A Summary of Adversarial Attacks and Defenses on ML- and Hardware-based IoT Device Fingerprinting and Identification

Pedro Miguel Sánchez Sánchez[1], Alberto Huertas Celdrán[2], Gérôme Bovet[3], Gregorio Martínez Pérez[1]
[1]Department of Information and Communications Engineering, University of Murcia, Spain
[pedromiguel.sanchez@um.es]
[2]Communication Systems Group (CSG), Department of Informatics (IfI), University of Zurich UZH, Switzerland
[3]Cyber-Defence Campus, armasuisse Science & Technology, Switzerland

***Abstract*—In response to the rapid expansion of Internet-of-Things (IoT) devices and associated cybersecurity threats, this work proposes a novel LSTM-CNN architecture for robust individual device identification, leveraging behavior monitoring and ML/DL advancements. Evaluated against a dataset from 45 Raspberry Pi devices, this model outperforms traditional ML/DL methods, achieving a +0.96 average F1-Score and demonstrating strong resilience to adversarial attacks, including context-based and ML/DL-specific evasion attempts. Through the application of adversarial training and model distillation defenses, the model vulnerability to the most effective attack was reduced from a 0.88 success rate to 0.17, maintaining high-performance integrity.**

***Index Terms*—Adversarial attacks, Device Identification, Artificial Intelligence, Internet of Things, Context Attack**

Tipo de contribución: *Investigación ya publicada*

I. Introduction

With the expansion of IoT and advancements in processing technologies, the deployment of IoT devices has surged, enriching sectors like Industry 4.0, Smart Cities, and Healthcare with diverse applications. This proliferation, however, raises significant cybersecurity challenges, particularly as more potent IoT devices, such as Single-Board Computers (SBCs), are prone to sophisticated cyber threats. Conventional static identifiers for device authentication are increasingly vulnerable, prompting a shift towards behavior and hardware-based identification methods. These methods offer enhanced security by distinguishing devices through unique hardware characteristics and performance patterns.

This work presents a summary of [1], which focuses on individual device identification through hardware performance, considering the challenges posed by adversarial attacks on data integrity and identification techniques. It proposes an LSTM-CNN architecture for identifying devices based on the unique behavior of their CPU, GPU, memory, and storage, utilizing a dataset from 45 Raspberry Pi devices for evaluation. This architecture demonstrates high accuracy, with an average F1-Score of +0.96 and a True Positive Rate (TPR) of +0.80.

Moreover, this work outlines a threat model for adversarial attacks affecting hardware behavior-based identification, analyzing both context-related and ML/DL-focused adversarial evasion attacks. It finds that while context attacks like temperature variations have minimal impact, ML/DL evasion techniques can significantly challenge identification accuracy. To counter these threats, the research applies adversarial training and model distillation defense techniques, effectively reducing the success rate of the most potent attacks to approximately 0.18, thereby enhancing the model robustness against such adversarial tactics.

II. Individual Identification

The section elaborates on an ML/DL framework for individual device identification via hardware performance, establishing a baseline for analyzing the impact of attacks and defenses. The dataset, named LwHBench, collects performance metrics from CPU, GPU, Memory, and Storage of 45 Raspberry Pi devices, under controlled conditions to ensure data integrity.

The LSTM-1DCNN architecture, designed for time series analysis, combines LSTM recursive pattern recognition with 1D-CNN spatial pattern extraction. TABLE I showcases the classification performance of various models. The LSTM-1DCNN model notably achieves the highest performance, highlighting its effectiveness in identifying individual devices based on hardware performance metrics.

TABLE I: Baseline classification model performance.

Model	Accuracy	Avg. Precision	Avg. Recall	Avg. F1-Score
Single vector approaches				
SVM	0.7838	0.7955	0.7829	0.7849
XGBoost	0.9059	0.9173	0.9056	0.9087
Random Forest	0.8549	0.8664	0.8542	0.8570
MLP	0.8895	0.8960	0.8880	0.8899
Time series approaches (10 values)				
1D-CNN	0.9428	0.9453	0.9428	0.9428
LSTM	0.9346	0.9430	0.9346	0.9346
LSTM_1D-CNN	**0.9602**	**0.9626**	**0.9602**	**0.9602**
Multi_1DCNN_LSTM	0.9535	0.9553	0.9535	0.9535

The performance of the LSTM-1DCNN model suggests its potential to enhance IoT security. This contrasts prior works that relied on aggregated features, indicating the advantage of using extensive datasets and time series DL models.

III. Threat Model

This section outlines the threat model for ML/DL-based device identification solutions that rely on monitoring internal hardware performance. It identifies potential vulnerabilities in both the data generation by hardware and the subsequent evaluation by ML/DL models, as illustrated in Fig. 1.

