OPEN ACCESS
Vietnam Journal of Computer Science
(2024) 1–19
© The Author(s)
DOI: 10.1142/S2196888824500052

Crossed-IoT Device Portability of Electromagnetic Side Channel Analysis: Challenges and Dataset

Tharindu Lakshan Yasarathna

School of Computer Science
University College Dublin, Belfield
Dublin 4, Ireland

Lojenaa Navanesan

School of Computing, University of Colombo
Colombo, Sri Lanka

Simon Barcque

School of Computer Science
University College Dublin, Belfield, Dublin 4, Ireland

Asanka Sayakkara

School of Computing, University of Colombo
Colombo, Sri Lanka

Nhien-An Le-Khac*

School of Computer Science
University College Dublin, Belfield, Dublin 4, Ireland
**an.lekhac@ucd.ie*

Received 4 October 2023
Revised 24 January 2024
Accepted 27 January 2024
Published 23 February 2024

Internet of Things (IoT) refers to the network of interconnected physical devices, vehicles, home appliances, and other items embedded with sensors, software, and connectivity, enabling them to collect and exchange data. IoT forensics aim to investigate cybercrimes, security breaches, and other malicious activities that may have taken place on these connected devices. In particular, Electromagnetic Side-Channel Analysis (EM-SCA) has become an essential tool for IoT forensics due to its ability to reveal confidential information about the internal functions of IoT devices without interfering with these devices or wiretapping their networks. However, the

*Corresponding author.

accuracy and reliability of EM-SCA results can be limited by device variability, environmental factors, and data collection and processing methods. In fact, very few studies have explored potential solutions to address the significant impact of these limitations on the accuracy of EM-SCA approaches applied to crossed IoT devices. Therefore, this paper examines the impact of device variability on the accuracy and reliability of machine learning (ML)-based approaches for EM-SCA. We first present the background, basic concepts and techniques used to evaluate the limitations of current EM-SCA approaches and datasets. Our study then addresses one important limitation, which is caused by the multi-core architecture of the processors (aka. System-On-Chip). We present an approach to collect the EM-SCA datasets and demonstrate the feasibility of using transfer learning to obtain more meaningful and reliable results from EM-SCA in IoT forensics of crossed-IoT devices. Moreover, this study contributes a new dataset for using deep learning (DL) models in analyzing Electromagnetic Side-Channel data with regard to the cross-device portability matter.

Keywords: IoT security and forensics; electromagnetic side-channel analysis; EM-SCA dataset; crossed-IoT devices; transfer learning.

1. Introduction

Internet of Things (IoT) is increasingly becoming a part of our daily life. IoT devices interact with each other, people, and systems over the internet.[1] With the continuous growth of IoT devices, it needs a more secure and efficient way of managing the data being collected, transmitted, and processed. IoT forensics[2] has emerged as an essential study area to investigate preventing malicious activities on IoT devices.[3] In particular, Electromagnetic Side-Channel Analysis (EM-SCA) is a promising approach for IoT forensics due to its ability to reveal confidential information about the internal functions of IoT devices without interfering with them.

EM-SCA is a method of analyzing the electromagnetic emanations that are produced by electronic devices. These emanations can provide information about the device's internal functions, including power consumption, encryption key, and processing speed.[4] This information can be used to perform a range of forensic activities, including data recovery, malware analysis, and key extraction. In the context of IoT forensics, EM-SCA can extract sensitive information from IoT devices, such as login credentials, private keys, and other confidential information.[5] By analyzing the electromagnetic emanations produced by IoT devices, it is possible to gain insight into the device's internal operations and identify potential security vulnerabilities.[6]

Machine learning (ML) and deep learning (DL) methods are being extensively used for data analytics. They can be applied to a variety of application domains such as agriculture,[7] renewable energies,[8] climate,[9] and health systems.[10] Moreover, ML/DL techniques have been widely used in cybersecurity[11] and digital forensics domains[12] to assist investigators in fighting against cybercrimes. As such, ML/DL techniques have been used in EM-SCA to enhance the capability of identifying application activities of mobile devices[13] and of IoT devices[14] to detect required artifacts as well as potential security issues. The creation of relevant EM-SCA datasets is crucial in enhancing the capability to identify activities of mobile devices[13] and IoT devices.[14] These datasets contribute to detecting necessary artifacts and potential security issues within the realm of ML/DL techniques applied to EM-SCA, supporting investigators in the fight against cybercrimes.

Recently, ML/DL-based approaches for EM-SCA have become a new essential research trend in cybersecurity and digital forensic domains, especially in the cross-device portability problems where ML/DL models for detecting activities trained from a given device could be used to detect relevant activities in other devices of the same kind.

Despite its usefulness, the accuracy and reliability of ML/DL models for EM-SCA depends on the EM datasets' quality and some limitations can be listed as follows:

- Device variability: Different devices produce different electromagnetic emanations.
- Environmental factors: The electromagnetic environment can also impact the results, affecting the strength and quality of the electromagnetic emanations.
- Data quality, reliability, and accuracy of the data collected by EM-SCA.

Among these limitations, the device variability is the most important factor that affects the overall accuracy of ML/DL models used for EM-SCA. This is more severe when EM data of the same activities captured from the different devices of the same kind are also different where are they should be at certain levels of similarity,[14] given these devices are running the same system-on-chip (SoC). This issue also limits the ability of applying ML/DL-based approaches for EM-SCA in the real-world applications. To the best of our knowledge, there is no research in literature that studied this limitation properly in terms of how and why EM data captured could be dissimilar from devices of the same kind and how to overcome this issue for ML/DL approaches.

