



HAL
open science

How to Design Channel-Resilient Database for Radio Frequency Fingerprint Identification?

Alice Chillet, Robin Gerzaguët, Karol Desnos, Matthieu Gautier, Elena Simona Lohan, Mikko Valkama

► **To cite this version:**

Alice Chillet, Robin Gerzaguët, Karol Desnos, Matthieu Gautier, Elena Simona Lohan, et al.. How to Design Channel-Resilient Database for Radio Frequency Fingerprint Identification?. IEEE International Conference on Communications (ICC), Jun 2024, Denver, United States. hal-04617952

HAL Id: hal-04617952

<https://inria.hal.science/hal-04617952>

Submitted on 20 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

How to Design Channel-Resilient Database for Radio Frequency Fingerprint Identification?

Alice CHILLET[†], Robin GERZAGUET[†], Karol DESNOS^{*}, Matthieu GAUTIER[†],
Elena Simona LOHAN[‡], Erwan NOGUES[§], and Mikko VALKAMA[‡]

[†] Univ Rennes, CNRS, IRISA, firstname.lastname@irisa.fr

^{*} Univ Rennes, INSA Rennes, CNRS, IETR - UMR 6164, Rennes, France, firstname.lastname@insa-rennes.fr

[‡] Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland

[§] DGA-MI

Abstract—This paper proposes to explore the Radio Frequency Fingerprint (RFF) identification with a virtual database generator. RFF is a unique signature created in the emitter transmission chain by hardware flaws. These flaws may be used as a secure identifier as they cannot be easily replicated for spoofing purposes. In recent years, the RFF identification relies mainly on Deep Learning (DL), and large databases are consequently needed to improve identification in different environmental conditions. In this paper, we introduce a virtual database and suggest utilizing it for the examination of three crucial aspects when creating a RFF database: the number of signals required to perform DL classification, the impact of RFF similarities between emitters, and the propagation channel impact in static and dynamic contexts. For instance, such analysis shows that data augmentation with 10 channels improves accuracy classification up to 65% in a scenario where RFFs are close from a transmitter to another.

Index Terms—Deep Learning, Radio Frequency Fingerprint, RF impairments models.

I. INTRODUCTION

In recent years, secure authentication has been the topic of many research papers [1], and particularly the Radio Frequency Fingerprint (RFF) identification has been largely studied [2]. The purpose of RFF is to uniquely identify a device by recognizing flaws in the emitted signal. These flaws are created by hardware impairments of the transmitter. The impairments create unique electromagnetic distortions in the transmitted signal [3], and these distortions are used to differentiate devices. While, in most telecommunication standards, identification methods are based on the meta-data of the communication protocols such as a Media Access Control (MAC) address, the RFF identification can be combined with such classic identification to improve and ensure identification without spoofing [1].

While RFF identification principle was born in the 2000's, using parametric models, the number of RFF identification by classification methods has recently exploded with the advent of Deep Learning (DL) [1], [4]. In particular, supervised DL is massively used in RFF classification, as it automatically learns how to classify radio transmitters by recognizing complex patterns from labeled signals. However, the classification accuracy of such methods dramatically depends on the database used to train the network. Therefore, there is a strong need for large and robust databases composed of raw labeled signals [2]

from different transmitters to ensure RFF recognition in many environmental conditions.

Many experimental studies have shown promising results in static propagation channel conditions [4]–[6]. However, in dynamic channel conditions, RFF classification is disturbed by the channel [4], [6]. To tackle this issue, three approaches are proposed in the state-of-the-art (SoA). First, [6] proposed to mitigate the channel effect through equalization, but this requires a complete demodulation chain. Second, some authors proposed to use a large database and physical or virtual data augmentation [5], [7]. The database used by Soltani et al. [7] is the largest experimental RFF database, composed of two datasets, with more than 5000 devices for WiFi and ADS-B protocols [7]. Unfortunately, this database has no public access. In [5], we propose to use a public database called WiSig [8], to present physical data augmentation. The dataset used was created by six WiFi transmitters over four different days of captures performed over a month, and captured by 12 devices with 20 MHz bandwidth from different references. Third, Sankhe et al. [6] suggested introducing controlled impairments at the transmitter side to enhance identification robustness. This recommendation arises from the fact that the transmitters featured in their database are produced with low variability components, resulting in minimal Radio Frequency Front-end (RFF) variations between two devices.

