Soteria: Detecting Adversarial Examples in Control Flow Graph-based Malware Classifiers

Hisham Alasmary†‡*, Ahmed Abusnaina†*, Rhongho Jang‡*, Mohammed Abuhmad†, Afsah Anwar†, DaeHun Nyang‡, and David Mohaisen†

†University of Central Florida  ‡King Khalid University  ⋆Ewha Womans University
hisham, ahmed.abusnaina, r.h.jang, abuhmad, afsahanwar @knights.ucf.edu; nyang@ewha.ac.kr; mohaisen@ucf.edu

Abstract—Deep learning algorithms have been widely used for security applications, including malware detection and classification. Recent results have shown that those algorithms are vulnerable to adversarial examples, whereby a small perturbation in the input sample may result in misclassification. In this paper, we systematically tackle the problem of adversarial examples detection in the control flow graph (CFG) based classifiers for malware detection using Soteria. Unique to Soteria, we use both density-based and level-based labels for CFG labeling to yield a consistent representation, a random walk-based traversal approach for feature extraction, and n-gram based module for feature representation. End-to-end, Soteria’s representation ensures a simple yet powerful randomization property of the used classification features, making it difficult even for a powerful adversary to launch a successful attack. Soteria also employs a deep learning approach, consisting of an auto-encoder for detecting adversarial examples, and a CNN architecture for detecting and classifying malware samples. We evaluate the performance of Soteria, using a large dataset consisting of 16,814 IoT samples, and demonstrate its superiority in comparison with state-of-the-art approaches. In particular, Soteria yields an accuracy rate of 97.79% for detecting AEs, and 99.91% overall accuracy for classification malware families.

Index Terms—Internet of Things; Adversarial Machine Learning; Malware Detection; Deep Learning

I. INTRODUCTION

The rising acceptance of IoT devices for different industrial and personal applications has been paralleled with a proportionally increase to their susceptibility to attacks. A major reason for their susceptibility to take is their use of vulnerable or insecure functions and services. As such, adversaries exploit these vulnerable services to deliver malware to launch orchestrated attacks. This makes malware detection an important issue. To address this issue, several prior works have leveraged different machine- and deep-learning algorithms for malware detection atop program analysis [1], [2], [3], [4], [5], [6]. Program analysis approaches utilized for malware include both static and dynamic analyses. Dynamic analysis requires executing malware for obtaining behavior features that are fed into machine learning algorithms for detection. Although the dynamic features are comprehensive, dynamic analysis techniques are subject to various shortcomings, and require executing malware for obtaining behavior features. As such, adversaries exploiting these vulnerable services to deliver malware and to launch orchestrated attacks. This makes malware detection an important issue. To address this issue, several prior works have leveraged different machine- and deep-learning algorithms for malware detection atop program analysis [1], [2], [3], [4], [5], [6]. Program analysis approaches utilized for malware include both static and dynamic analyses. Dynamic analysis requires executing malware for obtaining behavior features that are fed into machine learning algorithms for detection. Although the dynamic features are comprehensive, dynamic analysis techniques are subject to various shortcomings, and most importantly their complexity and time consumption, resulting in poor scalability. Static analysis, on the other hand, does not require running programs, but relies on programs contents, obtained from the static binary. A popular static analysis technique is using Control Flow Graph (CFG) to build a representative feature modality for malware detection, and is shown to be effective in various studies [7], [8]. Machine learning algorithms are typically implemented atop of the features extracted from the static (and dynamic) analysis techniques. However, these algorithms are susceptible to adversarial attacks, thereby circumventing such detection systems [9]. Therefore, it is essential to detect such attacks. To this end, this work proposes Soteria, a system to defeat the adversarial example attacks on CFG based classifiers for malware detection.

Given that ML models’ output depends on the input patterns, ML models can be prone to targeted attacks on their inputs. Particularly, an adversary may fool the models by applying perturbations to the input to generate adversarial examples (AEs) that have similar characteristics with the original sample. As such, recent works have examined the robustness of the machine learning models in general, and have demonstrated the generation of AEs using methods such as the fast gradient sign method [10], generative adversarial networks [11], the DeepFool method [12], and the graph-based adversarial learning [9], among others. Nevertheless, there have been several attempts to defend against the adversarial attacks on the machine learning models by including the adversarial examples in the training process [13]. Although prior works have shown the inefficiency of the malware detection models when subjected to adversarial examples, to the best of our knowledge, there is no work on defending such models from adversarial attacks. Identifying the research gap, with this work, we inch closer towards bridging the gap.

Adversarial attacks on malware detectors have recently been conducted by Abusnaina et al. [9], Kolosnjaji et al. [14], and Kreuk et al. [15]. While Abusnaina et al. [9] show the susceptibility of the CFG-based detectors, the other works append bytes to the binary file. Both of these methods change the files while preserving the practicality and functionality of the clean IoT malware. Additionally, the AE creation of malware is limited due to the risk of un-executability. Acknowledging the importance of having effective defense to detect adversarial examples, Soteria utilizes features from...
the CFG to detect them. Particularly, Soteria consists of two major components, the AEs detector and the IoT malware classifier. Soteria starts by labeling the CFG nodes based on two approaches: density-based labeling and level-based labeling. Then, Soteria applies a set of random walks, with a length proportional to the number of nodes in the CFG, on every labeling approach to deeply express and represent the behaviors of the software processes manifested in the CFG. The nodes making up the random walks are then used as the features for the operation of Soteria. In the first phase, the detection system that uses the deep features from the CFG to detect the AEs, thereby stopping their access to the malware classifier with an accuracy of 97.79%. In the next phase, with a flexibility to re-use the feature-set from the detection phase, it classifies the input file as benign or assigns an appropriate family label to the malware with an accuracy of 99.91%.