Therefore, this paper first studies the accuracy and reliability of using ML/DL models for EM-SCA in the context of crossed-IoT device portability. We then examine the ability of using transfer learning to address these issues. Finally, we create and validate a new EM-SCA dataset, which is ready for building transfer learning models for EM-SCA with regard to the crossed-IoT device portability. This dataset can be used in both cybersecurity and digital forensic domains. This paper makes the following contributions to the domain of IoT Forensics:

- Examination of the impact of device variability, environmental factors, and data collection and processing on the accuracy and reliability of EM-SCA results.
- Analysis the feasibility of using transfer learning in improving the performance of ML/DL models used in analyzing EM-SCA data.
- Discussion of the importance of addressing these limitations in order to obtain more meaningful and reliable results from EM-SCA in IoT forensics.
- A new public dataset for using DL models in analyzing EM-SCA data with regard to the cross-device portability.[a]

[a] https://aseados.ucd.ie/datasets/EMSCA-2023-Latest/

In the following section, we discuss some of the existing state-of-art research related to our topic in Sec. 2. In Sec. 3, we present our methodology with a detailed explanation followed by experimental analysis in Sec. 4. Section 5 is about the discussion of our findings. Finally in Sec. 6, we conclude this paper by addressing some open questions to the research community and, our future direction of the research.

2. Related Work

In recent years, there has been a growing interest in IoT forensics, particularly in EM-SCA. Researchers and practitioners have been exploring the potential of EM-SCA to analyze the electromagnetic emanations produced by IoT devices to reveal information about the device's internal operations. In this section, we will review some recent studies conducted in the field of EM-SCA for IoT forensics.

One of the earliest studies in this area was conducted by Ref. 15, who used EM-SCA to extract the encryption key from an IoT device. They found that the electromagnetic emanations produced by the device contained information about the encryption key, which could be used to decrypt the data transmitted by the device. This study demonstrated the potential of EM-SCA as a tool for IoT forensics and highlighted the need for better security measures in IoT devices to protect against attacks. Another study by Ref. 16 focused on using EM-SCA to detect malware in IoT devices. The authors used EM-SCA to analyze the electromagnetic emanations produced by a range of IoT devices and found that the emanations could reveal information about malware in the machine. They also found that the results were consistent with other malware detection methods, demonstrating the potential of EM-SCA as a complementary tool for detecting malware in IoT devices.

A study by Ref. 17 investigated the limitations of EM-SCA in IoT forensics by examining the variability of the electromagnetic emanations produced by different devices. They found that the electromagnetic emanations produced by different devices could vary widely, affecting the accuracy of the results obtained from EM-SCA. The authors also found that the variability was influenced by various factors, including the device's operating system, hardware, and environmental conditions. They also emphasize the importance of variability of electromagnetic emanations for EM-SCA-based IoT forensics. More recently, a study by Refs. 18 and 19 explored ML techniques in EM-SCA for IoT forensics. The authors found that ML algorithms could be used to improve the accuracy and reliability of the results obtained from EM-SCA. They also found that using ML algorithms could reduce the dependence on large and representative datasets, making EM-SCA more accessible and practical for IoT forensics.

EM-SCA is a powerful tool for IoT forensics that has the potential to reveal confidential information about the internal workings of IoT devices. Recent studies have demonstrated the potential of EM-SCA as a tool for extracting encryption keys, detecting malware, and improving the accuracy and reliability of the results obtained from EM-SCA. However, some limitations and challenges need to be addressed to fully realize the potential of EM-SCA for IoT forensics, such as the variability of the electromagnetic emanations produced by different devices, the dependence on large and representative datasets, and the need for improved data collection and processing methods. Therefore, further research is needed to address these limitations and challenges and improve the overall quality of using EM-SCA in IoT forensics.

3. Adopted Approach

Although ML/DL-based approaches for EM-SCA have become a new research trend in cybersecurity and digital forensic domains, the current approaches in literature are mostly working on the datasets generated from the same device. The problem of crossed-IoT device portability has not been studied in details. It limits the ability of deploying ML/DL approaches for EM-SCA in the real-world applications of cybersecurtiy and digital forensics. One of the key challenges is ML/DL methods perform poorly when applying crossed-IoT devices. To address this challenge, this paper focuses on three main research questions about the possible factors that affect the ML/DL models' accuracy and how to improve it. These research questions are listed as follows:

- **RQ1:** How does the multi-core architecture of SoC integrated on these devices affect the ML/DL models' accuracy?
- **RQ2:** How does the number of activities running on the IoT device affect the ML/DL models' accuracy?
- **RQ3:** Can transfer learning techniques be applied to improve the ML/DL models' accuracy?

We address these research questions through our empirical studies where we set up relevant experiments that we describe in this section. In addition, we deployed tests in varied scenarios. A detailed analysis of these test results is presented in the following section.

This section provides an overview of the essential technical background and the experimental platform for data generation.

3.1. *Experimental platform*

Acquiring EM radiation from a computing system requires multiple hardware and software components. In this work, we used Dragon board connected to a smart device through the wireless connection as a device-under-test (DUT).[14] Apart from that, signal-capturing equipment (HackRF One) is used with an antenna to capture the EM radiation of the DUT (cf. Fig. 1). The signal acquisition equipment is connected to a host computer that runs the necessary software to read the EM data samples and save them into trace files (cf. Fig. 2).

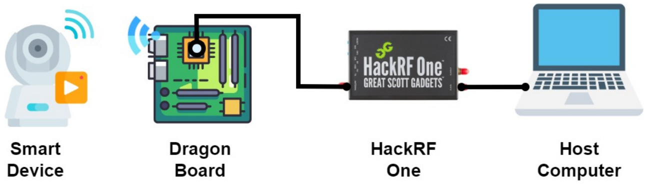
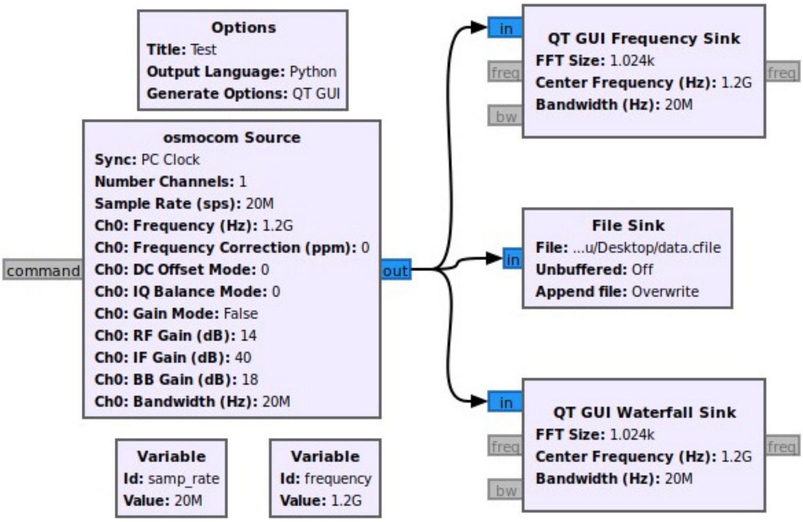


Fig. 1. The hardware setup for acquiring EM radiation from Dragon board. In the meantime, it connects with the smart device through the custom-built WiFi network.



