The Day-After-Tomorrow: On the Performance of Radio Fingerprinting over Time

Saeif AlHazbi
salhazbi@hbku.edu.qa
Division of Information and Computing Technology, College of Science and Engineering, Hamad Bin Khalifa University
Doha, Qatar

Savio Sciancalepore
s.sciancalepore@tue.nl
Eindhoven University of Technology
Eindhoven, Netherlands

Gabriele Oligeri
goligeri@hbku.edu.qa
Division of Information and Computing Technology, College of Science and Engineering, Hamad Bin Khalifa University
Doha, Qatar

ABSTRACT

The performance of Radio Frequency (RF) Fingerprinting (RFF) techniques is negatively impacted when the training data is not temporally close to the testing data. This can limit the practical implementation of physical-layer authentication solutions. To circumvent this problem, current solutions involve collecting training and testing datasets at close time intervals—this being detrimental to the real-life deployment of any physical-layer authentication solution. We refer to this issue as the Day-After-Tomorrow (DAT) effect, being widely attributed to the temporal variability of the wireless channel, which masks the physical-layer features of the transmitter, thus impairing the fingerprinting process.

In this work, we investigate the DAT effect shedding light on its root causes. Our results refute previous knowledge by demonstrating that the DAT effect is not solely caused by the variability of the wireless channel. Instead, we prove that it is also due to the power cycling of the radios, i.e., the turning off and on of the radios between the collection of training and testing data. We show that state-of-the-art RFF solutions double their performance when the devices under test are not power cycled, i.e., the accuracy increases from about 0.5 to about 1 in a controlled scenario. Finally, we show how to mitigate the DAT effect in real-world scenarios, through pre-processing of the I-Q samples. Our experimental results show a significant improvement in accuracy, from approximately 0.45 to 0.85. Additionally, we reduce the variance of the results, making the overall performance more reliable.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

Physical-Layer Security, Authentication, I-Q Data

ACM Reference Format:
Saeif AlHazbi, Savio Sciancalepore, and Gabriele Oligeri. 2023. The Day-After-Tomorrow: On the Performance of Radio Fingerprinting over Time. In *Annual Computer Security Applications Conference (ACSAC '23), December 4–8, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3627106.3627192

ACSAC '23, December 4–8, 2023, Austin, TX, USA

1 INTRODUCTION

Radio Frequency Fingerprinting (RFF) techniques have attracted the attention of researchers working in the wireless security domain, as a way of authenticating wireless transmitters by identifying unique patterns from received signals [34]. Indeed, RFF promises an effective and efficient way to authenticate the transmitting source without involving any crypto technique, thus being particularly suitable for scenarios where devices are characterized by strict battery constraints and high exposure to spoofing attacks. The transmitter does not require any additional computation or transmission, while the authentication process is completely offloaded to the receiver side, i.e., the receiver matches a pre-trained model of known transmitters to patterns extracted from the received signals, thus being able to authenticate the source. Such patterns are indeed unique, being the side-effect of unwanted and unpredictable phenomena such as manufacturing inaccuracies during the production process at the sub-millimetre level and electronic components' impurities. Such (small) differences eventually affect radio signals, which are detectable by receivers that combine Software-Defined Radio (SDR) capabilities with advanced Artificial Intelligence (AI) tools, such as the ones based on Machine Learning (ML) and Deep Learning (DL) [16]. At the time of this writing, extensive research is available on RFF. Some of the contributions focused on wireless communication technologies such as LTE [1], WiFi [17], Zigbee [5], Bluetooth [4], LoRa [31], and ADS-B [18], to name a few. Other works investigated the suitability of several AI-based techniques for addressing the RFF problem, either re-adapting well-known neural networks or casting ad-hoc solutions tailored to the specific technologies and data [35], [39], [12]. At the same time, a few works highlighted reliability issues pointing out phenomena that prevent the real-life deployment of such techniques. These works include the challenges of carrying out reliable training [14], use of multiple customized signal processing techniques by the devices under test [6], nonlinear characteristics of the power amplifiers making the fingerprint unpredictably dependent on the transmission power [19], interplay with heat dissipation, operations in different temperature conditions [25], aging of the devices and, last but not least, variable channel conditions [3]. In this context, some recent authoritative scientific contributions, such as [3] and [14], found that training RFF models on one day and testing on another day produces very poor performance, with a drop in the achieved classification accuracy of about 0.5. In this paper, we refer to such a phenomenon as the Day-After-Tomorrow (DAT) effect. Note that such studies adopt state-of-the-art classifiers based on the Convolutional Neural Network (CNN) *Resnet-50*, re-adapted to accept raw physical-layer signals (i.e., I-Q samples) as the input sequence, where the considered input size was either $N \times 1$ or $N \times 2$, being N the size of the input layer of the CNN. The RFF community attributes the performance drop *mainly* to the variability of the radio channel, thus proposing several techniques to mitigate the impact of radio channel impairments on classification accuracy.

Contribution. In this paper, we provide several contributions. First, we reproduce the DAT effect in the same setup of previous works, while we identify and expose another root cause of the performance loss behind the DAT effect itself by considering several experiments in different wired and wireless scenarios. Although our analysis confirms that channel impairments affect classification accuracy, we prove that the measurement methodology also has a significant impact on performance. Specifically, we show that the radio's power cycle, i.e., switching on and off both the transmitter and the receiver, significantly affects the classification accuracy when training and testing are performed on datasets collected before and after the power cycle of the radio itself. We verified our assumptions with a wired link between the transmitter and the receiver, thus excluding all the phenomena associated with radio channel variability. Subsequently, we consider a wireless link running for several days and propose a new methodology to mitigate the DAT effect, exploiting the pre-processing of the I-Q samples. Inspired by recent results in the area, we refined the technique of converting the I-Q samples into images, achieving accuracy values that are (on average) twice better than the ones experienced with raw I-Q samples, while halving the variability associated with the reported accuracy. The data used for our analysis are available on request.

Roadmap. The paper is organized as follows. Sect. 2 reviews some related work on the robustness of RFF, Sect. 3 introduces preliminary concepts, Sect. 4 provides the details of our measurement campaign, Sect. 5 provides an in-depth analysis of the DAT effect describing the impact of radio power cycle behind it. We also propose a pre-processing technique applying to the I-Q samples to mitigate the DAT effect when the power cycle is strictly required. Sect. 6 summarizes our findings and limitations and, finally, Sect. 7 concludes the paper and outlines future work.

2 RELATED WORK

RFF strategies include a set of techniques to identify and authenticate RF devices by using distinctive patterns in emitted signals [16]. Such patterns originate from hardware imperfections in the devices introduced during the manufacturing processes. In the early stage, research on RFF focused primarily on developing custom feature extraction methods using ML and DL techniques, as shown by [29], [30], [21], [37], and [7], to name a few. Although significant progress has been achieved in the development of highly accurate methods for extracting RF features from over-the-air signals, the deployment of RFF systems in the real world faces many challenges. One major limitation of RFF systems based on DL algorithms is their sensitivity to wireless channel variability, which can negatively affect their performance. This problem has been reported in several studies [3], where the authors found that training a DL model on data collected in one day and then testing it on data collected in a different day significantly reduces the classification accuracy. In their experiment, the authors trained three DL models on a dataset of 20 wireless transmitters collected over several days in different environmental settings, including a cable, an anechoic chamber, and in the wild. They showed that the performance of the models was not consistent on different days, indicating that the models were unable to generalize well to new environments or conditions. Throughout this paper, we refer to such phenomenon as the DAT effect, causing a dataset shift [23]. The findings cited were corroborated by the work of the authors in [15], who also found that changing the receiver used to collect RF signals during training can further degrade the performance of the models. Such findings highlight that the final captured RF fingerprint is a combination of three distinct factors: the transmitter's emitted signal, the channel, and the receiver's hardware. The drop in performance when training and testing on different datasets has also been observed by [14]. The authors explained such phenomenon as "changes in channel conditions", as per Fig.4 in [14]. We stress that neither [14] nor [3] provided the details on the power cycle of the considered devices, making the comparison with this work impossible. The authors in [9] studied the sensitivity of RFF systems for LoRa networks in a wide range of scenarios, including indoor and outdoor environments, wired and wireless setups, various distances, configurations, hardware receivers, and locations. According to previous research, they found that testing on different days and using different receivers can significantly impact the accuracy of the RFF process. Furthermore, they observed that the variability of protocol configuration and location also has notable effects on the achievable classification accuracy. To address the challenges cited, several mitigation techniques have been proposed in the literature. An approach is to augment the training data by exposing the fingerprinting process to a variety of channel conditions and environments, as demonstrated in [2, 13, 33]. In particular, [32] investigated how carrier frequency offset (CFO) affects RFF while proposing an ad-hoc classification algorithm to mitigate the phenomenon. Another approach uses the idea of injecting unique impairments into the transmitted signal, such as in [22, 26, 30]. Other methods involve using channel modelling and simulations, as in [36], or using digital signal processing techniques with specialized filters, as in [27, 28]. However, none of these works provides an in-depth analysis of the DAT effect; indeed, they do not consider the impact of the radio power cycle on the RFF and they do not provide any specific solution that can deal with a real deployment. In fact, although Radio Frequency (RF) fingerprinting is a promising technique for authenticating RF devices at the PHY layer, it still faces many limitations and challenges. In this work, we further advance the analysis of the performance of the RFF in real-world scenarios by shedding light on the root causes that affect its performance. Finally, we compare our contribution with the reference literature as per Table 1 in terms of the adopted communication technology, receiver radio hardware, adopted neural network, communication medium, measurement duration, and finally experienced phenomena.