Contributions. In this paper, we make two contributions:

1. Motivated by the recent body of work on developing adversarial examples on machine learning-based malware detection algorithms, we propose the design and implementation of Soteria, a system for detecting IoT malware. Similar to other efforts in this space, Soteria utilizes Control Flow Graph (CFG) based feature representations. Unique to Soteria, we use both density-based and level-based labels for CFG labeling, a random walk-based traversal approach for feature extraction, and n-gram based module for feature representation. End-to-end, Soteria’s representation ensures a simple yet powerful randomization property of the used classification features, making it difficult even for a powerful adversary to launch a successful attack. Soteria also employs a deep learning approach, consisting of an auto-encoder for detecting AEs, and eliminating them from the classification process, and a CNN architecture for detecting and classifying malware samples.

2. We evaluate the performance of Soteria, using a large dataset consisting of 16,814 IoT samples, and demonstrate its superiority in comparison with state-of-the-art approaches. Soteria yields an accuracy rate of 97.79% for detecting AEs, and 99.91% overall accuracy for classification malware families.

Organization. The rest of this paper is organized as follows.

We introduce our motivation in section II, including practical adversarial example manipulation, limitation of adversarial learning, Graph Embedding, and Augmentation approach, and the threat model. We describe the system design in section III. We analyze and evaluate Soteria in section IV and discuss the results in section V. We review the literature in section VI, and conclude our study in section VII.

II. BACKGROUND AND MOTIVATION

Adversarial examples (AEs) can be generated by slightly manipulating a sample to fool the classifier, and done in the context of malware on either the binary or the code level. 1 Binary-level AEs the generation of such AEs entails manipulating the bytes of the malware sample upon compilation, without any regard to the function and purpose of such bytes, as has been done in several works [16], [17], [18]. Another method for binary-level AEs generation would entail injecting a benign block of bytes into an unreachable part of the malware binary, e.g., by adding a new section or appending the benign bytes to the end of malicious code, thus altering the feature representation introduced by the AE. 2 Code-level AEs the generation of those AEs entails applying perturbation over the original code by either modifying the structure of the code or inserting an external code into it. For instance, augmenting or splitting functions results in a structure modification, thus altering the resulting feature space representation of the sample (e.g., CFG-based).

A. Practical Adversarial Examples

For adversarial attacks against machine learning-based malware detection models to be practical, adversaries must ensure the AEs resulting from the manipulation of a malware sample should still be executable (undamaged), making many of the algorithms proposed in the literature for AEs generation impractical for the malware detection domain. To this end, AEs can be categorized into impractical and practical AEs.

Impractical Adversarial Example. An AE is impractical if the injected code is compiled as an unused function during the compilation process. In the binary level, a sample that manipulated by any form of byte injection (e.g., adding a new section or appending at the end of file) is not considered as the practical adversarial example.

Practical Adversarial Example. A practical AE is a mixture of the benign and/or malicious functions where the manipulated components are reachable (part of the code flow) and executable (do not damage the code).

Both code- and binary-level approaches can be used for generating practical AEs, although binary-level approaches are difficult to apply for fine-grained perturbations. A recent study [9] showed that adding external code to the original one leads to a high misclassification rate of the model’s outputs while preserving the functionality of the original code. Such behavior can be critical as it results in changing the source code, execution flow, signature, and binaries, which reduces the performance of state-of-the-art classifiers. In this study, we focus on the injection of external code as a capability for creating AEs, since such an approach affects various representations of the original samples (See section II-D).

B. Limitation of Adversarial Learning

Adversarial training is a defense to increase the learning model’s robustness by training over clean and adversarial datasets. Per Table I, this technique is used for enhancing the robustness of image classifiers by perturbing the training data.

Drawbacks. A large number of studies on adversarial learning were implemented to generate AEs by perturbing the feature space, typically an image. Training a model over AEs generated by one method may not increase its robustness against other methods. This highlights the problem of adversarial learning, training against a set of methods does not guarantee the robustness against different attacks. This problem becomes critical with the existence of code-level manipulation. Where an adversary can change the outcome of the attack by changing...
TABLE I: Current adversarial attacks defenses. The focus of the adversarial attacks defenses is on adversarial examples in the context of image classification. Note: MLP is Multilayer Perceptron, DNN is Deep Neural Network, and RNN is Recurrent Neural Network.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Model</th>
<th>Dataset</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goodfellow et al. [19]</td>
<td>MLP</td>
<td>130,000</td>
<td>Image</td>
</tr>
<tr>
<td>Xu et al. [20]</td>
<td>DNN</td>
<td>6,000</td>
<td>Image</td>
</tr>
<tr>
<td>Meng et al. [21]</td>
<td>DNN</td>
<td>121,000</td>
<td>Image</td>
</tr>
<tr>
<td>Liao et al. [22]</td>
<td>DNN</td>
<td>280,000</td>
<td>Image</td>
</tr>
<tr>
<td>Dhillon et al. [23]</td>
<td>DNN</td>
<td>60,000</td>
<td>Image</td>
</tr>
<tr>
<td>Papernot et al. [24]</td>
<td>DNN</td>
<td>130,000</td>
<td>Image</td>
</tr>
<tr>
<td>Samangouei et al. [25]</td>
<td>DNN</td>
<td>70,000</td>
<td>Image</td>
</tr>
<tr>
<td>Miyato et al. [26]</td>
<td>RNN</td>
<td>805,753</td>
<td>Text</td>
</tr>
</tbody>
</table>

Fig. 1: GEA adversarial approach. The CFG in Fig. 1(c) is obtained by embedding the CFG of a selected sample in Fig. 1(b) into the CFG of the original sample in Fig. 1(a). It can be done by injecting code directly.

