PATRIoTA: A Similarity-based IoT Malware Detection Method Robust Against Adversarial Samples

1st József Sándor
Budapest University of Technology and Economics
Hungary
jsandor@crysys.hu

2nd Roland Nagy
Budapest University of Technology and Economics
Hungary
rnagy@crysys.hu

3rd Levente Buttyán
Budapest University of Technology and Economics
Hungary
buttyan@crysys.hu

***Abstract*—Detecting malware targeting IoT devices has became an important challenge with the recent emergence of IoT botnets. Gateways at the edge between the Internet and IoT devices deployed in the field are particularly well-positioned for the task of malware detection, as malware typically spreads over the Internet and resource-constrained field devices may not have the means to protect themselves. Hence, we believe that, among other things, edge intelligence should also include effective and efficient IoT malware detection. A recently proposed similarity-based IoT malware detection method, called SIMBIoTA, would be suitable in this context, but its robustness against adversarial malware samples has been shown to be rather weak. In this paper, we propose PATRIoTA, a similarity-based IoT malware detection method inspired by SIMBIoTA, but being significantly more robust than SIMBIoTA is. We describe the operation of PATRIoTA, and compare its malware detection performance and robustness against adversarial samples to that of SIMBIoTA. We show that PATRIoTA outperforms SIMBIoTA with respect to both measures.**

***Index Terms*—Internet-of-Things; malware detection; binary similarity; locality-sensitive hashing; robustness against adversarial samples.**

I. Introduction

The expansion of the Internet-of-Things (IoT) is unwavering: the number of installed IoT devices exceeds 15 billion and it is constantly growing[1]. At the same time, the security of IoT devices is notoriously weak [1], [2]. This poses a threat from two aspects: on the one hand, compromised IoT devices can potentially be used to build huge attacking infrastructures (e.g., botnets), with which Internet-based services can be effectively attacked (see e.g. the Mirai botnet and the DDoS attacks launched from it in 2016 [3]); on the other hand, in cyber-physical applications (e.g., industrial IoT systems, self-driving cars), the compromise of IoT devices can lead to physical damage of expensive equipment or even fatal accidents (see e.g. the proof-of-concept attack on a Jeep Cherokee carried out in 2015 [4] and its potential consequences). The problem is so significant that regulatory processes aiming at increasing the security of IoT systems have been initiated in both the US[2] and Europe[3].

An increasingly widespread method for compromising IoT devices at large scale is infecting them with malware (i.e., malicious programs). This is made possible by the fact that IoT devices are essentially embedded computers and malware can be installed on them just like in the case of traditional computers. As a result, several malware families targeting IoT devices have appeared in the past few years (e.g., Mirai, Gafgyt, Tsunami, Hajime). At the same time, traditional anti-virus solutions require too many resources (e.g., storage capacity and CPU cycles), and therefore, they cannot be applied directly on the typically resource-constrained IoT devices. Hence, there is a great demand for efficient and effective IoT malware detection methods.

Gateways placed at the edge between the Internet and the IoT devices deployed in the field are particularly well-positioned for protecting IoT devices against malware [5]. The main reasons are that malware typically spread over the Internet and that resource-constrained IoT devices may not have the means to protect themselves. Edge gateways, on the other hand, have more resources to support malware detection and, thanks to their placement, they can block malware to reach IoT devices. At the same time, malware detection on gateways must be extremely fast for not imposing delays in communications. This essentially excludes the outsourcing of the malware detecion task to some cloud-based backend, and requires performing all operations locally on the edge gateways themselves. This means that, among other things,

The research presented in this paper was supported by the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory. The presented work also builds on results of the SETIT Project (2018-1.2.1-NKP-2018-00004), which was implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the 2018-1.2.1-NKP funding scheme.

[1]https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/ (accessed on April 25, 2023)

[2]https://www.security.org/blog/california-passes-first-cybersecurity-law-iot/ (accessed on April 25, 2023)

[3]https://digital-strategy.ec.europa.eu/en/library/cyber-resilience-act (accessed on April 25, 2023)

edge intelligence should also include effective and efficient IoT malware detection.

A recently proposed similarity-based IoT malware detection method, called SIMBIoTA [6], fits this context perfectly: SIMBIoTA is lightweight and fast, it performs IoT malware detection entirely locally, yet it has a remarkable malware detection capability. However, its robustness against adversarial malware samples has been shown to be rather weak [7]. In the context of malware detection, adversarial samples are meant to be malware samples crafted specifically to evade detection by a given malware detection method. In the case of SIMBIoTA, such adversarial samples can be created easily by appending some extra bytes at the end of existing malware binaries, as shown in [7].

In this paper, we propose PATRIoTA (PArticle TRained IoT Antivirus), a similarity-based IoT malware detection method inspired by SIMBIoTA, but being significantly more robust than SIMBIoTA is. The main idea of PATRIoTA is to split malware samples known to the antivirus provider into multiple fixed-size parts, referred to as *particles*, and to perform the same operations on those particles as the operations performed by SIMBIoTA on entire samples. This means that the antivirus provider builds a similarity graph of known particles, computes its dominating set, and distributes similarity preserving hash values (in our case, TLSH [8] values) corresponding to the particles in the dominating set to the clients, which are the edge gateways in this work. The clients also split any file to be scanned (e.g., a binary extracted from network traffic) into particles, compute the similarity preserving hash values of them, and if a threshold number of those computed hashes are similar to the hashes in the dominating set, then the file is detected as malware.

PATRIoTA is robust against adversarial sample creating strategies that add extra bytes to an existing malware binary, because the sample created in this way will always contain in it the original binary, and, hence, all of its particles, which can be recognized by PATRIoTA despite the presence of the added extra bytes. In addition, PATRIoTA may be robust against even more sophisticated adversarial strategies that keep sizable chunks of the original malware binary intact within the created adversarial sample, as those chunks may result in particles that are similar to the particles of the original sample. Moreover, our measurement results indicate that, besides increased robustness, PATRIoTA also has better malware detection capabilities than SIMBIoTA has.

The structure of this paper is the following: In Section II, we present the operation of SIMBIoTA and introduce two simple adversarial sample creation strategies that mislead it. In Section III, we introduce PATRIoTA and present its design considerations in details. In Section IV, we compare PATRIoTA to SIMBIoTA in terms of malware detection capability and robustness against the adversarial sample creation strategies introduced in Section II. We discuss the robustness of PATRIoTA against another adversarial sample creation strategy, as well as some alternative ways of increasing robustness against adversarial samples in Section V. Finally, we present some related work in Section VI and conclude the paper in Section VII.

