4], which are also not explainable.

Fourth, every attack dataset comes with a specific implementation strategy. In reality, the attack strategy may be different from the one implemented in a dataset. The implications of such variations on the robustness of learning classifiers are untested.

1.3 Contributions

In this paper, we propose an explainable machine learning-based approach based on bio-inspired information theory for IOT device-level detection of direct and reflective DDoS attacks of various

attack strengths that generalize well across IoT device types. Specifically, we first convert IoT device-generated .pcap files to CSV that summarize the per-second traffic conversation of IoT devices with the rest of the internet (both local and remote). Due to the high dimensionality of the conversations, we calculate a reduced feature set that only takes the sum of the total number of bytes exchanged between the IoT device and the rest of the internet, over a strategic time window length.

Then, we construct a detection model architecture based on bio-inspired ecological information theory, which takes as input our reduced feature set and is parameterized by three parameters. We identify the importance of the time window parameter and the binning width parameter choices which affect the shape of the probability distributions of the feature set under attack and benign labels. Additionally, we identify a diversity order parameter that affects the accuracy of our explainable ML model. We discuss why the above can detect low-strength DDoS attacks injected in an ON-OFF manner, by being able to quantify subtle changes in the shape of the distributions of the feature set with and without attacks.

Finally, we take a supervised approach to learning the optimal parameters of our explainable model as a mixed integer problem which has a brute force exact solution. Then, we also discuss various pruning strategies to correctly reduce the search space of model parameters. To further enable an efficient solution to parameter learning, we implement a heuristic local neighborhood search technique to find a quick approximation of the optimal parameters.

For validation, we apply the learned parameter values of our model in a test set containing benign and attack datasets for four different IoT device types. We also ensure the test set attack datasets are not biased by randomly sampling and mixing different attack strengths and durations. We find that our model fitted with the learned parameters can readily distinguish benign versus attack. Specifically, our model produces a diversity index score in the test set, where the diversity score is significantly high under attacks compared to under benign , even when the attack dataset contains a mix of both benign and attack traffic.

2 SYSTEM AND THREAT MODELS

Here we describe smart home IoT, followed by the dataset/testbed details, reasons why we chose this dataset, steps we took to ensure unbiased evaluation, and discuss different attacks in the dataset.

2.1 Architecture and Operations

A typical smart home IoT network contains the following types of components: (1) IoT devices (e.g., Smart Light Bulb, Smart Thermostat) and IoT Hub devices (Samsung Smart Things, Hub, Amazon Echo) ; (2) non-IoT devices (laptops, tablets, smart phones); (3) smart home edge gateway router.

IoT Devices: An IoT device (e.g. Smart Camera) is one that is connected to a internet and runs an IP-protocol. Each IoT device's goal is to offer some services to customers (e.g., smart surveillance).

Smart Home Gateway (SHG): Each smart home has a gateway router (SHG), that connects the IoT devices to the rest of the internet. SHG can also host security middleware services by mirroring IoT traffic to a Raspberry Pi. Thus, SHG will be able to monitor SH network activity originating from a given smart home, where our proposed framework can be deployed for IoT device-level detection at the source IoT network.

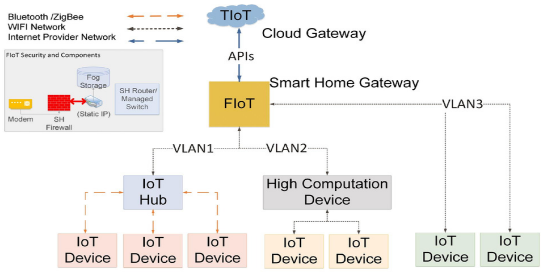

Figure 1: Smart Home Network Architecture

2.2 IoT Device and Dataset Details

We use the UNSW dataset [4] that contains labeled benign and attacked traffic collected from a smart home IoT testbed containing various IoT devices and some non-IoT devices. We chose this dataset because (1) this is one of the only datasets which contain varying DDoS attack volumes (ranging from as low as 1 pps, 10 pps, to some higher volumes (e.g. 100 pps). This allows us to test the sensitivity and specificity of detection to very low volume DDoS; (2) the attacks are implemented in an ON-OFF manner; (3) timestamps that correspond to an attack are also given, apart from just the attack and benign dates helping us to create and test different combinations of attack possibilities. In this way, we avoid sampling bias in the original experiments used to create the attacks.

