

Detecting Compromised IoT Devices Using Autoencoders with Sequential Hypothesis Testing

Md Mainuddin, Zhenhai Duan

Department of Computer Science

Florida State University

Tallahassee, FL 32306

Email: {mainuddi, duan}@cs.fsu.edu

Yingfei Dong

Department of Electrical Engineering

University of Hawaii

Honolulu, HI 96822 USA

Email: yingfei@hawaii.edu

Abstract—IoT devices fundamentally lack built-in security mechanisms to protect themselves from security attacks. Existing works on improving IoT security mostly focus on detecting anomalous behaviors of IoT devices. However, these existing anomaly detection schemes may trigger an overwhelmingly large number of false alerts, rendering them unusable in detecting compromised IoT devices. In this paper we develop an effective and efficient framework, named CUMAD, to detect compromised IoT devices. Instead of directly relying on individual anomalous events, CUMAD aims to accumulate sufficient evidence in detecting compromised IoT devices, by integrating an autoencoder-based anomaly detection subsystem with a sequential probability ratio test (SPRT)-based sequential hypothesis testing subsystem. CUMAD can effectively reduce the number of false alerts in detecting compromised IoT devices, and moreover, it can detect compromised IoT devices quickly. Our evaluation studies based on the public-domain N-BaIoT dataset show that CUMAD can on average reduce the false positive rate from about 3.57% using only the autoencoder-based anomaly detection scheme to about 0.5%; in addition, CUMAD can detect compromised IoT devices quickly, with less than 5 observations on average.

1. Introduction

In recent years Internet of Things (IoT) devices have been increasingly integrated into our daily lives and our society, with notable example environments such as smart homes, healthcare, transportation, and power grid. On one hand, this rapid development helps to improve the quality and efficiency of our daily lives. On the other hand, this same development also poses potentially unprecedented security and privacy challenges on the Internet, given that most of these IoT devices are low-cost systems with limited computation, memory, and energy resources. These devices often lack proper built-in security mechanisms to protect themselves and are vulnerable to various security attacks.

Many security attacks targeting or based on IoT devices have been reported in the past [1]. In response to the growing problems of IoT security, government agencies such as US NIST have developed many recommendations that manufac-

turers should adopt to mitigate the security risks associated with IoT devices [2]. In addition, many research efforts have been carried out to improve IoT security, including both proactive approaches to enhancing security mechanisms of IoT devices and more reactive solutions to monitor IoT device behaviors to detect rogue or infected IoT devices [3].

Although some of the recommendations, for example, avoiding default common credentials, are relatively easy to be incorporated into IoT device manufacturing and certainly help mitigate IoT security risks, IoT devices are still fundamentally vulnerable to security attacks. As low-cost systems, IoT devices are inherently constrained in resources to support advanced security mechanisms. In addition, from the perspectives of both manufacturers and users, there are often conflicting objectives of IoT device usability and security, which often discourage the adoption of advanced security mechanisms in IoT devices.

Given these constraints of deploying advanced security mechanisms on IoT devices, network-based solutions have attracted a great amount of research efforts in recent years [3]. In particular, many machine learning (ML) based methods have been developed in detecting anomalous network behaviors of IoT devices [3]. (In this paper we use the term ML to refer to both traditional machine learning algorithms such as SVM and deep learning (DL) algorithms such as RNN.) However, most existing solutions only targeted the problem of anomaly detection in IoT devices [4], instead of detecting compromised IoT devices. Although detecting individual anomalies is of critical importance in certain application domains [5], we note that these solutions may not be directly translated into the detection of compromised IoT devices. Given the large amount of network traffic, even a small false positive rate of an anomaly detection method can often translate into a large number of false alerts, rendering the detection method unusable in detecting compromised IoT devices in the real-world deployment.

In this paper we develop an effective and efficient framework to detect compromised IoT devices, named CUMAD (cumulative anomaly detection). In essence, CUMAD integrates an autoencoder-based anomaly detection subsystem with a sequential probability ratio test (SPRT)-based sequential hypothesis testing subsystem [6], [7]. In CUMAD, the

normal behavior of each IoT device is learnt and modeled by an autoencoder. During the training of an autoencoder model, it learns a latent space representation of the training data. More importantly, due to the nature of autoencoder, it excels at reconstructing inputs that are similar to the data used in training the model, but performs poorly when the new data is very different from the training data, manifested as large reconstruction errors. Although autoencoder has been mainly used in dimensionality reduction and feature learning in the past, in recent years it has also attracted a great amount of interests in anomaly detection in many different application domains.

Instead of focusing on individual anomalous events detected by autoencoder, CUMAD aims to accumulate sufficient evidence to detect if an IoT device has been compromised. In CUMAD, the output of the autoencoder-based anomaly detection subsystem is fed into an SPRT-based sequential hypothesis testing subsystem. Unlike traditional probability ratio test methods that require a pre-defined fixed number of observations to reach a decision, SPRT works in an online manner and updates as observations arrive sequentially. SPRT reaches a conclusion whenever sufficient evidence has been observed. Therefore, SPRT can make a decision quickly (and consequently, CUMAD can detect compromised IoT devices quickly).