All of these studies, underscore the significance of the database design, particularly the database size, the choice of transmitters, and the diversity of propagation channels. However, the creation of an experimental database for RFF requires the capture and labeling of a substantial volume of signals across various environmental configurations to mitigate the static bias inherent to the propagation channel. Furthermore, a difficult task is to pre-assess the required size of the database including i) the number of signals per transmitter essential for effective training, ii) the number of environmental configurations needed, and iii) the similarities between transmitter impairments i.e. how radio frequency impairments differ from one transmitter to another.

In this paper, we propose to perform a pre-assessment using a comprehensive and realistic simulation environment akin to a digital twin, allowing the exploration of degrees of freedom that are challenging to investigate through real acquisitions.

With this in mind, this paper proposes a new database generator based on a virtual model of both the impairments and the propagation channel scenarios and focuses on the three main issues highlighted in the SoA: the RFF similarity between transmitters, the number of signals required in the database, and the channel impact. The core contributions of this paper are as follows:

- 1) We present a virtual database complex baseband model of RFF with four impairments models.
- 2) We study the impact of the transmitters impairments similarities on classification accuracy.
- 3) We study the impact of the number of signals on the classification accuracy and demonstrate how the number of required signals is related to the similarities of RFF.
- 4) We study the impact of the propagation channel and demonstrate how the RFF identification is enhanced with channel augmentation.

The structure of this paper is as follows: In Section II, the transmission models, RFF models, and the DL classification are presented. Section III introduces the transmission and RFF models used in creating the virtual database. Following that, Section IV showcases our experiments and results, while also discussing the advantages of a virtual database. Finally, the paper concludes in Section V.

II. RFF CLASSIFICATION

A. RF transmission chain model data and RFF

Before the transmission, the source information data is converted into symbol sequences. Symbol sequences are converted into IQ samples and transmitted via the RF chain (Tx, Figure 1). A Digital Analog Converter (DAC) transforms the signal into the analog domain to yield $x(t)$. The Local Oscillator (LO) modulates it at carrier frequency f_{RF} , and the Power Amplifier (PA) amplifies the signal, creating $x_{ant}(t)$ for transmission via the antenna.

All those components distort the signal and create the signature called the RFF of the transmitter Tx denoted $\mathcal{F}_{\text{RFF}_{\text{Tx}}}$. The emitted signal could be modeled by:

$$x_{ant}(t) = \mathcal{F}_{\text{RFF}_{\text{Tx}}}(x(t)) = \mathcal{F}_{\text{PA}} \circ \mathcal{F}_{\text{LO}} \circ \mathcal{F}_{\text{DAC}}(x(t)), \quad (1)$$

where \circ represents the function composition, and \mathcal{F} represents a component behaviour, including its distortion.

In Figure 1, the channel block represents the wireless communication environment defined by the noise, interference signals, and the multi-path and fading channels that could impact the signal. The propagation channel is modeled by $\mathcal{F}_{\text{channel}}$.

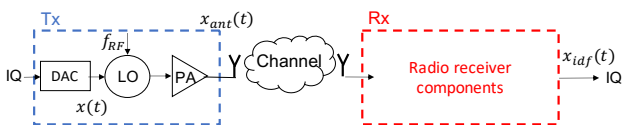


Fig. 1. Transmission and reception chain.

The receiver also has an RFF, but investigating its impact is beyond the scope of this paper. The received signal can thus be expressed as

$$x_{idf}(t) = \mathcal{F}_{\text{channel}} \circ \mathcal{F}_{\text{RFF}_{\text{Tx}}}(x(t)). \quad (2)$$

$\mathcal{F}_{\text{channel}}$ creates some disturbances that make difficult the Tx RFF identification. This issue is largely studied in the SoA [4], [7] and will be called the channel or environmental conditions impact in the rest of the paper. The environmental conditions represent an important bias in database construction for RFF identification with DL solutions.

B. RFF identification with deep learning

To perform identification with DL, the network is trained from labeled data. As classical methods, the database is split in training and test sets. The network predicts labels corresponding to the signals and then compares them to the actual labels using cross-entropy as the loss function for implementing back-propagation to update the network.

