Adversarial Machine Learning for Image-Based Radio Frequency Fingerprinting: Attacks and Defenses

Article *in* IEEE Communications Magazine · January 2024

DOI: 10.1109/MCOM.001.2300464

6 authors, including:

Savio Sciancalepore
Eindhoven University of Technology
110 PUBLICATIONS **1,433** CITATIONS

SEE PROFILE

Gabriele Oligeri
Hamad bin Khalifa University
102 PUBLICATIONS **1,164** CITATIONS

SEE PROFILE

Giuseppe Piro
Politecnico di Bari
132 PUBLICATIONS **4,609** CITATIONS

SEE PROFILE

Adversarial Machine Learning for Image-Based Radio Frequency Fingerprinting: Attacks and Defenses

Lorenzo Papangelo*, Maurizio Pistilli*, Savio Sciancalepore‡,
Gabriele Oligeri†, Giuseppe Piro*, Gennaro Boggia*,
* Politecnico di Bari, Italy; {l.papangelo1, m.pistilli1}@studenti.poliba.it, {giuseppe.piro, gennaro.boggia}@poliba.it
† Division of Information and Computing Technology (ICT), College of Science and Engineering (CSE), Hamad Bin Khalifa University (HBKU), Doha, Qatar; goligeri@hbku.edu.qa
‡ Eindhoven University of Technology, Eindhoven, Netherlands; s.sciancalepore@tue.nl

***Abstract*—[1] Image-based Radio Frequency Fingerprinting (RFF) is a promising variant of traditional RFF systems. As a distinctive feature, such systems convert Physical-layer signals into matrices resembling 2-D or 3-D images and consider the latter as the input for state-of-the-art image classifiers. Compared to traditional ones, image-based RFF systems have recently shown enhanced flexibility for device identification, as they can better mitigate channel conditions, devices movement, and power cycle. However, previous works have yet to investigate their performance when subject to Adversarial Machine Learning (AML) attacks using state-of-the-art techniques such as Generative Adversarial Networks and the Fast Gradient Sign Method. Similarly, there are no studies about their capability to integrate adversarial learning strategies for enhancing their robustness to such attacks. In this paper, we fill the gap by conducting an experimental analysis of the effectiveness of AML attacks and adversarial training techniques for image-based RFF systems. Using a state-of-the-art image-based RFF system and actual measurements, we show that adversarial samples can effectively degrade classification performance. At the same time, training the image-based RFF system with adversarial samples increases the reliability and robustness of such methods at the cost of a lower classification accuracy.**

***Index Terms*—Physical-Layer Security; Wireless Security; Artificial Intelligence for Security.**

I. Introduction

Radio Frequency Fingerprinting (RFF) solutions enable physical device authentication by analyzing the transmitted signal from the radio spectrum. The main idea behind RFF is that two perfectly identical devices do not exist, and even the smallest differences, such as imperfections at the hardware level, can bias the over-the-air signal (potentially) allowing a receiver to identify them uniquely. Traditional RFF systems acquire N raw samples of the signals at the Physical (PHY)-layer, namely I-Q samples, and feed them directly into a Deep Learning (DL) classifier, using a complex vector of size $1\times 2N$ as in [1], or a real matrix of size $2\times N$ as in [2]. Although this strategy has shown remarkable performance, the considered DL models are sensitive to external factors and cannot generalize to different channel conditions, movement of the transmitters, and power cycling of RF devices [3]. Conversely, image-based RFF systems are based on the intuition that DL classifiers are particularly successful in classifying images. Thus, they involve pre-processing the I-Q samples into data structures resembling either 2-D or 3-D images, thus translating the RFF problem into an image classification problem. In the scenarios described above, image-based RFF systems exhibit much more flexibility, being successful in identifying RF devices even under challenging channel conditions [4] and across various power cycles of devices [3].

At the same time, due to the relevance of cybersecurity attacks to Artificial Intelligence (AI)-based systems, we also need to investigate the robustness of such systems to adversarial attacks [5]. Since RFF leverages Neural Networks (NN) classifiers, RFF systems are vulnerable to Adversarial Machine Learning (AML) attacks, such as the ones performed through the use of the Fast Gradient Signed Method (FGSM) and Generative Adversarial Networks (GAN). In fact, by using such techniques, an attacker can generate adversarial samples that mislead NN classifiers, compromising the accuracy of such systems. At the same time, training NNs to recognize such adversarial samples can improve the reliability and robustness of the classification.