In this paper we develop the CUMAD framework, and we also evaluate the performance of CUMAD using a public-domain IoT dataset N-BaIoT [8], which contains both benign and (Mirai and Bashlite) attack traffic of IoT devices. Our evaluation studies show that CUMAD can greatly improve the performance in detecting IoT devices in terms of false positive rates, for example, compared to the simple autoencoder-based anomaly detection system, CUMAD on average reduces the false positive rate from about 3.57% to 0.5%, representing about 7 times performance improvement in terms of false positive rate of the systems. In addition, CUMAD can detect a compromised IoT device quickly, with less than 5 sequential observations on average. We note that although both autoencoder and SPRT have been proposed in developing anomaly detection systems before, to our knowledge, we are the first to integrate the two techniques to detect compromised IoT devices, instead of being used separately for anomaly detection. In addition, we are the first to introduce the notion of *cumulative anomaly* in detecting compromised IoT devices (see Section 2 for more details).

The remainder of the paper is organized as follows. In Section 2 we discuss related work. We present the background on autoencoder and SPRT in Section 3. We describe the design of CUMAD in Section 4, and evaluate its performance in Section 5. We conclude the paper in Section 6.

2. Related Work

The problem of anomaly detection has been studied in many different application domains and many techniques have been proposed, based on statistical inference, data mining, signal processing, and recently machine learning,

among others. We note that in the literature of anomaly detection, anomalies have been classified into three categories: point anomaly, contextual anomaly, and collective anomaly [5]. However, they are all concerned with the detection of individual anomalous events, which are different from the *cumulative anomaly* we consider in this paper. In cumulative anomaly we are more concerned with the cause of anomalous events (for example, compromised IoT device), instead of individual anomalous events. As a consequence, we need to accumulate sufficient evidence (individual anomalous events) to reach a conclusion (for example, if an IoT device is compromised) in cumulative anomaly detection.

Given the importance of improving IoT security, many security attack detection techniques have been proposed, including various ML-based solutions [9]. However, some of them required the training data of both benign and attack traffic. They cannot detect new security attacks. Others developed anomaly detection based schemes to detect anomalous traffic originated from IoT devices. However, as we have discussed in Section 1, they often trigger a large number of false alerts, rendering them unusable in detecting compromised IoT devices in the real-world deployment.

In [10], Gelenbe and Nakip developed an online scheme CDIS to detect compromised IoT devices based on auto-associative learning. However, the design of CDIS was tailored to Mirai botnet, and may not be effective to detect other types of compromised IoT devices. In addition, CDIS still only targeted individual anomalous events, instead of cumulative anomaly detection as we perform in this paper. The authors of [11] developed a federated self-learning based scheme D²IoT to detect compromised IoT devices, where local security gateways communicate with remote IoT Security Service to build a more comprehensive normal traffic model of IoT devices. In order to further reduce the false alerts generated by the aggregated anomaly detection model, a window-based scheme was adopted, where anomaly alarm was triggered only if the fraction of anomalous packets was greater than a pre-defined threshold value. In [8], Meidan *et al.* presented an autoencoder-based anomaly detection system N-BaIoT to detect compromised IoT devices. N-BaIoT also tried to reduce the number of false alerts triggered by the pure anomaly detection system using a window-based scheme with a majority vote to reach a decision.

3. Background on Autoencoder and SPRT

In this section we provide the necessary background on autoencoder and sequential probability ratio test (SPRT) for understanding the development of the proposed CUMAD framework. We refer interested readers to [6] and [7], respectively, for the detailed treatment on these two topics.

3.1. Autoencoder

Autoencoder is an unsupervised neural network that aims to reconstruct the input at the output. Figure 1 illustrates a simple standard (undercomplete) autoencoder.

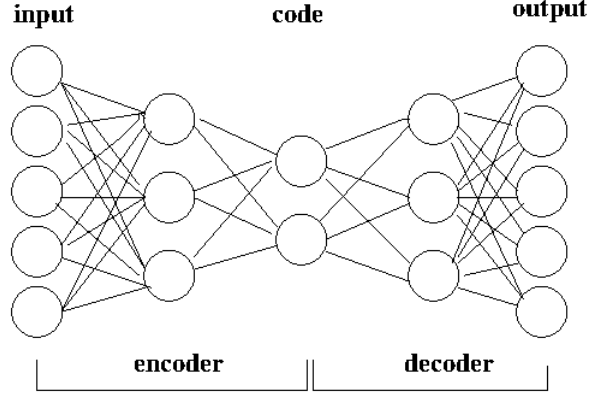


Figure 1. Illustration of Autoencoder.

