



Diversification des Données pour l'Identification d'Empreinte RF

Alice CHILLET¹ Robin GERZAGUET¹ Karol DESNOS² Paul BAZERQUE¹ Erwan NOGUES³ Matthieu GAUTIER¹

¹Univ Rennes, CNRS, IRISA, France

²Univ Rennes, INSA Rennes, CNRS, IETR - UMR 6164, Rennes, France

³DGA-MI

Résumé – Ce papier propose d’améliorer l’identification par empreinte Radio Frequence (RF) quelques soit l’environnement de propagation en diversifiant les données transmises. Une empreinte RF est une signature unique créée par les distorsions des différents composants matériels du dispositif radio. L’identification par empreinte RF est principalement basée sur des techniques d’apprentissage supervisées Deep Learning, qui nécessitent une Base De Données (BDD) d’entraînement permettant un apprentissage robuste. La conception de cette BDD étant chronophage, nous proposons d’ajouter la diversité aux données pour accélérer jusqu’à 16 fois le temps d’acquisition des données par rapport à la diversité de canal.

Abstract – This paper proposes to improve the resilience of the Radio Frequency Fingerprint (RFF) identification to the propagation channel effects by massively diversifying the transmitted data. RFF refers to a unique signature resulting from signal distortions caused by the imperfections of the hardware transmission chains. RFF identification mainly relies on supervised Deep Learning (DL) techniques, which require an unbiased training database to avoid over-fitting the propagation channel environment. This can be achieved during signal acquisition by acquiring many different channel realisations. However, designing such database is time-consuming, mainly due to the diversity process. In this paper, instead of adding diversity to the channel, we propose to add diversity to the data to improve the data acquisition time of up to 16 times.

1 Introduction

L’empreinte Radio Frequence (RF) d’un dispositif est définie par les distorsions de sa chaîne de transmission, et peut servir d’identifiant sécurisé. Récemment, le nombre de contributions utilisant des méthodes de classification a explosé avec l’essor des techniques d’apprentissage automatique [1, 5]. Cependant, la précision de la classification des méthodes Deep Learning (DL) dépend considérablement de la Base De Données (BDD) utilisée pour former le réseau. Des études expérimentales ont montré des résultats prometteurs lorsque les conditions de propagation sont identiques pour les données d’entraînement et d’inférence [1, 3, 8]. Cependant, lorsque les conditions changent, la classification devient plus difficile en raison des perturbations du canal de propagation [1, 8]. Trois approches sont classiquement proposées dans l’état de l’art pour résoudre ce problème : (i) l’égalisation des canaux [1], (ii) l’ajout d’imperfection contrôlées côté émetteur, (iii) l’entraînement avec une grande diversité de canaux de propagation, appelé augmentation de données. Morin et al. [7] proposent ainsi une BDD avec des données fixe, transmission de préambule avec un récepteur mobile et Sankhe et al. [8] quant à eux, ont enregistré des préambules à différentes distances. Cette technique permet d’améliorer la précision de la classification dans différents contextes de transmission, mais elle s’accompagne d’une forte complexité lors de la création de la BDD. Pour réduire le temps d’enregistrement des données, une augmentation virtuelle a été proposée dans [10, 9] en ajoutant un modèle de canal de propagation coté récepteur comme pré-traitement. Cependant, la méthode la plus populaire reste la diversification lors de l’enregistrement, ce qui permet d’améliorer la précision de la classification de 40 % environ [10, 7].

Dans cet article, nous proposons de modifier le paradigme

de la diversification des données : plutôt que d’introduire de la diversité en variant les conditions du canal, nous ajoutons de la diversité aux données transmises. Notre approche permet d’atteindre la diversité nécessaire sans avoir à modifier les conditions de transmission à plusieurs reprises.