Options
Title: Test
Output Language: Python
Generate Options: QT GUI

osmocom Source
Sync: PC Clock
Number Channels: 1
Sample Rate (sps): 20M
Ch0: Frequency (Hz): 1.2G
Ch0: Frequency Correction (ppm): 0
Ch0: DC Offset Mode: 0
Ch0: IQ Balance Mode: 0
Ch0: Gain Mode: False
Ch0: RF Gain (dB): 14
Ch0: IF Gain (dB): 40
Ch0: BB Gain (dB): 18
Ch0: Bandwidth (Hz): 20M

Variable
Id: samp_rate
Value: 20M

Variable
Id: frequency
Value: 1.2G

QT GUI Frequency Sink
FFT Size: 1.024k
Center Frequency (Hz): 1.2G
Bandwidth (Hz): 20M

File Sink
File: ...u/Desktop/data.cfile
Unbuffered: Off
Append file: Overwrite

QT GUI Waterfall Sink
FFT Size: 1.024k
Center Frequency (Hz): 1.2G
Bandwidth (Hz): 20M

Fig. 2. The GNU Radio Companion (GRC) flow graph for acquiring EM traces.

For this purpose, Dragon board 410c development board is used based on a Qualcomm APQ8016e application processor with a 1.2 GHz frequency (cf. Table 1). Linaro Linux distribution is used as the operating system of Dragon board and configured as a Routed Wireless Access Point. The new wireless network with the SSID DB used to connect smart devices to Dragon board. Besides, a HackRF One SDR is used as the EM radiation acquisition equipment. The device has a maximum sampling rate of 20 MHz. It supports a data acquisition frequency range from 1 MHz to 6 GHz, and the H-loop antenna is connected to HackRF One to acquire data from the DUT within close proximity. GNU Radio library is employed on the host computer to configure the SDR device and process the data it produces.[14]

The GNU Radio library provides a graphical interface called GRC, which facilitates creating visual flow graphs to build EM data processing pipelines. To capture EM traces from DUT, place the H-loop antenna on the optimal position of the CPU of the Dragon board using the GNU Radio library visualizations. The flow diagram created by GRC to record EM traces is shown in Fig. 2. Osmocom source depicts how the HackRF SDR device is set up to provide I/Q data samples. While the file sink

Table 1. Technical specifications of the smart devices used to create the EM dataset.

Devices	Processor	CPU frequency	WiFi bands	Scenarios used to collect data
Dragon board 410c (two devices)	Qualcomm® APQ8016e	1.2 GHz (4 core)	On-board Wi-Fi, Wi-Fi 802.11/g 2.4 GHz	(**EXP 1**) Connect one smart device to built WiFi network and collect EM traces from the CPU. (**EXP 2**) Consider Dragon board as DUT and run separate print, math, memory, and I/O activities on it. Here, we use only one core (use the same core for all activities) and collect CPU EM traces.
Amazon echo Show 5 (two devices)	MediaTek MT 8163	1.5 GHz (4 core)	Dual-band Wi-Fi 802.11 a/b/g/n/ac 2.4 GHz and 5 GHz	asking for time, (2) asking to play music, (3) asking for the weather today, (4) asking to play NRJ radio.

records the I/Q data stream into a raw data file, the frequency sink and waterfall sink are used to visualize data (see Fig. 3).

3.2. *Electromagnetic dataset generation*

To investigate the device variability, we used Amazon Echo Show 5 and Dragon board 410c for the experiments (cf. Table 1). We used two (02) similar devices to create a training dataset from one device and another to create a test dataset. The sample rate of the HackRF One SDR was set to 20 MHz, which is the device's maximum capacity. In experiment 1 (**EXP1**), capture EM traces from the Dragon Board CPU while the smart device is connected through our custom WiFi network DB. In experiment 2 (**EXP2**), we run different activities on Dragon board (without connecting other smart devices) utilizing single core (Core #1 of SoC) and capture EM traces. We do the same for all devices individually. For this purpose, it is required to set the center frequency to HackRF One SDR and put it as 1.2 GHz according to the Dragon board processor clock frequency. Using the pre-described experimental setup, we captured EM traces from each device for the activities shown in Table 1. We also disabled frequency sink and waterfall sink when file sink was writing data into a raw data file with *.cfile* extension in the host computer. Two sampling methods are available for analyzing EM radiation captured through signal acquisition equipment, (1) real-valued sampling and (2) complex In-phase and Quadrature-phase (I/Q) sampling.[14] Due to the high expense of data-collecting equipment and the overhead of storing and processing massive amounts of data, we used complex In-phase and Quadrature-phase (I/Q) sampling method to build

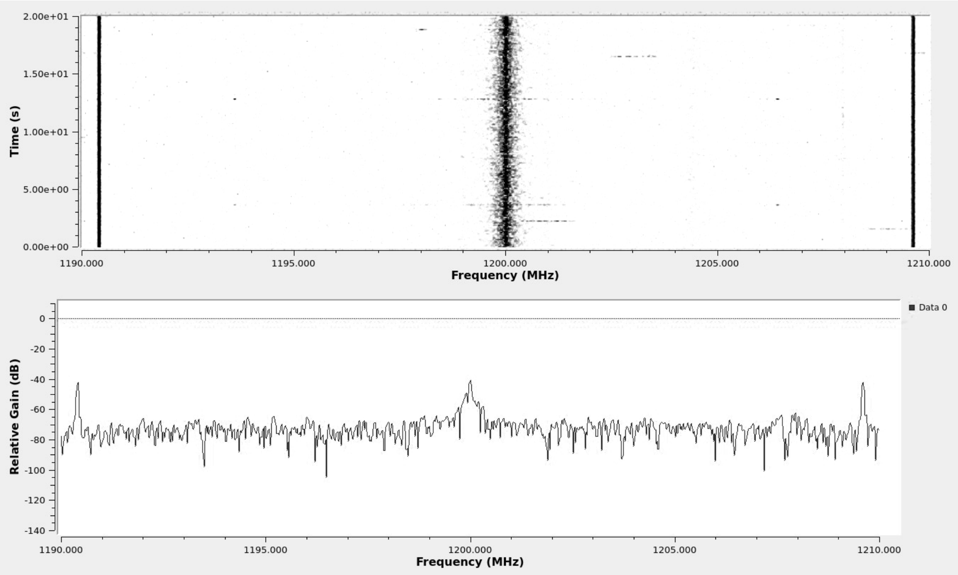


Fig. 3. The GNU Radio library visualization of EM radiation generating from Dragon board at 1.2 GHz processor clock frequency, captured with a sample rate of 20 MHz using an HackRF One.