Fig. 2: Architecture of Soteria. IoT samples are fed to the feature extraction process, where each sample is represented by multiple feature vectors. The feature vectors are forwarded to adversarial example detector. All non-AEs are then forwarded to the classifier to be classified into its corresponding family.

Fig. 3: Soteria feature extraction process. IoT samples binaries are disassembled to extract their corresponding CFGs. Then, two nodes labeling techniques are used (Dense-based and level-based), then, several random walks are done over each labeled graph. The trace of the random walk is then used for feature extraction by using n-grams with TF-IDF.

C. Graph Embedding and Augmentation

Graph Embedding and Augmentation (GEA) is an approach [9] that is shown to produce AEs that are executable, while allowing targeted adversarial attacks, addressing limitations of the literature. GEA inserts benign code into a target malware sample to generate an AE, with different feature representations. GEA applies direct modifications to the CFG of the sample, affecting the extracted features and resulting in a misclassification, both targeted and non-targeted.

Generating AEs with GEA. The process of generating adversarial examples is done by merging the code of the original sample with the code of a targeted sample. The targeted sample belongs to the class which the adversary desires to misclassify to. Fig. 1 shows the generated CFGs of each sample. Fig. 1(a) is the generated CFG of the original sample, whereas Fig. 1(b) represents the CFG of the embedded code. Combining the original and external code results in Fig. 1(c). The combination is done by creating a shared entry and exit blocks, where only one branch is executed. In this case, the left branch that belongs to the original sample will be executed. Note that changing in the structure of the code will change the extracted features, resulting in misclassification.

D. Threat Model

The adversary’s goal is to fool the classifier by misclassifying malicious samples as benign and vice versa, while preserving the functionality and practicality of the original sample. Our threat model focuses on AEs generated based on code-level manipulation using GEA (in section II-C).

We assume that the adversary has full access to the source code of benign and malicious samples. Moreover, the adversary can edit, compile, and merge samples, and knows the model’s design and its internal architecture. The adversary’s
goal is to conduct targeted and non-targeted misclassification. The objective of Soteria is to provide a robust and accurate classification in the presence of this model.

III. SYSTEM DESIGN

A. High-Level Architecture

To address the impracticability of modification-based adversarial examples, we propose Soteria, a malware classification framework that incorporates two modules: adversarial sample detection and malware classification. Soteria manifests the following advantage. It eliminates the cost of extracting new features, meaning that it can re-use the features generated during the detection of the AEs to classify a sample as benign or malicious. Alternatively, the user has the flexibility over the choice of classifier, meaning that the user can make use of different set of features, classifier parameters, or another classifier altogether. Figure 2 represents the high-level architecture of Soteria, comprising of three major components, feature extractor, AEs detector, and malware classifier.

**Feature Extractor.** Soteria utilizes the features from the graphical representation of a program’s flow execution, i.e., CFG. For a graph \( G \), such that, \( G = (V,E) \), and nodes \((V)\) and edges \((E)\) represent the basic blocks and the traversed paths, respectively. A critical advantage of the CFG is that it summarizes the control flow by connecting the entry block with reachable blocks directly or indirectly. Particularly, if a block of code is appended to an existing program, with an intention to fool the classifiers, knowing that the appended blocks are unreachable, our feature extraction methodology ignores such blocks, in contrast to binary- and image-base classifiers. The features driven from the CFG ignore the non-executable part of samples, eliminating the effect from noise injection and unused functions in the sample.

**AEs Detector.** The detector is a standalone component used prior to the classification process to filter out practical AEs. In this way, we eliminate the model’s vulnerability to AEs by forwarding only legitimate samples (i.e., benign or malicious) to the classifier that was trained using a non-adversarial dataset. Unlike the commonly used approaches in the literature, the detector is trained using only non-adversarial dataset, while maintaining a distinguishable feature representation that enables detecting potential AEs.

**Classifier.** Soteria requires a classifier that can accurately classify the samples into malicious and benign with the resistance towards the impractical AEs. For evaluation, we make use of an ensemble of CNN classifiers, however, it can be replaced with another desired method.

B. Adversarial Examples Detector

The purpose of the detector is to distinguish normal samples from adversarial ones, regardless of whether the sample is malicious or not. Fig. 3 shows the flow of the feature extraction, including sample pre-processing with CFG extraction and labeling, followed by feature extraction using \( n \)-gram of the obtained random walks on the labeled CFG.

![Diagram](image-url)

Fig. 4: Graph labeling using two approaches, density- and level-based. Each node has a label in \([0, |V| - 1]\), where \(|V|\) is the number of nodes in \( G \). Fig. 4(b) and Fig. 4(d) show the labeling of the GEA generated CFG over the original graphs in Fig. 4(a) and Fig. 4(c), respectively.

1) Sample Pre-processing: The pre-processing phase is concerned with nodes labeling. For a graph \( G = (V,E) \), we use two labeling approaches: density-based and level-based.