The dataset is available as daily .pcap files captured by mirroring the traffic at smart home gateway. The traffic contains the flow-specific source and destination IP, port nos, MAC addresses for local sources and destinations, and the number of packets and bytes per second per flow. We filtered the traffic using wireshark to study four different IoT devices; (i) Netatmo Welcome camera, (ii) Belkin Wemo Powerswitch, (iii) Dropcam, and (iv) Samsung Smart camera. We chose these devices because the previous works [4] reported that IoT devices with video/audio streaming applications have highly unpredictable traffic patterns. Therefore, device-level anomaly-based attack detection is more challenging in these devices.

2.3 Attack Types, Strengths, Durations

In this section, we briefly enumerate each device's attack types and briefly describe the attack types, attack strengths, attack strategies.

We collect the benign and attack datasets from [4]. Various specification compliant DDoS attack types on smart home IoT testbed is reported in [4].

Direct attacks: The direct attack types included in our labeled attack dataset for the devices include well-known attacks such as (1) Fraggle, (2) Ping-of-Death, and (3) TCP-SYN Flood, (4) UDP-Flood [1], (5) NTP amplification [4]. The direct attacks incapacitate the IoT device itself.

Reflective DDoS attacks: In reflective DDoS attacks, the attacker uses the compromised SH IoT device as an intermediate to source traffic attacking victims. The attack path is indicated with origin-intermediate IoT device-victim. In such a case, the attacker uses the IoT device as a weapon intermediate for targeting a victim - which could be another device/server in the local network or outside on the web. Such attacks have been on the rise due to the growing proliferation of millions of vulnerable IoT devices

which can generate the required amplification volumes necessary to overwhelm target victims. The devices we picked contained TCP SYN Reflection and SNMP Reflection attacks.

Three variations in DDoS attack strengths: (i) 1 packet per second (pps) (2) 10 pps (3) 100 pps (for direct attacks), which capture various low volume DDoS attack strengths. The attacks are implemented in an episodic manner (ON-OFF) where a a certain attack type and strength pair lasts for 10 minutes.

2.4 Attack Type to Device Mapping

Below we enumerate which of our chosen devices contain what kind of attacks:

Samsung Smart Camera: The attacks launched included TCP-SYN flood, UDP Flood, SNMP reflection, TCP-SYN-reflection, PING-of-Death, and SMURF.

NetAMO Welcome Camera: The attacks launched included TCP-SYN-reflection, TCP-SYN Flood.

Belkin Wemo Power Switch: The attack launched included TCP-SYN-reflection, TCP-SYN Flood, Ping-of-Death

Dropcam: The attacks launched was NTP amplication.

2.5 Removing Bias from Attack Dataset

Note that it is not necessary that all attackers will create attack episodes that last 10 minutes. Additionally, the attack volume per episode need not be constant. Hence, security evaluation will be biased because just using this attack dataset is not representative of various possible attack behaviors. *This is an often overlooked but important aspect of security research*; i.e., since the actual attacker's strategy may be different from the strategy intent manifested in the authors of a specific attack dataset and testbed. To avoid this problem, we created different "sampled" (removing/replacing parts) versions of the original attacked dataset, in the following manner:

Given that attack and benign samples are labeled at the timestamp level in our chosen dataset (with full details on attack type, strength, start, and stop times), we could reliably remove and replace parts of the attack samples and benign samples randomly such that there is no fixed pattern to the ON-OFF periods of each attack episode, and also no fixed pattern in the attack volume given one attack episode. Through the above, we created a more representative malicious dataset that contains different possibilities of attack strengths and attack-durations. This ensures that our method's success (or failure) is not dependent on the specifics of the attack episode as implemented by the dataset's authors. While sampling, we ensure that within one attack episode, the attack type and strength do not change for a given IoT device to preserve the intended impact of each attack episode.