To evaluate the network classification performance we use the F1 score. The F1 score is calculated on the batch sequences as follows:

$$F1 = \mathbb{E}_{c \in \mathcal{C}} \left(\frac{2}{\frac{1}{P(c)} + \frac{1}{R(c)}} \right), \quad (3)$$

$$\text{with } \begin{cases} P(c) = \frac{t_p(c)}{t_p(c) + f_p(c)} \\ R(c) = \frac{t_p(c)}{t_p(c) + f_n(c)} \end{cases},$$

where $\mathbb{E}[\cdot]$ stands for the expectancy operator applied here on all classes $c \in \mathcal{C}$. $P(c)$ is called the precision for the class c and is a function of the number of true positives $t_p(c)$ and false positives $f_p(c)$. $R(c)$ is the recall for the class c and is function of $t_p(c)$ and the false negatives $f_n(c)$.

The network chosen in this paper has been proposed and studied for RFF identification in [5], [9], which is one of the first contributions on RFF with DL solution and public database. It is a Convolutional Neural Network (CNN) inspired by AlexNet, with 4 convolutional layers. Each layer is composed of two blocks of 128 filters of size 7×1 and 5×1 and a max-pooling stage. The chosen activation function is ReLu and the optimizer is Adam. After the 4 convolutional layers, the CNN has 3 fully connected layers with 256 nodes, 128 and the number of transmitters to identify, in this case, 6 classes.

As inputs, the network takes complex baseband signals without pre-processing. These signals are split in two raw I and Q data, and the input size N is 256 samples defined as it was done in [5], [8]. This architecture is presented in Figure 2 and has 1,232,774 parameters. The CNN is trained on the GPU NVIDIA Quadro RTX 5000.

III. VIRTUAL DATABASE AND RADIO MODELS

Creating a virtual database requires wireless communication models, hardware impairments models, and wireless propagation channel models. This section describes models implemented in the virtual databases used in this paper. database.

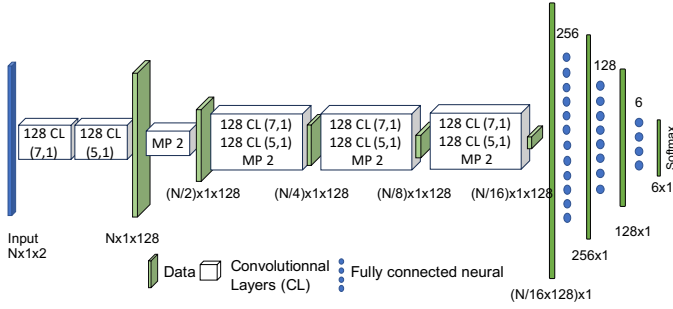


Fig. 2. Deep Learning Architecture.

A. IQ symbols

We consider that the binary sequence is modulated by Quadrature Amplitude Modulation (QAM) and then followed by an Orthogonal Frequency-Division Multiplexing (OFDM). The OFDM modulation is extensively employed in standard communications, such as WiFi. This is especially pertinent in the context of the RFF database, where the signal exhibits significant amplitude variations. This characteristic renders it particularly compelling for the analysis of non-linear imperfections.

B. Transmitter impairment models

The impairments modeling is described in Figure 3 and is based on SoA models. Without loss of generality, the models remain valid for low intermediate frequency conversion structures, assuming sufficient rejection of image frequencies. The signal is multiplied by a carrier generated from a LO, and different impairments occur in the chain. As Zhang et al. [10], we focus on the main features: CFO impairments, gain and phase IQ imbalance, Phase Noise (PN), and PA nonlinearity. The model of each impairment is described hereafter.