Table 1: Comparison with related work.

Reference	Comm. Tech.	Receiver Radio Tech.	Network	Comm. Medium	Measurement duration	Observed phenomena
[14]	LoRa	USRP B210	Self-designed CNN	Wireless	Multiple days	Impact of environment and measurement time
[32]	LoRa	USRP N210	Self-designed MLP/CNN/LSTM	Cable	Multiple days	Impact of measurement time
[3]	WiFi/ADS-B	USRP N210 / X310	Self-designed and ResNet50	Wireless	Multiple days	Impact of measurement time
[2]	LoRa	USRP N210	Self-designed LSTM/CNN	Wireless	Few days	Impact of measurement time
[27, 28]	WiFi/ADS-B	USRP X310	Self-designed CNN	Wireless	Multiple days	Impact of channel conditions
[29]	IEEE802.11ac	USRP B210	Inspired by AlexNet	Wireless	Same day	Impact of channel conditions and distance
[30]	WiFi	USRPX310	Self-designed CNN	Wireless	Same day	Impact of channel conditions
[21]	ZigBee	Rohde & Schwarz FSW67	Self-designed CNN	Wireless	Same day	Impact of channel conditions
[37]	ZigBee	USRP N210	Self-designed CNN	Wireless	Same day	Impact of channel conditions
[7]	PHY - QPSK	USRP N210 / B210	Self-designed CNN	Wireless	Same day	Impact of channel conditions
[32]	LoRa	USRP N210	Self-Designed CNN	Wireless	Multiple days	Impact of CFO
Our contribution	**PHY - BPSK**	**USRP X310**	**ResNet50**	**Wireless and Cable**	**Multiple days**	**Impact of channel conditions and power cycle**

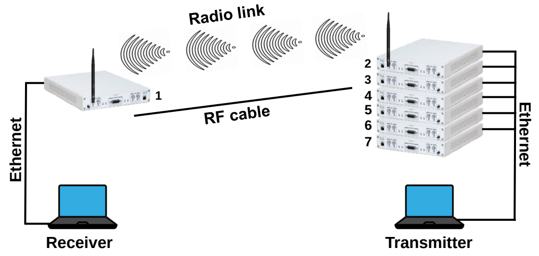


Figure 1: Hardware setup: our measurement setup consists of 1 receiver and 6 transmitters communicating via either a wireless or a wired link, depending on the scenario.

3 PRELIMINARIES

In this section, we introduce the hardware and software setup adopted in this work and the required background concepts related to RF communication and DL for radio fingerprinting.

Hardware Set-up. The hardware considered in this paper includes seven (7) SDRs USRP X310 [11], featuring the UBX160 daughterboard and the VERT900 antenna [10]. The general setup is depicted in Fig. 1, showing that the radios are connected to two laptops HP EliteBook I7, featuring 32GB of RAM. All the considered scenarios involve radio 1 (on the left side in Fig. 1) as the receiver, while we consider the other ones as the transmitters (only one active for each experiment). Depending on the specific scenario, we connected the transmitter and the receiver via either an RF (wireless) or a wired link. In the former, we considered a transmitting power of 35mW and a normalized receiver gain of 1, while when using a wired link we set the transmission power to 1 mW and the normalized receiver gain to 0.8, where the normalized receiver gain is defined according to the logic in the USRP Source block provided by GNURadio (see software setup below). Finally, for the wired link, we considered a coaxial cable type RG58A/U. We do note that other work in the RFF research area might consider a larger number of transmitter radios. However, our findings confirm their performance, while paving the way for future research in the area.

Software Setup. We adopted GNURadio v3.8 for the measurements, defining a transmission chain with the following blocks:

- *File source.* We generated a message including a string of 256 bytes with incremental value, i.e., $[0, \ldots, 255]$. Note that specifying a valid message is necessary to draw any conclusions on the effectiveness of the TX-RX chain, as well as on the bit-error rate experienced by the communication link.
- *Constellation modulator.* We configured the modulator according to the Binary Phase Shift Keying (BPSK) modulation.
- *UHD: USRP Sink.* We set the transmission frequency to 900 MHz (carrier), with a sample rate of 1M samples per second, and finally, the transmission power of either 35mW or 1mW for the wireless and the wired link, respectively.

We configured the receiver according to the following chain:

- *UHD: USRP Source.* We set the receiving frequency at 900 MHz (carrier), with a sample rate of 1M samples per second and a normalized receiver gain of 1 or 0.8 depending on the adopted link, being wireless or wired, respectively. Note that when receiving complex I-Q values, the modified Nyquist theorem requires that the sample rate at the receiver should be at least equal (or higher) to the sample rate at the transmitter [8]. This is due to the fact that at each sampling instant, we acquire two samples (I and Q). Note that this is the standard configuration to receive and demodulate a 1M bandwidth signal transmitted at the frequency of 900MHz.
- *AGC.* We included the Adaptive Gain Control block to mitigate channel fluctuations.
- *Symbol Sync.* We included a symbol synchronizer to receive and decode digital signals based on the maximum likelihood estimation (mle) criterion.
- *Costas Loop.* We adopted the Costas loop to mitigate the phase offset and residual frequency offsets.
- *File Sink.* The output of the receiving chain is stored inside a file, for follow-up analysis.

I-Q Samples. Given the carrier frequency f_0, which in our case is 900 MHz, a digital modulation scheme can be described through Eq. 1 [20].

$$x(t) = I\cos(2\pi f_0 t) + Q\sin(2\pi f_0 t), \tag{1}$$

where $x(t)$ is the transmitted modulated signal, I is the *in-phase* component, while Q is the *quadrature* component. For the specific modulation scheme considered in this work, that is, BPSK, the

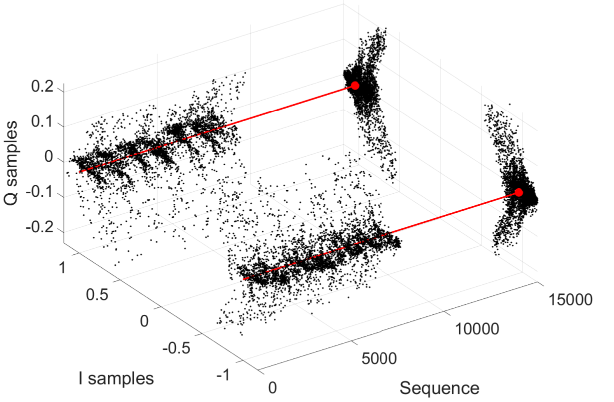


Figure 2: I-Q samples: solid red lines represent the theoretical position of the I-Q samples for a BPSK modulation scheme, i.e., [-1, 0] and [1, 0], while the dispersion of their values is mainly caused by the transducers' imperfections—these samples are coming from a wired measurement—thus representing the fingerprint of the radio.

quadrature component always has a null value ($Q = 0$), while the in-phase component is used to translate the value of the bit b, that is, $I = -1$ and $Q = 0$ when $b = 0$, and $I = +1$ and $Q = 0$ when $b = 1$, or vice versa, as shown by Eq. 2.

$$x(t) = \begin{cases} +1\cos(2\pi f_0 t), & \text{if b == 1} \\ -1\cos(2\pi f_0 t), & \text{if b == 0} \end{cases} \\ = \cos(2\pi f_0 t + \phi), \qquad (2)$$

where ϕ takes on the value of either 0 or π as a function of the value of the bit. Since I and Q constitute an alternative way of representing the magnitude and phase of the modulated signal $x(t)$, it is natural to consider the components I and Q as the real and imaginary parts of a complex number, respectively. In particular, for the BPSK, there is no imaginary part ($Q = 0$), while we only consider the real component, that is, $I = \pm 1$. Given a sequence of bits, the transmitter implements Eq. 2 to translate bits into I-Q samples, while the receiver takes on the challenge of reversing each IQ sample to the original value of the bit. As a toy example, we consider Fig. 2, which represents a sequence of I-Q samples over time. The measurement consists of 8000 I-Q samples (black dots) collected by using a wired link between the transmitter and the receiver, while the red lines and the associated red dots (at the end of the lines) represent the theoretical (ideal) position of the I-Q samples, that is, [1, 0] and [-1, 0]. Due to radio imperfections, the I-Q samples spread over the I-Q plane with a specific pattern peculiar to the adopted combination of transmitter and receiver radios. Finally, we report the projection of the I-Q samples at the bottom of the figure to provide a more concrete representation of the actual spreading of the I-Q samples, i.e. the *radio fingerprint.*