A few contributions already investigated AML-inspired attacks and mitigation strategies for RFF systems. For example, the authors in [6] launched several AML attacks against RFF systems based on the analysis of raw I-Q samples, demonstrating the theoretical vulnerability of such systems when the attacker can bias with outstanding accuracy the value of I-Q samples at the receiver side. The authors in [7] reported similar findings analyzing synthetic data, while the authors in [8] validated such results on an actual deployment using GANs. However, to our knowledge, none of the available contributions investigated AML attacks and mitigation strategies for image-based RFF systems. As such systems are particularly promising, evaluating their robustness to AML attacks in various configurations

is critical to further pushing their adoption into real application scenarios.

Contribution. In this manuscript, we investigate the effectiveness of AML techniques to attack and improve the robustness of image-based RFF systems. We set up an experimental testbed comprising seven Software-Defined Radios (SDRs) USRP X310, we used it to gather actual Radio Frequency (RF) data (I-Q samples) emitted by such devices, and we also implemented a reference image-based RFF system using Matlab R2022b. Then, we carried out various experiments to test the robustness of such a system against adversarial attacks. We show that, when not trained to reject adversarial attacks, such systems can hardly reject adversarial examples generated through the FGSM and GANs. We also applied adversarial training techniques to enhance the robustness of image-based RFF systems. Our results demonstrate that, when training on adversarial images, image-based RFF systems can reject such AML-inspired attacks at the cost of a slight decrease in the achieved classification accuracy.

Roadmap. The rest of this paper is organized as follows. Sect. II introduces preliminaries, Sect. III illustrates our scenario and attacker model, Sect. IV outlines our measurements and methodology, Sect. V reports experimental results and, finally, Sect. VI concludes the paper and outlines future work.

II. Background

A. Radio Frequency Fingerprinting

RFF systems uniquely identify RF devices using their emitted radio signals, avoiding using cryptography. RF signals emitted by these devices are characterized by patterns due to hardware imperfections at the micrometric scale connected to industrial manufacturing processes [9]. The differences among the I-Q samples transmitted by various devices are hard to detect, thus requiring the usage of DL-based techniques.

State-of-the-art RFF systems require *training* a NN model with chunks of I-Q samples and then *testing* a sequence (of I-Q samples) from the wild to identify the transmitter. Among the several options, Residual NNs (RNNs) are often preferred, being a good trade-off between training speed and classification performance. To fit the use-case of RFF systems, the input layer of the RNN is usually adapted to match the structure of the I-Q samples, while the output layer is typically matched to the number of transmitters; then, the network is re-trained on a large set of I-Q samples. Current RFF systems comprise two main families, i.e., the ones considering raw I-Q samples and the image-based ones. The first ones consider an input of interleaved raw I-Q samples consisting of a vector of dimensions either $1\times 2N$ as in [1], or $2\times N$ as in [2]. The latter ones, instead, consider images where the input is a matrix, e.g., $X\times Y\times 3$, where $X\times Y$ is the size of the image, as in [4] and [3]. Although several contributions focus on the former, recent research has shown the enhanced robustness and reliability of the second when considering mobile devices [4], channel unpredictability, and devices' power cycle [3]. In this manuscript, we consider a DL-based image-based RFF system inspired by the one in [4]. We consider as input the regular representation of raw digital wireless signals into I and Q components, and a reference chunk of $100,000$ I-Q samples. For each symbol, considering minimum and maximum values $I_{Max} = Q_{Max}$ and $I_{Min} = Q_{Min}$, the RFF system divides the I-Q plane into a fixed number of tiles $N \times N$, each tile being a square with side $l = \frac{I_{Max}-I_{min}}{N}$. In line with [4], we consider the *resnet18* network. Thus, we divide the I-Q plane of each symbol into 224×224 tiles. The RFF system evaluates how many of the received I-Q samples per symbol fall into each tile (bi-variate histogram); then, it truncates such values to 255, according to the maximum value of a pixel. According to [3], we consider three layers, i.e., one layer for each primary color component (red, green, and blue). Thus, the generated structure has size $224\times 224\times 3$, consistently with a 3-D color image. The images generated from the I-Q samples of each device, including only valid wireless messages at the receiver ($100,000$ samples per image), are used to train the classifier and generate the corresponding profile. At runtime, the RFF system acquires the same set of I-Q samples from an RF device, generates the images as described above, and finally, it tests if such an image aligns with the known profiles of the legitimate RF devices.