Before the LO, the analytical signal is modeled as:

$$\underline{x}(t) = x_I(t) + jx_Q(t), \quad (4)$$

where x_I and x_Q represent the real part and imaginary part of the complex signal (all complex variables will be underlined in the rest of the paper). The LO allows modulating the signal

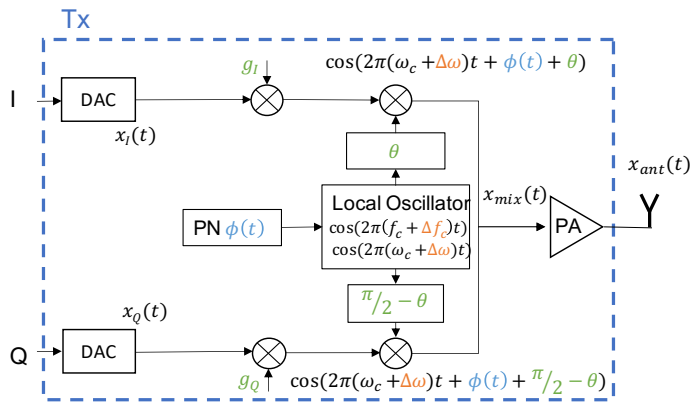


Fig. 3. Transmitter chain architecture and impairments.

to a carrier frequency noted f_c .

CFO impairments: The LO modulates the signal at the ideal carrier frequency, f_c . However, CFO impairments introduce a frequency offset Δf , resulting in the effective carrier frequency, f_0 , noted as:

$$f_0 = f_c + \Delta f. \quad (5)$$

For the sake of brevity, the models are expressed in terms of angular frequency with $\omega_* = 2\pi f_*$. With such impairment, the modulated signal $x_{mix}(t)$ is expressed as a gain and phase error by:

$$x_{mix}(t) = x_I(t) \cos((\omega_c + \Delta\omega)t) - x_Q(t) \sin((\omega_c + \Delta\omega)t), \quad (6)$$

which can be equivalently written as:

$$\underline{x}_{mix}(t) = \underline{x}(t)e^{j(\omega_c + \Delta\omega)t}.$$

IQ imbalance impairments: In the presence of IQ imbalance, the LO can be expressed according to Figure 3 in the form:

$$\underline{X}_{LO}(t) = g_I \cos(\omega_0 t + \theta) - jg_Q \sin(\omega_0 t - \theta),$$

where θ is the phase impairment considering balanced, and g_I and g_Q the gain impairments. The expression can be simplified as done by Valkama et al. [11]:

$$\underline{X}_{LO}(t) = K_1 e^{-j\omega_0 t} + K_2 e^{j\omega_0 t}, \quad (7)$$

$$\text{where } K_1 = \frac{g_I e^{-j\theta} + g_Q e^{j\theta}}{2}, \quad K_2 = \frac{g_I e^{j\theta} - g_Q e^{-j\theta}}{2}.$$

The signal $\underline{x}_{mix}(t)$ at the output of the LO in the presence of IQ imbalance is expressed:

$$\begin{aligned} \underline{x}_{mix}(t) &= \underline{x}(t) \times \underline{X}_{LO}(t), \\ &= \underline{x}(t)K_1 e^{-j\omega_0 t} + \underline{x}(t)K_2 e^{j\omega_0 t}. \end{aligned} \quad (8)$$

In our model, as it is often done in the SoA, a balanced IQ mismatch is considered with $g_I = g_Q = \frac{g}{2}$.

Phase noise impairments: The PN has been modeled in the literature with different models, like Gaussian, Wiener or Lorentz and we focus on the Wiener model as it is a commonly used case in the literature to model free oscillator [12]. The LO PN $\phi(t)$ may be modeled [13]:

$$\phi(t) = \sqrt{c}B(t), \quad (9)$$

where $B(t)$ denotes a standard Wiener process and parameter c describes the LO quality called diffusion rate [13]. $B(t)$ is defined as $B(t_2) - B(t_1)$ with t_1 and t_2 correspond to the duration that forms the noise of variance $\sqrt{t_2 - t_1}\mathcal{N}(0, 1)$. In the rest of the paper, we consider the digital Wiener PN model parameterized by its state noise variance σ_ξ^2 [14].