Deep Learning. We denote *radio fingerprinting* as the task of identifying unique features at the physical-layer (I-Q samples) that can be used to discriminate a radio transmitter. A radio fingerprint is a unique pattern in the I-Q samples, like the one depicted in Fig. 2. Many researchers have taken on the challenge of identifying a robust fingerprint and developing a methodology to use it efficiently and effectively. State-of-the-art solutions involve a two-stage process: (i) *training* a neural network model with chunks of I-Q samples, and (ii) *testing* a sequence (of I-Q samples) from the wild to identify the actual transmitter. Among the several neural networks proposed in the literature, the family of Residual Neural Networks (RNN) emerged as a good trade-off between training speed and classification performance. In particular, *resnet50* is the one adopted in this work, in line with other contributions which highlighted the temporal variation of the fingeprint [3], [14], [24], while many other contributions adopt similar networks, changing a subset of its layers. The choice of *resnet50* (as also discussed in [24]) is based on an empirical verification of its performance while considering different input configurations. Specifically, we consider the *resnet50* implemented in MatLab R2022b®, constituted by 50 layers and pre-trained on the ImageNet database. The neural network, as originally designed, cannot be used for the training and classification of the I-Q samples, and researchers developed different ways to modify it to fit the radio fingerprinting problem. Indeed, *resnet50* is designed to classify images from the ImageNet dataset, thus requiring some modifications at both the input and output layers. The input layer is usually adapted to fit raw I-Q samples (and not images), while the output layer is changed to fit the number of transmitters, being different from the 1000 classes of ImageNet. Moreover, the network itself is already trained to classify images such as dogs, cats, flowers, etc.—these being different from the input adopted for the RF fingerprinting; therefore, the network requires a partial re-train to expose the model to the new input. In detail, the output layers (*fullyConnectedLayer* and *classificationLayer*) should be re-adapted to take into account the number of classes in the radio fingerprinting problem, i.e., the number of radio transmitters to identify. Regarding the input layers, two main configurations have been considered: (i) input of interleaved raw I-Q samples consisting of a vector of dimensions either $N \times 1 \times 1$, as in [3], or $N \times 2 \times 1$, as in [14], and (ii) images where the input is a matrix of size $224 \times 224 \times 3$, where 224×224 is the size of the images in the ImageNet dataset, as in [24].

In the remainder of this paper, we will consider both the approaches of raw I-Q samples and images, to compare the performance and highlight their limitations. Figure 3 shows how we deployed *resnet50* to identify the transmitter given the I-Q samples collected by the receiver, considering the pre-processing approaches discussed above.

4 MEASUREMENT COLLECTION

We define *measurement* a continuous flow of I-Q samples between the transmitter and the receiver. As will be clarified in the following, we consider two types of measurements: (i) short measurements lasting 300 seconds (5 minutes) and (ii) long-term measurements lasting 4 days. In the remainder of this work, we refer to I-Q samples as complex-valued data; thus, each radio sample (bit) is constituted by one I and one Q sample. Given the availability of 6 transmitters, we define *run* of measurements as the sequence of 6 consecutive

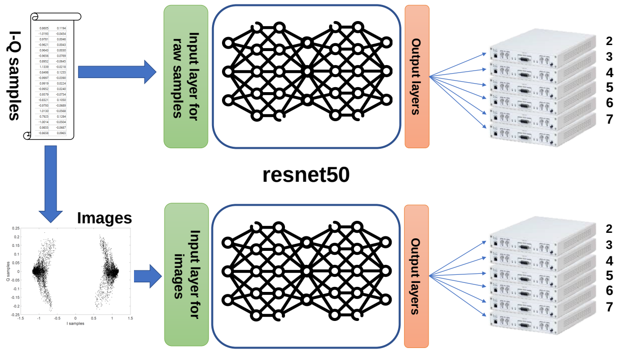


Figure 3: The RFF process: a neural network (*resnet50*) is re-adapted to fit either raw I-Q samples or images (input layers) and to identify the correct number of radios (output layers).

measurements where the receiver is the same while the transmitter changes every time among all the available ones. Finally, we define *dataset* as a sequence of several runs of measurements. All the datasets characterized by a wireless link share the same scenario, that is, an indoor environment where the transmitter is located 10 meters from the receiver, without the line-of-sight (nLoS). During our measurement campaign, we collected three datasets of measurements, listed below and summarized in Tab. 2. All the measurements can be requested by email to the authors of the paper.

(1) *Dataset 1 (DS1).* This dataset is constituted of 78 measurements, organized in 13 runs, collected by using a wired link and power-cycling the radios (both the transmitter and the receiver) after each measurement. We kept the same receiver for all measurements while we changed the transmitter (using 6 different radios, in total).
(2) *Dataset 2 (DS2).* This dataset is constituted by 6 non-stop measurements collected by using a wired link for 3 days for each measurement ($6 \times 3 = 18$ days, 265B+ samples). All the measurements share the same receiver, while we changed the transmitters (using 6 different radios in total, as for the previous dataset). Given the duration of the measurements, we had to decrease the sample rate to 256K samples per second.
(3) *Dataset 3 (DS3).* This dataset is constituted by 72 measurements, collected using the wireless radio link (using the carrier frequency $f = 900$ Mhz) over 4 days with a distance between the transmitter and the receiver of 10 meters. The dataset has been collected on the ground floor of a villa where the transmitter has been set up in the living room and the receivers in the kitchen (no line of sight), with 2 people randomly crossing the measurement set-up. For each day, we collected 3 runs of measurements (6 measures each) early in the morning, noon, and late in the evening, power-cycling the devices in-between the measurements. To be consistent with the other measurements, we considered the same radio as the receiver, while we swapped the transmitters among the 6 available radios.

Table 2: Measurements description: We collected three datasets over wired and wireless links, lasting multiple days.

	Link	Sample Rate [Msps]	Duration [Days]	Runs	Samples per Measurement
DS1	Wired	1	3	13	144M+
DS2	Wired	0.256	18	1	33B+
DS3	Radio	1	4	12	144M+

Note that we consider multiple transmitters, but only a single receiver. This is done on purpose to replicate the traditional setup adopted in the literature for RF fingerprinting. Analyzing the impact of a different receiver on the performance of RFF solutions is out of the scope of this contribution, as well as part of our future work.

5 THE DAY-AFTER-TOMORROW EFFECT

In this section, we first summarize the methodology followed in our investigation (Sect. 5.1), and subsequently, we provide an in-depth analysis of the DAT effect (Sect. 5.2). Moreover, in Sect. 5.3, we introduce a mitigation technique for the DAT effect leveraging the pre-processing of I-Q samples into images and, finally, in Sect. 5.4, we deploy our solution to a real wireless radio link. In the following, we use the qualitative terms *low* and *high* associated with the accuracy of the classifier just for the sake of simplicity, as well as to introduce the problem discussed in our manuscript. A quantitative analysis of the phenomena involving real measurements, taken with both wireless and wired links, will be presented in the subsequent sections to support our claims.

5.1 Methodology

This section discusses the *Day-After-Tomorrow* effect, i.e., a common problem affecting all the approaches dealing with RFF presented in the literature and mentioned explicitly in some of them, e.g., [3] and [14]. We describe the phenomenon through real measurements, whose logic is depicted in Fig. 4 using two experiments (E1 and E2). We highlight that experiments E1 and E2 depict a radio link, but, in the following, we might consider either a wireless or a wired link, depending on the objectives of the experiments. For both experiments (E1 and E2), two measurements (red boxes) are collected. During the first experiment (E1), the measurements (M1 and M2) are taken on two different days (Day 1 and Day 3). In contrast, during the second experiment (E2), only one measurement is taken (one red box) and then split into two chunks (blue boxes), M3 and M4 respectively, taken on Day 1 without interrupting the measurement process. For both experiments, we use one measurement to train a neural network model, i.e., M1 in E1 and M3 in E2, while we consider the second measurement for testing, i.e., M2 in E1 and M4 in E2. We refer to the *Day-After-Tomorrow* (DAT) effect as the phenomenon where the first experiment E1 is characterized by low classification accuracy compared to the second experiment E2, which experiences a high classification accuracy. This is a well-known effect in the literature, reported by recent authoritative contributions such as [3] and [14], to name a few. Such works explain the DAT effect by referring to the unpredictability of the wireless radio channel and its associated phenomena, e.g., multipath and fading. We highlight that the DAT effect significantly hinders the deployment of

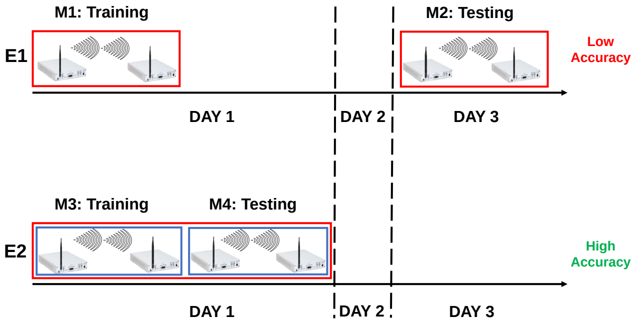


Figure 4: The *Day-After-Tomorrow* effect is constituted by two experiments (E1 and E2): training a model on a measurement (M1) taken one day (DAY 1) and then testing on a measurement (M2) taken on another day (DAY 3) gives low accuracy performance. Conversely, training and testing on chunks (M3 and M4) taken from the same measurement (DAY 1) gives high accuracy performance.

radio fingerprinting techniques in any real-world scenarios since the model trained at a specific time cannot be effectively re-used in the future without reporting significant performance loss. While we acknowledge the impact of the radio channel in the fingerprinting process, in the following, we formulate a different hypothesis rooted in real experiments involving both wireless and wired links. Our intuition is that the loss of performance described above is not caused by the channel variations only—although the channel might play an important role—but by another phenomenon, i.e., the *power cycle* of the radios in-between the measurements used for the training and the testing tasks.