An autoencoder can be considered as consisting of two components: an encoder f and a decoder g . Given input data $\mathbf{x}$, the encoder function f maps $\mathbf{x}$ to a latent-space representation, or code $\mathbf{h}$, that is $\mathbf{h} = f(\mathbf{x})$. Using the corresponding code $\mathbf{h}$ as the input, the decoder function g tries to reconstruct the original input $\mathbf{x}$ at its output $\mathbf{x}'$, that is, $\mathbf{x}' = g(\mathbf{h})$. Combining both the encoder function and decoder function together, we have $\mathbf{x}' = g(f(\mathbf{x}))$. Let $L(\mathbf{x}, \mathbf{x}')$ be the reconstruction error, that is, the difference between $\mathbf{x}$ and $\mathbf{x}'$. The autoencoder aims to minimize $L(\mathbf{x}, \mathbf{x}')$. We note that there are different definitions of $L(\mathbf{x}, \mathbf{x}')$ and one of the most common definitions is the *mean squared errors* (MSE). We note that in the example autoencoder of Figure 1, both the encoder and decoder have only one hidden layer. This is only for illustration purpose. In reality they can have many hidden layers, depending on the specific application requirement.

Autoencoders have been traditionally used in applications of dimensionality reduction and feature learning, by focusing on the compressed code of an autoencoder, which holds the latent-space representation of the original data. On the other hand, autoencoders also possess a few desired properties, making them an attractive candidate for anomaly detection. For example, an autoencoder is able to extract the salient features of the original data to remove dependency in the original data. More importantly, an autoencoder can only learn the properties or distributions of the data that it has seen during the training stage, that is, the data points in the training dataset. It excels at reconstructing data that are similar to the training data, but performs poorly on data that are very different from the training data, in terms of the reconstruction error $L(\mathbf{x}, \mathbf{x}')$.

This is an appealing property of autoencoders in the application of anomaly detection. For example, in the context of detecting compromised IoT devices, we can establish the normal behavioral model of an IoT device using an autoencoder by training it with benign network traffic before the device has been compromised. We can continue monitoring the IoT device by passing the corresponding network traffic of the device into the trained model. If the reconstruction error is no greater than a pre-specified

threshold, we consider the corresponding network traffic to be benign. When the reconstruction error is greater than the threshold, we claim that the network traffic is anomalous.

3.2. Sequential Probability Ratio Test

Sequential probability ratio test (SPRT) is a simple yet powerful statistical tool that has found applications in many different domains, in particular, fault detection or quality control [7]. SPRT is a variant of the traditional probability ratio test for testing under what distribution (or with what distribution parameters), it is more likely to have the observed sequence of samples. Unlike traditional probability ratio test that requires a pre-defined fixed number of samples to carry out the test, SPRT works in an online fashion; it updates the corresponding statistical measure as samples arrive sequentially, and can conclude when sufficient samples have arrived to reach a decision. In its simplest form, SPRT is a statistical method to test a simple null hypothesis against a simple alternative hypothesis. In the following we will more formally describe the operation of SPRT.

Let $\mathbf{y}$ denote a Bernoulli random variable with an unknown parameter θ , and let $\mathbf{y}_i$, for $i = 1, 2, \dots$ denote the corresponding successive observations of $\mathbf{y}$. SPRT can be used to test a simple null hypothesis H_0 that $\theta = \theta_0$ against a simple alternative hypothesis H_1 that $\theta = \theta_1$, more specifically,

$$\begin{aligned} Pr(\mathbf{y}_i = 1|H_0) &= 1 - Pr(\mathbf{y}_i = 0|H_0) = \theta_0 \\ Pr(\mathbf{y}_i = 1|H_1) &= 1 - Pr(\mathbf{y}_i = 0|H_1) = \theta_1. \end{aligned}$$

As observations $\mathbf{y}_i$ for $i = 1, 2, \dots, n$ arrive one by one, SPRT maintains a probability ratio measure, namely,

$$\frac{Pr(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n|H_1)}{Pr(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n|H_0)}$$

to test under which hypothesis (H_1 or H_0) it is more likely to observe the sequence of samples $\mathbf{y}_i$, for $i = 1, 2, \dots, n$. In order to simplify exposition and computation, in the following we will compute the logarithm of the probability ratio and denote it as Λ_n :

$$\Lambda_n = \ln \frac{Pr(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n|H_1)}{Pr(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n|H_0)}.$$

Assume that $\mathbf{y}_i$'s are independent (and identically distributed), we have

$$\Lambda_n = \ln \frac{\prod_{i=1}^n Pr(\mathbf{y}_i|H_1)}{\prod_{i=1}^n Pr(\mathbf{y}_i|H_0)} = \sum_{i=1}^n \ln \frac{Pr(\mathbf{y}_i|H_1)}{Pr(\mathbf{y}_i|H_0)} = \sum_{i=1}^n z_i \quad (1)$$

where $z_i = \ln \frac{Pr(\mathbf{y}_i|H_1)}{Pr(\mathbf{y}_i|H_0)}$, which can be considered as the step in the random walk represented by Λ . When the observation is one ($\mathbf{y}_i = 1$), the constant $\ln \frac{\theta_1}{\theta_0}$ is added to the preceding value of Λ . When the observation is zero ($\mathbf{y}_i = 0$), the constant $\ln \frac{1-\theta_1}{1-\theta_0}$ is added.