II. Background

Let us start with a more detailed introduction of SIMBIoTA and the lack of its robustness against some simple adversarial sample creation strategies.

A. SIMBIoTA

SIMBIoTA [6] (SIMilarity-based IoT Antivirus) is a lightweight malware detection method tailored to resource constrained IoT devices. It detects malware by checking the similarity of scanned files to known malware samples, but it does this efficiently. In particular, SIMBIoTA exploits the fact that malware samples that belong to the same malware family are typically similar to each other, while samples from different families, as well as benign programs are dissimilar. This means that the similarity graph of the malware samples known to the antivirus provider is clustered and it is disconnected from the similarity graph of benign programs. This phenomenon is illustrated in Figure 1. Here, the similarity graph of a set of binary files is defined as a graph whose vertices represent the binaries and two vertices are connected with an edge if the corresponding files are similar according to some similarity measure. Typically, each cluster in the similarity graph can be represented by a few representative samples such that all members of the cluster are similar to at least one of the representative samples. It is then sufficient for malware detection to know only about the representative samples of the clusters: any scanned file that is similar to any of these representative samples are likely to be malware, while files that are not similar to any of the representative samples are likely to be benign.

More specifically, SIMBIoTA assumes that the antivirus provider continuously collects malware samples from its malware intelligence network (e.g., honeypots and various malware feeds), stores them in a database, and also creates their similarity graph. The (dis)similarity measure applied by SIMBIoTA is called TLSH difference [8] and two vertices of the similarity graph are connected if the TLSH difference between the corresponding malware samples is below the threshold 40. The selection of this particular threshold value is explained in [9]. The antivirus company then calculates a dominating set of the current similarity graph. A dominating set is a subset of the graph's vertices such that each vertex of the graph is either included in the dominating set or it is adjacent to a dominating vertex. The vertices in the dominating set are the representatives of all malware in the database of the antivirus provider. Finally, the antivirus provider delivers the TLSH hash values of the samples belonging to the current dominating set to all clients.

When scanning any file, the client examines how close the TLSH hash value of the scanned file is to the TLSH hash values of the samples belonging to the dominating set. It detects the scanned file as malware if it is similar to any of the samples in the dominating set, and as benign otherwise.

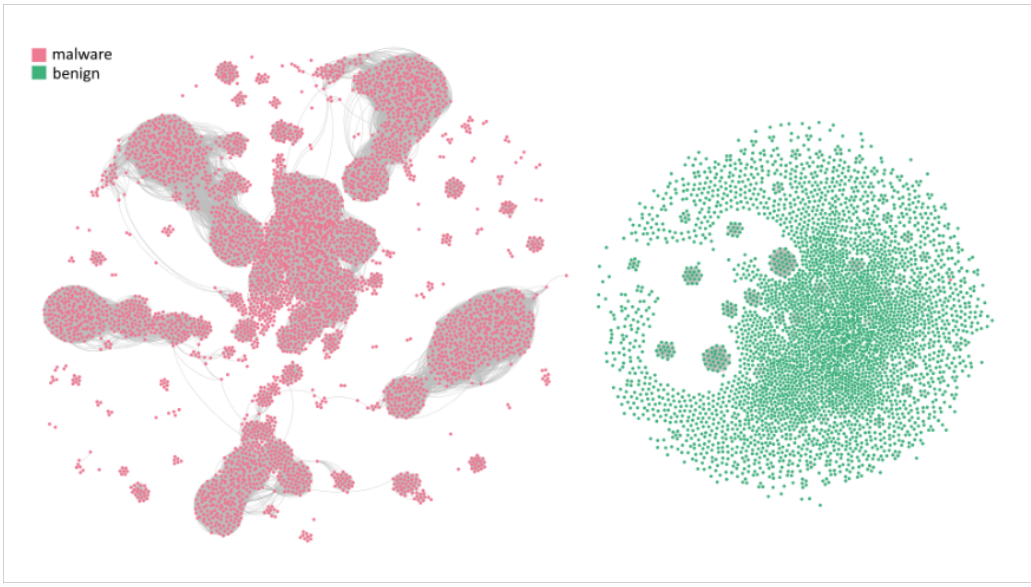

Fig. 1. Illustration of a similarity graph of a set of binaries. Vertices represent the binaries and two vertices are connected with an edge if the corresponding files are similar according to some similarity measure. In this figure, two files are considered to be similar, if their TLSH difference score is below the threshold 40, where TLSH is a locality sensitive hash function and TLSH difference is a TLSH-based dissimilarity metric. Malware binaries are represented by red nodes and benign binaries are represented by green nodes. As it can be seen, there is no similarity between malware and benign files and the malware similarity graph is strongly clustered.

This procedure guarantees that the client detects all malware samples seen by the antivirus company (and included in the similarity graph), and can also detect previously unseen malware that is similar to the samples belonging to the dominating set. In addition, efficiency stems from the facts that the size of the dominating set is just a small fraction of the size of the entire similarity graph, thanks to the high clusteredness of the latter, and the TLSH value and TLSH difference calculations are very fast. Indeed, according to the evaluation in [6], SIMBIoTA required only 6-8 KB of storage capacity and it could decide about any file if it was malicious or benign in 0.12-0.14 ms. In addition, its malware detection capability proved to be surprisingly accurate: it achieved approximately 95% true positive detection rate even on previously unseen malware samples, while its false positive rate remained at 0% throughout the experiments.

B. Adversarial strategies

The malware detection operation in SIMBIoTA is simple, hence, it may be easy to mislead. In order to achieve that, one has to manipulate the TLSH hash value of a malware sample by modifying the sample without harming its malicious functionality. As SIMBIoTA uses the TLSH difference 40 as the similarity threshold, the intuition is that if the TLSH difference between the original sample and the modified sample becomes larger than 40, then SIMBIoTA will likely misclassify it. A modified malware sample with the same functionality as the original one is called an adversarial sample if it is likely to be misclassified as benign by the malware detection mechanism.

One approach for creating adversarial samples is to append extra bytes at the end of a malware binary such that those bytes are never executed, but they affect the calculation of the TLSH hash value. Two specific adversarial sample creating strategies, following this approach, have been proposed in [7]. They are called Chunker and Disguiser. Chunker appends a carefully chosen chunk of the original sample to itself with the goal of increasing the TLSH difference between the modified and the original samples above 40 (or beyond). Disguiser appends an appropriately chosen benign file to the malware binary and its goal is to decrease the TLSH difference between the modified malware and the benign file below 40 (i.e., to make the modified malware similar to the added benign file). These strategies are simple enough to be easily implemented by a real-world attacker. In addition, as shown in [7] (and later replicated in Section IV of this paper), SIMBIoTA can be completely misled by these simple adversarial sample creating strategies such that its detection rate on the adversarial samples created by them is close to 0%.