Definition 1. *We define* power cycle *as the process involving the software (re-)initialization of the radio. This takes place by applying the power-off/power-on of the radio.*

Figure 5 summarizes the experiments we conducted to expose the phenomenon described above. Although the experiments depict a radio link, in the following, we might consider either a wireless or a wired link depending on the objectives of the experiments. Specifically, we consider two additional experiments (E3 and E4): during E3, two measurements (M5 and M6) are collected by power-cycling the transmitter and the receiver in-between the measurements, i.e., switching off and on the radios. We stress that M5 and M6 are taken one immediately after the other, i.e., very close in time, by only switching off and on the radios (power-cycling). We use M5 for training and M6 for testing. We observed a low classification accuracy for both the wireless and the wired scenario. Finally, we consider another experiment (E4) constituted by a long measurement (spanning 3 consecutive days) where no power cycle is performed in between the measurements. From this experiment, we extract two chunks, i.e., M7 and M8. When considering M7 for training and M8 for testing, the resulting classification accuracy is high. In the following sections, we also show that it is possible to mitigate this phenomenon and significantly increase the classification accuracy for measurements separated by a power cycle (E3).

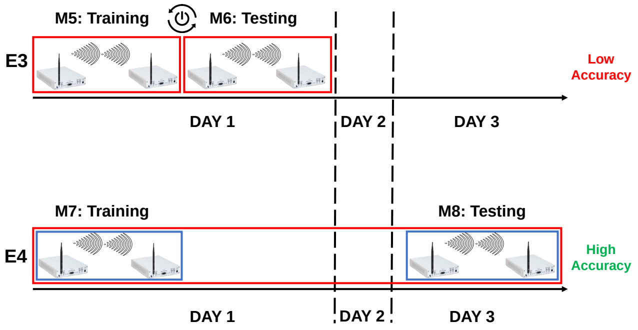


Figure 5: The *power cycle* problem: consecutive measurements (M5 and M6) experiencing similar channel conditions achieve low classification accuracy if radios are power cycled in-between the measurements. Conversely, a long-lasting measurement (E4) experiencing very different channel conditions achieves high accuracy if the two chunks (M7 and M8) are from the same measurement, i.e, they are not separated by a power cycle.

This is achieved by pre-processing the I-Q samples and converting them into images, adopting these later ones as the input to the CNN.

5.2 DAT: In-depth analysis

The vast majority of the literature explains the DAT effect with the unpredictability of the radio channel, considering the changes that affect the environment surrounding the transmitter and receiver. As an example, we consider Fig. 6, consisting of 7 sub-figures (100, 000 I-Q samples each) taken from subsequent chunks in an actual radio measurement. Fig. 6(a) and (g) represent the steady state, before and after the perturbation event, that is, a person walking close to the radio link. It is worth noting that the I-Q samples, and therefore the transmitter's fingerprint, are strongly affected by the considered event. This is evident when looking at the figures from Fig. 6(c) to (e), and in particular, Fig. 6(d), showing a set of I-Q samples completely different from the ones at steady state, i.e., Fig. 6(a) and (g). Phenomena such as the one in Fig. 6(d) completely change the I-Q displacement at the receiver, making the retrieval of the fingerprint (almost) impossible. To discuss in more depth the DAT effect and its causes, we consider a wired link between radios. We consider this scenario as it excludes any dependencies on the multipath and other effects due to RF propagation while keeping the noise due to the communication channel at a minimum. We will extend our results to the RF link in Sect. 5.4.

power cycle. We consider the dataset *DS1* (from Section 4), and we run the experiment E1. Given the 13 runs of *DS1*, we randomly choose an increasing subset of the runs for the training process (from 1 to 12) and only one run (for all the considered cases) for testing—the testing run is always mutually exclusive with respect to the runs constituting the training set. We repeated this procedure 20 times. Figure 7 shows the results of our experiments considering a CNN structure similar to the one of [3] and [14], i.e., *ResNet50*, providing as input raw I-Q samples in the form $N \times 1$. Figure 7(a)

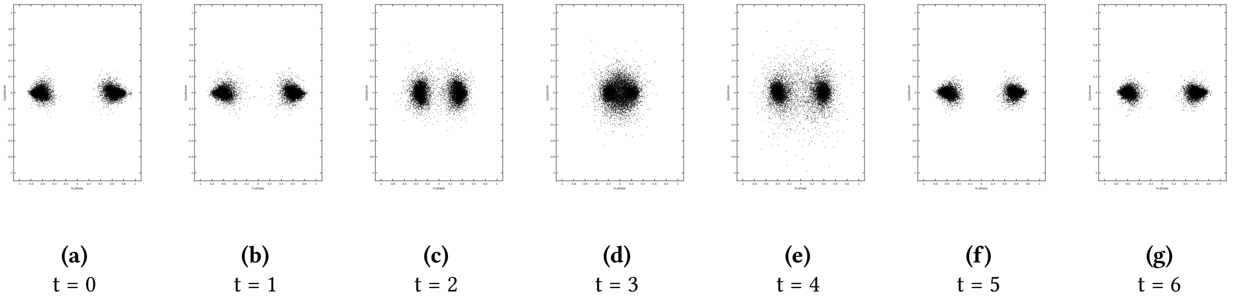

(a) t = 0 **(b)** t = 1 **(c)** t = 2 **(d)** t = 3 **(e)** t = 4 **(f)** t = 5 **(g)** t = 6

Figure 6: The effect of multi-path to the transmitter-receiver link across seven time windows. Compared to an ideal (static) scenario (Figs. (a), (b), (f), and (g)), the shape of the I-Q samples at the receiver (modulated through the BPSK scheme), is significantly affected (see Figs. (c), (d), and (e)) when an event is affecting the scene, e.g., moving objects.

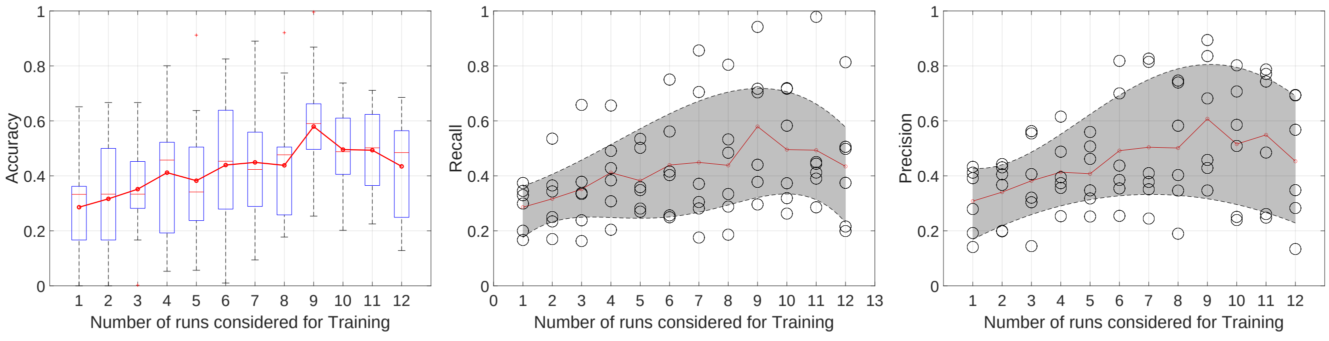
shows the accuracy of the classifier when the training process is exposed to an increasing number of runs, i.e., from 1 to 12. For each considered number of runs in the training process, the central mark in the box plots indicates the median value, while the bottom and top edges of the box indicate the percentiles 0.25 and 0.75, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the "+" marker symbol. The solid red line represents the mean value of the accuracy samples. The accuracy spans approx. between 0.3 and 0.6 (on average) while being characterized by a relatively high variance (≈ 0.3). These results are consistent with the ones reported in [3], where the authors considered the same scenario (wired) and a similar neural network structure. We stress that all the measurements have been taken by power-cycling the radios (both the transmitter and the receiver) every time (E1). Thus, the low performance of the classifier cannot be attributed to the radio channel unpredictability—indeed, we are using a cable, attenuating the dependency on the RF channel propagation. For the sake of completeness, we report recall and precision considering the average value (solid red line), and the region comprised between quantile 10 and 90. Finally, black circles represent the actual outcomes of each specific experiment. Finally, we highlight that Fig. 7 is also taking into account experiment E3. Indeed, since the 13 runs of *DS1* are collected over 3 days, i.e., 3 on the first day, 5 on the second one, and finally, 5 on the third one, there are high chances to have adjacent measurements (one after the other, like in E3) when considering a high number of runs, e.g., 12 runs. Thus, Fig. 7 captures two distinct phenomena: (i) when the number of runs is small (left side of the x-axis), the training and testing are likely performed on datasets that are temporally far away from each other (E1); and (ii) considering the right part of the x-axis, the training and testing are more and more likely to be temporally close each other (E3)—still being separated by a power cycle. Both experiments (E1 and E3) confirm that the power cycle strongly affects the classifier performance.