**Density-based Labeling (DBL).** The density of a node is defined as the summation of in- and out-edges over the total number of edges in the graph. DBL sorts all nodes according to their density, where the most dense node is labeled as 0 and the least dense node is labeled as \(|V| - 1\), and the centrality factor of a node is used to rank nodes with tied density \( CF_{vi} \). If two or more nodes still have the same centrality factor, we assign labels based on their levels, considering the main or entry block function as the entry node. We notice some cases where two nodes with equal values are at the same level (symmetric nodes), and label them in ascending order since switching their labels will not affect the consistency of labeling. Fig. 4(a) shows the result of the density-based labeling. As shown, node 0 and 1 are the most dense nodes because they are connected to four blocks, and node 0 has a higher centrality factor value. The labeling ends by assigning label 10 to the entry block as

\[ CF_{vi} = \frac{1}{t} \sum_{j=1}^{t} \sum_{m=1}^{t} \Delta(v_i) \Delta(m) \]

where \( \Delta(v_i) \) is the count of shortest paths travel through \( v_i \) and connecting nodes \( v_j \) and \( v_k \), for all \( i \neq j \neq k \) and \( \Delta(m) \) is the total number of shortest paths between such nodes. The closeness centrality \( CF_{vi} \) of a node is defined as the average shortest path between node \( v_i \) and all other nodes in the graph \( G \).

---

1Centrality factor of a node is the sum of node’s betweenness and closeness centrality values, \( CF_{vi} = B_{vi} + C_{vi} \). The betweenness centrality \( (B_{vi}) \) of a node \( v_i \) is defined as \( \Delta(v_i) \), where \( \Delta(v_i) \) the number of shortest paths travel through \( v_i \) and connecting nodes \( v_j \) and \( v_k \), for all \( i \neq j \neq k \) and \( \Delta(m) \) is the total number of shortest paths between such nodes. The closeness centrality \( C_{vi} \) of a node is defined as the average shortest path between node \( v_i \) and all other nodes in the graph \( G \).
the least dense node with the lowest centrality factor.

**Level-based Labeling (LBL).** The level of a node $v_i$ is defined by the smallest number of steps $S_{v_i}$ from the entry node to reach $v_i$, where the level of a node is equal to $1 + S_{v_i}$. In LBL, we consider the main or entry block function in the CFG as the first level layer, and follow (in breadth-first search manner) other levels for labeling them. For nodes at the same level, we follow the same labeling mechanism in DBL. Fig. 4(c) shows an example for the result of LBL, where the entry block is assigned with label 0. In the second level, there are two nodes with the same density values, and the centrality factor values are used. The process ends by labeling the last level nodes. Note that the entry block will always have the label 0 when using the LBL method.

Both density- and level-based labeling follow the strict predefined rules to guarantee consistency of representation and ensures that any modification applied to the graph will be reflected in the labels’ assignment. Fig. 4(b) and Fig. 4(d) show the labeling of the generated graphs using GEA. It is worth noting that the labels’ assignment varies for each graph, even when they share a sub-graph. Labels’ assignment over GEA results in changing the labels, and the feature extraction process, hence affecting the detector’s behavior.

2) **Feature Representation:** For feature generation and representation, we apply a random walk and use a method based on the $n$-gram model to approximate the graph.

**Random Walk:** A random walk describes random steps in the graph space, and is used to estimate the graph state space. Let $G$ be an undirected graph with a marker placed at $v_i$, initially the entry block. At each step, the marker moves to an adjacent vertex $v_j$ with probability $\frac{1}{\text{deg}(v_j)}$, where $\text{deg}(v_j)$ is the degree of $v_j$. The marker keeps track of the visited vertices’ labels as it moves, e.g., random walk over the original sample graph in Fig. 4 may generate $W = "10 9 2 1 2 \ldots "$ when using DBL and $W' = "0 2 4 3 4 \ldots "$ when using LBL. We define the length of the random walk as $|W|$ (the number of labeled nodes collected by a random walk of length $|W|$ is $(|W|+1)$. In Soteria, $W = 5 \times |V|$, and repeat the walk ten times over DBL and ten times over LBL, resulting in 20 vectors. The use of random walk helps to randomize the feature extraction process, making it difficult to generate practical AEs. We observed that the repetition of the process improves the quality of the random walks’ feature representation, corresponding to the underlying graph.

**n-grams:** The $n$-gram technique can be used in different models for feature representation of text, documents, graphs, etc. Unique terms or $n$-gram are extracted from the entire corpus before counting the frequencies in individual samples. Inspired by node2vec [27], we use $n$-gram representation of the graphs from the sequences of nodes obtained by the random walk. From the derived random walks with the lengths specified above, we extract $n$-grams of lengths 2, 3, and 4 as a feature representation of the CFG. Given that, the number of $n$-grams is large even for small graphs. We select and use the top 500 discriminative features for each LBL and DBL (thus, 1,000 features in total). The selection of the top discriminative feature is based on the frequency of $W$.

3) **Building Detection Model:** The core of the detection model is an auto-encoder that consists of five fully connected dense layers (as shown in Fig. 5). The auto-encoder reconstructs the given input at the output layer, and consists of four main blocks: an input layer, an output layer, hidden layers, and a validation unit, which are described in the following.

• **Input Layer.** This layer is a one-dimensional vector of size $1 \times 1000$ fed by the density- and level-based features vectors.

• **Hidden Layers.** These layers consist of three fully connected dense layers, and extract a deep representation of the features. The design is based on decoding the features from $1 \times 1000$ to $1 \times 2000$ and $1 \times 3000$. Afterward, a third layer encodes the features presentation to $1 \times 2000$. This structure eliminates the features dependencies in the reconstruction process, as the extracted features are mutually independent.

• **Output Layer.** This layer is fully connected to the third hidden layer. With a shape of $1 \times 1000$, the output layer reconstructs the features seen at the input as its output, which is then returned as a density- and level-based vector.