B. Generative Adversarial Networks

GANs have recently gained momentum as a powerful tool for testing the robustness of classification systems based on NNs [10]. In summary, GANs can generate synthetic data that closely resemble some input samples and can be used to test the reliability of NN classifiers. To do so, GANs leverage two Deep NN (DNN). The first network, namely, the *generator*, captures the distribution of the data. It is fed with random noise and trained to produce data with a distribution that mimics the statistical properties of the original samples. The second network, the *discriminator*, estimates the probability that samples belong to the train set or to generator-made data. The two networks compete in a *minimax two-player game*, where the generator is trained to maximize the probability of the discriminator making classification mistakes. Interested readers can find more details on GANs' rationale in [11].

C. Fast Gradient Sign Method

The FGSM is a popular method for deploying evasion attacks against DL models. Specifically, the FGSM is a one-step gradient-based technique developed to find the scaled sign of the gradient of a cost function via perturbations of such a function, to minimize the strength of the applied perturbation [12]. Through the FGSM, it is possible to generate an adversarial input sample x' starting from a legitimate sample x by adding a perturbation to such input sample in the direction of the gradient of the loss function to the input, as formalized in [6]. The objective is achieved iteratively by tuning a parameter ϵ, which, in turn, controls the intensity of the applied perturbation. Interested readers can find more details on FGSMs' rationale in [12].

D. Adversarial Training

AML techniques can also be used to enhance the robustness of a classifier based on NNs. Indeed, by exposing the classifier

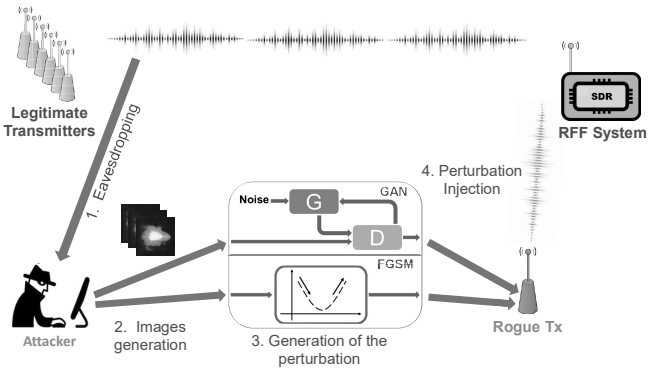

Fig. 1. Reference Scenario.

to the adversarial samples, we can make the system aware of the existence of such samples, e.g., by denoting them through a specific label and submitting them at the training time, to let the model identify such samples correctly. Such a technique, known in the literature as *adversarial training*, allows the system to mitigate and possibly reject evasion attacks. More details on adversarial training techniques can be found in [13].

III. Reference Scenario and Adversary Model

System Model. Fig. 1 shows our reference scenario. We consider a wireless network where K devices transmit RF signals. The devices communicate via a n-Phase Shift Keying (PSK) modulation scheme, being $n = 2$ the number of adopted symbols [14]. We assume the network uses RFF techniques to authenticate transmitting devices. Thus, in line with typical RFF systems, a dedicated receiver detects all RF data exchange, stores the related raw I-Q samples, and performs techniques based on DL to authenticate transmitting devices. The RFF system features an image-based approach, as discussed above. We do not make assumptions about the devices' movement and noise profile affecting the wireless channel, since such factors are orthogonal to our problem.

In line with the literature, we assume a *closed set* scenario, i.e., the RFF system knows in advance which devices can communicate over the network. Thus, the RF profiles of the transmissions from such devices are acquired offline and deployed into the RFF system before deployment. At runtime, the receiver acquires the RF data and associates the current transmission with the (known) device whose RF profile matches its features.