No power cycle. We now consider *DS2*, i.e., the long measurements over 3 days, and the same methodology as before. We run experiments E1 and E2 over the long measurements to investigate if the performance of the classifier is affected by either the power cycle (absent) or the temporal distance between two chunks extracted from the measurements. To this aim, we split the 3-day measurements into 10 chunks, obtaining a total of 60 chunks. Starting from 6 measures characterized by the same receiver and 6 different transmitters (recall the general set-up from Fig. 1), we randomly selected one or more chunks (up to 9) for training and one for testing (not belonging to the training set). Figure 8 shows the accuracy (quantiles 0.25, 0.75, median, and outliers) of the classifier based on the *resnet-50* network. The differences between Fig. 7(a) and Fig. 8 are striking: given the same scenario, i.e., a wired link between the transmitter and the receiver, the different performance of the classifier are due to how the measurements have been collected. If the measurements adopted for training and testing are separated by a power cycle, the performances of the classifier are negatively affected (Fig. 7(a)); conversely, if the training and testing datasets are not separated by a power cycle, the fingerprint is consistent and can be identified with overwhelming probability (Fig. 8). Finally, Fig. 9 provides the normalized confusion matrix considering all the experiments performed for Fig. 8: it is worth noting that only about 1% of the samples are misclassified.

Wrap-up. Although we acknowledge that the radio channel variations affect the fingerprinting process, we showed that the power-cycling of the radio plays a major role, as well. Indeed, while common knowledge assumes that low classification accuracy is due to different channel conditions between the measurement adopted for training and the one for testing (Experiment E1 in Fig. 4), we proved that two consecutive measurements are affected by the same effect when a power cycle is performed in-between (Experiment E3 in Fig. 5)—we observed this phenomenon by considering a wired link, so being independent of the multipath fading. Moreover, we also confute the common assumption that measurements taken during the same day (Experiment E2 in Fig. 4) do not suffer the DAT effect, due to the high correlation of the experienced radio channels. Indeed, by considering experiment E4 in Fig. 5, the two chunks (M7 and M8) are separated by a long time but still allow a high classification accuracy due to the absence of the power cycle between M7 and M8.

5.3 DAT mitigation

In this section, we introduce a possible solution to mitigate the DAT effect through a dedicated pre-processing of the I-Q samples. Our intuition is rooted in the observation that raw I-Q samples are too noisy, even when considering controlled scenarios (wired).



(a) (b) (c)

Figure 7: Experiment E3, wired scenario, and raw I-Q samples: the power cycle affects the performance of the classifier in terms of Accuracy (a), Recall (b), and Precision (c). The number of runs (with a power cycle in between) considered for the training process does not affect the performance.

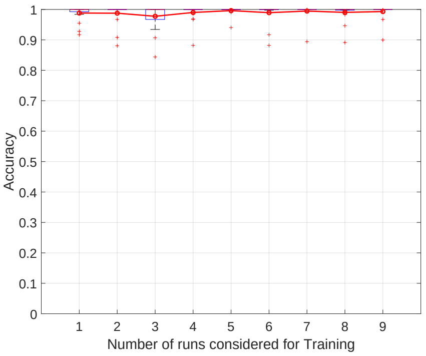


Figure 8: Experiment E4, wired scenario, and raw I-Q samples. The absence of power cycles boosts the performance of the classifier although the runs have been taken on different days. Note how the number of runs considered in the training process does not affect the performance of the classifier.

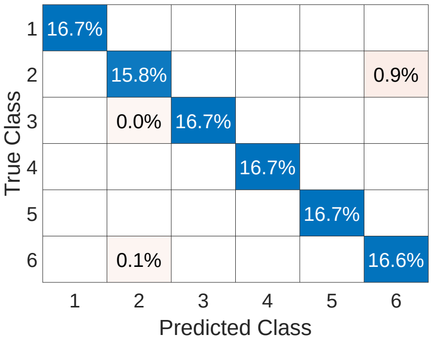


Figure 9: Experiment E4, wired scenario, and raw I-Q samples. The confusion matrix shows outstanding performance for all the considered devices.

Therefore, we consider a de-noising technique based on the spatial and temporal averaging of the I-Q samples. Inspired by [24], we propose a modified pre-processing technique as depicted in Fig. 10. Our methodology considers raw I-Q samples as input from the modulated signal, in this case, BPSK. The raw I-Q samples are mainly organized in two clouds, i.e., one cloud represents the bits equal to 1, while the other cloud represents the zeros—similar considerations can be extended for more complex modulation schemes. The first step consists of defining a chunk size to process the I-Q samples: we consider a chunk size equal to 100, 000 I-Q samples, already taken into account in [24]. The subsequent step differentiates from the reference approach by splitting the original chunk into two chunks containing the left and right clouds, respectively. We trim each cloud and compute a bi-variate histogram, by dividing the I-Q plane into 224 × 224 tiles (the size of the images to be used as input to the *resnet50* network). Then for each tile, we count the number of received I-Q samples. Contrary to [24], we consider three layers for the generation of the images, i.e., one layer for each primary colour component (red, green, and blue). Therefore, assuming an image constituted by a three-layer matrix, i.e., [224 × 224 × 3] (one layer for each primary colour), and the pixel value between 0 and 255, we assign each value of the tile through the following rule.

- $0 \leq x_T \leq 255$, then $p_R = 0, p_G = 0, p_B = x_T$,
- $256 \leq x_T \leq 511$, then $p_R = 0, p_G = x_T - 255, p_B = 255$,
- $x_T > 511$, then , then $p_R = x_T - 510, p_G = 255, p_B = 255$,

where x_T represents the value of the tile from the bi-variate histogram, while p_R, p_G and p_B are the pixel values, i.e., red, green and blue, respectively. Finally, we observe that if $x_T > 767$, it is clipped to 767—this issue can also be controlled by properly adjusting the chunk size. We stress that the boundaries of x_T are derived from pixel values. Since the pixel value is comprised between 0 and 255, we attribute the colors red ([0, 255]), green (256 + [0, 255]), and blue (256*2 + [0, 255]) as functions of such boundaries. In fact, the output of the bivariate histogram is assigned to a color according to the interval defined previously. Figure 10 summarizes the image generation process considering the three image's components, i.e., red, green and blue. Note that the mentioned modification compared

