Monotonic-HMDs: Exploiting Monotonic Features to Defend Against Evasive Malware

Md Shohidul Islam , Behnam Omidi , Khaled N. Khasawneh

2021, 22nd Int'l Symposium on Quality Electronic Design

Problem Addressed in the Paper :

Malware attacks are the main threats to the security of the systems. Attacker exploits vulnerabilities of the system and develop numerous malware to fulfill their malicious goals. Preventing malware from malware attacks requires detection, which is mainly of two types, static and dynamic deteciton. In static detection, malware detects by matching their signature with the previously stored database. This technique is ineffective in detecting obfuscated/metamorphic/polymorphic malware and zero-day attacks whose signature have not yet been encountered. Dynamic malware detects by monitoring their run-time behaviour, which makes the technique more robust in detecting unseen signatures. However, dynamic detection, when done with software implementation, poses significant overhead on system performance and power.

To make detection more efficient, Hardware Malware Detection(HMD) comes into play. HMD are machine learning classifier that detect malware by using low-level hardware features. However, HMD gaining popularity in malware detection, attacker gains upper hand by adapting malware so that they can bypass malware detectors. The attacker does not know the victim HMD model internal details therefore attacker query the victim HMD and observe model input and output. Due to unavailibility of victim model parameters, black-box approaches takes two steps,

1. Reverse Engineering the victim HMDs, which build proxy models
2. Generate evasive malware based on proxy model.

There exist another threat model called white-box attack, which assume that attackers have access to victim HMDs model parameters and thus attacker follow only the secound step.

The key contributions of this paper are as follows:

- They propose Monotonic-HMDs as a defence against evasive malware attacks. Our security analysis shows that Monotonic-HMDs are robust against against both black-box and white-box attacks scenarios.
- The evaluate the Monotonic-HMDs model overhead, in terms of inference performance and model size.
- The evaluate the hardware complexity of integrating Monotonic-HMDs into an open core.

Motivation for the problem :

To defend against evasive malware attacks resilient hardware malware detector (RHMDs) is proven to be effective against black-box attacks. It uses multiple diverse detectors and randomly switches between them at run-time. Such random switching between the base detectors leads to increase the unpredictablility of the model which makes the reverse engineering harder. But RHMDs are found to be ineffective against the white-box attacks as the attackers have exact knowledge of the victim HMD models and can exploit them in generating evasive malware samples. RHMD increased the hardware complexity in implementation.

Motivation for the solution :

Adversaries only targets the benign features which contribute to increase the probability of dectecting the input as benign by adding more of them to craft evasive samples of the malware that can evade detection. They does not touch the malware features. Therefore, they introduced Monotonic-HMDs which uses only the malicious features for the detection. If the attacker tries to evade by adding more bening features it will not affect the detection and unable to bypass the detection which makes Monotonous-HMD more robust against both black-box and white-box attacks.

Theory :

Attackers uses benign features to develop evasive malware samples that can evade HMDs detection. The rational behind targeting benign features is that it is easy to manipulate without changing the malicious funcitonality of the malware. All proposed adversarial attacks target increasing the benign features values to craft evasive malware samples. Monotonic classifier can enforce monotonic constraints such that inceasing the value of a feature would result in a higher probability of getting detected. Therefore, to build such a classifier, we first need to identify the benign features that can be an easy target for the attackers to create evasive malware. After identifying the benign features, we train monotonic constrained models using the selected features only.

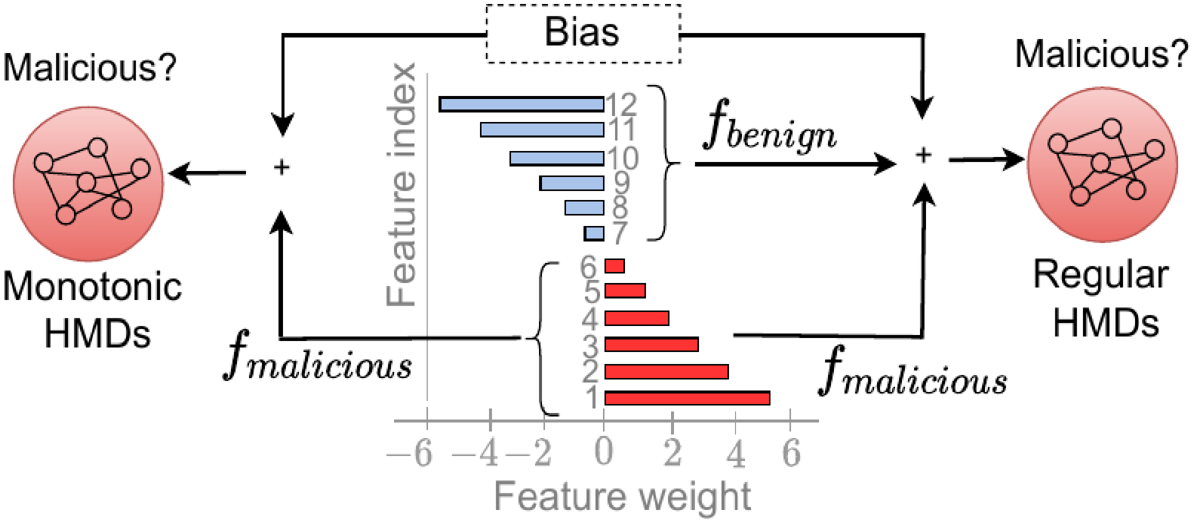

Figure 1: Monotonic-HMDs

At first, they train a logistic regression classifier using the collected dataset. Then they identify the malicious and benign features; malicious features contribute to the HMDs decision making toward detecting a running process as malware, and benign features contribute to the probability of a process being benign. Features having positive weights are identified as malicious and features with negative weights are identified as benign. As shown in figure 1, Monotonic-HMD discard benign features but comprise only the malicious features and bias. Specific to our evaluation, the total number of features in our dataset is 50. After training a logistic regression classifier using the 50 features, the classifier contains 25 positive weights, i.e, malicious features, and 25 negative weights, i.e., benign features.

Implementation :

Their dataset comprising benign programs and malware programs. They executed 600 benign program of numerous types such as text editing tools, browser programs, system programs etc. For malware program 3000 malware samples of different types are collected from a malware database MalwareDB. They were executed on a 32-bit windows 7 virtual machine. Note that in order to allow the malware to run freely and perform their intended malicious activities we disabled Windows security and Firewall services. To collect dynamic profiling information and low-level hardware features of a training program, we used Intel Pin instrumentation tool. Dynamic traces of running program were collected for a duration of 5000 system calls or 15 million committed instructions, after the warm-up period. They divided the dataset into victim training (50%), attacker training (25%) and testing (25%).

Your own Critique :

For this approach to work, classifier need to be monotonic such that they impose monotonic constraint on model. Suppose x_i and x_j are two different values of i-th feature. Then classifier f is increasingly monotonic if $x_i \leq x_j$ implies $f(x_1,x_2,...,x_i,...x_d) \leq f(x_1,x_2,..,x_j,..x_d)$.

Regarding related works :

- Nd-hmds: Non-differentialbe hardware malware detectors against evasive transient execution attacks (2020).
- RHMD: evasion resilient hardware malware detectors (2017).
- Ensemble learning for low-level hardware supported malware detection (2015).