Attacker Model. We assume an adversary $\mathcal{A}$ able to receive and transmit radio signals. $\mathcal{A}$ is an omnidirectional eavesdropper, capable of detecting and decoding all RF transmissions in the network. Also, $\mathcal{A}$ can inject RF signals into the channel, replaying previously eavesdropped messages or transmitting forged ones, pretending to be a legitimate RF device. To increase the effectiveness of the impersonation attack, we assume $\mathcal{A}$ can use channel equalization techniques to compensate for random fluctuations in the wireless channel due to its specific location [15]. Although this is hardly achievable in practice, this technique allows us to model the worst-case scenario, where the attacker has complete control of the channel. Moreover, $\mathcal{A}$ features a SDR, being able to control (with a given maximum accuracy) the displacement of the I-Q samples. Combined with previous channel compensations capabilities, this feature allows $\mathcal{A}$ to generate I-Q samples potentially appearing at the receiver very close to the values of legitimate RF devices. Since the considered RFF system leverages an image-based approach, to be successful, $\mathcal{A}$ has to place the I-Q samples at the center value of the tile, with a maximum error equal to half of the size of the tile used to compute the bi-variate histogram, i.e., $\frac{l}{2}$. Such a requirement makes the described attack to image-based RFF systems much more realistic than the ones usually considered for RFF systems leveraging raw I-Q samples [6], and thus, particularly worthy of investigation. A successful attack against RFF systems leveraging raw I-Q samples should adopt a SDR capable of placing the I-Q samples with very high precision, and such hardware is currently not available. Conversely, when image-based systems are adopted, the precision the attacker requires to execute the attack successfully is much smaller. Moreover, $\mathcal{A}$ knows the PHY-layer model (algorithm and corresponding hyper-parameters) used for RFF. The receivers use this information to identify RF transmissions; thus, this is publicly available or generated by the adversary after training.

$\mathcal{A}$ can carry out two different attacks through the capabilities and tools described above. First, $\mathcal{A}$ can perform a targeted *spoofing attack* to appear as a legitimate RF device to the RFF system. To do so, through objective function minimization, $\mathcal{A}$ can use GANs to generate images that mislead the RFF system and, in turn, obtain the distribution of the I-Q samples to be generated to deliver a successful attack.
Moreover, $\mathcal{A}$ can carry out untargeted Denial of Service (DoS) attacks, appearing as a random legitimate device to the RFF system. To do so, through gradient minimization, $\mathcal{A}$ can use FGSMs to generate I-Q samples and thus images to mislead the RFF system. Compared to jamming attacks, FGSM-based DoS attacks are smarter, as they generate noise specifically to deceive the RFF system. As discussed in [6], generating adversarial samples through GANs and FGSM requires determining statistical distributions used to extract random power values to inject in the channel to achieve the mentioned attacks.

IV. Measurements and Methodology

Measurements. We acquired measurements matching the scenario described in Sec. III. Our dataset comprises real-world data acquired through seven SDRs USRP X310, featuring the UBX160 daughterboard and the VERT900 antenna. The SDRs were connected to two laptops, HP EliteBook I7, featuring 32GB of RAM. We considered the radio 1 as the receiver, while the other ones (2, 3, 4, 5, 6, 7) as the transmitters (only one being active for each experiment). To match the adversary model, the transmitter and the receiver have been connected via a wired link using a coaxial cable type RG58A/U. Such a setup allowed us to avoid the random fluctuations of the wireless channel and, thus, to model the worst-case adversary model capable of controlling the channel. We set the transmission power of the radios at 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 below). Our dataset includes 78 different measurements, organized in 13 runs. For each run, we kept the same receiver while we swapped the 6 transmitters. We used the carrier frequency 900 MHz, with a sample rate of 1 Msa/s at both the transmitter and at the receiver.

As for the software, we adopted GNURadio v3.8 for controlling the SDRs. For the transmitters, we set up a standard transmission chain implementing the modulation Binary Phase Shift Keying (BPSK). At the receiver, we deployed the receiving chain of the BPSK, and we saved the received I-Q samples immediately after their reception for further processing. Note that the BPSK is currently used in many real-life communication scenarios, e.g., IEEE 802.11 a/b/g/n, LEO satellites, and WiMax, to name a few.

Methodology. We specifically focus on the image-based RFF system introduced by the authors in [3], which is derived, in turn, from the one used by the authors in [4]. As a first step, we verified the results reported in the cited papers using the exact configuration of the RFF system to have a reliable baseline. Then, we set up the attacks described in Sec. III. We first deployed multiple spoofing attacks targeted to specific RF transmitters, using GANs. To this end, we used an ad-hoc Wasserstein GAN with Gradient Penalty (WGAN-GP). Specifically, the GAN generator takes random noise as input and applies the following operations: (i) it inputs the noise to a dense layer; (ii) it reshapes the output to have three dimensions, representing the length, the width, and the number of filters, respectively; (iii) it adopts a *Conv2DTranspose* layer to perform deconvolution, reducing the number of filters by half and using a stride of 2; and (iv) in the final layer, it upsamples the features to the size of the training images, which in our case is $224 \times 224 \times 3$. We also perform batch normalization, except for the final deconvolution layer. The discriminator uses stridden convolutions to reduce the dimensionality of the input images, activated, as best practice, by *LeakyRELU*. The output features are flattened and fed to a 1-unit dense layer without activation. We trained the GAN to generate adversarial images by training six distinct generators, each with a model of the specific legitimate RF transmitter. Then, we used each of the six generators to create 350 images, used to challenge the RFF system described in Sec. III.