3 TECHNICAL FRAMEWORK

In this subsection, we provide a theoretical intuition of why our framework is required and at a high level why it works, followed by the model architecture, parameters involved, search space pruning, and finding optimal values of the parameters.

3.1 Reducing Feature Space and Pre-processing

In the actual dataset, for every IoT device, there is a multitude of features such as source IP, list of all destination IPs with which there was communication in the given duration of the dataset, the source port, destination port, source MAC addresses, and destination MACs in the local network. For each of these, the traffic dataset contains packet/byte numbers.

However, there is scope to be smart in understanding what is it that we want to learn. i.e. whether the IoT device is under the influence of a direct or reflective volumetric DDoS attack. In that regard, whether local or remote destinations or a wide area botnet is involved in orchestrating the attack, looking at the ports and flows separately is not necessary. What is relevant is that is there an anomalous change in the total traffic volume exchanged between the IoT device and the rest of the internet (including both local and remote web). Tracking the total traffic volume exchanged rather than tracking volumes per destination and per port is the model that is simplest and has the most reduced complexity. The above is aligned with *Occam's razor* and a very fundamental but often ignored aspect of machine learning. It says that amongst all the information available, picking the simplest or most reduced model is the best way to achieve high prediction accuracy.

Now, the term traffic volume in itself is vague and requires further elucidation. Using Wireshark software, we extracted the total byte volume per second exchanged between an IoT device (by filtering via its MAC address) and the rest of the internet. Formally, let us denote this by $b^{(j)}(t)$; which is the total number of bytes exchanged between the j-th device and the rest of the internet at the t-th time slot (slotted as per second).

Note that DDoS detection will be easy near the periphery of a victim network, due to a tangible increase in the DDoS traffic that is enough to take down the victim server. However, the same IoT reflectors can be used to target many other victim servers. Therefore, detecting the IoT devices used to reflect DDoS traffic is equally important at the source IoT network. At the IoT device level though, the differences in byte volume distributions are very subtle in low-intensity DDoS attacks (See Fig. 2 for a Dropcam device).

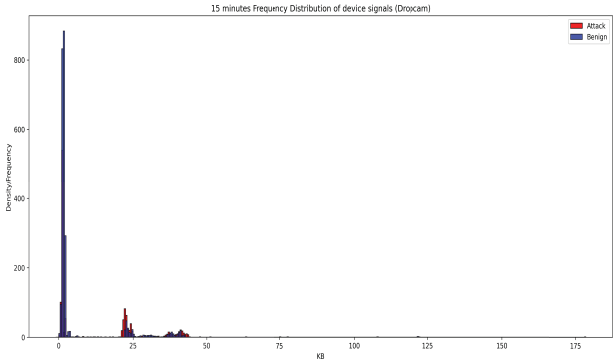

Figure 2: Subtle Differences in Byte Volume $b^{(j)}(t)$

3.2 Analogy with DDoS Attacks and Ecological Disruption

Now, we relate our problem of detecting small-volume attacks at the IoT device level with ecological information theory. Some information theoretic models in ecology study the relative level of abundance/rarity of a certain species over all the species (i.e., probability of a species given all numbers of organisms) across different time frames. It helps to detect subtle events that disrupt

ecological balance and uncover underlying causes. Importance is provided on rarer and endangered species, to quickly understand the true impact of an underlying cause on the species diversity.

Our intuition is that a change in the rarer bins will be seen when comparing the correctly constructed byte volume probability distributions. This is because DDoS attacks are volumetric attacks. Hence, asymmetric and exponentially more importance given to the rarer bins in the probability distributions will enable reliable detection of low-intensity attacks. This intuition can be verified from Fig. 2, where a minor change is seen in low probability bins, while high probability bins are identical across benign and attack, and there is no difference in the range of data ranges under benign and attack; proving the low-intensity nature of the DDoS attack.

Therefore, given the parallelism between these two different problems, we hypothesize that diversity index scoring on probability distributions of $b^{(j)}(t)$ can help identify low-volume attacks at the IoT device level. The diversity index score is inspired from Renyi entropy rather than the more popular Shannon's entropy. We discuss these preliminaries next.