Table 2. Statistics of datasets.

Device	Number of classes	Samples per class	Total samples per device
Dragon Board 1	4	30,000	120,000
Dragon Board 2	4	30,000	120,000
Echo Show 1	4	30,000	120,000
Echo Show 2	4	30,000	120,000

separate data files for each device activity shown in Table 1 and combine them to build the final dataset for each device in the latter stage (cf. Table 2). We keep each data file size under 3 GB due to processing limitations.

4. Experiments and Findings

To apply ML methods for data analytics, we used Short-time Fourier transform (STFT) for features extraction from the EM trace data.[14] In the experiments, we used 4096 I/Q samples as Fast Fourier Transform(FFT) size and an FFT overlap of 512 I/Q samples. The STFT converted dataset belongs to the label of the original smart device's software activity and it can be used for ML tasks.

From each class, 30,000 samples were used to build the final datasets for all devices. MinMaxScaler is used to normalize the data that transform numerical data to a standard scale, which can improve algorithm performance, data visualization, and comparability of different datasets. In the current analysis, one dataset from each device was first split into training and test with the ratio of 70:30. Keras sequential model was used to build a deep neural network with six (06) hidden layers, as this architecture proves effective in capturing intricate data patterns. The model was trained for 30 epochs to achieve convergence while maintaining computational efficiency. Allocating 10% of the dataset to the validation set facilitates a comprehensive assessment of the model's generalization capabilities. The created model from one device is used to test with the dataset of another device.

The accuracy, precision, recall, and F1-scores[14] were used to evaluate the performance of the models. We used a previous study[14] as a baseline to assess the results obtained from our tests on Dragon Board and Echo Show devices. The previous study was quite similar to ours in that it also involved capturing electromagnetic traces. The results of this study show 99% accuracy with a higher precision and F1-score for all classes of the Echo Show device. Moreover, they captured EM traces from other IoT devices like Google Home, Samsung SmarThings Hub, and several Smart Phones from different vendors. Their deep neural network-based models built using those EM traces captured from each smart device archived an accuracy of over 98.0%. Therefore, in our experiments on Dragon board, we use this observation as the baseline model.

4.1. *Challenges of using ML techniques in analyzing EM-SCA data in the crossed-IoT device portability context*

In **EXP1** Amazon Echo Show device, captured data belonging to four (04) activities. The final dataset contains 120,000 data samples for each amazon show device (cf. Table 2). Table 3 shows the classification report of the model build using the Echo Show device 1 dataset. This classifier achieved 91.0% training accuracy, and test accuracy of 89.0% with 90.50% macro precision, 89.0% macro recall, and 89.5% macro F1-score for the same device (Echo Show 1) data used to train the model. There is a significant difference between the training and test accuracy achieved by our classifier for the Echo Show 1 device (91% and 89%, respectively) compared to the accuracy achieved by the baseline study (99.66%). On the other hand, our Echo Show 1 model test with the Echo Show device 2 data shown in Table 7 only got the

Table 3. Classification report of the sequential Keras model for Amazon Echo Show device 1.

	Training data				Test data			
Label	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
0	0.96	0.83	0.89	0.91	0.95	0.80	0.87	0.89
1	1.00	1.00	1.00		1.00	1.00	1.00	
2	0.95	0.88	0.92		0.94	0.86	0.90	
3	0.77	0.93	0.84		0.73	0.90	0.81	

Notes: Dataset was split with a 70:30 ratio to training (70%) and test (30%) the model (0-asking to play music, 1-asking to play radio, 2-asking for time, 3-asking for weather today).

test accuracy of 25.0% with extremely low precision, recall, and F1-scores for most of the classes. These results suggest that the model trained on one device may not generalize well to other devices, highlighting the need for further analysis.

To address the first research question i.e. "How the multi-core architecture of SoC integrated on these devices affects the ML/DL models' accuracy?" (cf. **RQ1**, Sec. 3), we carried out the experiment **EXP2**, which investigates the issues identified in **EXP1**. Dragon boards were used as the Device Under Test (DUT), and separate Print, Math, Memory, and I/O activities were run on it while collecting CPU EM traces. **RQ1** assumed that using multiple cores in SoC architecture can affect EM emissions and make it more difficult to identify the right source of emissions for a particular activity. Therefore, it affects the quality of EM data captured and the overall accuracy of ML model dropped significantly (cf. Table 7). To prove this hypothesis, in **EXP2**, a single core of DUT is used for all activities in the experiment. The objective is to isolate the EM emissions from each activity. Moreover, the second research question assumed that the number of activities, which is also a factor that affects the ML model's accuracy (cf. **RQ2**, Sec. 3). To address this hypothesis, in **EXP2** we varied the number of activities in three (03) tasks as follows:

- **Task 1.** Involved in building a model for Dragon Board 1 for four (04) activities: Print, Math, Memory, and I/O. The model achieved 96% of training accuracy and test accuracy of 95% with 95% macro precision, 95% macro recall, and macro 94.75% F1-score on the training dataset (cf. Table 4). The model was tested with the Dragon Board 2 dataset for the same four activities. The model achieved an accuracy of 38% on the test dataset shown in Table 8.
- **Task 2.** Involved in building a model for Dragon Board 1 for three (03) activities: Print, Math, and Memory. The model achieved 97% of training accuracy and test accuracy of 96% with 96.30% macro precision, macro recall, and macro F1-score on the training dataset (cf. Table 5). The model was tested with the Dragon Board 2 dataset for the same three activities. The model achieved an accuracy of 52% on the test dataset shown in Table 9. The final dataset used for this task contained 90,000 data points.