Now we are ready to describe the operation of SPRT for testing H_0 against H_1 . As a sample $\mathbf{y}_i$ arrives, depending on the value of the sample, the statistical measure Λ_n is updated

according to Eq. (1). The measure Λ_n is then compared against two user-defined thresholds A and B ($A < B$) to check if SPRT can reach a decision or more samples are needed:

$$\begin{aligned} \Lambda_n \leq A &\implies \text{accept } H_0 \text{ and terminate test,} \\ \Lambda_n \geq B &\implies \text{accept } H_1 \text{ and terminate test,} \\ A < \Lambda_n < B &\implies \text{take an additional observation.} \end{aligned} \quad (2)$$

The two thresholds A and B in the operation of SPRT can be approximated using the user-desired false positive rate α and false negative rate β [7]:

$$A \approx \ln \frac{\beta}{1-\alpha}, \quad B \approx \ln \frac{1-\beta}{\alpha}. \quad (3)$$

As a simple and powerful statistical tool, SPRT possesses a few critical and desired properties that lead to the wide-spread application of the technique in many different domains. First, the false positive and false negative rates of SPRT can be specified by user-desired error rates, which in turn control the thresholds of the model. Second, it has been proved that, among all sequential and non-sequential probability ratio testing algorithms, SPRT minimizes the expected number of observations to reach a decision with no greater errors. Put in another way, on average SPRT can reach a conclusion quickly compared to other probability ratio testing algorithms.

In the context of anomaly detection, we can consider $y_i = 1$ as an anomalous sample, and $y_i = 0$ as a normal sample. Instead of simply relying on a single sample to conclude the nature of the sample's origin (being normal or abnormal), SPRT continues observing additional samples $y_{i+1}, y_{i+2}, \dots$ until sufficient evidence is collected to reach a conclusion. In essence, SPRT infers the distribution or the parameters of the distribution of a random process, based on sequential observations of samples drawn from the random process. This matches our objectives of CUMAD well, where we would like to accumulate sufficient evidence to detect if an IoT device is compromised, instead of an individual anomalous sample.

4. Design of CUMAD

In this section we will first discuss the considered network model, where CUMAD will be deployed, and then we will present the design of the CUMAD framework.

4.1. Network Model

Figure 2 illustrates the conceptual network model, where CUMAD is deployed. As shown in the figure, in order for CUMAD to carry out its task to detect compromised IoT devices in a network, CUMAD needs to have access to the network traffic associated with the IoT devices in the network. Depending on the deployment scenarios of CUMAD in the network and the corresponding network architecture, there can be a few different ways for CUMAD to obtain the corresponding network traffic of IoT devices. In essence,

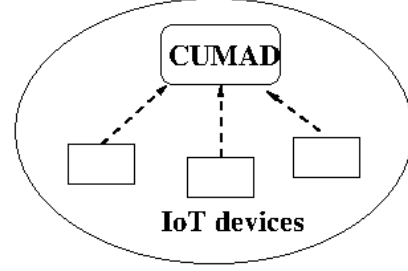


Figure 2. Conceptual network model.

CUMAD as a network-based solution can be deployed in a similar way as network-based intrusion detection systems.

In the current design of CUMAD, (statistical) features from raw network traffic will be extracted and fed to CUMAD for detecting compromised IoT devices. Each input data point fed to CUMAD comprises these extracted features, and can be summarized at different levels of granularity of network traffic, such as packets, flows, and time windows. These features will capture the network behavioral characteristics of the corresponding IoT devices. In Section 5 we will discuss the network traffic features contained in the public-domain N-BaIoT dataset when we perform evaluation studies on CUMAD [8].

4.2. CUMAD: Cumulative Anomaly Detection

Figure 3 illustrates the high-level architecture of the CUMAD framework. CUMAD consists of two main components: an anomaly detection component (ADC) and a cumulative anomaly component (CAC). Assuming the model has been properly trained (will be discussed shortly), given an input data point with the corresponding features, the main responsibility of ADC is to classify an input data point as either normal or anomalous. After the classification of the input data point, the result is passed to the second component (CAC), which will maintain a cumulative view of the network traffic behavior of the corresponding IoT device, by sequentially merging the individual classification results into the view. When sufficient evidence on an IoT device has been collected to indicate that it has been compromised, an alert will be generated. In the following we will describe each component in details, both in model training and deployment to detect compromised IoT devices.

We note that different types of IoT devices perform drastically different functionalities, and in addition, we would like to detect which IoT device is compromised, we need to develop a separate CUMAD model for each IoT device and monitor their network traffic behaviors separately using their corresponding CUMAD models. Therefore, the following discussions are for one IoT device. We note that, although there are vastly diverse types of IoT devices on the Internet, autoencoder is a powerful neural network that is capable of learning different models. Therefore, we are able to build diverse autoencoder models, one for each IoT device, despite their vastly different network traffic behaviors of

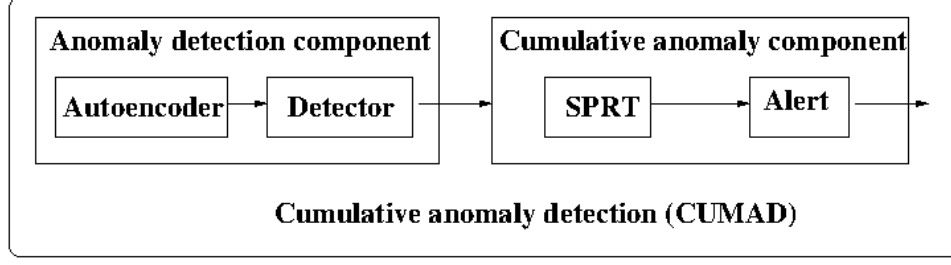


Figure 3. Illustration of CUMAD architecture.

these IoT devices. In addition, IoT devices also provide us with unique opportunities in establishing the models of normal behaviors, compared to traditional computer systems. In particular, each IoT device only performs a few well-defined simple functionalities in an autonomous or semi-autonomous fashion, with very limited user interactions after the initial device configuration and setup. This makes it simpler to establish a model of normal behaviors in carrying out anomaly detection.