3.3 Renyi Entropy and Diversity Index Model

The Renyi entropy provisions for an additional parameter 'diversity order of entropy' denoted by q. While quantifying uncertainty in information via information theory, the parameter 'q' in Renyi Entropy allows to adaptively and asymmetrically control the relative importance of each bin in a discretized probability distribution $\boldsymbol{p}$ in the entropy calculated.

When the $q \rightarrow 1$, Renyi Entropy is approximately equal to Shannon entropy. Mathematically, Renyi entropy is defined as:

$$H_q(\boldsymbol{p}) = \frac{1}{1-q} ln\Big(\sum_{s=1}^{R} p_{(s)}^{q}\Big) \tag{1}$$

In ecological studies, [13] shows that instead of Renyi Entropy, the Hill's Diversity Index derived from Renyi Entropy is more useful, since this gives control on the importance of each bin type. Hill's Diversity Index can quantify how diverse is the uncertainty in the information content. Mathematically, the Hill's Diversity index of the order $q \in \mathbb{R}$ is defined as:

$$e^{H_q} = D^{(q)} = \Big(\sum_{i=1}^{R} p_{(i)}^{q}\Big)^{\frac{1}{1-q}} \tag{2}$$

where q is a 'diversity order'. The higher the value of q, the more asymmetric importance is given to the higher probability bins (most common occurrences) compared to the lower probability bins while the diversity of the information is being quantified via e^{H_q}.

3.4 Parameterizing Model Architecture

Time Windowing Parameter: Observing smart home IoT datasets reveal that most of the time a device is not actively used. Therefore, the $b^j(t)$ contains a lot of intervals with little or no traffic exchanged. Therefore, we further reduce the feature space by the sum of the byte volume exchanged between the j-th IoT device and rest of the internet over a time window of length T, such that $B^{(j)} = \sum_{t=1}^{T} b^{(j)}(t)$ is the random variable for which we need to construct probability distributions. Please note that the random variable $B^{(j)}$ is a function of the *parameter* T i.e., the window length that will directly affect the parameters of a probability distribution of $B^{(j)}$. Hence, we need to find the optimal time windowing parameter T. Since our mechanism separately works in the same manner for each device j, we drop the suffix j from our notations, to keep notational simplicity.

Binning Width Parameter: The continuous range of data under each class label needs to be binned into several discrete partitions of bin width (denoted by s bytes). Given a bin width of s bytes, the number of bins will change and so will be the quantized probability distribution of attack and benign byte volume data.

Furthermore, the ranges of benign data distributions can greatly vary from IoT device to device due to its specific functionality, which dictate traffic volumes. This also indicates that the probability distribution of $B^{(j)}$ r.v. will depend on the *binning width parameter* s, which again changes the shape of the distribution. Hence, we need to find the optimal s.

Diversity Order Parameter Finally, it is evident from the Eqn. 2, that we need to learn the optimal value of the diversity order parameter q for modeling the diversity scores' and maximize its discriminative power for classifying attacks from benign. To conclude we have the following model architecture where the function we need to learn is specified by the diversity index model. The model is parameterized by three parameters: (i) q is the diversity order; (ii) s the bin width; (iii) T the time window length. Therefore, formally we can write the following equation that specifies are simple explainable ML architecture:

$$DI(q, s, T) = \Big(\sum_{i=1}^{R(s,T)} p_i^q(s, T)\Big)^{\frac{1}{1-q}} \tag{3}$$

The good thing is that our model just needs to learn three parameters. In contrast, popular ML detectors like an LSTM has $(m * n + n^2 + n)$ parameters, where m = the input dimension and n = Number of activations in each gate.

3.5 Pruning Parameter Search Space

We discuss some key design issues which we found result in *biased and unpredictable outcomes* that prevent unified treatment.

Search Space of S: First, the comparison between attack and benign data probability distribution cannot happen on even terms, since their benign data ranges greatly vary from IoT device to IoT device. Wherever binning is involved in information theory applied to computer security, this issue is not addressed in a scientifically generic manner. Furthermore, fixing the number of bins while constructing the probability distributions produces different bin widths, which in turn means a drastically different number of bins; producing incomparable diversity index model scores.