• **Validation Unit.** The validation unit computes the reconstruction error (RE) by calculating the Root Mean Square Error (RMSE) between the original input $x$ and the reconstructed output $\hat{x}$. If the RMSE exceeds the threshold, set to be 50%, the sample $x$ is labeled as AE.

**C. Classifier**

As the detector distinguishes between adversarial and clean samples, the classifier distinguishes clean samples into benign or one of three malicious families: Gafgyt, Mirai, and Tsunami. For this purpose, two CNN classifiers are utilized to incorporates separately the density- and level-based features.

1) **CNN Classifiers:** The input to the classifier in Soteria is a one dimensional (1D) vector of size $1 \times 500$ representing the density- and level-based extracted features. Fig. 7 shows the structure of the classifier, which consists of three blocks: convolutional blocks (ConvB) 1 and 2 and a classification block (CB). All layers use the Rectified Linear Units (ReLU)
and a dropout with
Similar to ConvB1, except for the number of
ConvB1.
For each sample, we perform ten
as the dropout probability,
1
as the max-pooling size.
ConvB1’s input is the extracted features, and
CB’s input is the flattened feature maps of ConvB2,
498
ConvB2 = 1
3
&EZ
ĞŶƐĞ;ϱϭϮͿ
^ŽĨƚŵĂdž>ĂLJĞƌ
ϭdžϮϰϵ
ƌŽƉŽƵƚƉ;Ϭ͘ϱϬͿ
&WZ
3,342
14.33%
3,016
13,472
66.18%
# Test
% of Samples
262
52
600
To determine the family label
8,911
# Total
16,814
17.94%
1,935
210
1.55%
A VClass
VirusTotal
2,217
2,408
that Gafgyt, Mirai, and Tsunami have 2,217, 473,
# of Samples
is used as the sample’s label (see Fig.
58x166
the majority voting unit, where the class with the highest vote
corresponding CNN classifiers. The final output is based on
DBL and LBL). These feature vectors are forwarded to their
classification decision is the majority vote of the CNN classifiers
output probabilities over the feature vectors.

activation function, and dropout regularization to prevent
model over-fitting. We describe the CNN structure in the
following using the notation $p$ as the dropout probability, $s$
as the stride, $m$ as the max-pooling size.

- ConvB1. ConvB1’s input is the extracted features, and
consists of two consecutive convolutional layers with 46
filters of size $1 \times 3$, that operate convolutions with $s = 1$
with no padding to generate feature maps of size $46 \times 498$.
Each convolutional layer is followed by a max-pooling with
$s = m = 2$ and a dropout with $p = 0.25$.

- ConvB2. Similar to ConvB1, except for the number of
filters. ConvB2 consists of two convolutional layers with 92
filters of size $1 \times 3$, followed by max-pooling and dropout.

- CB. CB’s input is the flattened feature maps of ConvB2,
 fed to a fully connected layer of size $512$ with a dropout
$p = 0.5$. The output of the fully connected layer is fed to a
softmax layer for the classification.

2) Majority Voting: For each sample, we perform ten
random walks and generate 20 feature vectors (from both
DBL and LBL). These feature vectors are forwarded to their
 corresponding CNN classifiers. The final output is based on
the majority voting unit, where the class with the highest vote
is used as the sample’s label (see Fig. 6).

IV. DATASET AND EVALUATION
A. Dataset
To evaluate Soteria, we assembled a dataset of IoT benign
samples and IoT malware. We collected 13,798 malware
samples, randomly selected from CyberIOCs [28] during the
period of January 2018 to late February of 2019. For the
benign samples, we manually assembled a dataset of 3,016
samples from source-code projects available on GitHub [29].
Next, we used radare2 [30] to obtain the CFGs of the samples.
Throughout the study, wherever required, we use 80% our
dataset for training and validation, and 20% for evaluation.

Malware Family (Class). To determine the family label
of the malware, we inspect the malware samples through
VirusTotal [31]. The scan results from the VirusTotal are then
passed through AVClass [32] to label them with their family
class. VirusTotal scans include scan results from multiple anti-
 virus software, each of which assign a family name to the
malware. AVClass further uses majority vote to determine
the family label. Soteria classifies the samples into different
classes, i.e., family labels and benign. Table II shows the IoT
samples’ distribution across classes.

Adversarial Dataset. Recall that we utilize the GEA to the
generate the AEs to evaluate Soteria’s robustness. These AEs
are generated from the test dataset (20% of the samples per
class). Towards this, we start by selecting three samples from
each class, i.e., one from each sizes, small, medium, and large.
We define small, medium, and large by minimum, median, and
maximum number of nodes in the dataset. Taking a sample
from a class for each size, as the targeted sample, we generated
adversarial examples by applying GEA over every sample in
the test dataset of all the classes except for the targeted sample
class. For example, if we select a sample of size Small from
the benign dataset, we then apply GEA over this sample and
each of the samples in the test dataset of Gafgyt, Mirai, and
Tsunami, giving us a total of 2,742 AEs (it can be seen in
Table II that Gafgyt, Mirai, and Tsunami have 2,217, 473,
and 52 samples, respectively, aggregating to 2,742 AEs). The
number of generated AEs of each class is in Table III.

B. Feature Analysis
We extract features from 200 random samples from each
class. Recall that we extract density- and level-based features
from each sample. We use both of these feature vectors
together to create a combined feature vector of size $1 \times 1000$.
We used Principal Component Analysis (PCA) [33] with a
dimension of two. PCA converts a set of observations of
possibly correlated variables into a set of values of linearly

---

Table II: Distribution of IoT samples across the benign and malicious families. Gafgyt is the most popular IoT family with 66.18% of the dataset samples, while Tsunami is the least popular with only 262 samples (1.55% of the samples). The dataset is split into the train (80%) and test (20%) subsets.