The list of identified threats is the following one. *TH1. Fingerprint eavesdropping and hijacking*: Attackers intercept and misuse fingerprint data to impersonate devices by exploiting data transmission and storage vulnerabilities. *TH2. Fingerprint*

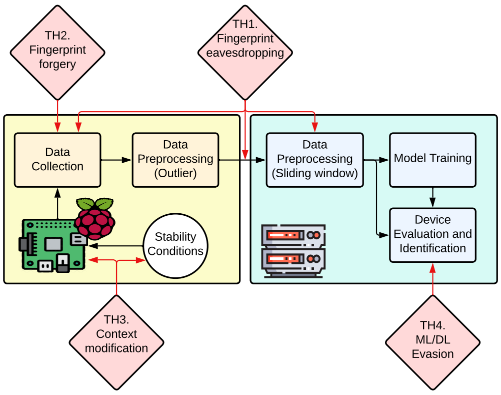
TH2. Fingerprint forgery
TH1. Fingerprint eavesdropping
Data Collection
Data Preprocessing (Outlier)
Data Preprocessing (Sliding window)
Model Training
Device Evaluation and Identification
Stability Conditions
TH3. Context modification
TH4. ML/DL Evasion

Fig. 1: Threat impact on the steps of the identification process.

forgery: With detailed knowledge of the fingerprint generation process, adversaries craft counterfeit fingerprints to mimic legitimate devices. *TH3. Context modification*: External or operational condition manipulations are used to disrupt the accurate generation or recognition of device fingerprints. *TH4. ML/DL evaluation evasion*: Knowledgeable attackers create malicious data samples to fool the ML/DL evaluation models, aiming to impersonate specific devices.

IV. Adversarial Attacks and Defenses

This section evaluates the vulnerability of an ML/DL-based device identification model to adversarial attacks designed for device impersonation or identification disruption. The model faces threats from TH2 (Fingerprint forgery), TH3 (Context modification), and TH4 (ML/DL evaluation evasion), with TH1 (Fingerprint eavesdropping and hijacking) neutralized through encryption and high-privilege process isolation.

1) Identification Disruption Attack (TH3): The DL model resilience to context modification, particularly temperature changes, was tested. Despite the temperature variation during data collection, attacking the model with new temperature conditions only slightly affected its performance. The experiments showed a minor 0.03 decrease in average metrics for descending order temperatures and a negligible impact on ascending order, albeit with a decrease in minimum TPR to about 0.65 for two devices. This minimal performance degradation suggests robustness to context modifications, with all devices still identifiable above a 0.50 TPR threshold.

2) Device Spoofing Attacks (TH2, TH4): Device spoofing tested the model susceptibility to ML/DL evaluation evasion. White-box attacks such as the Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), and Momentum Iterative Method (MIM) demonstrated significant success rates, pointing to a glaring susceptibility. TABLE II shows the success rate for the main attacks tested. Notably, the BIM and Projected Gradient Descent (PGD) attacks achieved success rates above 0.85, underlining a pronounced risk of malicious device impersonation. This analysis highlights the model exposure to targeted evasion attacks, underscoring the necessity for enhanced security measures.

In response to the vulnerabilities identified, adversarial training and model distillation were applied as defense mechanisms. These strategies aimed to enhance the model robustness against ML/DL evasion attacks, with a focus on device spoofing threats. The integration of adversarial training and distillation notably reduced the success rate of the most effective attacks to below 0.18, significantly bolstering the model defenses without impacting overall performance metrics. TABLE III shows the Attack Success Rate (ASR) after applying the defense mechanisms. The combination of these defenses proved most effective, marking a pivotal advancement in safeguarding device identification models against sophisticated adversarial tactics.

TABLE II: Adversarial attack results.

Attack	Attack Success Rate	Time
FGSM, $\epsilon = 0.05$	0.3056	8.79 s
BIM, $\epsilon = 0.5$	0.8823	752.64 s
MIM, $\epsilon = 0.05$	0.8537	793.97 s
PGD, $\epsilon = 0.6$	0.8823	748.06 s

TABLE III: Attack ASR on the robust models.

Attack	Baseline Model	Distilled Model	Adversarial Training	Adversarial Training + Distilled
FGSM, $\epsilon = 0.05$	0.3056	0.2725	0.2704	0.1561
BIM, $\epsilon = 0.5$	0.8823	0.3024	0.1482	0.1631
MIM, $\epsilon = 0.05$	0.8537	0.7950	0.1918	0.1784
PGD, $\epsilon = 0.6$	0.8823	0.2741	0.1155	0.1235

V. Conclusions and Future Work

The rapid proliferation of IoT devices has necessitated the development of novel identification techniques leveraging hardware behavior and ML/DL methods. This work utilized performance data from 45 Raspberry Pi devices, to evaluate ML/DL classifiers for device identification. A DL model integrating LSTM and 1D-CNN layers emerged as the most effective, achieving an average F1-Score of 0.96 and successfully identifying devices with a minimum TPR of +0.80. Despite its robustness against temperature changes, the model was vulnerable to certain ML/DL evasion attacks, which achieved up to a 0.88 success rate. Implementing model distillation and adversarial training significantly enhanced resilience against these attacks, with minimal impact on accuracy and identification thresholds. Future work will explore additional adversarial attacks and defense strategies, including generative models. Incorporating trust metrics and evaluating the fairness and robustness of predictions will also be a focus, alongside investigating federated learning for distributed model generation.

Acknowledgment

This work has been partially supported by *(a)* the Swiss Federal Office for Defense Procurement (armasuisse) with the CyberMind and DATRIS projects (CYD-C-2020003) and *(b)* CDL-TALENTUM project from the Spanish National Institute of Cybersecurity (INCIBE) by the Recovery, Transformation, and Resilience Plan, Next Generation EU.

References

[1] Sánchez Sánchez, P. M., Huertas Celdrán, A., Bovet, G., & Martínez Pérez, G.(2024). Adversarial attacks and defenses on ML-and hardware-based IoT device fingerprinting and identification. *Future Generation Computer Systems*, 152, 30-42.