If we do not prune the search space of the bin width, the differences in the shape of the distributions between attack and benign will not be apparent for many parameter combinations. This is not smart since we know this explodes the search space unnecessarily. To solve this, we applied Scott's rule [9], as a scientifically sound way of assigning an upper bound on the bin width parameter, that prunes the search space of s. Next we discuss how we did this:

Scott's Normal Reference Rule: Scott's binning rule is a recommended method for determining the optimal number of bins in a histogram which is Gaussian distributed. Scott's rule takes into

account the sample size and the standard deviation of the data, to determine the appropriate number of bins. Mathematically, it is denoted by:

$$h = 3.49\frac{\sigma}{\sqrt[3]{n}} \tag{4}$$

where σ is the estimate of the standard deviation and n is the total number of observations and h is Scott's number.

However, we *cannot apply Scott's rule directly* in our problem to get the optimal number of bins because, our feature distributions are not Gaussian. Unpredictable human behavior causes even benign byte volumes to have high levels of variability. Moreover, there are variations in the strength of DDoS attacks across attack episodes. The above reasons result in a higher sample standard deviation violating the model assumptions of Scotts' Reference Rule. So with the high sample standard deviation, Scott's rule will give an inflated bin width rather than an optimal one. However, this inflated bin width can be utilized as *statistically sound upper bound over the unknown bin width parameter s*. Therefore, to include such potential scenarios we defined a range of values for s ranging from 0.1 to $s_{max} = h$ to ensure that we constrained the search space of the binning parameter.

Search Space of T: The time window search space is a discrete integer ranging from 1 to T_{max} minutes, with a step size of 1 minute. We kept the T_{max} as 30 minutes since too large T_{max} would count the bursty benign byte volume usage too, making discriminative classification difficult, especially under on-off attacks.

Search Space of q: The order of diversity parameters directly affects the diversity index scores. The more negative the q, the more importance the rarer bins get in the resulting $DI(q, s, T)$. The more positive the q, the more abundant or high probability bins get more weight in the final DI score. Since over a time window, greater than normal amounts of traffic are exchanged during DDoS attacks, we posit that the optimal value of q that should give the best results should be in the negative real domain (i.e. $-\infty < q < 1$), because with a more negative q, rarer bins are given more importance while calculating scores. Later, we verify that our hypothesis matches with the outcome of the optimal q in the optimization problem and this explains why the optimal parameter values fitted to the model produce linearly separable class-conditioned DI scores.

3.6 Parameter Optimization

In this section, we put forward the supervised learning approach to learn the parameters of the diversity index model by formulating the objective function that would enable us to learn optimal values of the parameters

$$C = \left(DI^{attack}(q, s, T) - DI^{benign}(q, s, T)\right)$$

$$\max_{s,q,T} \quad C \tag{5}$$

$$\text{s.t.} \qquad 0.1 < s \leq s_{max} \qquad s \in \mathbb{R}$$
$$1 < T < T_{max} \qquad T \in \mathbb{I}$$
$$-\infty < q < 1; \qquad \mathbb{R}$$

The optimal solution of the above can be represented as

$$s^*, q^*, T^* = \arg\max_{s,T,q}(C) \tag{6}$$

The above problem is a mixed integer problem, since the T is strictly integer while parameter constraints s and q are real valued. This partial integer nature of the constraints prevents us from applying gradient-based approaches popular in machine learning. Therefore, we need an approach to solve Eqn. 5 to an approximate solution for the Eqn. 6. We do a neighborhood search to find solution in the following way:

We visualize the parameters (s, T) on a 2D coordinate system. For each point, we find the optimal value of q that maximizes the objective function C. Given this, we can perform a local search with (s, T) and search the entire parameter space of q per point.

Given neighborhood search techniques are highly influenced by the choice of start point, we had random start points in the (s, T) space, such that r is the r-th random start point. For each r, we start an iterative process indexed by l, such that r_l, is the l-th iteration of the local search corresponding to the r-th random start point; the $Z_r(l)$ is the set of neighborhood points in (s, T) including the random start point. For each point in $Z_r(l)$, we find the value of the objective function. Then, we pick that combination of parameters from the set Z_r^l, which maximizes the local estimate of C after iterating over the search space of q.