Following, we set up DoS attacks against the target RFF systems using the FGSM. To this aim, in line with the rationale of FGSM (see Sect. II-C), we applied a perturbation ϵ with increasing intensity, from 1 to 5, to the images used for testing the RFF system. More in detail, for each element of a given layer of the input image, we applied a random integer perturbation in the range $[-\epsilon, \epsilon]$, so that it changes the number of samples falling in a given tile, possibly misleading the RFF model. We also enforced a few constraints to make the attack stealthy and hard to detect at the receiver. Specifically, we: (i) ensured that the values of the elements in each of the layers after perturbation always fall in the same interval $[0, 255]$, (ii) enforced that the sum of the pixel values for a given image is consistent with the number of I-Q samples used to generate the images by the RFF system, and finally (iii) verified that the generated adversarial images are consistent with the ones obtained as a result of the reception of a signal modulated through the BPSK scheme, in line with the regular input to the system. The enforcement of such requirements ensures that the generated adversarial images fully adhere to the input that the RFF system expects. Then we evaluated the accuracy of the RFF system under investigation to identify the legitimate transmitting device despite injection of the perturbation.

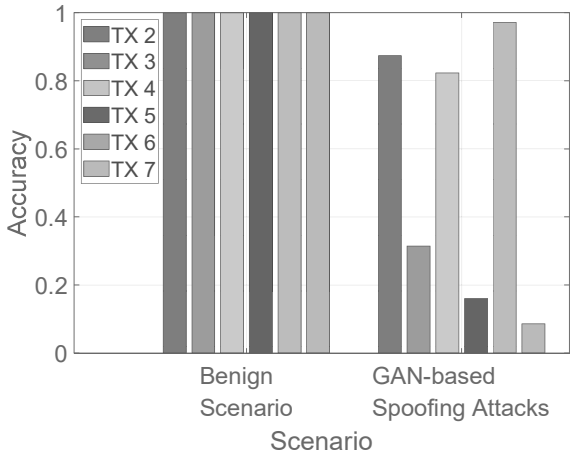

Fig. 2. Performance of the image-based RFF system in a benign scenario and with spoofing attacks via images generated through the WGAN-GP.

Finally, we set up experiments to investigate the effectiveness of adversarial training techniques to mitigate the described attacks. For GAN-based attacks, we used the images generated by the WGAN-GP to train the classifier of the RFF system. For every attack scenario, we created a new class *adversary* and trained the RFF system on a subset of the images. We also considered various batch size values. Similarly, to mitigate DoS attacks, we trained the image-based RFF system with adversarial images generated by applying $\epsilon = 5$, and we evaluated the accuracy of the classifier in rejecting the FGSM-based DoS attack.

Finally, we implemented and launched all experiments using Matlab R2022b, using a server featuring 64 cores, 512GB RAM, and 4 GPUs Nvidia Tesla M40.

V. Experimental Results

A. Adversarial Attacks to Image-based RFF

We first investigate the effectiveness of GAN-based spoofing attacks against image-based RFF systems. We report in Fig. 2 the performance of the considered RFF system in a benign scenario and under spoofing attacks using images generated through the WGAN-GP, as described in Sec. IV.

The RFF system performs very well when tested on images generated from I-Q samples received from legitimate RF devices, i.e., it reports an accuracy of 1 for all test cases (Benign scenario). However, the performance reported in Fig. 2 also highlights the vulnerability of the RFF system to GAN-based spoofing attacks. Indeed, our tests indicate that

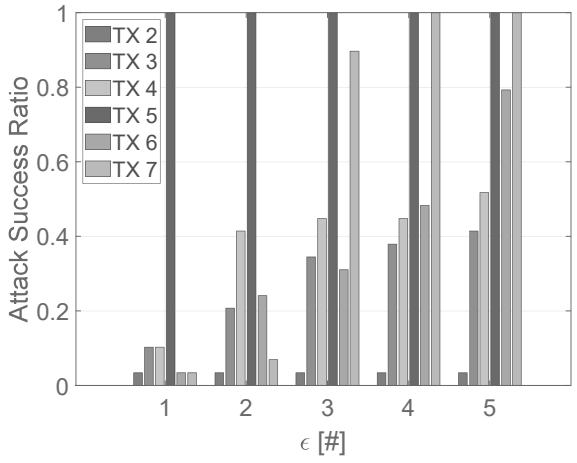

Fig. 3. Performance of the image-based RFF system under a DoS attack performed using the FGSM.