III. PATRIoTA

In Section II, we mentioned that SIMBIoTA is not robust against the adversarial samples created with the Chunker and Disguiser strategies. Both attack strategies append some bytes at the end of an existing malware binary in such a way that those bytes are never executed, while the TLSH value of the modified sample becomes dissimilar to that of the original malware, and therefore, SIMBIoTA has a good chance of misclassifying it. In the case of these, and similar, append attacks, the original malware binary can be found in the adversarial sample. Hence, in order to detect such an adversarial sample as malware, we need a method that identifies the original malware inside the adversarial sample. PATRIoTA, the method

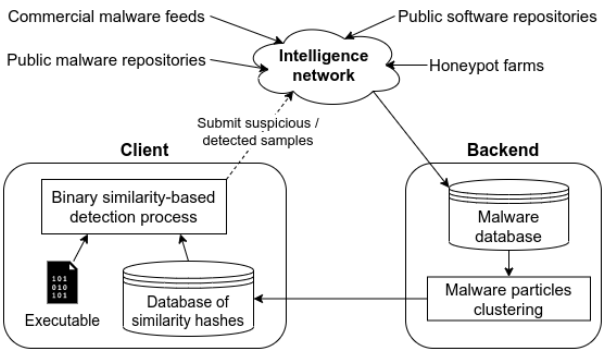

Fig. 2. High-level overview of PATRIoTA.

we propose and describe in this section, will do exactly this: it identifies parts of known malware samples inside any file being checked with it. PATRIoTA can be viewed as a general method of defense against adversarial samples created with append attack strategies.

A. Overview

The design of PATRIoTA was inspired by SIMBIoTA (and their similarity is also reflected in their names). Basically, PATRIoTA is a modified version of SIMBIoTA where the difference is that PATRIoTA works with fixed size parts of malware samples instead of entire malware binaries. Not surprisingly, the architecture of PATRIoTA is also almost the same as that of SIMBIoTA, as it is illustrated in Figure 2. Malware samples are continuously collected from the intelligence network of the antivirus provider, and the PATRIoTA backend splits them into fixed-size parts, which we call *particles* in the sequel. Similar to SIMBIoTA, a similarity graph is built by the backend, but in this case, this graph is built from the malware particles. In addition, PATRIoTA uses a different similarity threshold to build the similarity graph. In Subsection III-D, we explain how to determine the optimal values for the particle size and the similarity threshold used by PATRIoTA. Again similarly to SIMBIoTA, the backend computes a dominating set of the current similarity graph and makes the list of TLSH hash values of the dominating vertices available to clients.

The detection method on the client side is somewhat different in the case of PATRIoTA: the client splits the file to be checked into particles (of the same size used by the backend); calculates the TLSH hashes of the particles; and compares these TLSH hashes with those of the current dominating set. A file is considered malicious if it contains a threshold number of particles that are similar to known malicious particles. The selection of this threshold is discussed in Subsection III-C.

B. Particle size and similarity threshold

PATRIoTA uses some special parameters, which we have already mentioned in the previous subsection, including the size of the particles and the similarity threshold used during the graph construction from the TLSH hashes. Finding the optimal configuration of these parameters is not a trivial task. We can state that the optimal configuration (if it exists at all) is highly context dependent; for example, we can imagine a situation where the low latency of the detection process is more critical than its memory usage.

Despite all this, we developed an iterative methodology to determine a recommended parameter configuration. We performed measurements to determine the optimal parameters on a smaller data set, not the one presented in Subsection IV-A. This data set consisted of 2000 malware and 2000 benign samples for both the ARM and the MIPS architectures.

When we were designing PATRIoTA, the first question was the size of the particles. We first considered the values of 1 kB, 2 kB, 4 kB, 8 kB, 12 kB and 16 kB, but later excluded 1 kB and 2 kB, because the number of graph nodes built from particles of those sizes grew unmanageably large.

PATRIoTA builds a graph from the TLSH values of malware particles, where the TLSH hash values are the nodes and there is an edge between two nodes if the TLSH difference of the two hash belonging to the nodes is below a certain similarity threshold value. SIMBIoTA uses 40 as TLSH similarity threshold, because the average clustering coefficient of the built graph is the highest in that case [9]. The same value cannot be used for PATRIoTA, because it does not build the graph from the TLSH hashes of entire malware samples, but from its particles. To determine the optimal value of the TLSH similarity threshold for different particle sizes, we used the same technique as for SIMBIoTA. In Figure 3, we measure the average clustering coefficient of the graph built from the particles of the 2000 malware samples using different TLSH similarity thresholds. We select the TLSH similarity threshold that gives the highest average clustering coefficient for each particle sizes (e.g., for particles of size 4 kB, the selected TLSH similarity threshold is 65 in the case of ARM samples).

C. Detection threshold

A suspicious sample is considered malicious if it contains at least a threshold number of particles that are similar to known malware particles. If this threshold is set to 1, the true positive detection rate (TPR) of malware will be as high as possible, but the unwanted effect may occur that even benign files are considered malicious (e.g. a benign and a malicious program use the same statically compiled library, therefore, both contain the same sequence of bytes). The consequence is that the larger the detection threshold is, the lower the false positive rate (FPR) and, unfortunately, the lower the TPR will be. So, we choose the smallest possible value where the FPR is still below 1%, which is 2 in the case of ARM samples and 4 in the case of MIPS samples (see Section IV).

D. Optimal configurations

At this point we have 4 possible particle size and TLSH similarity threshold configurations, but which of them is the best?

Before answering this question let's take a look at Figure 4, where we examine the number of similar particles between malware samples with the configuration of 4k particle size and 65 similarity threshold for ARM samples. For that, we