Considering all impairments described from now, the output of the LO could be expressed by:

$$\underline{x}_{mix}(t) = \underline{x}(t)K_1 e^{-j(\omega_0 t + \phi(t))} + \underline{x}(t)K_2 e^{j(\omega_0 t + \phi(t))}. \quad (10)$$

Power amplifier impairments: To model the memoryless nonlinear effect of the power amplifier, the Saleh model used

in SoA [10] is chosen. The non-linearity is modeled as amplitude/amplitude (AM/AM) denoted $A(t)$ and amplitude/phase (AM/PM) distortions denoted $\xi(t)$. they are modeled as:

$$A(t) = \frac{\alpha_{AM} |\underline{x}_{mix}(t)|}{1 + \beta_{AM} |\underline{x}_{mix}(t)|^2}, \quad (11)$$

$$\xi(t) = \frac{\alpha_{PM} |\underline{x}_{mix}(t)|^2}{1 + \beta_{PM} |\underline{x}_{mix}(t)|^2},$$

where $|\cdot|$ denoted L1 norm. $\alpha_{AM}, \alpha_{PM}, \beta_{AM}, \beta_{PM}$ are the parameters of Saleh model [10].

Finally, the signal $\underline{x}_{PA}(t)$ after the PA is modeled as

$$\underline{x}_{PA}(t) = A(t)e^{j(\angle \underline{x}_{mix}(t) + \xi(t))}, \quad (12)$$

where \angle represent the angle of $\underline{x}_{mix}(t)$.

C. Channel model

The channel models implemented in our database generator are multipath fading channel models. Two different models are specified with different delay profiles. The models are the Extended Vehicular A (EVA) model and the Extended Typical Urban model (ETU) [15]. EVA model represents a medium delay spread environment while the ETU model is a low delay spread environment. The signal obtained after the channel is modeled as:

$$\underline{x}_{channel}(t) = \underline{x}_{PA}(t) * h(t) + n(t), \quad (13)$$

where $*$ is the convolution operator and $h(t)$ represents the propagation channel and $n(t)$ a Gaussian additive white noise of 20dB.

IV. EXPERIMENT AND RESULTS

A. Databases

The database generator allows to explore many scenarios and meta-parameters. In this paper, we only focus the exploration on a few:

- the similarities or dissimilarities between RFF emitters,
- the number of signals per device, $N_{signals}$, in the database and used to train our network,
- the number of propagation channel scenarios.

A signal is a sequence of 256 complex IQ samples sampled at 5.26 MHz, corresponding to the input size of the network [5], [8]. The baseline of the signal is created with OFDM symbols composed of 548 IQ samples with an FFT size of 512 and a cyclic prefix size of 36. The database used in this paper simulates a preamble scenario for which all transmitters transmit the same sequence of 64 signals (so that only the RFF can be learned).

The impairment models described in Section III-B, are defined for each transmitter. The similarity between two emitters RFF is modeled by different gap between the modeling RFF parameters and is defined by a percentage around a mean value. Finally, the effect of the wireless transmission is simulated by the channel model introduced in Section III-C. We consider channel realizations generated from the power delay profile, with a random Rayleigh distribution applied to

TABLE I
MEAN VALUE CHOSEN FOR IMPAIRMENT PARAMETERS.

Impairment	Parameters	Mean value
CFO	Δf	300 Hz
	g	1.5 dB
Imbalance	θ	2.5°
	σ_ξ^2	10^{-7}
PA	α_{AM}	2.1587
	β_{AM}	1.1517
	α_{PM}	4.0033
	β_{PM}	9.104

each path. Generated database is finally split into the training set and test set.

The database simulates $N_{Tx} = 6$ transmitters. The transmitters impairments are defined in a confidence interval of $i\%$ around the mean values defined in Table I. Except for θ , the parameter values $P_{Tx_k}^i$ for device $k \in [1, N_{Tx}]$ are computed as:

$$P_{Tx_k}^i = P_{min}^i + k \frac{(P_{max}^i - P_{min}^i)}{N_{Tx}}, \quad (14)$$

$$\text{with } P_{min}^i = \text{Mean Value}(1 - i\%), \quad (15)$$

$$P_{max}^i = \text{Mean Value}(1 + i\%), \quad (16)$$

with P_{min}^i the minimum of the impairment parameter in the $i\%$ similarities scenario and P_{max}^i the maximum. Four different similarity scenarios are studied in this paper, 5%, 3%, 2% and 1%. For θ parameters, $P_{Tx_k}^i$ follows (18) but P_{min}^i and P_{max}^i depend of the similarity scenario. For $i = 5\%$ and 3% , we set:

$$P_{min}^i = 0^\circ \quad P_{max}^i = 5^\circ, \quad (17)$$

while for $i = 2\%$ and 1% , we set:

$$P_{min}^i = 1^\circ \quad P_{max}^i = 4^\circ. \quad (18)$$

B. How close can the RFF of 6 devices be?

In this section, the impact of the impairment distribution (i.e. the similarity scenarios) on the training phase is studied. In this first simulation, a database is created for each similarity scenario: 5%, 3%, 2% and 1%. The databases are composed of 6 emitters with 1000 signals per emitters, and the propagation channel is not modeled. Each database is split into 90% and 10% to create training and test sets, respectively. Table II presents the F1 score evolution for both training and test sets, during the training phase for the different similarity scenarios. The training is ended when the network obtains a F1 score of 98% on the training set. The network has no difficulty to classify the 6 emitters in the 5% scenario. As the similarity increases between emitters impairments, the network needs more time to learn and classify the devices. The complexity of the classification problem increases as device impairments become closely situated, making it more challenging for the network to distinguish between them. Furthermore, in the case of the 1% similarity scenario, while the training accuracy presents a good progression, the test F1 score remains at approximately 17%, close to random value $1/N_{Tx}$. This suggests that the network struggles to learn RFF

TABLE II
MEAN F1 SCORE EVOLUTION DURING TRAINING PHASE FOR DIFFERENT SIMILARITY SCENARIOS.

F1 score at	200s, 315 epochs		400s, 630 epochs		600s, 940 epochs	
	Train	Test	Train	Test	Train	Test
5%	90%	86%	98%	95%	-	-
3%	66%	61%	92%	85%	98%	90%
2%	42%	20%	67%	56%	86%	70%
1%	35%	17%	71%	17%	89%	17%

due to the proximity of impairments and just overfits on the training set. For the rest of the paper, we no longer take into account the 1% scenario. However it reveals a countermeasure to RFF identification: by using emitters with very similar (or close) impairments, it will be very difficult for the network to separate and classify them.

C. How many signals do we need?

In this section, the impact of the number of signals in the training database is studied for a static propagation channel scenario. The databases are composed of 6 emitters with N_{signals} signals per emitter, the database is split in 90% and 10% to create training and test sets. So, for $N_{\text{signals}} = 1,000$, the number of signals in the training set is 900. This study is done for the three remaining similarity scenarios.

Figure 4 illustrates the F1 score obtained during testing as a function of the database training set size. The training stops when the F1 score on the training set reaches 98%. The training set and test set have been created with the same ETU channel models and therefore simulates the identification of 6 static emitters in different locations.

Table III presents the time and the number of epochs required to reach 98% of F1 score on the training set. For 1,000 signals, the network takes 839 epochs, 540 seconds and reaches 82% with the test set, shown on Figure 4, for 5% scenario compared to Table II where only 630 epochs, 400 seconds were required, to reaches 95% with the test set. We can therefore infer that the channel interference disrupted the network, creating challenges for the receiver in learning the transmitter characteristics.

Figure 4 shows the importance of the number of signals in the database, especially when the similarity between the device impairments is increased. If transmitters have significantly different RFF signatures e.g., low-quality devices or devices from different manufacturers, the database designer can limit the number of capturing signals to create the database. For example, in 3% scenarios, the use of 4,500 signals is not useful because the F1 score does not increase but the training time

TABLE III
TIME IN SECOND AND NUMBER OF EPOCH (EPC) REQUIRE TO REACH 98% OF F1 SCORE IN TRAINING SET.

	900 sig		1,800 sig		2,700 sig		4,500 sig	
	Time	Epc	Time	Epc	Time	Epc	Time	Epc
5%	540	839	627	481	642	324	652	200
3%	1864	1720	1333	977	1436	729	1659	508
2%	1260	1945	2270	1593	2792	1424	3433	1025