<table>
<thead>
<tr>
<th>Class</th>
<th># Train</th>
<th># of Samples</th>
<th># Total</th>
<th>% of Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benign</td>
<td>2,416</td>
<td>600</td>
<td>3,016</td>
<td>17.94%</td>
</tr>
<tr>
<td>Gafgyt</td>
<td>8,911</td>
<td>2,217</td>
<td>11,128</td>
<td>66.18%</td>
</tr>
<tr>
<td>Mirai</td>
<td>1,935</td>
<td>473</td>
<td>2,408</td>
<td>14.33%</td>
</tr>
<tr>
<td>Tsunami</td>
<td>210</td>
<td>52</td>
<td>262</td>
<td>1.55%</td>
</tr>
<tr>
<td>Overall</td>
<td>13,472</td>
<td>3,342</td>
<td>16,814</td>
<td>100%</td>
</tr>
</tbody>
</table>
TABLE III: GEA selected targeted samples. These samples are used to generate the AEs to evaluate Soteria. Three samples from each class are selected of different sizes (number of nodes), i.e., small, medium, and large.

<table>
<thead>
<tr>
<th>Class</th>
<th>Size</th>
<th># Nodes</th>
<th># AEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benign</td>
<td>Small</td>
<td>10</td>
<td>2742</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>50</td>
<td>2742</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>443</td>
<td>2742</td>
</tr>
<tr>
<td>Gafgyt</td>
<td>Small</td>
<td>13</td>
<td>1125</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>64</td>
<td>1125</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>133</td>
<td>1125</td>
</tr>
<tr>
<td>Mirai</td>
<td>Small</td>
<td>12</td>
<td>2869</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>48</td>
<td>2869</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>235</td>
<td>2869</td>
</tr>
<tr>
<td>Tsunami</td>
<td>Small</td>
<td>15</td>
<td>3290</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>46</td>
<td>3290</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>79</td>
<td>3290</td>
</tr>
</tbody>
</table>

Fig. 8: The PCA comparison between the benign and malware families using features used in Alasmary et al. [3].

uncorrelated variables called principal components.

Baseline Comparison. Prior works, like, Alasmary et al. [3] and Abusnaina et al. [9] use graph theoretic features extracted from the general structure of the CFG. With the comparative analysis of such features with our feature considerations, we exhibit our feature sets to be more discriminative. Fig. 8, and Figures 9(a), 10(a), and 11(a) show the PCA visualization of the feature vectors between the classes of features considered in the prior works and our features design, respectively. Notice that our feature representation is more discriminative of the classes. Additionally, we notice that the malicious classes in the figures are indistinguishable using the graph theoretic features. Table IV shows the distribution of the discriminative features over the four classes with 51 and 129 density-based and level-based features, respectively, shared between classes.

AE vs. Clean Features. To detect AEs, i.e., distinguish the AEs from the clean samples, understanding the differences in feature representation between clean and AEs is important. To examine this, we applied PCA on the clean and adversarial feature vectors, the results of which are shown in Fig. 9(b), 10(b), and 11(b). Notice that the clean and AEs are distinguishable, particularly when using the combined feature vectors.

C. Evaluation and Analysis

Recall that Soteria has two major functionality, AE detection and classification. Below, we present the evaluation of Soteria’s performance and also compare it with the baseline.

1) Adversarial Example Detector: We evaluated AE detector of Soteria by its ability to detect adversarial examples and distinguish them from the clean samples, regardless of their class. Fig. 9(b), 10(b), and 11(b) show the spatial differences between clean and adversarial samples.

Training Parameters. We trained Soteria on reconstructing the training data in Table II. The reconstruction error (RE) is the RMSE between the original and reconstructed samples, we set the number of epochs to 100 with a batch size of 128.

Testing. Given the trade-off between adversarial detection sensitivity (false negatives) and the clean samples misdetection (false positive), setting a proper RE threshold is essential. We calculate the RE and set the threshold \( T_h = \mu(R_E) + \sigma(R_E) \), where \( R_E \) is a vector of all RE values of the training samples, and \( \mu \) and \( \sigma \) are the mean and standard
deviation of the training samples RE, respectively. Fig. 12 shows the RE distribution over the clean and adversarial features vectors. To consider a sample as adversarial, half of its feature vectors should have a RE higher than the threshold.

**Performance.** Table V shows Soteria’s performance against AEs. Overall, the detector detects 97.79% of the AEs. In most cases (9 out of 12), the detector was able to detect AEs with an accuracy greater than 99%. Furthermore, Table VI shows the detection performance against clean samples. Notice that only samples from Gafgyt family were misdetected as AEs, mainly because of the high number of discriminative features associated with this family, as shown in Table IV. In conclusion, we detected AEs and distinguish them from the clean samples with high accuracy. Detected samples are labeled as adversarial and will not be forwarded to the classifier.

**Analysis.** To show the importance of setting the right threshold, we re-implement the threshold as $T_b = Mean(R^T_E) + \alpha \times SDV(R^T_E)$, where $\alpha$ is an arbitrary value. We test the detector performance against the clean and adversarial samples by varying $\alpha$ from 0 to 2.0. Fig. 13 shows the effect of $\alpha$ on the detection error. With $\alpha = 0$, all AEs were detected, although more than 60% of the clean samples were classified as AEs. With $\alpha = 2.0$, all clean samples were correctly detected and no AEs were detected by Soteria. Note that our selected threshold was chosen without access to the test dataset.