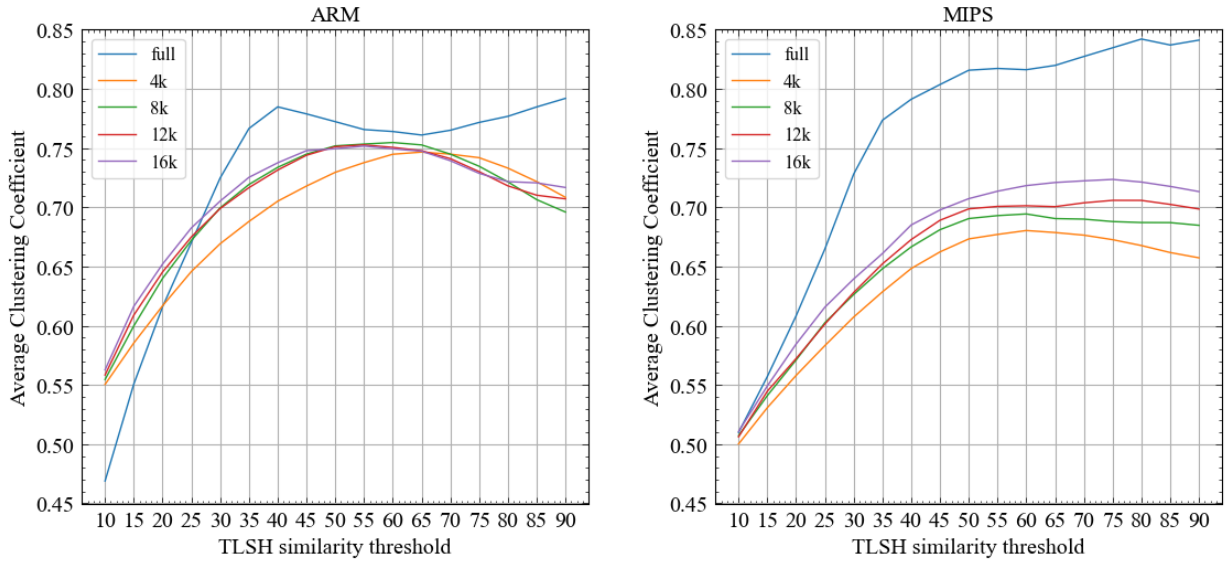

Fig. 3. Average clustering coefficient as a function of the TLSH similarity threshold in the case of ARM (left) and MIPS (right) architectures. Different curves belong to different particle sizes, including the case where the particle size and the file size are equal (full).

divide the 2000 malware samples to 10% train and 90% test set, we split to particles the items in the train set and we build the graph from their TLSH hashes. Finally, we iterate over the elements of the test set, split each file, and count how many similar particles there are between them and the particles in the graph. In other words, we simulate the operation of PATRIoTA on a small data set and repeat this simulation ten times (just like in Section IV, in the case of the large data set). There are ca. 40 samples in Figure 4 that do not contain any similar malware particles to the particles in the train set, therefore, we cannot detect these. Furthermore, in Figure 4 we see how many malware samples would not be detected depending on the selected detection threshold. For instance, if we required that at least 3 particles of the file should be similar to some known malware particle to detect the file as malicious (i.e., detection threshold 3), then ca. 40+50+30=120 malware samples would not be detected.

To select the best particle size and TLSH similarity threshold configuration, we examine how it changes the TPR and FPR values depending on the detection threshold. Figure 5 shows the ROC (Receiver Operating Characteristic) curves of the different configurations, where each jump in the step function corresponds to a certain detection threshold value between 1 and 10. According to our expectations, as the detection threshold increases, the FPR decreases, but so does the TPR. We choose the configuration with the highest AUC (Area Under the ROC Curve) value. This is 4k particle size and 65 TLSH similarity threshold for ARM samples and 8k particle size and 60 TLSH similarity threshold for MIPS samples.

IV. Evaluation

In Section III, we presented PATRIoTA, including its architecture and operating principles, as well as the selection of its parameters (particle size, similarity threshold, and detection threshold), as the main contributions of this paper. It is time to evaluate PATRIoTA's performance, especially its ability to detect adversarial samples. However, before doing that, we present the data set and methodology used for the performance measurements.

A. Experiment design

In this work, we perform all experiments using the same dataset as used for the evaluation of SIMBIoTA [6]. This dataset is called CrySyS-Ukatemi benchmark dataset of IoT malware 2021 (or CUBE-MALIoT-2021 for short). The dataset consists of 29,209 malicious ARM samples and 18,715 malicious MIPS samples, extended with 4,727 benign ARM samples and 9,392 benign MIPS samples. For malicious samples, metadata is also available, which details, among others, the date the sample was first seen in the wild (i.e., submitted to VirusTotal). CUBE-MALIoT-2021 is publicly available[4] for use by the IoT malware research community.

As a first step for testing PATRIoTA, we split the malware samples into a 10% train set and a 90% test set. To do this, we use K-folds cross-validation [10], which is a reliable and frequently used model checking technique. We use K-folds with 10 folds and repeat each measurement 10 times, where the samples of each fold belong to the train set once, and the test set consists of the samples of the other 9 folds. This ensures that each malware appears exactly once in the train set and 9 times in the test set.

PATRIoTA does not need benign samples for training, so we add benign samples only to the test set. Moreover, we extend the test set with adversarial samples for measuring the robustness of the system. These adversarial samples are

[4]https://github.com/CrySyS/cube-maliot-2021 (accessed: June 5, 2023)

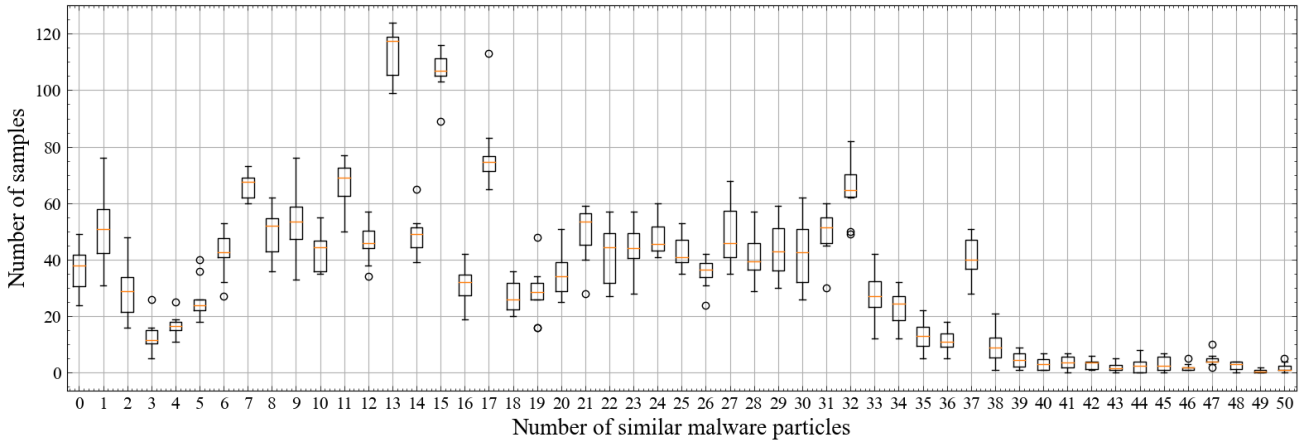

Fig. 4. The x axis shows the number of particles in a sample from the test set that are similar to items in the train set (from 0 to 50), while the y axis shows the number of these samples in the case of 4k particle size and 65 TLSH similarity threshold configuration for ARM samples.