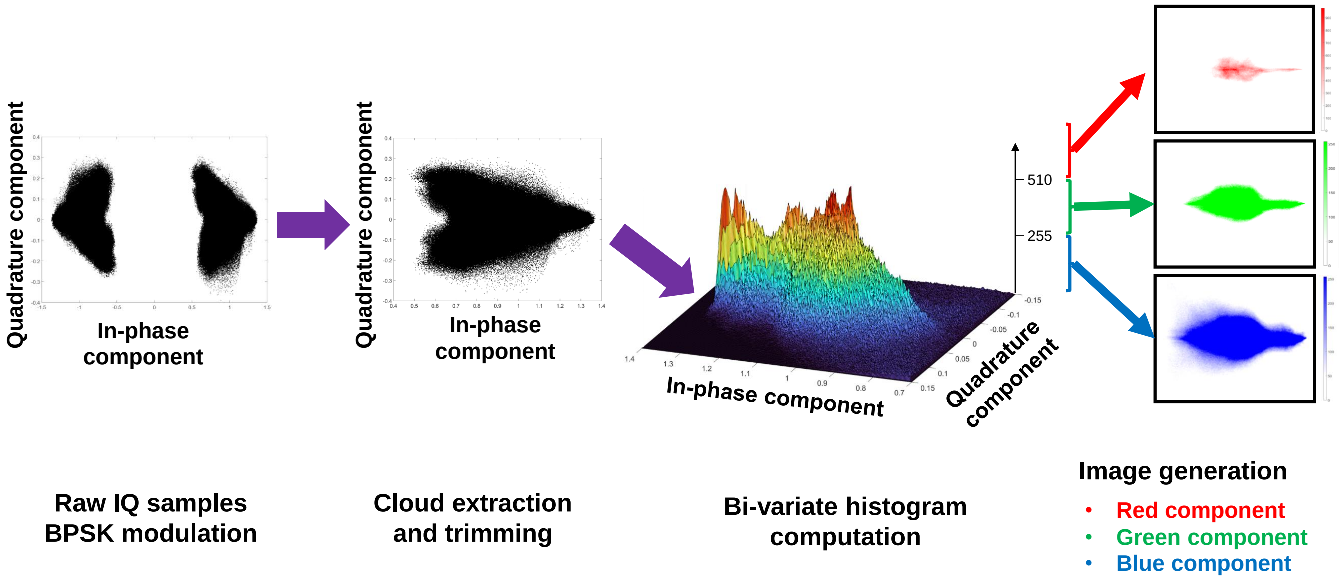


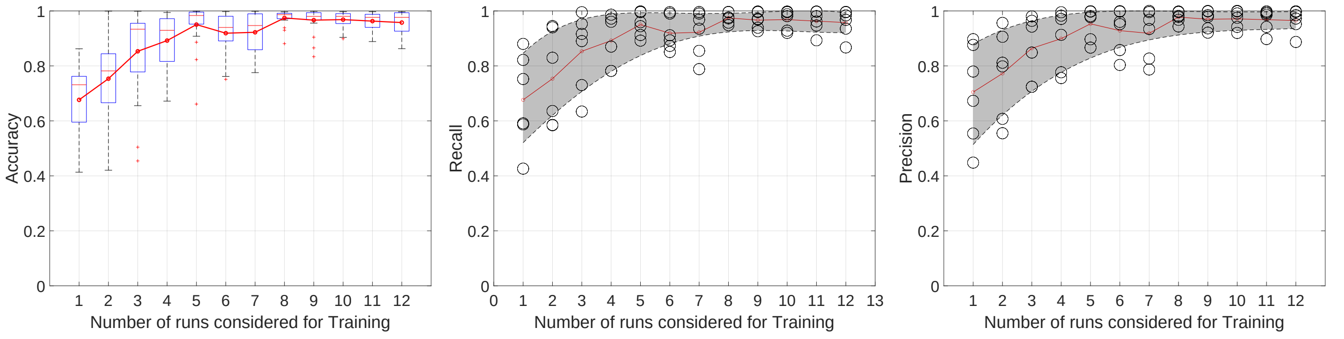
Figure 10: Our solution to mitigate the DAT effect: I-Q samples are pre-processed to generate images.

to the solution in [24] improves the ability of the CNN to adapt to multiple power cycles of the devices. Moreover, as previously demonstrated [24], converting raw I-Q samples to images is a smart way to average input data over time and space. Previous works have already proved that the (unavoidable) loss of information caused by such techniques does not significantly affect the classifier performance while it mitigates the noise. Indeed, random noise can be filtered out by averaging over time and space: I-Q samples are considered per groups (tiles) in the bivariate histogram.

Power Cycle. We now reconsider the Dataset 1 (*DS1*) from Sect. 4 and experiments E1 and E3 from Fig. 4 and Fig. 5, respectively. Moreover, we repeated the same experiments as we did before (Fig. 7) by only considering the different procedure presented in Fig. 10, i.e., we considered the images as input to the CNN in place of raw I-Q samples. Figure 11 shows the accuracy, recall and precision associated with our tests. We considered an increasing number of randomly chosen runs, from 1 to 12, for the training process, and we tested each configuration 20 times. For the testing, instead, we considered only one run (random and disjoint from the training set). The performance of the proposed classification methodology is much better compared to the ones of raw I-Q samples adopted by [3] (recall Fig. 7). Indeed, when training with 5 randomly chosen runs of measurements, the accuracy is higher than 0.9 while it is about 0.35 when considering raw I-Q samples. Moreover, exposing the model to more and more runs significantly increases the performance—this is not happening when considering raw I-Q samples. Finally, recall and precision metrics, reported in Figs. 11(b) and (c), confirm the quality of our proposed classification algorithm when considering false negatives and false positives, respectively.

5.4 Real scenario: radio link

In this section, we consider a real scenario constituted by the measurement set-up considered in Fig. 1, with a wireless link working at the frequency $f_0 = 900MHz$. In this context, we compare the performance of the classifiers considering raw I-Q samples ([3]) and images from dataset *DS3*, as described in Sect. 4. We recall that the measurements in *DS3* have been taken to expose as much as possible the DAT effect: indeed, each measurement lasts for 5 minutes and it is collected at random for 4 days, with several power cycles in-between. Figure 12 shows the comparison between the raw I-Q samples and the images when considering *resnet50* and dataset *DS3*. We adopted the same methodology described before, selecting from 1 to 11 runs for training and a random run for testing, disjoint from the training set. Each black circle and cross in Fig. 12 represents the outcome of a training/testing process for the images and raw I-Q samples, respectively. The shaded green and red areas refer to the quantiles 0.2 and 0.8 calculated on the measured accuracy for both the images and raw I-Q samples. Finally, the solid red and green lines interpolate the accuracy outcomes from the raw I-Q samples and the images, respectively. We observe that the performance of our proposed image-based technique outperforms those of raw I-Q samples, i.e., the green shaded area is always well on top of the red one, with a minor overlapping area in the region comprised between 1 and 10 runs. We also highlight the variance associated with the accuracy: raw I-Q samples are characterized by a very high variance (≈ 0.6) independently of the number of considered runs. This is a critical issue for the repeatability of the experiments when adopting raw I-Q samples: in many cases, single runs based on raw I-Q samples might experience high accuracy, e.g., higher than 0.8, leading to exceptional claims about the feasibility of using such a technique for fingerprinting. However, a systematic replication of the training/testing procedure exposes the issue of



(a) (b) (c)

Figure 11: Experiment E3, wired scenario, and raw images: the power cycle affects the performance of the classifier in terms of Accuracy (a), Recall (b), and Precision (c) when the number of runs (with a power cycle in-between) is low as per Fig. 7. When the number of runs increases, the performance is much better (> 0.9 on average) when using images as input of the classifier.

flat (and poor) performance, which is (almost) independent of the training set size. Images (shaded green area) behave much better. The average accuracy spans between ≈ 0.6 and ≈ 0.85, depending on the training set size, and our results indicate that extending the training process to a higher number of runs might increase the performance even more. Moreover, we observe that the variance associated with the results is still an issue: even assuming the best configuration (high number of runs), the variance associated with the accuracy is ≈ 0.2, although it decreases when more data is considered for training. In our vision, such a variance might be still relatively high for many real-life applications, requiring smaller confidence. At the same time, we observe that the usage of our technique based on images significantly improves the performance of the RFF process, taking a significant step toward the deployment of RFF techniques in the wild. Finally, we acknowledge that the DAT phenomenon still affects the classification process even when pre-processing I-Q samples into images. In this context, our work shows that techniques based on raw I-Q samples analysis are less robust to power-cycling, whereas solutions based on images generated from I-Q samples have the potential to overcome such an issue, showing promising results. For the sake of completeness, we also report the performance of other DL networks with higher complexity, which require a longer training time. We considered 10 runs and two networks, i.e., *inceptionv3* and *inceptionresnetv2*, and experienced an average accuracy of 0.89 and 0.89 with variances of 0.013 and 0.010, respectively. Thus, further research is likely required in this area to increase accuracy while reducing associated variance. However, to the best of our knowledge, this manuscript is the first to perform a systematic analysis of the DAT effect and demonstrate experimentally the impact of devices' power cycle on the accuracy of RFF techniques.

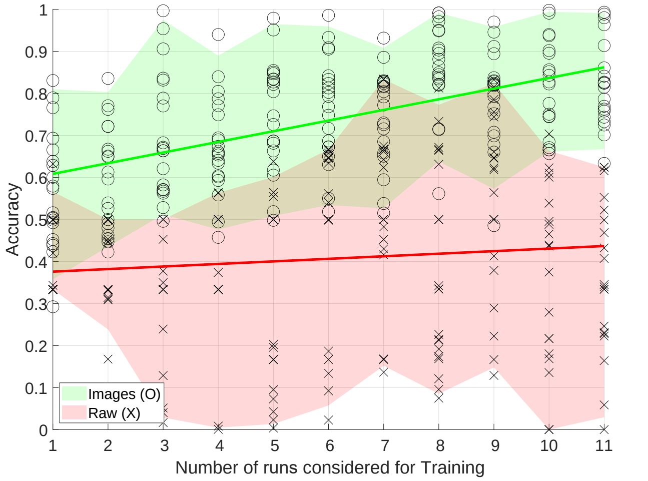


Figure 12: Comparison of raw I-Q samples and images on measurements taken on a wireless link with power cycle.