Now let $Z_r^*(l)$, be the point maximizing C for a certain value of $q(l)$. Now this point forms the new pivot for the second iteration of the local search (i.e l=2) is initiated where the neighborhood set is selected such that $Z_r(l) \setminus Z_r(l-1)$. The above means that the new neighborhood around the pivot excludes overlap between this local neighborhood and the neighbor set of the previous iteration.

The process goes on until $C(l)$ stops changing, and then whichever point (s, T) the process converges to is the local optimal parameter choice $s(r), T(r), q(r)$, which is entirely dependent on the random start point r. We performed this process for 50 different random start points. At the end, across random start points the highest local maximum is treated as the approximate global maxima and the corresponding parameter values at that point is the optimal solution s^*, T^*, q^*.

We do the above for each IoT device and then do a model averaging of each of the parameters across devices, and across all sampled versions of the training set; and finally, we get the following average optimal parameter values $T^* = 17$ minutes, $q^* = -20$, and $s^* = 2.3$ of the universal model that should apply for any device for any realization of attack and benign data. Later we plug these optimal parameter values for diversity index calculation in the test set. The test set contains both benign and attack data, whose ground truths are known.

4 EXPERIMENTAL RESULTS

In all, we used .pcap files from specific days of May, June and October of 2018, with 26 days worth of benign data, and 16 days of data containing attack samples introduced in an ON-OFF manner. We created 30 versions of benign and attack dataset as mentioned in Sec. 2.5, for training and testing in an unbiased manner. The number of versions we tried was 30 because it is typically considered the 'minimum sampling bound for central limit theorem'. This would mean that reporting average performance by taking the average of the class conditional DI scores will be statistically sound and unbiased.

During training, we need data from both benign and attack days for supervised learning. We used 70% of all available data for training and the remaining 30% for the purpose of testing. Furthermore, we sampled this 70% of training and 30% of test data randomly from the available labeled attack and benign days by sampling rows randomly based on the timestamps and created 30 different realizations of training and test datasets as mentioned earlier.

4.1 Detection Threshold Identification

To obtain the threshold, we retrofit the optimal parameter values s^*, T^*, q^* into the parameterized model (i.e. in Eqn 3) to get an optimized DI value for each device for the attack and the benign part of the training data respectively. These DI values can be seen in Fig. 3, and as evident from the picture when the optimal parameters are retrofitted to each IoT device's model, we get linearly separable score embeddings under benign and attack. These label and score embeddings can now be given as input to a simple linear two-class SVM, whose optimal hyperplane can serve as a classification threshold. Therefore, we did the same and obtained the threshold as seen in Fig. 3. We will use this learned threshold for distinguishing between an attack and a benign during the test set.

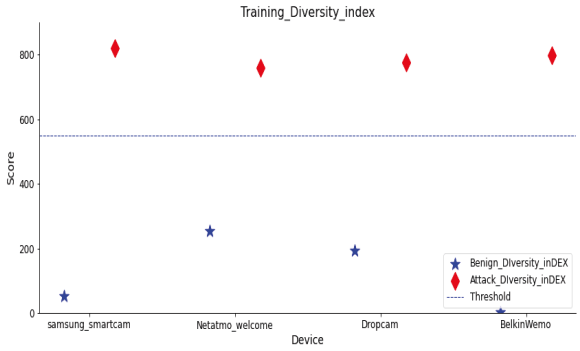

Figure 3: Sample Averaged Training Set Model Scores and SVM Hyperplane as Threshold

4.2 Testing Set Performance

30% of all the data were used for the test set. The benign part of test set contains only benign data while the attack part of test set contains a mix of both benign and attack samples emulating the ON-OFF strategy. We had 30 different realizations of benign and attack subsets.

Performance in the test set is depicted in Fig. 4. We can observe that on average, the DI scores under benign (indicated by blue markers) remain significantly below the threshold identified during training. In contrast, the average diversity index scores for the same devices under data that contains benign and attack traffic in the test set, are much above the classification threshold, proving that our model detects low-volume attacks.