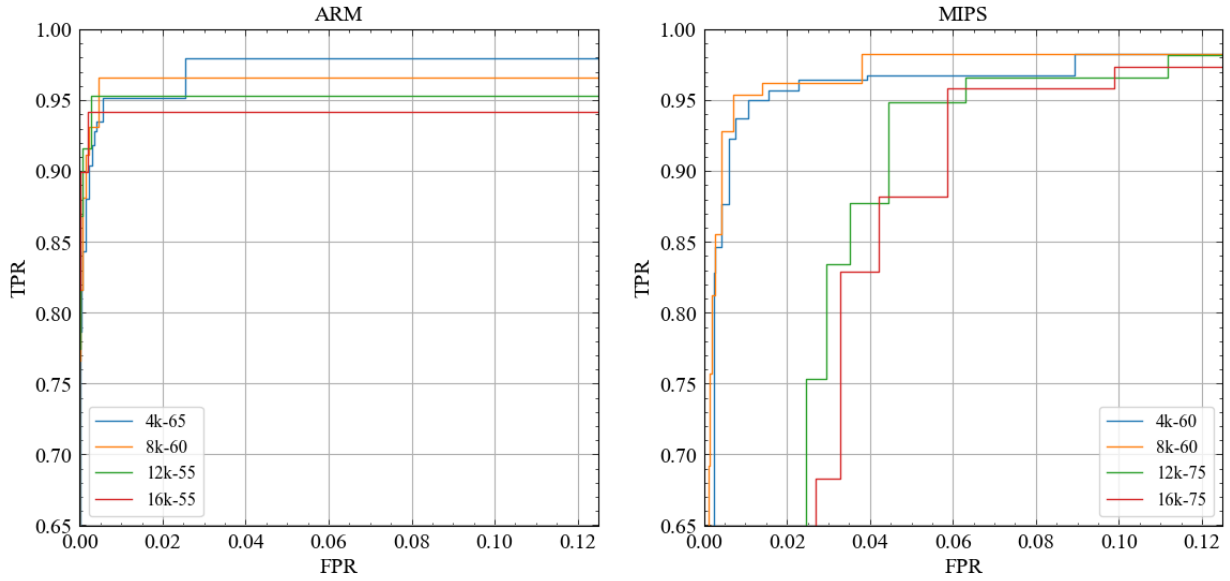

Fig. 5. ROC curves of different configurations, where each jump in the step function corresponds to a certain detection threshold value between 1 and 10 in the ARM and MIPS cases.

created using the two strategies mentioned in Subsection II-B: Chunker and Disguiser. We create these adversarial samples from the malware binaries in the test set, simulating that an attacker has malware samples unknown to the antivirus company and can create adversarial samples from them. Table I shows the exact number of samples in the train and test set.

With the presented construction, we simulate the operation of PATRIoTA: we build the model properly from the train samples, and then give samples from the test set to the model for detection (see Section III). Furthermore, since PATRIoTA was inspired by SIMBIoTA, we compare their performances in all aspects. Indeed, we train and test the two systems on the same samples and measure the same performance metrics. In the next 3 subsections, we present the results of the performed simulation.

TABLE I
NUMBER OF SAMPLES IN THE TRAIN AND TEST SET, IN THE ARM AND MIPS CASES.

ARM

	Malware	Benign	Chunker	Disguiser
Train	2,921	–	–	–
Test	26,288	4,727	24,285	26,288

MIPS

	Malware	Benign	Chunker	Disguiser
Train	1,872	–	–	–
Test	16,843	9,392	13,862 - 13,916	16,813 - 16,819

B. Detection capability

Using the experimental setup presented in the previous subsection, we measure the detection accuracy of SIMBIoTA and PATRIoTA in 3 different cases: on unmodified malware

TABLE II
STORAGE REQUIREMENT OF SIMBIOTA AND PATRIOTA ON THE CLIENT SIDE IN THE ARM AND MIPS CASES.

	SIMBIoTA	PATRIoTA
ARM	8,365 - 9,030 B	333,585 - 371,805 B
MIPS	5,775 - 6,440 B	111,965 - 139,370 B

and benign files, on adversarial samples of the Chunker strategy, and on adversarial samples of the Disguiser strategy. Our goal is to achieve the highest possible accuracy in all 3 cases, while keeping the FPR of benign files below 1%. To do this, we try PATRIoTA with different detection thresholds (i.e., minimum number of particles of a file that need to be similar to known malware particles in order for the file to be classified as malware). The smaller the detection threshold is, the higher the TPR will be, but the FPR will increase too. According to our measurements, the optimal value for the detection threshold is 2 for ARM samples and 4 for MIPS samples, as for smaller values, the FPR exceeds 1%.

In Figure 6, we compare the detection accuracy of the two system. PATRIoTA drastically outperforms SIMBIoTA in terms of accuracy in all test cases! On the sample set of unmodified benign and malicious programs, PATRIoTA has an impressive 98.5% accuracy in the case of ARM samples and 98.2% in the case of MIPS samples. Moreover, it performs extremely well even on adversarial samples of both the Chunker and Disguiser strategies, with 98% accuracy on ARM samples and 95% on MIPS samples.

C. Storage requirement

IoT devices are usually limited by resources, including available memory and storage capacity. Therefore, we measure the storage space requirement of PATRIoTA on the client side (i.e., on the IoT device), which in our case is the size of the dominating set multiplied by the size of the TLSH hash. Compared to SIMBIoTA, unfortunately, the higher accuracy and robustness of PATRIoTA comes with a higher storage requirement, due to the increased number of nodes in the dominating set. In Table II, we present the required memory sizes of the two system.

D. Run time performance

Another price we have to pay for the increased detection accuracy and robustness of PATRIoTA is the increased processing time compared to SIMBIoTA. By detection time, we mean the time that elapses from the beginning of the binary scan of any file to the decision whether it is malicious or not. In the case of SIMBIoTA, this time consists of the TLSH hash computation time of the binary and the decision time of the model. For PATRIoTA, the detection time consists of the sum of 3 components: the time required to split the binary into fixed-size particles, the sum of TLSH hash computation time of the particles, and the sum of decision times required for particles. Basically, PATRIoTA performs SIMBIoTA's detection method multiple times, more precisely for each particle, until the number of particles considered malicious reaches the value of the detection threshold parameter. Therefore, PATRIoTA requires more time for detection than SIMBIoTA, as shown in Figure 7.