6 WRAP-UP AND DISCUSSION

Following the in-depth investigation described in Sect. 5, the DAT effect can be re-defined and summarized as the twist of performance affecting the RFF process when the samples used for the training and testing datasets do not belong to the same measurement. A common (and accepted) explanation involves the wireless channel variability, i.e., different measurement times involve a different radio channel, thus causing different performance. The investigation carried out in this work proves this to be a partial (and in some cases ineffective) explanation of the problem. Although we acknowledge the impact of channel variability, we identified the power cycle of the radios as the main cause of performance loss. This phenomenon is independent of the status of the radio channel and detrimental to the identification of the transmitter, also considering measurements very close in time, and thus experiencing the same channel conditions. Although some authors have already observed it (e.g., [3] and [14]), to the best of our knowledge, no one has previously provided a detailed analysis of the DAT effect, shedding a light on its causes and possible mitigation strategies. In our manuscript, we first reproduced the DAT effect in a controlled scenario (a wired link between the transmitter and the receiver), by power-cycling the radios under test. Indeed, the power cycle changes the fingerprint enough to significantly reduce the accuracy of the identification process—halving its accuracy in many cases. The DAT effect disappears when considering the same measurement set-up, but taking chunks of data from a long measurement not being affected by a power cycle. In this scenario, standard techniques (adopting raw I-Q samples) easily achieve an accuracy close to 1. These findings show that the DAT effect exists independently of the wireless channel, providing new information on the current state of the art. The causes behind the fingerprint's change are likely linked to the software re-initialization affecting the SDRs after a power cycle. Indeed, at a high-level, SDRs involve mainly two blocks: (i) the FPGA and (ii) the radio-frequency (RF) module. The former (FPGA module) implements the network interface drivers, the data flow command and control, the decimation/interpolation, and finally, the digital up/down conversion (DUC/DDC), while the latter (RF module) implements the analogue transmission of the signal. An important intermediate block between the FPGA and the RF module performs the analogue-to-digital (ADC)/digital-to-analogue (DAC) conversion of the signal. We believe that the power cycle might affect the internal parameters of those blocks, changing the behavior of the radio at the physical-layer.

DAT analysis. The DAT effect refers to the change of the RFF of a device over time, leading to a significant performance drop of the classifier when training and testing on measurements taken across different days. The DAT effect sums up different factors, such as different signal processing techniques [6], non-linear characteristics of power amplifiers that depend on the average output [19], heat dissipation and device temperature [25], and clock skews [38]. Nevertheless, we proved that the devices' power cycle significantly affects the performance as well—this being a major cause, since in our analysis we excluded channel impairments and local oscillator drifts (the radios have been calibrated before each measurement), while making our analysis and the associated parameters consistent with other recent work. Finally, we believe that the software nature of the SDR—being used as the receiver in all related works—can introduce random artifacts into the data collection process every time they are power cycled, thus preventing transmitter identification.

Limitations. Our analysis involved only specific radios, that is, USRP Ettus X310, and we cannot infer the impact of the DAT effect on other devices. However, our analysis is in line with the findings of the authors in [3], where they observed the same phenomenon considering 20 USRP N210 and USRP X310. Similarly, the authors in [14] observed the phenomenon considering USRP B210 and 25 Pycom Lora-enabled devices. Therefore, we are confident in the general validity of our results. Finally, we are confident that the DAT effect has been observed by many other researchers, but since its cause can be easily attributed to the wrong source (channel variability), it did not receive the attention it deserves.

Robustness to Spoofing Attacks. Recent scientific contributions, such as [24], experimentally investigated the robustness of RF fingerprinting techniques against spoofing attacks. They found that, independently of the chosen approach (being either raw I-Q samples or images), an attacker can defeat RF fingerprinting solutions if it injects less than $\frac{N}{2}$ samples into the flow of the samples, being N the number of I-Q samples adopted to create an image. The findings reported in our manuscript are orthogonal to these considerations in spoofing attacks. Indeed, in line with the current literature and considering the approach presented in Sect. 5.3, adversaries injecting less than $\frac{N}{2}$ samples ($N = 100,000$) are likely not to be detected. However, it should be noted that assuming the same number of I-Q samples of competing solutions, such as [3] and [14], our proposed DAT mitigation strategy achieves significantly higher accuracy. In fact, our proposed solution requires fewer I-Q samples to generate reliable RF fingerprinting solutions, thus reducing the maximum number of I-Q samples that an adversary can inject without being detected. In this context, we also highlight that, at the time of this writing, adversaries have been able to inject only fully-formed packets (either replayed or fully-crafted), while on-the-fly modification of I-Q samples emitted by legitimate transmitters is out of the technological capabilities of currently-available hardware. Therefore, RF fingerprinting solutions are considered effective and robust mainly when applied to high-bandwidth communication links, such as WiFi. For other communication technologies, for example, Internet of Things (IoT) protocols involving low data-rates, RF fingerprinting should only be considered as an additional layer of security and coupled with other cryptography-based solutions.

7 CONCLUSION AND FUTURE WORK

A critical aspect in the successful deployment of Radio Frequency Fingerprinting for physical-layer device authentication is to improve its reliability and robustness. In this paper, we have identified and characterized a new factor that affects the performance of RFF, that is, the power cycle of the devices under test. Our results show that power cycling the radios has a negative impact on the performance of RFF. To mitigate such an issue, we used a technique based on pre-processing raw I-Q samples collected at the PHY layer. The samples are transformed into images and fed into a *ResNet* Convolutional Neural Network. Through a comprehensive performance evaluation, we have shown that such an approach mitigates the impact of both power-cycling and wireless channel fluctuations, resulting in an average classification accuracy of 0.85. This is a significant improvement compared to the average accuracy of approximately 0.5 obtained using state-of-the-art techniques based on raw I-Q samples. However, the variance observed in our results (smaller than competing solutions) suggests that a larger dataset may be necessary to achieve more reliable testing results. Our future work will focus on evaluating the performance of our technique with additional communication technologies, devices, networks, and its associated parameters, as well as its robustness to distance and noise, while also considering the power cycle of either the transmitter or the receiver (independently).

ACKNOWLEDGMENTS

Authors would like to thank the anonymous Reviewers for their constructive comments on the article. This publication was made possible by the GSRA7-1-0510-20045 and NPRP12C-0814-190012-SP165 awards from Qatar National Research Fund (a member of Qatar Foundation). This work has also been partially supported by the INTERSECT project, Grant No. NWA.1162.18.301, funded by the Netherlands Organisation for Scientific Research (NWO). The content herein are solely the responsibility of the authors.

REFERENCES

[1] Sohail Abbas, Qassim Nasir, Douae Nouichi, Mohamed Abdelsalam, Manar Abu Talib, Omnia Abu Waraga, et al. 2021. Improving Security of the Internet of Things via RF fingerprinting based device identification system. *Neural Computing and Applications* 33, 21 (2021), 14753–14769.

[2] Amani Al-Shawabka, Philip Pietraski, Sudhir B Pattar, Francesco Restuccia, and Tommaso Melodia. 2021. DeepLoRa: Fingerprinting LoRa devices at scale through deep learning and data augmentation. In *Proceedings of the Twenty-second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*. 251–260.

[3] Amani Al-Shawabka, Francesco Restuccia, Salvatore D'Oro, Tong Jian, Bruno Costa Rendon, Nasim Soltani, Jennifer Dy, Stratis Ioannidis, Kaushik Chowdhury, and Tommaso Melodia. 2020. Exposing the Fingerprint: Dissecting the Impact of the Wireless Channel on Radio Fingerprinting. In *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications* (Toronto, ON, Canada). IEEE Press, 646–655. https://doi.org/10.1109/INFOCOM41043.2020.9155259

[4] Aysha M Ali, Emre Uzundurukan, and Ali Kara. 2019. Assessment of features and classifiers for Bluetooth RF fingerprinting. *IEEE Access* 7 (2019), 50524–50535.

[5] Trevor J. Bihl, Kenneth W. Bauer, and Michael A. Temple. 2016. Feature Selection for RF Fingerprinting With Multiple Discriminant Analysis and Using ZigBee Device Emissions. *IEEE Transactions on Information Forensics and Security* 11, 8 (2016), 1862–1874. https://doi.org/10.1109/TIFS.2016.2561902

[6] Vladimir Brik, Suman Banerjee, Marco Gruteser, and Sangho Oh. 2008. Wireless Device Identification with Radiometric Signatures. In *Proceedings of the 14th ACM International Conference on Mobile Computing and Networking* (San Francisco, California, USA) *(MobiCom '08)*. Association for Computing Machinery, New York, NY, USA, 116–127. https://doi.org/10.1145/1409944.1409959

[7] Lida Ding, Shilian Wang, Fanggang Wang, and Wei Zhang. 2018. Specific emitter identification via convolutional neural networks. *IEEE Communications Letters* 22, 12 (2018), 2591–2594.

[8] DSP StackExchange. 2017. Bandwidth with complex sampling. https://dsp.stackexchange.com/questions/36927/bandwidth-with-complex-sampling. (Accessed: 2023-Sep-28).