4.2.1. Model Training and Setup. Before CUMAD can be used to monitor network traffic to detect compromised IoT devices, we need to train a CUMAD system for each IoT device so that it can learn the normal model of the device. During the training stage of a deployed CUMAD system, it is critical that we should only feed normal (benign) network traffic of the device to the system. This can be done, for example, when an IoT device is first deployed in the network. In order to minimize false positives during the detection stage, it is also important that CUMAD has a reasonably complete view of all the normal network traffic behavior of the device.

The training of the autoencoder module in ADC (see Figure 3) follows the basic standard steps in ML training [12]. Given a set $\mathbf{D}$ of normal input data points to the autoencoder of ADC, we partition it into two subsets: the training set $\mathbf{D}_t$ and the validation set $\mathbf{D}_v$. We use $\mathbf{D}_t$ to train the model and use $\mathbf{D}_v$ to obtain the optimal value of the hyperparameters including learning rate and epochs to optimize the performance of the final model in terms of reconstruction errors. After the optimal values of these hyperparameters are obtained, we perform the final training of the model using $\mathbf{D}$. After the training is complete, the autoencoder obtains the latent-space representation of the normal network traffic behavior of the corresponding IoT device, which can capture the salient properties of the device's network traffic.

As discussed above, the premise of using an autoencoder as an anomaly detection mechanism is that, although it can effectively reconstruct data points that are similar to the data points that it has seen previously during the training stage, it in general performs poorly to reconstruct data points that substantially differ from the training data. This is manifested in large reconstruction errors. Therefore, we will use the reconstruction error as the anomaly score, and when the

anomaly score is greater than the pre-defined threshold, we classify the corresponding input data point as an anomalous sample.

Clearly, the choice of the anomaly score threshold T_{as} will greatly affect the performance of the resulting CUMAD system. In the current design, we choose T_{as} in the following manner. Let μ_D and σ_D denote the average and standard deviation of the mean squared errors for data points in the dataset $\mathbf{D}$, respectively, then $T_{as} = \mu_D + \sigma_D$. This threshold value is maintained by the Detector module in ADC (see Figure 3).

The training of the autoencoder and the choice of T_{as} also provide us with some hints on the parameters of SPRT in CAC. In particular, based on the value of T_{as} we can compute the proportion of input data points in $\mathbf{D}$ that will be (mistakenly) classified as being anomalous, and we can treat this proportion as the value for θ_0 , that is, the probability that we will have an anomalous sample when the device is not compromised. In addition to the parameter θ_0 , we also need to set up a few other parameters for SPRT to work, including θ_1 , α , and β . θ_1 is the probability that an input data point is anomalous when the hypothesis H_1 is true, that is, when the IoT device is compromised. During the time when a compromised IoT device is used to launch a security attack, this probability should be much higher than θ_0 . Although in general we cannot obtain the precise value of θ_1 , a value sufficiently greater than θ_0 should work. We note that imprecise values of θ_1 (and θ_0) may result in a larger number of samples for SPRT to reach a decision.

The parameters α and β are the user-desired false positive rate and false negative rate, respectively. They normally have small values for all practical applications, for example, in the range 0.01 to 0.05. The initial value of Λ_n in Eq. (1) is set to 0 during the setup stage of the system. The functionality of the Alert module is to generate proper alert to inform system administrators of the detection of a compromised IoT device. Other actions can also be taken based on the local security policies, for example, informing proper agents to isolate the compromised IoT device.

4.2.2. Detection. After the model has been trained and the required parameters have been set for the CUMAD system, it can be used to monitor network traffic to detect if the corresponding IoT device has been compromised. In the following we describe the basic steps of a CUMAD system