V. DISCUSSION

In this work, we were concerned with increasing the robustness of binary similarity-based malware detection methods against adversarial samples that are crafted specifically to mislead a given malware detector. More specifically, we proposed PATRIoTA, a robust, similarity-based antivirus solution, which was inspired by SIMBIoTA [6]. It turns out that SIMBIoTA has a machine-learning based variant, called SIMBIoTA-ML, proposed in [11], and the robustness of SIMBIoTA-ML has already been studied in [12], where an adversarial training approach was proposed as a solution. So a natural question, at this point, is whether an adversarial training approach could have increased the robustness of SIMBIoTA as well. If so, then the need for a new approach, i.e., our PATRIoTA, would be much weaker.

It is clear what adversarial training means in case of a machine learning-based method: the training set is expanded with adversarial samples created by various known adversarial strategies. But SIMBIoTA is not a machine learning-based method. Nevertheless, we can define adversarial training quite intuitively for SIMBIoTA too: the antivirus provider extends the similarity graph of known malware samples with adversarial samples and computes the dominating set of this extended graph. One can then check the detection performance of this modified SIMBIoTA on adversarial samples to determine how robust this approach is.

We performed adversarial training of SIMBIoTA by extending the similarity graph of known malware samples with adversarial samples created from those known malware by the Chunker and Disguiser strategies introduced in [12], and computed the dominating set of the extended graph. We then measured the detection performance on adversarial samples created from malware unknown to the antivirus provider by the same Chunker and Disguiser strategies. The results we got were not so promising: adversarial ARM and MIPS samples created by the Chunker strategy were detected with 92% and 90% accuracy, respectively, while adversarial ARM and MIPS samples created by the Disguiser strategy were detected only with 86% and 67% accuracy, respectively. These results confirm the raison d'être of PATRIoTA.

In addition, while PATRIoTA was designed to be robust against adversarial samples that were created from existing malware samples by appending extra bytes to them, we have the intuition that it is also robust against other strategies that create adversarial samples that contain chunks of the original sample, as those chunks may result in particles that are similar to the particles of the original sample. In order to test this intuition, we measured the robustness of PATRIoTA against such a strategy. In particular, a very clever adversarial sample creation strategy against similarity-based malware detection was proposed in [13] that consists in modifying a few unused portions of a malware binary (e.g., the section header tables

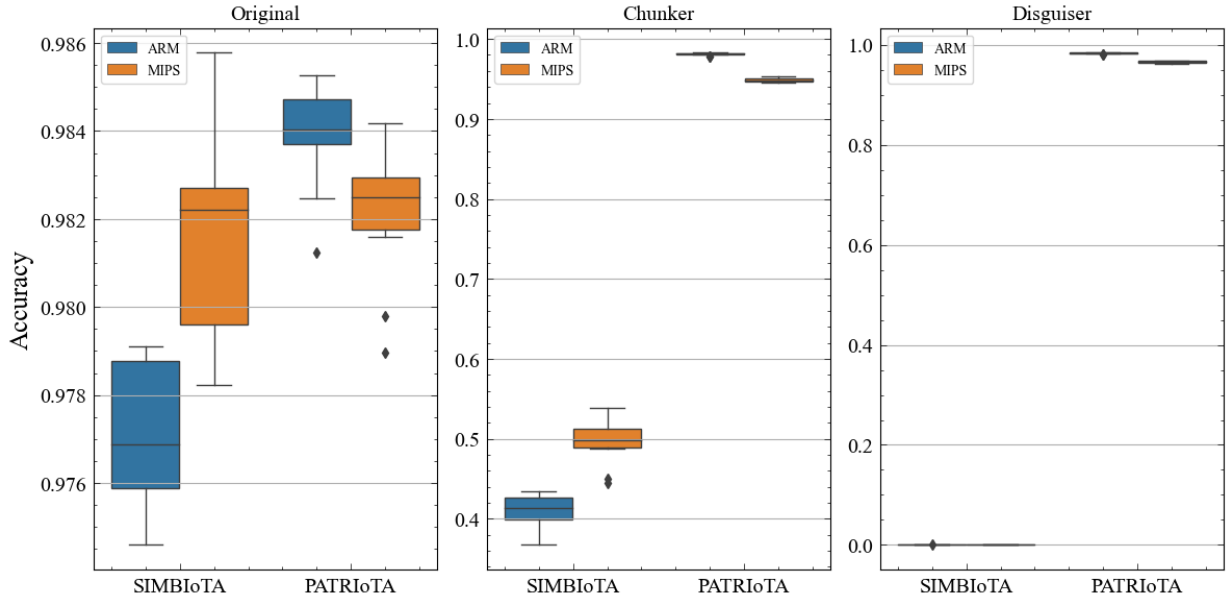

Fig. 6. Comparison of the detection accuracy of SIMBIoTA and PATRIoTA on unmodified malicious and benign samples (Original), adversarial samples created with the Chunker strategy, and adversarial samples created with the Disguiser strategy, in the ARM and MIPS cases.

were modified in [13]) such that the TLSH difference between the modified and the original files is maximized, while the functionality of the original binary is fully preserved, the size of the modified file remains the same as that of the original one, and even the binary content is only slightly changed. As reported in [13], it is rather easy to create adversarial samples in this way that are misclassified by SIMBIoTA: out of 2000 randomly chosen ARM malware samples, 1779 samples were suitable for such kind of modification, and 1465 samples could be created with a TLSH difference of at least 40 (the threshold used by SIMBIoTA) between the modified and original files. We tested both SIMBIoTA and PATRIoTA with those samples, and SIMBIoTA recognized only 17% of them as malware, while PATRIoTA detected a remarkable 98% of them as malware!

One may wonder whether statically linked libraries decrease the detection accuracy of PATRIoTA. Such libraries may be included in both malware and benign binaries, so actually, some portions of statically linked malware and benign samples that include the same libraries can be identical. This may lead to multiple similar particles in them, potentially above the threshold number used by PATRIoTA. In other words, benign files may contain particles resulting from linked libraries that are similar to particles seen in malware binaries using the same libraries. Such benign files may be classified as malware by PATRIoTA, which leads to an increased false positive rate. Indeed, we tested PATRIoTA on 118 statically linked ARM and 64 statically linked MIPS benign binaries and it misclassified 15% and 6% of them, respectively, as malware. This misclassification rate is not really acceptable, therefore, further research is needed to reduce it. We note that SIMBIoTA had a false positive detection rate of 0% throughout our experiments.

VI. Related Work

Although SIMBIoTA, and thus PATRIoTA as well does not belong to ML-based malware detectors, adversarial examples and adversarial robustness, the main focus of this work are closely tied to machine learning, therefore we include an outlook on ML-based malware detection in this section.