[9] Abdurrahman Elmaghbub and Bechir Hamdaoui. 2021. LoRa device fingerprinting in the wild: Disclosing RF data-driven fingerprint sensitivity to deployment variability. *IEEE Access* 9 (2021), 142893–142909.

[10] Ettus Research. 2020. UBX160 Daughterboard. https://www.ettus.com/product/details/UBX160. (Accessed: 2023-Sep-28).

[11] Ettus Research. 2020. USRP X310. https://www.ettus.com/all-products/x310-kit/. (Accessed: 2023-Sep-28).

[12] Jialiang Gong, Xiaodong Xu, and Yingke Lei. 2020. Unsupervised specific emitter identification method using radio-frequency fingerprint embedded InfoGAN. *IEEE Transactions on Information Forensics and Security* 15 (2020), 2898–2913.

[13] Omer Melih Gul, Michel Kulhandjian, Burak Kantarci, Azzedine Touazi, Cliff Ellement, and Claude D'Amours. 2022. Fine-grained Augmentation for RF Fingerprinting under Impaired Channels. In *2022 IEEE 27th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*. 115–120. https://doi.org/10.1109/CAMAD55695.2022.9966888

[14] Bechir Hamdaoui and Abdurrahman Elmaghbub. 2022. Deep-Learning-Based Device Fingerprinting for Increased LoRa-IoT Security: Sensitivity to Network Deployment Changes. *IEEE Network* 36, 3 (2022), 204–210. https://doi.org/10.1109/MNET.001.2100553

[15] Samer Hanna, Samurdhi Karunaratne, and Danijela Cabric. 2022. WiSig: A Large-Scale WiFi Signal Dataset for Receiver and Channel Agnostic RF Fingerprinting. *IEEE Access* 10 (2022), 22808–22818.

[16] Anu Jagannath, Jithin Jagannath, and Prem Sagar Pattanshetty Vasanth Kumar. 2022. A Comprehensive Survey on Radio Frequency (RF) Fingerprinting: Traditional Approaches, Deep Learning, and Open Challenges. *arXiv preprint arXiv:2201.00680* (2022).

[17] Tong Jian, Yifan Gong, Zheng Zhan, Runbin Shi, Nasim Soltani, Zifeng Wang, Jennifer Dy, Kaushik Chowdhury, Yanzhi Wang, and Stratis Ioannidis. 2021. Radio Frequency Fingerprinting on the Edge. *IEEE Transactions on Mobile Computing* 21, 11 (2021), 4078–4093.

[18] Tong Jian, Bruno Costa Rendon, Emmanuel Ojuba, Nasim Soltani, Zifeng Wang, Kunal Sankhe, Andrey Gritsenko, Jennifer Dy, Kaushik Chowdhury, and Stratis Ioannidis. 2020. Deep learning for RF fingerprinting: A massive experimental study. *IEEE Internet of Things Magazine* 3, 1 (2020), 50–57.

[19] Dae Hyun Kwon, Hao Li, Yuchun Chang, Richard Tseng, and Yun Chiu. 2010. Digitally Equalized CMOS Transmitter Front-End With Integrated Power Amplifier. *IEEE Journal of Solid-State Circuits* 45, 8 (2010), 1602–1614. https://doi.org/10.1109/JSSC.2010.2048140

[20] B. P. Lathi, Adel S. Sedra, and M.E. Van Valkenburg. 1995. *Modern Digital and Analog Communication Systems* (2nd ed.). Oxford University Press, Inc., USA.

[21] Kevin Merchant, Shauna Revay, George Stantchev, and Bryan Nousain. 2018. Deep learning for RF device fingerprinting in cognitive communication networks. *IEEE Journal of Selected Topics in Signal Processing* 12, 1 (2018), 160–167.

[22] Subhramoy Mohanti, Nasim Soltani, Kunal Sankhe, Dheryta Jaisinghani, Marco Di Felice, and Kaushik Chowdhury. 2020. AirID: Injecting a custom RF fingerprint for enhanced UAV identification using deep learning. In *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 1–6.

[23] Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. 2012. A Unifying View on Dataset Shift in Classification. *Pattern recognition* 45, 1 (2012), 521–530.

[24] Gabriele Oligeri, Savio Sciancalepore, Simone Raponi, and Roberto Di Pietro. 2022. PAST-AI: Physical-layer Authentication of Satellite Transmitters via Deep Learning. *IEEE Transactions on Information Forensics and Security* (2022), 1–1. https://doi.org/10.1109/TIFS.2022.3219287

[25] Adam C. Polak and Dennis L. Goeckel. 2015. Identification of Wireless Devices of Users Who Actively Fake Their RF Fingerprints With Artificial Data Distortion. *IEEE Transactions on Wireless Communications* 14, 11 (2015), 5889–5899. https://doi.org/10.1109/TWC.2015.2443794

[26] Sekhar Rajendran, Zhi Sun, Feng Lin, and Kui Ren. 2020. Injecting reliable radio frequency fingerprints using metasurface for the Internet of Things. *IEEE Transactions on Information Forensics and Security* 16 (2020), 1896–1911.

[27] Francesco Restuccia, Salvatore D'Oro, Amani Al-Shawabka, Mauro Belgiovine, Luca Angioloni, Stratis Ioannidis, Kaushik Chowdhury, and Tommaso Melodia. 2019. DeepRadioID: Real-time channel-resilient optimization of deep learning-based radio fingerprinting algorithms. In *Proceedings of the Twentieth ACM International Symposium on Mobile Ad Hoc Networking and Computing*. 51–60.

[28] Francesco Restuccia, Salvatore D'Oro, Amani Al-Shawabka, Bruno Costa Rendon, Stratis Ioannidis, and Tommaso Melodia. 2021. DeepFIR: Channel-Robust Physical-Layer Deep Learning Through Adaptive Waveform Filtering. *IEEE Transactions on Wireless Communications* 20, 12 (2021), 8054–8066.

[29] Shamnaz Riyaz, Kunal Sankhe, Stratis Ioannidis, and Kaushik Chowdhury. 2018. Deep learning convolutional neural networks for radio identification. *IEEE Communications Magazine* 56, 9 (2018), 146–152.

[30] Kunal Sankhe, Mauro Belgiovine, Fan Zhou, Shamnaz Riyaz, Stratis Ioannidis, and Kaushik Chowdhury. 2019. ORACLE: Optimized radio classification through convolutional neural networks. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 370–378.

[31] Guanxiong Shen, Junqing Zhang, Alan Marshall, and Joseph R Cavallaro. 2022. Towards scalable and channel-robust radio frequency fingerprint identification for LoRa. *IEEE Transactions on Information Forensics and Security* 17 (2022), 774–787.

[32] Guanxiong Shen, Junqing Zhang, Alan Marshall, Linning Peng, and Xianbin Wang. 2021. Radio Frequency Fingerprint Identification for LoRa using Deep Learning. *IEEE Journal on Selected Areas in Communications* 39, 8 (2021), 2604–2616.

[33] Nasim Soltani, Kunal Sankhe, Jennifer Dy, Stratis Ioannidis, and Kaushik Chowdhury. 2020. More is better: Data augmentation for channel-resilient RF fingerprinting. *IEEE Communications Magazine* 58, 10 (2020), 66–72.

[34] Naeimeh Soltanieh, Yaser Norouzi, Yang Yang, and Nemai Chandra Karmakar. 2020. A Review of Radio Frequency Fingerprinting Techniques. *IEEE Journal of Radio Frequency Identification* 4, 3 (2020), 222–233.

[35] Weidong Wang and Lu Gan. 2022. Radio Frequency Fingerprinting Improved by Statistical Noise Reduction. *IEEE Transactions on Cognitive Communications and Networking* (2022).

[36] Wenqing Yan, Thiemo Voigt, and Christian Rohner. 2022. RRF: A Robust Radiometric Fingerprint System that Embraces Wireless Channel Diversity. In *Proceedings of the 15th ACM Conference on Security and Privacy in Wireless and Mobile Networks*. 85–97.

[37] Jiabao Yu, Aiqun Hu, Guyue Li, and Linning Peng. 2019. A robust RF fingerprinting approach using multisampling convolutional neural network. *IEEE Internet of Things Journal* 6, 4 (2019), 6786–6799.

[38] Davide Zanetti, Boris Danev, and Srdjan Capkun. 2010. Physical-Layer Identification of UHF RFID Tags. In *Proceedings of the Sixteenth Annual International Conference on Mobile Computing and Networking* (Chicago, Illinois, USA) *(MobiCom '10)*. Association for Computing Machinery, New York, NY, USA, 353–364. https://doi.org/10.1145/1859995.1860035

[39] Zhen Zhang, Aiqun Hu, Wei Xu, Jiabao Yu, and Yang Yang. 2022. An Artificial Radio Frequency Fingerprint Embedding Scheme for Device Identification. *IEEE Communications Letters* 26, 5 (2022), 974–978. https://doi.org/10.1109/LCOMM.2022.3148037