5 CONCLUSIONS

We conclude that an information diversity index-based model is an appropriate latent space embedding for smart home IoT device level low volume on-off DDoS attack detection due to its explainability, simplicity, and generalization power, if modeled in the way

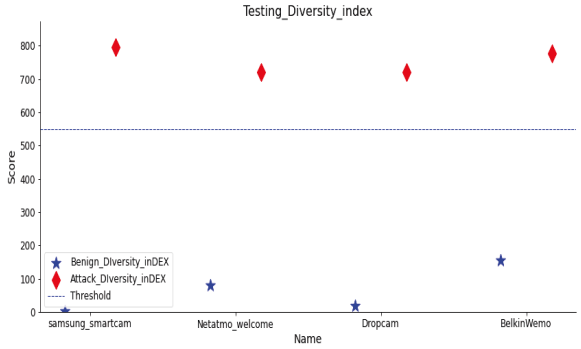

Figure 4: Classifier Results: Sample Averaged Test Set Scores

this paper proposes. We observe from our study that despite high variability in traffic, the classification scores in either label are very distinguishable. We took steps to avoid sampling bias by sampling different combinations of attacks instead of passively using one realization of attacks by just passively using the dataset. This allows machine learning for cybersecurity researchers a way to re-think how attack datasets are used for validating machine learning-based cybersecurity methods and why they could be biased. We also indicated that if model architecture is explainable, the search space can be pruned via reasoning about how the architecture relates to the problem.

ACKNOWLEDGMENTS

Research supported by NSF-USA grants SATC-2030611,OAC-2320951, SATC-2335824, and WMU's FRACAA and LSAMP funding supporting undergraduate student research.

REFERENCES

[1] M. Antonakakis et. al. "Understanding the Mirai Botnet", *USENIX Security Symposium*, 2017.
[2] M. Lyu, D. Sherratt, A. Sivanathan, H. Gharakheili, A. Radford, V. Sivaraman. "Quantifying the reflective DDoS attack capability of household IoT devices" *ACM WiSec*, pp. 46–51, 2017.
[3] M. Nobakht, C. Russell, W. Hu, A. Seneviratne, "IoT-NetSec: Policy-Based IoT Network Security Using OpenFlow" *IEEE PerCom Workshops*, pp. 955-960, 2019.
[4] A. Hamza, H. Gharakheili, T. Benson, V. Sivaraman, "Detecting Volumetric Attacks on IoT Devices via SDN-Based Monitoring of MUD Activity", *ACM SOSR*, 2019.
[5] O. Alrawi, C. Lever, M. Antonakakis, F. Monrose, "SoK: Security Evaluation of Home-Based IoT Deployments," *IEEE Symposium on Security and Privacy (SP)*, pp. 1362-1380, 2019.
[6] A. Sivanathan, D. Sherratt, H. H. Gharakheili, A. Radford, C. Wijenayake, A. Vishwanath, V. Sivaraman, "Characterizing and classifying IoT traffic in smart cities and campuses," *IEEE INFOCOM Workshops*, pp. 559-564, 2017.
[7] [Online] Available at: https://www.cloudflare.com/learning/ddos/what-is-a-ddos-attack/ [Accessed 24 February 2020].
[8] A. Parmisano, S. Garcia, M. Erquiaga, 'IoT-23 Dataset: A labeled dataset of Malware and Benign IoT Traffic.', *Avast-AIC laboratory, Stratosphere IPS, Czech Technical University (CTU)*, Prague, Czech Republic, 2019.
[9] D. Scott, "Multivariate Density Estimation: Theory, Practice, and Visualization" *Wiley*, 1992.
[10] Y. Amar, H. Haddadi, R. Mortier, A. Brown, J. Colley, A. Crabtree, "An Analysis of Home IoT Network Traffic and Behaviour" *arXiv:1803.05368*, 2018.
[11] Y. Mirsky, T. Doitshman, Y. Elovici, A. Shabtai, "Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection", *Network and Distributed System Security Symposium*, 2018.
[12] H. AlSheakh, S. Bhattacharjee, " Towards a Unified Trust Framework for Detecting Smart IoT Devices Under Attacks", *IEEE MASS*, 2020.
[13] L. Jost, "Entropy and diversity", Wiley Oikos, Vol. 112, pp. 363-375, 2006.