Table 4. Classification report of the sequential Keras model for Dragon Board 1 for four (04) activities.

	Training data				Test data			
Label	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
0	0.97	0.98	0.97	0.96	0.96	0.98	0.97	0.95
1	0.98	0.96	0.97		0.97	0.95	0.96	
2	0.90	0.99	0.94		0.89	0.98	0.93	
3	0.99	0.90	0.94		0.98	0.89	0.93	

Notes: The dataset was split with a 70:30 ratio to training (70%) and test (30%) of the model (0-Print Task, 1-Math Task, 2-Memory Task, 3-I/O Task).

Table 5. Classification report of the sequential Keras model for Dragon Board 1 for three (03) activities.

	Training data				Test data			
Label	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
0	0.98	0.94	0.96	0.97	0.97	0.93	0.95	0.96
1	0.94	0.98	0.96		0.93	0.96	0.95	
2	0.99	1.00	1.00		0.99	1.00	0.99	

Notes: The dataset was split with a 70:30 ratio to training (70%) and test (30%) of the model (0-Print Task, 1-Math Task, 2-Memory Task).

- **Task 3.** Involved in building a model for Dragon Board 1 for two (02) activities: Math Task and Memory Task. The final dataset used for this task contained 60,000 data points. The model achieved training and test accuracy of 99% with 99.0% macro precision, macro recall, and macro F1-score on the training dataset (cf. Table 6). The Dragon Board 1 model was tested with the Dragon Board 2 dataset for the same two activities, Math and Memory tasks, and the accuracy dropped to 77% on the test dataset shown in Table 10.

The above results showed that the device variability is directly affected on the ML/DL model performance in EM-SCA (cf. Table 7). Moreover, we identified the number of cores in a SoC as well as the number of running activities can significantly affect the ML/DL model performance (cf. Tables 8–10). These findings consolidate our assumptions in **RQ1** and **RQ1**. Further evaluation and discussion are in Sec. 5.

4.2. *Analyzing the feasibility of using transfer learning*

Furthermore, to address the last research question **RQ3** (cf. Sec. 3) we employed transfer learning techniques to enhance the performance of our models developed on the Echo Show 1 and Dragon Board 1 devices. Transfer learning is a widely used approach in ML that enables us to leverage the knowledge gained from one task to improve the performance of another related task.[20] Specifically, we utilized a pre-trained model as the starting point and fine-tuned it on a new dataset or task, which allowed the model to benefit from the features learned by the pre-trained model and achieve superior performance on new datasets. To achieve this, we utilized a

Table 6. Classification report of the sequential Keras model for Dragon Board 1 for two(02) activities.

	Training data				Test data			
Label	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
0	1.00	0.99	0.99	0.99	1.00	0.99	0.99	0.99
1	0.99	1.00	0.99		0.99	1.00	0.99	

Notes: The dataset was split with a 70:30 ratio to training (70%) and test (30%) the model (0-Math Task, 1-Memory Task).

Table 7. Classification report of the Amazon Echo Show 1 model test with the Amazon Echo Show 2 dataset (0-asking to play music, 1-asking to play radio, 2-asking for time, 3-asking for weather today).

	Test data (crossed-devices)			
Label	Precision	Recall	F1-score	Accuracy
0	0.05	0.00	0.00	0.25
1	0.33	0.00	0.00	
2	0.25	1.00	0.40	
3	0.07	0.00	0.00	

Table 8. Classification report of the Dragon Board 1 model test with the Dragon Board 2 dataset for four(04) activities. (0-Print Task, 1-Math Task, 2-Memory Task, 3-I/O Task).

	Test data (crossed-devices)			
Label	Precision	Recall	F1-score	Accuracy
0	0.72	0.09	0.16	0.38
1	0.38	0.85	0.52	
2	0.35	0.57	0.44	
3	0.85	0.01	0.02	

pre-trained model and froze the weights of the pre-trained layers up to the first dense layer. Subsequently, we added a transfer dense layer on top of the pre-trained model to build a transfer learning Keras model. We then trained this model on the respective datasets for each device (Echo Show 2 and Dragon Board 2) in our experiments. We repeated this process for all three models developed in three tasks of **EXP2** (cf. Sec. 4.1).

For Echo Show 2 data, we significantly improved our models' accuracy by using a transfer learning approach, with training and test accuracy reaching 53% (cf. Table 11). Previously, Echo Show 2 data was archived with only 25% accuracy when tested on the Echo Show 1 model. This result shows the effectiveness of transfer

Table 9. Classification report of the Dragon Board 1 model test with the Dragon Board 2 dataset for three (03) activities. (0-Print Task, 1-Math Task, 2-Memory Task).

	Test data (crossed-devices)			
Label	Precision	Recall	F1-score	Accuracy
0	0.70	0.05	0.10	0.52
1	0.42	0.93	0.58	
2	0.82	0.56	0.67	

Table 10. Classification report of the Dragon Board 1 model test with the Dragon Board 2 dataset for two (02) activities. (0-Math Task, 1-Memory Task).

	Test data (crossed-devices)			
Label	Precision	Recall	F1-score	Accuracy
0	0.70	0.93	0.80	0.77
1	0.89	0.61	0.73	

Table 11. Classification report of the **transfer learning** Keras model for the Amazon Echo Show 2 dataset, after fine-tuning the pre-trained Amazon Echo Show 1 model.