ML-based malware detection solutions, unlike their traditional counterparts, are highly automated [14], thus they can keep pace with the increasing amount of malware. They use static, dynamic or hybrid program analysis techniques to extract information from samples which they use to construct feature vectors [15].

Statically obtained features could include opcode-based solutions, where the samples' instructions [16] are used to construct such a vector, gray scale images created from binaries [17] or solutions built to perform detection based on the control-flow graph of samples [18]. Dynamic analysis-based solutions can rely on API or system call traces [19] or network traffic, observed during the execution of the sample [20].

Solutions, where ML-based and cloud-based approaches are combined, can be highly advantageous in the IoT domain, since these solutions can relieve resource constrained devices from performing computationally heavy tasks [21], thus more complicated models can be used as well, like convolutional neural networks [22] and recurrent neural networks, while they can also boost the performance of more light-weight models, like random forests [16] and fuzzy pattern trees [23].

Adversarial attacks against malware detectors can also use multiple approaches [24]; here we highlight two of these, called *append* and *slack* attacks [25]. Append, as the name suggests, works by adding extra bytes to the end of samples,

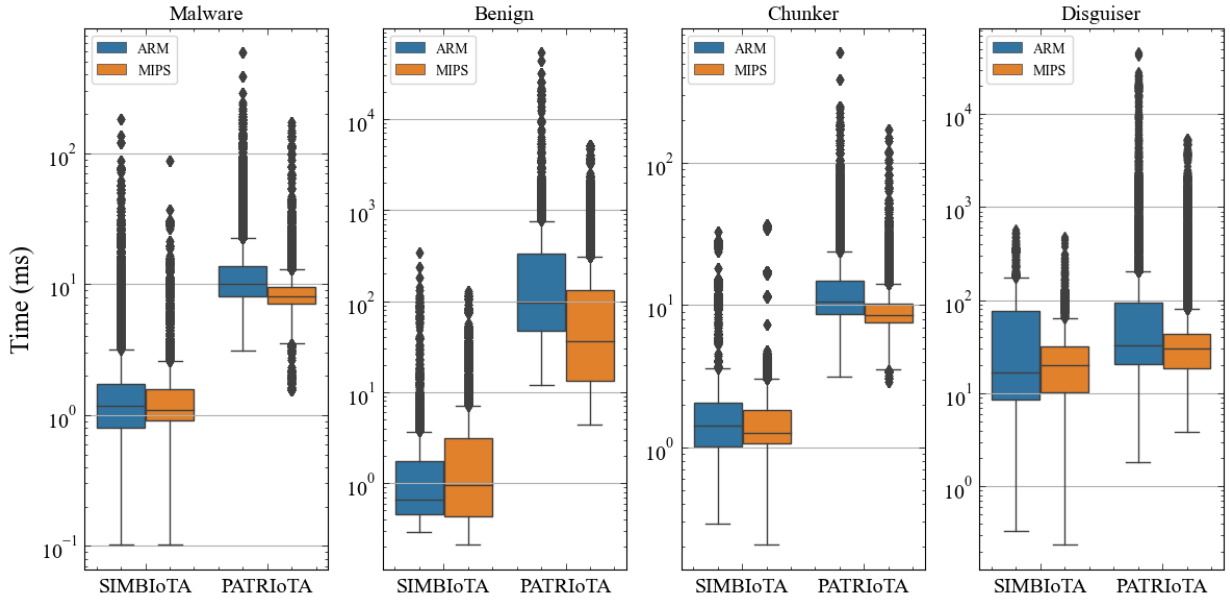

Fig. 7. Comparison of the required detection times (on logarithmic y-axis scale) of SIMBIoTA and PATRIoTA, separately for malware samples, benign samples, adversarial samples of Chunker strategy, and adversarial samples of Disguiser strategy, in the ARM and MIPS cases.

that will never be executed, thus they don't have any affect on the functionality of the modified sample. The strategies Chunker and Disguiser implement this approach. Slack attacks, on the other hand modify the content of so-called slack spaces in binary files. These are regions that contain no useful data, and usually exists because of alignment-related reasons: for example, the size of a section cannot be divided by the page size, but the next section's beginning must be aligned to another page, thus creating a slack space between the end of the section and the beginning of the next one. A special case of this approach is the strategy where the section header table is overwritten; as it is not required for loading and executing ELF files, it can be categorized as slack space. To use these strategies to fool the ML model, solutions like a gradient-based approach can be used [26], or the feature extraction process can be attacked as well [27].

More advanced techniques, like program obfuscation can be used as well, to change the binary file while preserving its original functionality; to do so, one could use reinforcement learning [28], like Recurrent Neural Networks (RNNs) or Generative Adversarial Networks (GANs) [29].

Increasing the adversarial robustness of ML-based malware detectors is a logical next step in the decades old arms race between malware developers and antivirus vendors. One such attempt was to improve SIMBIoTA-ML by applying adversarial training [12], meaning that the training set was extended with adversarial examples. Our solution aims to achieve the same goal (i.e. increasing adversarial robustness), but using another approach. This method works on SIMBIoTA as well, which wasn't suitable for adversarial training. We also believe that this approach is superior, since it does not require adversarial examples and only employs more general assumptions regarding the strategy of the attacker.

VII. Conclusion

In this paper, we proposed PATRIoTA, a similarity-based IoT malware detection method, and showed that it has outstanding malware detection capabilities, while being robust against various adversarial sample creation strategies too. More specifically, we compared the performance and robustness of PATRIoTA to those of SIMBIoTA, an IoT malware detection method in a similar vein as PATRIoTA. PATRIoTA has a higher true positive detection rate, but it also has a higher false positive rate, it requires more storage capacity, and it has longer detection time than SIMBIoTA has. Its true advantage is its strong robustness against adversarial samples: indeed, SIMBIoTA can be completely misled by adversarial samples created from existing malware by appending extra bytes to them, whereas PATRIoTA detects those samples with very high accuracy. In addition, PATRIoTA proved to be robust against another adversarial sample creation strategy too that produces adversarial samples by modifying unused portions of an existing malware binary such that the modified binary becomes dissimilar to the original one, and hence, likely misclassified by SIMBIoTA.

We argued that edge gateways are well-positioned for performing malware detection and protecting resource constrained IoT devices behind such gateways. They can identify the transfer of executable files in the network traffic, check those executables with a malware detection mechanism, and block any traffic carrying malware. PATRIoTA can be used for such malware detection on edge gateways. Although it has a larger storage requirement than SIMBIoTA has, edge gateways also offer more storage space than typical IoT field devices do. PATRIoTA also has an increased running time, in particular

when checking benign files. We believe that this does not hinder the use of PATRIoTA on edge gateways, but this requires further study and more measurements.