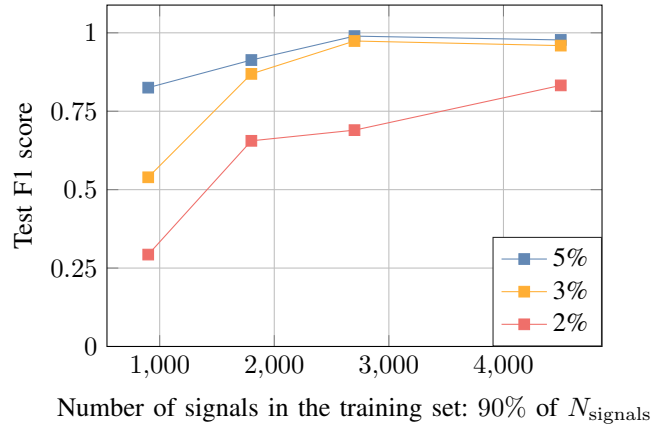


Fig. 4. F1 score obtained in test in function of the number of signals used to train the network when training has reached 98% of F1 score.

has increased compared to the scenario with 2,700 signals. However, emitters from the same reference, from the same manufacturer have a strong probability to have close RFFs and in this case, the database designer has to increase the number of signals, which must be taken into account in the automation of signal capture.

D. How many channel realizations should we have?

In this section, we study the propagation channel variations by applying data augmentation on the database. The database is extended by adding signals with impairments by different numbers of channels. To stress the capacity of the network to cope with different propagation channel, we consider that the ETU model is applied to the training set, while the EVA model is applied on the test set. The test set is composed of 100 signals per device and per EVA channel, and we simulate 100 different EVA channels for each device. The training set is composed of 900 signals per device and per channel.

Two different scenarios are investigated: i) the static scenario, in which the training database consists of only one channel; ii) the dynamic scenario, involving varying numbers of channels in the training set (2, 3, and 10). Figure 5 illustrates the mean F1 scores obtained from 100 EVA channels during testing as a function of the number of ETU channels used for network training. It is worth noting that the data augmentation offers two distinct advantages: it enriches the database with more signals and it enhances its diversity. To distinguish between these two advantages, we also plot in Figure 5 the results obtained by the configuration (1 channel, 1,800 signals represent by circles) alongside (1 channel, 900 signals).

Figure 5 demonstrates that the data augmentation in terms of increasing the number of signals, as indicated by the circular data points improves the test accuracy for example from 78% to 86% for 5% scenario similarity. However, it also highlights that simply increasing the number of signals is insufficient, as it does not achieve the same performance level as the configuration with 2 distinct channels (and thus 1,800 signals), here 94% for the 5% scenario. In terms of the number of channels, increasing the number during the training phase aids

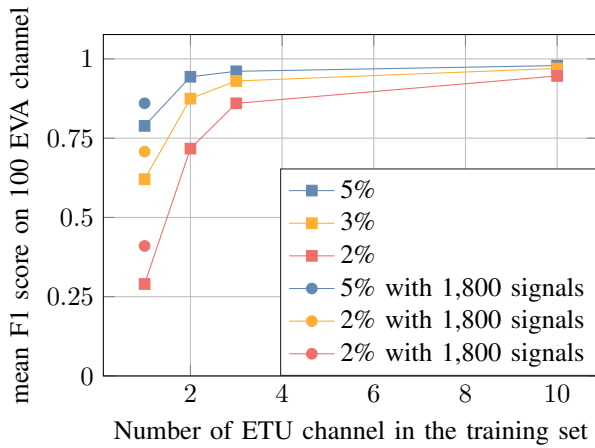


Fig. 5. Mean F1 score obtained in test in function of the number of channels used to train the network when training has reached 98% of F1 score.

the network in classifying devices during testing. Figure 5 can assist in determining the number of channel situations required to perform effectively under varying environmental conditions. This required number is also depending on the similarity between emitters.

V. CONCLUSION

This paper proposes to use a virtual database generator based on wireless transmission and RFF models to explore database design for RFF identification with DL. Our analysis showed that channel presence can destroy the RFF features of the emitters by deteriorating up to 35% the classification accuracy. We found that we need a very large number of signals per radio at least 2,700 for the scenarios studied. We also reveal that having similar RFF devices can be a countermeasure to avoid RFF identification. The virtual database generator can help to pre-assess the needed database design. The authors commit to release this generator as an open source tool immediately after acceptance of the paper.