	Training data				Test data			
Label	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
0	0.56	0.59	0.58	0.53	0.56	0.59	0.57	0.53
1	0.46	0.38	0.42		0.46	0.38	0.42	
2	0.65	0.70	0.67		0.65	0.69	0.67	
3	0.44	0.47	0.46		0.44	0.46	0.45	

Notes: The dataset was split with a 70:30 ratio to training (70%) and test (30%) the model (0-asking to play music, 1-asking to play radio, 2-asking for time, 3-asking for weather today).

learning in improving model performance on new devices. Similarly, for Dragon Board 2, we achieved remarkable improvements in the accuracy of our models by using transfer learning techniques. The training and test accuracy of the new model are 57% (cf. Table 12) for the four (04) activities run on the Dragon Board 2. The model improved from 38% tested using the Dragon Board 1 model (cf. Table 8). Besides, the training and test accuracy for the three (03) activities reached 68% and 67%, respectively (cf. Table 13). Previously, Dragon Board 2 data was achieved with only 52% accuracy when tested on the Dragon Board 1 model. We also achieved high accuracy of 88% and 87% for two activities on Dragon Board 2, which improved from 77% tested using the Dragon Board 1 model (cf. Table 14). Our approach of fine-tuning pre-trained models and adding new layers on top allowed us to extract

Table 12. Classification report of the **transfer learning** Keras model for the Dragon Board 2 dataset, after fine-tuning the pre-trained Dragon Board 1 model.

	Training data				Test data			
Label	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
0	0.53	0.54	0.54	0.57	0.53	0.55	0.54	0.57
1	0.62	0.60	0.61		0.62	0.60	0.61	
2	0.53	0.44	0.48		0.53	0.43	0.47	
3	0.59	0.69	0.63		0.58	0.69	0.63	

Notes: The dataset was split with a 70:30 ratio to training (70%) and test (30%) of the model (0-Print Task, 1-Math Task, 2-Memory Task, 3-I/O Task).

Table 13. Classification report of the **transfer learning** Keras model for the Dragon Board 2 dataset, after fine-tuning the pre-trained Dragon Board 1 model.

	Training data				Test data			
Label	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
0	0.60	0.50	0.55	0.68	0.60	0.52	0.56	0.67
1	0.63	0.71	0.67		0.64	0.70	0.67	
2	0.77	0.81	0.79		0.77	0.81	0.79	

Notes: The dataset was split with a 70:30 ratio to training (70%) and test (30%) of the model (0-Print Task, 1-Math Task, 2-Memory Task).

Table 14. Classification report of the **transfer learning** Keras model for the Dragon Board 2 dataset, after fine-tuning the pre-trained Dragon Board 1 model.

	Training data				Test data			
Label	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
0	0.86	0.90	0.88	0.88	0.86	0.89	0.87	0.87
1	0.89	0.86	0.87		0.89	0.85	0.87	

Notes: The dataset was split with a 70:30 ratio to training (70%) and test (30%) the model (0-MathTask, 1-Memory Task).

valuable features from pre-trained models and adapt them to new tasks, resulting in superior performance.

The above results proved **RQ3**'s assumption that transfer learning is an effective technique for analyzing EM-SCA datasets and overcomes device variability problem with the performance gain from 10% to 19% in terms of accuracy of ML/DL models used (cf. Tables 12–14). We will discuss these findings further in the following section.

5. Evaluation and Discussion

The results of our study demonstrate that ML methods can effectively analyze EM-SCA data to identify the software activities of smart IoT devices. In addition, our study also highlights the limitations of the ML-based approaches, including the need for careful data preprocessing and the potential impact of device variability and environmental factors on the model performance.

In this research, we identified device variability significantly impacts the model's performance in EM-SCA datasets. As a result, we observed low precision, recall, and F1-scores obtained when testing the models with different devices' data. It is essential to highlight that low precision means that the model identifies many false positives. In other words, the model incorrectly classifies data from one activity as belonging to another. Similarly, low recall means that the model needs many true positives. In other words, the model is not recognizing data from one activity as belonging to that activity. Last, a low F1-score means that the model has an overall low accuracy in correctly identifying data belonging to a specific activity. In contrast, the results obtained using our experiments significantly differ from the baseline study we used to validate our results. In the baseline study,[14] authors captured electromagnetic traces directly from the Echo Dot/Echo Show/Smart devices, while we used a Dragon board and Echo Show as a DUT in **EXP1** to capture the network traffic through the Dragon board as the EM traces. This difference in the data source could have led to variations in the quality and quantity of the captured data, which could impact the classification accuracy, precision, recall, and F1-scores.

Besides, most smart devices use multi-core CPUs to divide the workload and complete tasks more quickly. Using multiple cores can affect electromagnetic (EM) emissions, as the increased processing power can generate more electromagnetic interference (EMI). This interference can make it more difficult to identify the source of emissions for a particular activity, as the signals from multiple cores may overlap and create a complex and unpredictable electromagnetic environment. In **EXP2** we used only one (1) core out of four (4) cores in the Dragon Board CPU and ran different activities to investigate these problems. The results of **EXP2** showed a significant drop in accuracy when models trained on one device were tested on another, even if they were of the same model. In contrast, we observed a high correlation between the different activities, as the EM emissions from one activity can be mistaken for another (cf. Fig. 4). As the number of activities increases, the correlation between them

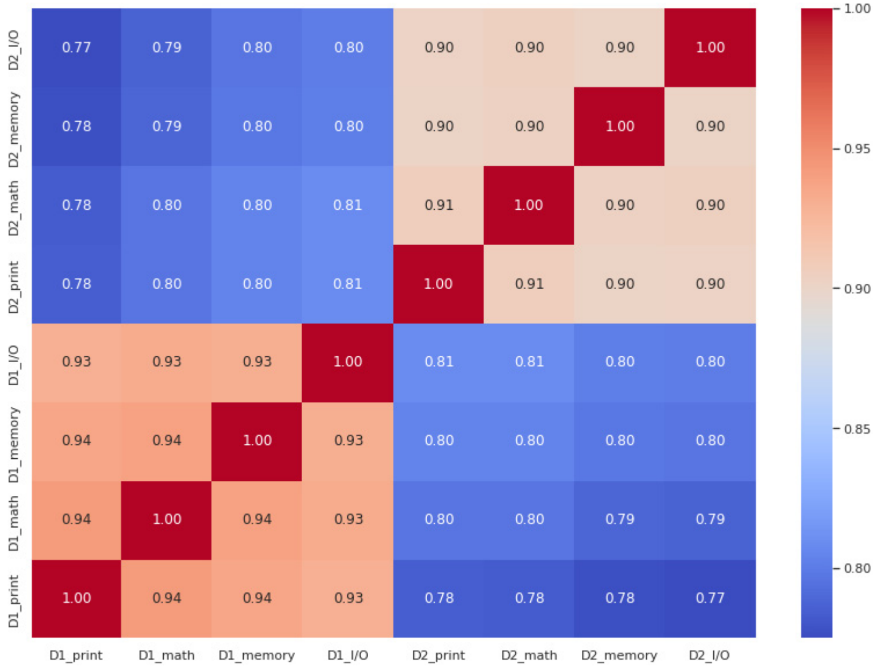


Fig. 4. A heat map was generated to perform correlation analysis for activities on Dragon Board 1(D1) and Dragon Board 2(D2). Only one core (core 1) was utilized for all the activities on both Dragon Boards.