2) **Classifier:** The classifier aims to correctly distinguish a sample into the aforementioned classes (Benign, Mirai, Gafgyt, or Tsunami). We evaluate the performance of Soteria alongside the existing approaches.

**Training Parameters.** We set the number of epochs to 100 with a batch size of 128 and evaluated the performance of each model individually and against the majority voting.

**Performance.** We evaluated Soteria’s classifier’s performance against two existing models: 1) Graph-based: Alasmary et al. [3] propose a malware detector based on features extracted from the general structure of the CFG. and 2) Image-based: Cui et al. [5] use image-based design where each sample is represented as an image of a fixed size to detect malware. We implemented the above two systems. Table VII shows the performance of the models for the different classes. The model accuracy over each class is measured by the number of samples correctly classified over the total number of samples that belong to that class. For image-based model, we implemented the four models mentioned in their design. The evaluation of 96 $\times$ 96 and 192 $\times$ 192 based models shows poor performance, with an overall accuracy rate of 66.37%. Therefore, we did not
TABLE VI: GEA: Detector Performance over clean samples. Only 6.16% of the clean samples were misclassified as AEs. All Benign clean samples passed the detector. DE refers to the detected samples (lower is better).

<table>
<thead>
<tr>
<th>Class</th>
<th># Samples</th>
<th># DE</th>
<th>% DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benign</td>
<td>600</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Gafgyt</td>
<td>2,217</td>
<td>206</td>
<td>9.29%</td>
</tr>
<tr>
<td>Mirai</td>
<td>473</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Tsunami</td>
<td>52</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Overall</td>
<td>3,342</td>
<td>206</td>
<td>6.16%</td>
</tr>
</tbody>
</table>

Table VII shows the classifier’s operation mode for detector.

V. DISCUSSION

System Robustness. Our evaluation shows that Soteria is robust, with the ability to detect AEs with an accuracy of 97.79%, and a trade-off of detecting 206 Gafgyt samples as adversarial. Moreover, Soteria outperforms other systems over the same training and testing datasets. The compared systems had an overall low Tsunami classification accuracy, due to the small dataset. Soteria, on the other hand, and using a majority voting system, achieved an accuracy of 100% in classifying Tsunami sample. In fact, the majority voting classifier only failed in classifying three Mirai samples in the evaluation, classifying them as benign samples.

Operation Mode for Detector. Soteria is used to distinguish AEs and detect them. To enable Soteria’s operation, the extracted features distribution of normal and AEs should be different. Moreover, we argue that the detector should not be aware of the AEs and their patterns in the training process, as this will bias the detector’s performance towards specific attacks, decreasing the robustness against other attacks.

Adversarial Capabilities. In section II-D, we discuss the threat model and adversarial capabilities. We assumed that the adversary can access the source code of the samples, and can modify and merge them. Moreover, he has prior knowledge of the design and its internal architecture. Soteria’s success implies that the adversary cannot generate practical AEs. What the adversary does not have in Soteria is the ability to know in advance what features are being used for the classifier, since those features are randomized for every run of the system. For instance, inserting a single block with a low density near the exit block will not highly affect the labeling of the sample, and will not be detected as an AE by Soteria. However, Soteria can classify the sample to its original class, since the labels are intact. Moreover, the adversary needs to ensure that the labels change in such a way the classification decision will be toggled, without being detected by the AE detector, which include it in the comparison. Our evaluation shows that Soteria outperforms the existing systems, as particularly shown in the Tsunami classification and overall accuracy rate.

Analysis. Recall that the accuracy of our AE detector was 97.79%, meaning that 2.21% of the AEs were not detected by Soteria, and were forwarded to the classifier. Given it’s application, it is important to understand the classifier’s behavior against those samples. Table VIII shows the classifier’s behavior over these samples. The classifier detects them as benign or Gafgyt, with a large percentage (76.1%) of the samples classified as benign. It is worth noting that the targeted classification is not valid in this design, as Fig. 9(b), Fig. 10(b), and Fig. 11(b) show a clear difference in the feature representation between clean and AEs. However, due to the variety in the benign samples’ features distribution, the adversarial examples that pass the detector are likely to be classified as benign. This can be critical, even with a detection rate of as high as 97.79%, given the application domain.

Classification performance of Soteria dense-, level-, and voting-based classification systems in classifying normal (non-adversarial) samples.

<table>
<thead>
<tr>
<th>Class</th>
<th>Soteria DBL</th>
<th>LBL</th>
<th>Voting</th>
<th>[9]</th>
<th>[5]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benign</td>
<td>99.45</td>
<td>99.70</td>
<td>100</td>
<td>99.00</td>
<td>99.50</td>
</tr>
<tr>
<td>Gafgyt</td>
<td>99.70</td>
<td>97.00</td>
<td>100</td>
<td>98.55</td>
<td>98.87</td>
</tr>
<tr>
<td>Mirai</td>
<td>99.49</td>
<td>98.73</td>
<td>99.36</td>
<td>97.67</td>
<td>92.81</td>
</tr>
<tr>
<td>Tsunami</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>84.61</td>
<td>32.69</td>
</tr>
<tr>
<td>Overall</td>
<td>99.63</td>
<td>97.77</td>
<td>99.91</td>
<td>98.29</td>
<td>97.01</td>
</tr>
</tbody>
</table>

Fig. 13: Effect of varying the detector threshold (α) on the detection error. The selected α in Soteria is the intersection between the error rates of normal and adversarial samples.