the adversary is successful when impersonating devices with IDs 3, 5, and 7, degrading the classifier's performance to less than 0.32 in all cases. The accuracy of the RFF system remains high for the other three devices, with an accuracy drop of no more than 0.18. Such behavior is strictly related to the distribution of the I-Q samples over time, leading to the generation of images used for training. Consistent I-Q values generate models that reject even small differences, enabling attack detection. Devices characterized by I-Q samples whose values change over time more significantly generate models that tolerate more anomalies, thus preventing attack detection. We also investigate the effectiveness of FGSM-based DoS attacks against image-based RFF system. Fig. 3 summarizes the results of our investigation.

For all the considered transmitters, increasing the intensity of the perturbation ϵ leads to a higher attack success ratio. In fact, the higher the intensity of the perturbation (noise), the higher the chances that the RFF system does not successfully classify the generated image. At the same time, for a given perturbation value, the classifier's performance changes significantly, given the considered legitimate RF device. For example, a minimal perturbation value $\epsilon = 1$ is enough to prevent the identification of the legitimate RF device with ID 5 (attack success ratio of 1), while the adversary achieves minimal effects with the other legitimate transmitters. When $\epsilon = 5$, the attack is effective for transmitters with ID $5, 6, 7$ (attack success ratios higher than 0.79), while it has limited effectiveness for the remaining ones. Such results confirm the intuition obtained from the previous results, i.e., some of the models constructed by the image-based RFF system are more robust than others to adversarial attacks, based on the *consistency* over time of the I-Q samples profile.

We also notice that the RFF model built for a specific device might be robust against a given adversarial attack, but not against another one. For example, the RFF model of the legitimate device with ID 6 is less robust to FGSM-based DoS attacks (attack success rate of 0.79) than to GAN-based spoofing attacks (success rate of ≈ 0.5). This result is also reasonable, as each adversarial technique applies a specific rationale to find an input that can mislead the classifier of the RFF system.

To provide a comparison in such a setup, we report in Fig. 4 the performance of the RFF system working on raw I-Q samples proposed by Hamdaoui et al. in [2] against FGSM-based DoS attacks, considering our same dataset. Similarly to

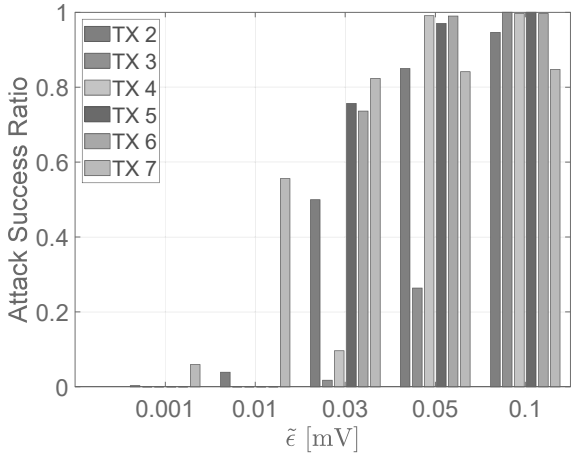

Fig. 4. Performance of the RFF system based on raw I-Q samples analysis in [2] under a DoS attack performed using the FGSM.

the results in Fig. 3, RFF systems' performance on raw I-Q samples decreases when subject to attacks using increasing perturbation $\tilde{\epsilon}$. Also, some models (e.g., the one for TX7) are more vulnerable than others. Note that, for the system proposed in [2], $\tilde{\epsilon}$ is measured in mV, as the perturbation applies to raw I-Q samples and not to pixel values. Thus, both image-based RFF systems and RFF systems using raw I-Q samples are vulnerable to AML-inspired attacks.

B. Defense Strategies for Image-based RFF

We applied adversarial training to evaluate the ability of the image-based RFF system to mitigate AML attacks. We first focus on the GAN-based spoofing attack. Fig. 5 reports the results of our investigation, considering two reference batch size values of 18 and 60, affecting the memory requirements of the solution.