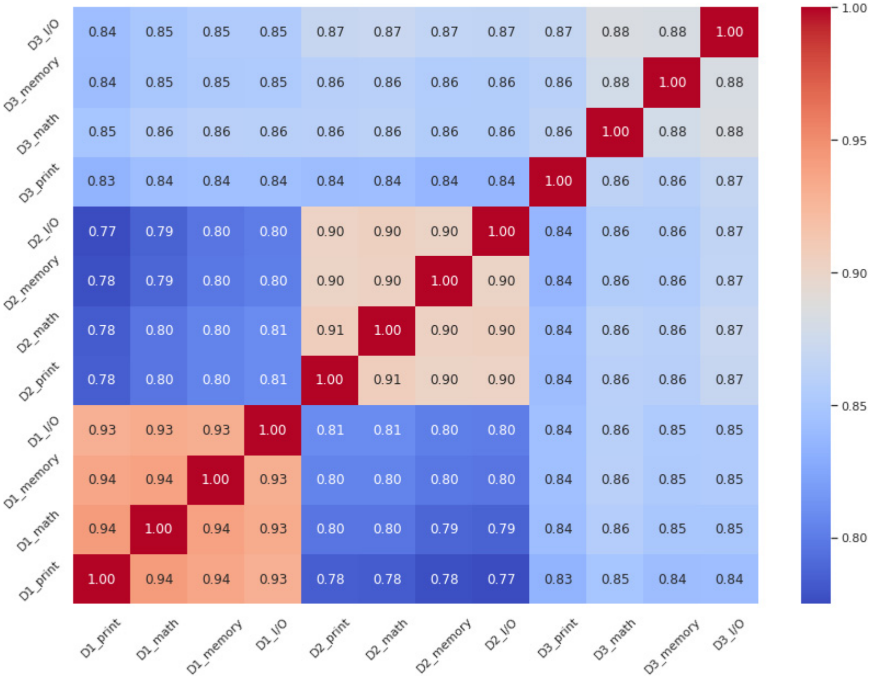


Fig. 5. A heat map was generated to perform correlation analysis for activities on Dragon Board 1(D1), Dragon Board 2(D2), and Dragon Board 3(D3). Only one core (core 1) was utilized for all the activities on all Dragon Boards.

also increases, making distinguishing between activities based on EM emissions more challenging. In contrast, if the number of devices increase may lead to a high correlation problem between similar activities (cf. Fig. 5). This can lead to more false positives, where the model identifies an activity as present when it is not. The higher number of false positives can further reduce the model's accuracy. Additionally, the experiment demonstrated that device variability significantly impacts the accuracy of the models, and test the models on datasets from multiple devices is essential to ensure their robustness.

Despite these differences, we still found the baseline study[14] to be a valuable point of comparison for our results. This outcome suggests that the models built with one device's data may not generalize well to other devices, and their accuracy may significantly decrease. The summary of results shows that EM-SCA datasets can vary significantly between devices, and therefore, it is essential to consider device variability when analyzing EM data. This finding is significant in practical applications where the goal is to identify and classify the activity of a device in a diverse range of real-world scenarios.

Our study also identified that environmental factors significantly affect the electromagnetic emissions of devices and the quality of the data collected. The low precision, recall, and F1-scores in these cases highlight the importance of training the models with data representative of the device and its operating environment. Any changes in the device or its environment can significantly affect the electromagnetic emissions of the device, leading to data that may not be representative of the device's normal operation. Therefore, environmental factors also affect models with low accuracy, as seen in our study. Further investigation is needed to optimize our data collection and preprocessing methods to improve the accuracy, precision, recall, and F1-scores of our classifier.

One limitation of our study is that we focused on a limited set of smart devices and software activities. Future research could explore the performance of ML models on a broader range of devices and activities and the impact of other factors, such as device age and usage patterns, on model performance. Another limitation of smart devices is the lack of control over the hardware level of each device. For example, the Echo Show has a 4-core CPU, and difficult to configure it to use a specific core, such as the Dragon board. This can have consequences in terms of electromagnetic (EM) emissions, as increased processing power can generate more EMI. This makes it more difficult to identify the source of emissions for a particular activity. With multiple cores in a CPU, signals can overlap and create a complex and unpredictable electromagnetic environment. The lack of control over the hardware level of smart devices can limit the ability to mitigate any potential EMI and cause the device variability problem.

Transfer learning involves using pre-trained models on similar tasks to improve the performance of a new model. In the context of our study, transfer learning shows promising outcomes on EM trace data from similar smart devices to improve the accuracy and generalizability of the models developed in this work. Also, transfer learning could be a promising approach for the future use of this work. Another potential direction for future research is to explore other data preprocessing techniques, such as wavelet transforms,[21] to improve feature extraction from EM trace data. In addition, further investigation is needed to identify the environmental factors that can affect the quality of the EM trace data and how to mitigate these effects in data collection and analysis.

Based on the analysis presented in this paper, several new research directions have been identified,

- **Multi-core problem for EM-SCA data:** This is an area of research that could explore ways to optimize the processing of EM-SCA data using multi-core architectures. Additionally, an important direction to investigate the impact of using multiple cores on EM emissions and how this affects the identification of the source of emissions. The number of cores in smart device CPU is potentially significant factor that may affect the accuracy of current EM analysis techniques.
- **High correlation between similar activities in different devices from the same model:** This could be another interesting research direction that examines the extent of correlation between activities across devices of the same model. It leads to insights into the consistency of sensor measurements across different

devices and potentially informs efforts to improve data quality. It highlights the need for more extensive data collection and analysis to accurately identify the source of EM emissions.