ACKNOWLEDGEMENT

This work is funded by DGA and Brittany region under the Creach Lab founding and by the French National Research Agency (ANR) under the grant number ANR-22-CE25-0007-01 (RedInBlack project).

REFERENCES

- [1] X. Guo, Z. Zhang, and J. Chang, "Survey of Mobile Device Authentication Methods Based on RF Fingerprint," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. Paris, France: IEEE, Apr. 2019, pp. 1–6.
- [2] A. Jagannath, J. Jagannath, and P. S. P. V. Kumar, "A Comprehensive Survey on Radio Frequency (RF) Fingerprinting: Traditional Approaches, Deep Learning, and Open Challenges," *arXiv:2201.00680 [cs]*, 2022.
- [3] V. Brik, S. Banerjee, M. Gruteser, and S. Oh, "Wireless device identification with radiometric signatures," in *2008 Proceedings of the 14th ACM international conference on Mobile computing and networking - MobiCom 08*. San Francisco, California, USA: ACM Press, p. 116.
- [4] A. Al-Shawabka, F. Restuccia, S. D'Oro, T. Jian, B. Costa Rendon, N. Soltani, J. Dy, S. Ioannidis, K. Chowdhury, and T. Melodia, "Exposing the Fingerprint: Dissecting the Impact of the Wireless Channel

- on Radio Fingerprinting," in *2020 INFOCOM IEEE Conference on Computer Communications*, Toronto, ON, Canada, pp. 646–655.
- [5] A. Chillet, B. Boyer, R. Gerzaguet, K. Desnos, and M. Gautier, "Tangled program graph for radio-frequency fingerprint identification," in *2023 Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (IEEE PIMRC)*, 2023.
- [6] K. Sankhe, M. Belgiovine, F. Zhou, S. Riyaz, S. Ioannidis, and K. Chowdhury, "ORACLE: Optimized Radio Classification through Convolutional neural networks," in *2019 INFOCOM IEEE Conference on Computer Communications*, Paris, France, pp. 370–378.
- [7] N. Soltani, K. Sankhe, J. Dy, S. Ioannidis, and K. Chowdhury, "More Is Better: Data Augmentation for Channel-Resilient RF Fingerprinting," in *2020 IEEE Communications Magazine*, vol. 58, no. 10, pp. 66–72.
- [8] S. Hanna, S. Karunaratne, and D. Cabric, "Wisig: A large-scale wifi signal dataset for receiver and channel agnostic rf fingerprinting," *IEEE Access*, vol. 10, pp. 22 808–22 818, 2022.
- [9] K. Sankhe, M. Belgiovine, F. Zhou, L. Angioloni, F. Restuccia, S. D'Oro, T. Melodia, S. Ioannidis, and K. Chowdhury, "No Radio Left Behind: Radio Fingerprinting Through Deep Learning of Physical-Layer Hardware Impairments," in *2020 IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 1, pp. 165–178.
- [10] J. Zhang, R. Woods, M. Sandell, M. Valkama, A. Marshall, and J. Cavallaro, "Radio Frequency Fingerprint Identification for Narrow-band Systems, Modelling and Classification," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 3974–3987, 2021.
- [11] M. Valkama, M. Renfors, and V. Koivunen, "Blind I/Q signal separation-based solutions for receiver signal processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 16, pp. 1–11, 2005.
- [12] Z. Zhang, A. Hu, W. Xu, J. Yu, and Y. Yang, "An Artificial Radio Frequency Fingerprint Embedding Scheme for Device Identification," *IEEE Communications Letters*, vol. 26, no. 5, pp. 974–978, May 2022. [Online]. Available: <https://ieeexplore.ieee.org/document/9698034/>
- [13] T. Pollet, M. Van Bladel, and M. Moeneclaey, "BER sensitivity of OFDM systems to carrier frequency offset and wiener phase noise," *IEEE Transactions on communications*, vol. 43, no. 2/3/4, pp. 191–193, 1995.
- [14] R. Gerzaguet, L. Ros, F. Belvèze, and J.-M. Brossier, "On the performance of digital adaptive spur cancellation for multi-standard radio frequency transceivers," *Digital Signal Processing*, vol. 33, pp. 83–97, 2014.
- [15] *3GPP TS 36.104. Base Station (BS) radio transmission and reception. 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA)*.