The results highlight that adversarial training is effective in mitigating GAN-based spoofing attacks. Indeed, more than 99% of the attacks are successfully rejected. Comparing Fig. 5 with Fig. 2, adversarial training causes a slight degradation of performance, as some images are incorrectly classified, leading to an accuracy of 0.98. The result stays stable even with the larger batch size. Although this is a minimal performance drop, the result is consistent with the rationale of adversarial training. Finally, we evaluate in Fig. 6 the effectiveness of adversarial training against FGSM-based DoS attacks. Adversarial training improves the system's robustness to adversarial attacks. The results stay stable for low values of ϵ. The RFF system can reject all the attacks with an accuracy of 1, except

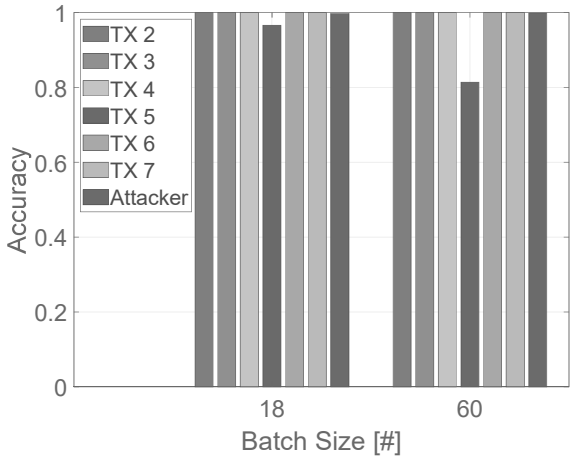

Fig. 5. Performance of the image-based RFF system after adversarial training with images generated from the WGAN-GP, with samples batch size of 18 and 60.

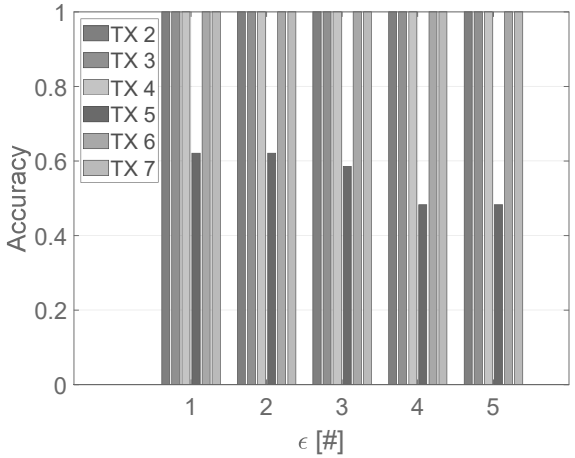

Fig. 6. Performance of the image-based RFF system after adversarial training with images generated through the FGSM with $\epsilon = 5$.

the ones affecting the device with ID 5, reporting a lower success rate. However, we observe a lower accuracy of about 0.81 with $\epsilon = 5$, due to the low performance associated with the device with ID 5 (0.52). The model built for TX5 is the least robust, similarly to Fig. 5. This result highlights that, besides randomness implicit in model construction, the I-Q samples collected for TX5 require building less robust models, more vulnerable to interference and channel fluctuations.

Overall, the results show that the effectiveness of adversarial training, especially towards FGSM-based attacks, depends significantly on the model built for the specific transmitter, while still achieving a general improvement compared to the baseline RFF system. Such improvements come at the cost of a performance drop in the classification of legitimate transmitters. Although this drop is very limited in our use case, it could become problematic for massive deployments.

VI. Conclusions and Future Work

In this paper, we carried out an experimental evaluation of the effectiveness of AML attacks and mitigation strategies for image-based RFF systems. We demonstrated that current state-of-the-art image-based RFF systems are prone to attacks based on FGSM and GANs. At the same time, such systems can be enhanced to detect such attacks at the expense of a slightly reduced classification accuracy. In the future, we plan to investigate further the effectiveness of AML attacks on RFF systems by evaluating the capability of modern SDRs to mitigate wireless channel fluctuations and precisely control I-Q samples at the receiver in various real-world scenarios.

Acknowledgments

This work was partially supported by the INTERSECT project, Grant No. NWA.1162.18.301, funded by the Netherlands Organisation for Scientific Research (NWO), the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, with reference to the partnership on ”Telecommunications of the Future” (PE00000001 - program ”RESTART”, CUP: D93C22000910001) and the national center on ”Sustainable Mobility” (CN00000023 - program “MOST”, CUP: D93C22000410001), the PRIN project no. 2017NS9FEY entitled ”Realtime Control of 5G Wireless Networks: Taming the Complexity of Future Transmission and Computation Challenges” and the PON AGREED (ARS01 00254) funded by the Italian MUR, and by ”The house of emerging technologies of Matera (CTEMT)” project funded by the Italian MIMIT.