Algorithm 1 CUMAD procedures in detection.

```

1: Let  $\mathbf{x}$  be an input data point
2: procedure AUTOENCODER( $\mathbf{x}$ )
3:    $\mathbf{x}' \leftarrow g(f(\mathbf{x}))$   $\triangleright$  Reconstruct  $\mathbf{x}$  in autoencoder
4:    $s \leftarrow L(\mathbf{x}, \mathbf{x}')$   $\triangleright$  Compute anomaly score
5: procedure DETECTOR( $s$ )
6:   if  $s > T_{as}$  then
7:     return 1  $\triangleright$  Anomalous input data point
8:   else
9:     return 0  $\triangleright$  Normal input data point
10: procedure SPRT( $o$ )
11:    $o$ : output of Detector
12:    $\triangleright$  Update  $\Lambda_n$  based on Eq. (1)  $\triangleleft$ 
13:   if ( $o == 1$ ) then
14:      $\triangleright$  anomalous  $\triangleleft$ 
15:      $\Lambda_n + = \ln \frac{\theta_1}{\theta_0}$ 
16:   else
17:      $\triangleright$  normal  $\triangleleft$ 
18:      $\Lambda_n + = \ln \frac{1-\theta_1}{1-\theta_0}$ 
19:    $\triangleright$  If decision can be made based on Eq. (2)  $\triangleleft$ 
20:   if ( $\Lambda_n \geq B$ ) then
21:     Device is compromised.
22:     Test terminates for the IoT device
23:     Calling Alert module to generate alert
24:   else if ( $\Lambda_n \leq A$ ) then
25:     Device is normal. Test is reset for the device
26:      $\Lambda_n \leftarrow 0$ 
27:     Test continues with new observations
28:   else
29:     Test continues with an additional observation

```

in carrying out the detection task (see Algorithm 1).

The input data point $\mathbf{x}$ is fed into the trained autoencoder model of CUMAD (lines 2 to 4 in Algorithm 1). The autoencoder will try to reconstruct the input and create the output $\mathbf{x}'$. The anomaly score s is then computed (it is the mean squared error in the current design). The anomaly score is then passed to the Detector module of ADC, which maintains an anomaly score threshold to distinguish between normal and anomalous data points (lines 5 to 9 in the algorithm). If the anomaly score is greater than the threshold T_{as} , an anomalous data point is identified, and an output value 1 is generated in the Detector to indicate the detection of the anomalous data point. Otherwise, an output value 0 is generated to indicate a normal data point.

The output of the Detector module is then passed to the SPRT module to determine if sufficient evidence has been accumulated to make a decision regarding the nature of the IoT device (compromised or normal; line 10 of the algorithm). SPRT updates the probability ratio measure Λ_n according to Eq. (1), as the 0 (normal data point) and 1 (anomalous data point) output sequence of the Detector module arrives (lines 13 to 18). After the value of Λ_n is updated for each input data point, SPRT compares the value of Λ_n with the two boundaries A and B to determine if a decision can be made (lines 20 to 29). When the value of Λ_n hits or crosses the upper bound B , SPRT will conclude that

the alternative hypothesis H_1 is true, that is, the IoT device has been compromised. In this case, SPRT will inform the Alert module the detection of an compromised IoT device. Proper alert will be generated and corresponding system administrators will be informed. In addition, from this time on, it is not necessary to monitor the IoT device anymore, until proper actions have been taken to clean up or remove the device.

When the value of Λ_n is equal to or smaller than the lower bound A , SPRT reaches the conclusion that H_0 is true, that is, the IoT device is not compromised. From the viewpoint of detecting compromised IoT devices, this conclusion is less interesting in that we cannot terminate the monitoring of the device as we have done when a compromised IoT device is detected. A normal IoT device may become compromised at a later time. Therefore, in this case, we will reset the state of SPRT to restart the monitoring of the IoT device, in particular, we will reset the value of Λ_n to zero. If a decision cannot be reached at this time (line 28), SPRT will simply wait for additional input data points and repeat the same procedure.

5. Evaluation Studies

In this section we perform evaluation studies to investigate the performance of CUMAD using the public-domain N-BaIoT dataset [8]. In order to better understand the evaluation studies, we will first describe the dataset, in particular, the features of the data points contained in the dataset. We will also compare the performance of CUMAD with that of the N-BaIoT scheme (which is the name for both the dataset and the corresponding scheme on detecting compromised IoT devices) [8].

5.1. Dataset, Features, and CUMAD System Setup

TABLE 1. N-BaIoT FEATURE EXTRACTION AGGREGATIONS.

Aggregation level	# features	Pkt attributes and measures
Source IP	3	# pkts, Mean/Variance of pkt sizes ¹
Source MAC-IP	3	# pkts, Mean/Variance of pkt sizes ¹
Channel	10	# pkts, Mean/Variance of pkt sizes ¹ , Mean/Variance/Count of IATs, Magnitude/Radius/Covariance/Correlation Coefficient of pkt sizes ²
Socket	7	# pkts, Mean/Variance of pkt sizes ¹ , Magnitude/Radius/Covariance/Correlation Coefficient of pkt sizes ²

¹ Outgoing packets only

² Both incoming and outgoing packets

N-BaIoT contains both benign and (Mirai and Bash-lite) attack traffic of 9 commercial IoT devices, including two doorbells (Danmini and Ennio), an Ecobee thermostat, three baby monitors (different models from Provision and Philips), two SimpleHome security cameras, and a Samsung webcam. Benign IoT device traffic was collected immediately after the corresponding IoT device was connected to the experimental testbed. Care was taken to ensure that various representative normal operations and behaviors of IoT devices were collected into the benign dataset.