- **Transfer learning methods:** The study highlighted the potential for transfer learning to improve activity recognition performance, particularly in situations where training data not performing well on the ML models. Future research could focus on exploring the effectiveness of transfer learning in different scenarios, as well as developing new transfer learning techniques that are tailored to specific applications.

Overall, our study highlights the potential of ML for analyzing EM trace data in the context of side-channel analysis while also underscoring the need for careful consideration of data preprocessing, device variability, and environmental factors in model development and evaluation.

6. Conclusion

In conclusion, the growth of IoT devices has led to the emergence of IoT forensics, which plays a critical role in investigating and preventing malicious activities on IoT devices. EM-SCA has become an essential tool for IoT forensics due to its ability to reveal confidential information about the internal workings of IoT devices. However, the accuracy and reliability of EM-SCA results can be limited by several factors, including device variability, environmental factors, and data collection and processing methods. Our study provides insights into the challenges of developing ML models for the side-channel analysis of smart devices. Device variability and environmental factors may significantly affect the performance of those models.

Further research is required to improve the accuracy and reliability of EM-SCA results for IoT forensics and to overcome the limitations and challenges associated with EM-SCA datasets. This study provides a foundation for future work in this area. It presents an opportunity for the research community to address some critical issues associated with using EM-SCA in IoT forensics.

ORCID

Tharindu Lakshan Yasarathna https://orcid.org/0000-0002-7459-0600
Lojenaa Navanesan https://orcid.org/0009-0005-3071-6792
Asanka Sayakkara https://orcid.org/0000-0001-9558-7913
Nhien-An Le-Khac https://orcid.org/0000-0003-4373-2212

References

1. P. Gokhale, O. Bhat and S. Bhat, Introduction to IoT, *Int. Adv. Res. J. Sci., Eng. Technol.* **5**(1) (2018) 41–44.
2. H. F. Atlam, E. E.-D. Hemdan, A. Alenezi, M. O. Alassafi and G. B. Wills, Internet of things forensics: A review, *Internet Things* **11** (2020) 100220.

3. M. Conti, A. Dehghantanha, K. Franke and S. Watson, Internet of things security and forensics: Challenges and opportunities (2018).
4. A. Ghosh, M. Nath, D. Das, S. Ghosh and S. Sen, Electromagnetic analysis of integrated on-chip sensing loop for side-channel and fault-injection attack detection, *IEEE Microw. Wirel. Compon. Lett.* **32**(6) (2022) 784–787.
5. A. Sayakkara, N.-A. Le-Khac and M. Scanlon, Facilitating electromagnetic side-channel analysis for IoT investigation: Evaluating the EMvidence framework, *Forensic Sci. Int.: Digit. Investig.* **33** (2020) 301003.
6. A. Sayakkara, N.-A. Le-Khac and M. Scanlon, A survey of electromagnetic side-channel attacks and discussion on their case-progressing potential for digital forensics, *Digit. Investig.* **29** (2019) 43–54.
7. A. Kamilaris and F. X. Prenafeta-Bold, Deep learning in agriculture: A survey, *Comput. Electron. Agric.* **147** (2018) 70–90.
8. B. Sri Revathi, A survey on advanced machine learning and deep learning techniques assisting in renewable energy generation, *Environ. Sci. Pollut. Res.* **30** (2023) 93407–93421.
9. F. Zennaro, E. Furlan, C. Simeoni, S. Torresan, S. Aslan, A. Critto and A. Marcomini, Exploring machine learning potential for climate change risk assessment, *Earth-Sci. Rev.* **220** (2021) 103752.
10. R. Bhardwaj, A. R. Nambiar and D. Dutta, A study of machine learning in healthcare, *2017 IEEE 41st Annual Computer Software and Applications Conf.*, 4–8 July 2017, Turin, Italy, pp. 236–241.
11. K. Shaukat, S. Luo, V. Varadharajan, I. A. Hameed and M. Xu, A survey on machine learning techniques for cyber security in the last decade, *IEEE Access* **8** (2020) 222310–222354.
12. F. Shahzad, A. R. Javed, Z. Jalil and F. Iqbal, Cyber forensics with machine learning, in *Encyclopedia of Machine Learning and Data Science*, eds. D. Phung, G. I. Webb and C. Sammut (Springer, US, New York, NY, 2020), pp. 1–6.
13. A. P. Sayakkara and N.-A. Le-Khac, Forensic insights from smartphones through electromagnetic side-channel analysis, *IEEE Access* **9** (2021) 13237–13247.
14. A. P. Sayakkara and N.-A. Le-Khac, Electromagnetic side-channel analysis for IoT forensics: Challenges, framework, and datasets, *IEEE Access* **9** (2021) 113585–113598.
15. S. Hong, D.-H. Kim and H. Lee, Extracting encryption key from an IoT device using electromagnetic side-channel analysis, *J. Ambient Intell. Humaniz. Comput.* **8**(1) (2017) 129–142.
16. D.-H. Kim, S. Hong and H. Lee, Detecting malware in IoT devices using electromagnetic side-channel analysis, *J. Ambient Intell. Humaniz. Comput.* **9**(1) (2018) 69–80.
17. L. Zhao, X. Wang, W. Li and M. Gao, Limitations in electromagnetic side-channel analysis datasets for IoT forensics, *J. Ambient Intell. Humaniz. Comput.* **10**(1) (2019) 55–66.
18. J. Chen, S. Lu, W. Zhang and X. Wang, Machine learning techniques in electromagnetic side-channel analysis for IoT forensics, *J. Ambient Intell. Humaniz. Comput.* **11**(1) (2020) 43–54.
19. A. Levina, D. Sleptsova and O. Zaitsev, Side-channel attacks and machine learning approach, in *2016 18th Conf. Open Innovations Association and Seminar on Information Security and Protection of Information Technology (FRUCT-ISPIT)* (IEEE, 2016), pp. 181–186.
20. F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong and Q. He, A comprehensive survey on transfer learning, *Proc. IEEE* **109**(1) (2020) 43–76.
21. D. Zhang and D. Zhang, Wavelet transform, in *Fundamentals of Image Data Mining: Analysis, Features, Classification and Retrieval* (2019), pp. 35–44.