References

[1] A. Al-Shawabka, et al., “Exposing the Fingerprint: Dissecting the Impact of the Wireless Channel on Radio Fingerprinting,” in *IEEE INFOCOM 2020*. IEEE Press, 2020, p. 646–655.

[2] B. Hamdaoui, et al., “Deep-learning-based device fingerprinting for increased lora-iot security: Sensitivity to network deployment changes,” *IEEE Network*, vol. 36, no. 3, pp. 204–210, 2022.

[3] S. Alhazbi et al., “The Day-After-Tomorrow: On the Performance of Radio Fingerprinting over Time,” *ACSAC 23 - Annual Computer Security Applications Conference*, 2023.

[4] G. Oligeri, et al., “PAST-AI: Physical-Layer Authentication of Satellite Transmitters via Deep Learning,” *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 274–289, 2022.

[5] X. Lu, et al., “Reinforcement Learning Based Physical Cross-Layer Security and Privacy in 6G,” *IEEE Commun. Surveys Tuts.*, 2022.

[6] L. Sun, et al., “Robustness of Deep Learning-Based Specific Emitter Identification under Adversarial Attacks,” *Remote Sensing*, vol. 14, no. 19, p. 4996, 2022.

[7] B. Liu, et al., “Robust Adversarial Attacks on Deep Learning Based RF Fingerprint Identification,” *IEEE Wireless Communications Letters*, 2023.

[8] D. Roy, et al., “RFAL: Adversarial Learning for RF Transmitter Identification and Classification,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 2, pp. 783–801, 2020.

[9] Jagannath, A. et al., “A Comprehensive Survey on Radio Frequency (RF) Fingerprinting: Traditional Approaches, Deep Learning, and Open Challenges,” *Computer Networks*, p. 109455, 2022.

[10] Y. Shi, et al., “Generative adversarial network in the air: Deep adversarial learning for wireless signal spoofing,” *IEEE Trans. on Cognitive Commun. and Netw.*, vol. 7, no. 1, pp. 294–303, 2020.

[11] J. Liu, et al., “Adversarial Machine Learning: A Multilayer Review of the State-of-the-Art and Challenges for Wireless and Mobile Systems,” *IEEE Communications Surveys & Tutorials*, vol. 24, no. 1, pp. 123–159, 2021.

[12] D. Adesina, et al., "Adversarial Machine Learning in Wireless Communications Using RF Data: A Review," *IEEE Communications Surveys & Tutorials*, 2022.
[13] J. Han, et al., "Adversarial Training in Affective Computing and Sentiment Analysis: Recent Advances and Perspectives [Review Article]," *IEEE Computational Intelligence Magazine*, vol. 14, no. 2, pp. 68–81, 2019.
[14] T. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. USA: Prentice Hall PTR, 2001.
[15] K. Burse, et al., "Channel Equalization using Neural Networks: A Review," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 3, pp. 352–357, 2010.

BIOGRAPHIES

Lorenzo Papangelo received his Master's degree in Telecommunication Engineering in 2023 from Politecnico di Bari. His research interests are in Wireless Network security, networking and privacy.

Maurizio Pistilli received his Master's degree in Telecommunications Engineering in 2023 from Politecnico di Bari, Italy. His research interests include the domain of security in Wireless Networks, Machine Learning, networking and privacy.

Savio Sciancalepore is Assistant Professor at TU/e, Eindhoven, Netherlands. He received the PhD degree in 2017 from Politecnico di Bari, Italy. From 2017 to 2020, he was Postdoctoral researcher at HBKU, Doha, Qatar. His research interests are in network security and privacy issues in IoT, Mobile and Wireless Networks.

Gabriele Oligeri is Associate Professor at Hamad bin Khalifa University, College of Science and Engineering, Qatar. He received his Ph.D. degree in Computer Engineering from the University of Pisa. His research interests include security and privacy of cyber-physical systems.

Giuseppe Piro is an Associate Professor at Politecnico di Bari (Italy). His main research interests include wireless networks, simulation tools, 5G and beyond, security, nano-scale communications, Internet of Things, and Software-Defined Networking. He serves as Associate Editor for Internet Technology Letter (Wiley), Wireless Communications and Mobile Computing (Hindawi), Sensors (MDPI).

Gennaro Boggia received the Dr.Eng. and Ph.D. degrees (with Hons.) in electronics engineering from the Politecnico di Bari (Italy). He is a Full Professor of Telecommunication with the Politecnico di Bari. His research interests include wireless and cellular communications, protocol stacks for industrial applications, Internet measurements, and network performance evaluation.