TABLE 2. PERFORMANCE RESULTS

Device	CUMAD				N-BaIoT			
	Mean Size	Accuracy	Recall	F1 Score	Window Size	Accuracy	Recall	F1 Score
Danmini Doorbell	5.19	0.979	1.000	0.979	82	0.995	1.000	0.995
Ecobee Thermostat	4.18	0.988	1.000	0.988	20	0.995	1.000	0.995
Philips B120N10 Baby Monitor	5.62	0.955	0.994	0.957	65	0.982	0.992	0.983
Provision PT 737E Security Cam	4.21	0.992	1.000	0.992	32	1.000	1.000	1.000
Provision PT 838 Security Cam	4.79	0.957	1.000	0.958	43	0.978	1.000	0.978
SimpleHome XCS7 1002 WHT Security Cam	4.08	0.995	0.999	0.995	23	0.998	0.998	0.998
SimpleHome XCS7 1003 WHT Security Cam	4.11	0.993	1.000	0.993	25	0.995	1.000	0.995

These IoT devices were later infected with Mirai and Bashlite malware in the controlled environment, and the attack traffic was also collected and added into the dataset. The dataset does not contain Mirai attack data for two of the devices (Ennio Doorbell and Samsung Webcam). We exclude these two IoT devices from our evaluation studies, therefore, we have 7 IoT devices in our evaluation studies. In our evaluation studies, we use the benign traffic and the scan attack traffic from the remaining 7 IoT devices. We equally partition the benign data of an IoT device into three datasets: the training dataset $\mathbf{D}_t$, the validation dataset $\mathbf{D}_v$, and the remaining 1/3 of benign data is merged with the same number of attack data points to form the balanced test dataset $\mathbf{D}_{tst}$.

In the N-BaIoT dataset, each data point corresponds to an arrived packet, and contains 115 statistical features, which together represent a behavioral snapshot that describes the context of the corresponding packet when it arrives at the data collection point. The snapshot contains the source and destination device information, the protocol information, among others. More specifically, the 115 features were extracted in the following manner. For each arriving packet, a total of 23 features were collected at different levels of aggregation (see Table 1), including features aggregated at source IP address level, at source MAC and IP addresses level, at level of channel (source and destination IP addresses), and at socket level (source and destination IP addresses and port numbers). These 23 features were extracted in a sliding window fashion, over 5 time windows of 100ms, 500ms, 1.5sec, 10sec, and 1min, respectively, generating a total of 115 features for each data point.

We use the Keras sequential model as the foundation for our development of the autoencoder [12]. The model's input dimension is set to match the number of features in the dataset (that is, 115). To ensure effective compression, we implement three hidden layers within the encoder. These layers progressively reduce the dimensions to 87, 58, 38, and 29, respectively, with the last one (29) being the dimension of the output layer of the encoder, that is, the dimension of the obtained code. Conversely, the decoder component mirrors the dimensions of the encoder layers in the reverse order, starting from 38. By employing compression and decompression in the encoder and decoder layers, we effectively eliminate redundant information from the features

of the input data points. To optimize training performance, we utilize the Adam optimizer, and the mean square error is used as the reconstruction error (objective function of the model).

SPRT requires four user-defined parameters in order to compute the upper and lower bounds A and B (see Eq. (3)), as well as the step function for computing Λ_n following each observation (see Eq. (1)). The desired values for both the false positive rate and the false negative rate (represented by α and β , respectively) are typically very small. In this study we set both α and β to 0.01. Ideally, the parameter θ indicates the true probability of an observation being classified as an anomaly, from either a benign or compromised IoT device. We determine the values for θ_0 and θ_1 through our preliminary studies, and set them to 0.2 and 0.8, respectively.

5.2. Performance Results

Table 2 shows the performance of CUMAD in detecting IoT devices, in terms of accuracy, recall, and F1 score [12]. From the table we can see that CUMAD achieves superior performance in all three metrics. For example, for 5 of the IoT devices, CUMAD is able to detect all the compromised cases (see the column of Recall). CUMAD is also able to detect vast majority of the compromised cases for the remaining two of the IoT devices, with recall scores of 0.999 and 0.994. Considering both detection precision of attack and benign traffic, we can see that CUMAD also performs very well, with an accuracy score ranging from 0.955 to 0.995 for all 7 IoT devices. The F1 scores, which is a weighted average of the precision and recall scores of a model, also confirm that CUMAD performs well in detecting compromised cases.

Figure 4 shows the false positive rates of an autoencoder-based anomaly detection scheme and CUMAD. As shown in the figure, the false positive rates of the autoencoder-based anomaly detection scheme for the 7 IoT devices range from 0.77% to 11.22%, while the false positive rates of CUMAD range from 0.014% to 2.067%. On average the autoencoder-based anomaly detection scheme has about 3.57% false positive rate, while the false positive rate of CUMAD is about 0.5%, which represents about 7 times performance improvement in terms of false positive rate for CUMAD over the autoencoder-based anomaly detection scheme.

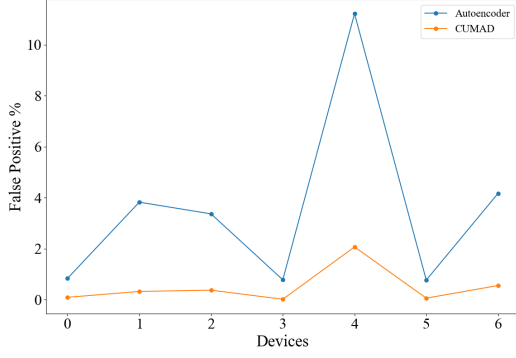


Figure 4. False positive rates.