TABLE VIII: Soteria’s classifier predictions over AEs misdetected by the detector. Most of the misdetected samples are generated using GEA with large size selected samples.

<table>
<thead>
<tr>
<th>Class</th>
<th>Size</th>
<th># AE</th>
<th>Benign</th>
<th>Gafgyt</th>
<th>Mirai</th>
<th>Tsunami</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benign</td>
<td>Small</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>402</td>
<td>287</td>
<td>115</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Gafgyt</td>
<td>Small</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>100</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mirai</td>
<td>Small</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>181</td>
<td>145</td>
<td>36</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Tsunami</td>
<td>Small</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>42</td>
<td>39</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
happened in our evaluation in 2.21% of the generated AEs. Finally, and due to the change in the labeling, the adversary cannot force the classifier into a targeted misclassification. **Alternative Features for Classifier.** In Soteria, we built a classifier that is based on the utilized features from the detector design process. However, the classifier can be replaced, with some caveats. The detector decision is based on the extracted CFG. Appended binaries at the end of the file will not affect the detector decision. Clean samples with adversarial binaries appended to them will not be detected as AEs by Soteria. While this is an advantage of Soteria classifier, it is equally a serious shortcoming with other approaches, such as image-based malware classifiers [34]. Ideally, the classifier should be at least as good as the classifier proposed in this paper, meaning that it should only consider the executable binaries in the classification process. Moreover, the discriminative features are highly distinguishable among classes, and the feature extraction process is immune to feature space manipulation. **Limitations.** Our work has two major limitations. 

1. **CFG-based Features:** CFG-based features are effective compared to other feature designs. However, CFG does not necessarily reflect the actual code. Editing the code without even changing the functionality (by creating an equivalence) would affect the structure of the CFG, which might be exploited by the adversary to evade detection in the first place. For example, an adversary may inject a sample of code that would not result in a new branching, but would still affect the structure of the CFG. While such an event is well within the scope of our adversary model, and would not affect the classification results, it would only affect the feature space, requiring us to retrain Soteria to capture the new feature space.

2. **Binary Obfuscation:** Obtaining a representative CFG would not be possible under obfuscation, typically done using string obfuscation, resulting in hiding parts of the code, or function obfuscation, resulting in an incomplete CFG. An incomplete CFG may result in an incomplete feature representation of the sample, and thus a misclassification. Obfuscation is a shortcoming of our work, and deobfuscation is an active research area in its own right, where developed tools can be used as the basis for our work to obtain representative CFGs.

V. RELATED WORK

Machine and deep learning algorithms are widely leveraged towards securing software against adversaries in general and detecting malware in particular. For instance, Alasmary et al. [3] analyzed two prominent malware, IoT and Android, based on the CFG-graph representation of the malicious software. Moreover, Alam et al. [35] analyzed the malware and proposed a malware detection system to detect malware with even small CFGs and then addressing the changes occurred in the frequencies of opcodes. Bruschi et al. [36] proposed a malware detection method that uses two CFG techniques to compare and detect malware based on two CFGs of malware code and other known malware.

Several research works have been proposed to defend against adversarial machine learning. Most of these approaches are image-based methods. For example, Goodfellow et al. [10] proposed to train the model with a set of AEs to minimize the test error between the real and AEs of the model’s result. Papernot et al. [37] designed a network distillation model to defend against adversarial attacks such as fast gradient sign method [10] and L-BFGS attack [38]. Cui et al. [39] introduced a malware detection method for malicious codes using deep learning by transferring the malicious code into grayscale images. Ni et al. [40] proposed a malware family classification system that converts malicious codes of nine different malware families into grayscale images. Metzen et al. [33] proposed a detection method for adversarial perturbation over trained AEs. Moreover, Rozsa et al. [41] proposed a machine learning model that tested the adversarial examples. They correlate their robustness of the three adversarial attacks to the accuracy of eight deep network classifiers. In addition, Miyato et al. [26] proposed a detection method on the text domain. They trained the model over adversarial examples that apply small perturbation to the word that is embedded in RNN.

Several methods have been proposed to generate adversarial examples that can manipulate the desired output to fool the classifiers [9], [10], [11], [12]. Adversaries can make small modifications to the malware to misclassify them as benign, yet they remain malware files [42], [43]. Other methods apply and add small noise or perturbation to optimize the images to generate the adversarial examples [44], [10], [45]. For example, Carlini and Wagner [45] proposed three adversarial attacks against distillated neural networks that break many defenses models. Moreover, Moosavi-Dezfooli et al. [12] proposed a DeepFool method that generates minimal perturbation to change the classification labels based on iterative linearization of the classifiers. Recently, Abusnaina et al. [9] proposed adversarial attacks over the CFGs of malware binaries through designing two adversarial attacks to craft the IoT detector.

VII. CONCLUSION

In this paper, with Soteria, we address the need to detect adversarial machine learning attacks by proposing an adversarial machine learning detector for IoT malware. Particularly, Soteria defends the CFG-based classifiers for malware detection against the AEs. The first component, the AE detector, is a Control Flow Graph (CFG)-based model that can detect adversarial samples without training the model over adversarial samples (as shown by prior works). The model computes the reconstruction error between the input data and the reconstructed output of the auto-encoder, and uses a threshold to detect the adversarial samples with an overall accuracy of 97.79%. Additionally, the second component of Soteria performs a family-based classification with an accuracy of 99.91% on the clean samples. These two models operate independently, increasing the robustness of Soteria. **Acknowledgement.** This work was supported by Cyber-Florida Collaborative Seed Award and NRF under grant 2016K1A1A2912757.
REFERENCES