In Section V, we discussed that PATRIoTA may make false positive decisions on statically linked benign binaries if they include libraries that have also been used in malware. This issue also needs further study and it is on our future research agenda.

REFERENCES

[1] S. Greengard, "Deep insecurities: The Internet of Things shifts technology risk," *Communications of the ACM*, May 2019.

[2] O. Alrawi, C. Lever, M. Antonakakis, and F. Monrose, "SoK: Security evaluation of home-based IoT deployments," in *IEEE Symposium on Security and Privacy*, 2019.

[3] M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman, L. Invernizzi, M. Kallitsis, D. Kumar, C. Lever, Z. Ma, J. Mason, D. Menscher, C. Seaman, N. Sullivan, K. Thomas, and Y. Zhou, "Understanding the mirai botnet," in *26th USENIX Security Symposium (USENIX Security 17)*. Vancouver, BC: USENIX Association, Aug. 2017, pp. 1093–1110.

[4] C. Miller and C. Valasek, "Remote exploitation of an unaltered passenger vehicle," in *Black Hat USA*, 2015.

[5] Y. J. Kim, C.-H. Park, and M. Yoon, "Film: Filtering and machine learning for malware detection in edge computing," *Sensors*, vol. 22, no. 6, 2022.

[6] C. Tamás, D. Papp, and L. Buttyán, "SIMBIoTA: Similarity-based malware detection on IoT devices," in *Proceedings of the 6th International Conference on Internet of Things, Big Data and Security (IoTBDS)*. SCITEPRESS, 2021, pp. 58–69.

[7] J. Sándor, "Robustness Against Evasion of Similarity-based IoT Malware Detection Methods," Budapest University of Technology and Economics, Scientific Students' Association Report, 2022.

[8] J. Oliver, C. Cheng, and Y. Chen, "Tlsh – a locality sensitive hash," in *2013 Fourth Cybercrime and Trustworthy Computing Workshop*, 2013, pp. 7–13.

[9] L. Buttyán, R. Nagy, and D. Papp, "SIMBIoTA++: Improved Similarity-based IoT Malware Detection," in *2022 IEEE 2nd Conference on Information Technology and Data Science (CITDS)*. IEEE, 2022, pp. 51–56.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[11] D. Papp, G. Ács, R. Nagy, and L. Buttyán, "SIMBIoTA-ML: Lightweight, machine learning-based malware detection for embedded IoT devices," in *Proceedings of the 7th International Conference on Internet of Things, Big Data and Security (IoTBDS)*. SCITEPRESS, 2022, pp. 55–66.

[12] J. Sándor, R. Nagy, and L. Buttyán, "Increasing the robustness of a machine learning-based IoT malware detection method with adversarial training," in *Proceedings of the 2023 ACM Workshop on Wireless Security and Machine Learning*, 2023.

[13] G. Fuchs, "Adversarial example creation strategies for defeating similarity-based IoT malware detection algorithms," BSc Thesis, Budapest University of Technology and Economics, 2023.

[14] D. Ucci, L. Aniello, and R. Baldoni, "Survey of machine learning techniques for malware analysis," *Computers & Security*, vol. 81, pp. 123–147, 2019.

[15] S. W. Soliman, M. A. Sobh, and A. M. Bahaa-Eldin, "Taxonomy of malware analysis in the IoT," in *2017 12th International Conference on Computer Engineering and Systems (ICCES)*, 2017, pp. 519–529.

[16] H. Takase, R. Kobayashi, M. Kato, and R. Ohmura, "A prototype implementation and evaluation of the malware detection mechanism for IoT devices using the processor information," *International Journal of Information Security*, vol. 19, pp. 71–81, 2019.

[17] L. Nataraj, S. Karthikeyan, G. Jacob, and B. S. Manjunath, "Malware images: visualization and automatic classification," in *Proceedings of the 8th international symposium on visualization for cyber security*, 2011, pp. 1–7.

[18] J. Sahs and L. Khan, "A machine learning approach to android malware detection," in *2012 European intelligence and security informatics conference*. IEEE, 2012, pp. 141–147.

[19] M. F. B. Abbas and T. Srikanthan, "Low-Complexity Signature-Based Malware Detection for IoT Devices," in *Applications and Techniques in Information Security*, L. Batten, D. S. Kim, X. Zhang, and G. Li, Eds. Singapore: Springer Singapore, 2017, pp. 181–189.

[20] Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, A. Shabtai, D. Breitenbacher, and Y. Elovici, "N-BaIoT—Network-Based Detection of IoT Botnet Attacks Using Deep Autoencoders," *IEEE Pervasive Computing*, vol. 17, pp. 12–22, 07 2018.

[21] H. Sun, X. Wang, R. Buyya, and J. Su, "CloudEyes: Cloud-Based Malware Detection with Reversible Sketch for Resource-Constrained Internet of Things IoT Devices," *Softw. Pract. Exper.*, vol. 47, no. 3, p. 421–441, mar 2017.

[22] J. Su, D. V. Vasconcellos, S. Prasad, D. Sgandurra, Y. Feng, and K. Sakurai, "Lightweight Classification of IoT Malware Based on Image Recognition," in *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, vol. 02, 2018, pp. 664–669.

[23] E. M. Dovom, A. Azmoodeh, A. Dehghantanha, D. E. Newton, R. M. Parizi, and H. Karimipour, "Fuzzy pattern tree for edge malware detection and categorization in IoT," *Journal of Systems Architecture*, vol. 97, pp. 1–7, 2019.

[24] K. Aryal, M. Gupta, and M. Abdelsalam, "A Survey on Adversarial Attacks for Malware Analysis," *ArXiv*, vol. abs/2111.08223, 2021.

[25] O. Suciu, S. E. Coull, and J. Johns, "Exploring Adversarial Examples in Malware Detection," *2019 IEEE Security and Privacy Workshops (SPW)*, pp. 8–14, 2019.

[26] B. Kolosnjaji, A. Demontis, B. Biggio, D. Maiorca, G. Giacinto, C. Eckert, and F. Roli, "Adversarial Malware Binaries: Evading Deep Learning for Malware Detection in Executables," in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 533–537.

[27] G. Fuchs, "Adversarial example creation strategies for defeating similarity-based IoT malware detection algorithms," Bachelor's Thesis, Budapest University of Technology and Economics, 2022.

[28] H. Anderson, A. Kharkar, B. Filar, D. Evans, and P. Roth, "Learning to Evade Static PE Machine Learning Malware Models via Reinforcement Learning," *ArXiv*, vol. abs/1801.08917, 2018.

[29] W. Hu and Y. Tan, "Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN," *CoRR*, vol. abs/1702.05983, 2017.