For performance comparison, we also include in the table the performance results of the N-BaIoT scheme, with the same evaluation studies setup. We can see from the table that CUMAD and N-BaIoT performs comparably in terms of all three performance metrics. However, N-BaIoT works on a fixed window size. Table 2 shows that N-BaIoT requires a relatively large window size, ranging from 20 to 82 (column with name *Window Size*). In contrast, CUMAD works in an online fashion and does not requires such a fixed window size. Table 2 shows the average number of observations required for CUMAD to reach a detection (column with name *Mean Size*); we can see from the table that it takes on average less than 5 observations for CUMAD to make a detection of a compromised case, much quicker than N-BaIoT. In order to have a better understanding of the number of observations for CUMAD to make a detection of a compromised case, Figure 5 shows the cumulative distribution function (CDF) of required observations for CUMAD to make a detection for all the 7 IoT devices. We can see from the figure that the vast majority of detection requires less than 10 observations for all 7 IoT devices.

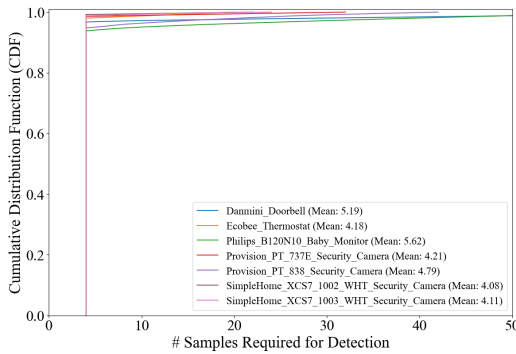


Figure 5. # of observations for detection in CUMAD.

In summary, compared to simple anomaly detection schemes such as the ones only based on autoencoders, CUMAD can greatly reduce the false positive rates, making CUMAD much more attractive than simple anomaly detection schemes in the real-world deployment. Compared with window-based schemes such as N-BaIoT, CUMAD requires

much less observations to reach a detection, and thus can detect compromised IoT devices much quicker.

6. Conclusions

In this paper we have developed CUMAD, a cumulative anomaly detection framework for detecting compromised IoT devices. CUMAD employs an unsupervised neural network autoencoder to classify whether an individual input data point is anomalous or normal. CUMAD also incorporates a statistical tool sequential probability ratio test (SPRT) to accumulate sufficient evidence to detect if an IoT device is compromised, instead of directly relying on individual anomalous input data points. CUMAD can greatly improve the performance in detecting compromised IoT devices in terms of false positive rate compared to the methods only relying on individual anomalous input data points. In addition, as a sequential method, CUMAD can quickly detect compromised IoT devices. Evaluation studies based on public-domain IoT dataset N-BaIoT confirmed the superior performance of CUMAD.

References

- [1] Yin Minn Pa Pa, Shogo Suzuki, Katsunari Yoshioka, Tsutomu Matsumoto, Takahiro Kasama, and Christian Rossow. Iotpot: analysing the rise of iot compromises. In *Proceedings of the 9th USENIX Conference on Offensive Technologies*, pages 9–9, 2015.
- [2] Michael Fagan, Katerina Megas, Karen Scarfone, and Matthew Smith. Foundational cybersecurity activities for iot device manufacturers. Technical report, National Institute of Standards and Technology, May 2020.
- [3] Mohammed Ali Al-Garadi, Amr Mohamed, Abdulla Khalid Al-Ali, Xiaojang Du, Ihsan Ali, and Mohsen Guizani. A survey of machine and deep learning methods for internet of things (iot) security. *IEEE Communications Surveys & Tutorials*, 22(3):1646–1685, 2020.
- [4] Andrew A Cook, Göksel Mısırlı, and Zhong Fan. Anomaly detection for iot time-series data: A survey. *IEEE Internet of Things Journal*, 7(7):6481–6494, 2019.
- [5] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.
- [6] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [7] Abraham Wald. *Sequential Analysis*. John Wiley & Sons, Inc, 1947.
- [8] Yair Meidan, Michael Bohadana, Yael Mathov, Yisroel Mirsky, Asaf Shabtai, Dominik Breitenbacher, and Yuval Elovici. N-baiot—network-based detection of iot botnet attacks using deep autoencoders. *IEEE Pervasive Computing*, 17(3):12–22, 2018.
- [9] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38, 2021.
- [10] Erol Gelenbe and Mert Nakip. Traffic based sequential learning during botnet attacks to identify compromised iot devices. *IEEE Access*, 10:126536–126549, 2022.
- [11] Thien Duc Nguyen, Samuel Marchal, Markus Miettinen, Hossein Fereidooni, N Asokan, and Ahmad-Reza Sadeghi. Diot: A federated self-learning anomaly detection system for iot. In *2019 IEEE 39th International conference on distributed computing systems (ICDCS)*, pages 756–767. IEEE, 2019.
- [12] Francois Chollet. *Deep learning with Python*. Simon and Schuster, 2021.