La section 2 présente l’identification d’empreinte RF, la résilience au canal, et motive le changement de paradigme par une simulation. Puis, la section 3 présente la diversification des données à l’aide d’un outil de simulation. Enfin, la section 4 présente les résultats obtenus grâce à nos BDD expérimentales. Le document se termine par la section 5.

2 Classification d’empreinte RF

Lors de la transmission d’information, les données binaires sont d’abord converties en séquences de symboles, puis transformées en échantillons IQ avant d’être transmises au convertisseur numérique-analogique qui transforme le signal dans le domaine analogique pour obtenir le signal $x(t)$. L’oscillateur local module ensuite le signal à la fréquence porteuse, puis l’amplificateur de puissance l’amplifie. Les imperfections des composants déforment le signal et créent la signature appelée empreinte RF de l’émetteur Tx, notée $\mathcal{F}_{\text{RFF}_{\text{Tx}}}$. Le signal est ensuite transmis par un canal de propagation qui altère plus ou moins le signal en raison du bruit additif, des interférences ou des canaux à trajets multiples et à évanouissements. Le canal est modélisé par $\mathcal{F}_{\text{canal}}$. L’empreinte du récepteur n’est pas prise en compte dans cet article. Le signal reçu $x_{\text{idf}}(t)$ peut donc s’écrire comme suit :

$$x_{\text{idf}}(t) = \mathcal{F}_{\text{canal}} \circ \mathcal{F}_{\text{RFF}_{\text{Tx}}}(x(t)), \quad (1)$$

où $\circ$ représente l'opérateur de composition de fonctions, avec $h(x) = g \circ f(x)$ désigne $h(x) = g(f(x))$. Les perturbations introduites par $\mathcal{F}_{\text{canal}}$ rendent difficile la reconnaissance de l'empreinte de l'émetteur. Il faut donc adopter une approche d'identification résistante au canal pour garantir la reconnaissance de l'émetteur dans différentes conditions de propagation.

Depuis 2019, plusieurs travaux se sont intéressés à l'amélioration de la résilience au canal de propagation des réseaux de neurones [10, 8, 7, 3]. Par exemple, Morin et al. [7] présentent leur dispositif expérimental appelé Cortex-Lab, et montrent que l'ajout de divers canaux à trajets multiples dans l'ensemble de données d'entraînement améliore la résilience du réseau dans différents scénarios de transmission, avec une amélioration de 40 % par rapport à l'entraînement statique, appelé « simple » dans leurs expériences. Toutefois, la phase d'enregistrement des signaux est une étape complexe, chronophage et nécessite du matériel spécifique (chambre anéchoïque, robot mobile, émulateur de canal...).

2.1 Entraînement et Diversité de canal

Cette sous-section présente la diversification des données basée sur la diversité des canaux de propagation, afin de mettre en évidence les limites de cette méthode et de motiver le changement de paradigme vers une diversité des données.

2.1.1 Réseau de neurone et BDD virtuelles avec RiFiFi

Le réseau utilisé dans cet article a déjà été proposé et étudié pour l'identification d'empreintes dans [8, 3]. Il s'agit d'un Réseau de neurones convolutif (CNN) avec 4 couches convolutives (CL). Chaque couche est composée de deux blocs de 128 filtres de taille 7×1 et 5×1 et d'un maxpool. La fonction d'activation est ReLu et l'optimiseur est Adam. Après les 4 CL, le CNN comporte 3 couches entièrement connectées avec, respectivement, 256 nœuds, 128 et le nombre d'émetteurs à identifier N_{Tx} (ici 5). Le réseau prend en entrée des signaux complexes en bande de base sans pré-traitement. Ces signaux sont divisés en deux données brutes I et Q, avec une taille d'entrée N de 256 échantillons, comme défini dans [6, 3]. Les données expérimentales étant très riches il est souvent difficile de comprendre ce que le CNN reconnaît, l'utilisation de données simulées à partir de modèle, comme c'est le cas avec RiFiFi [2], peut aider à la compréhension.

RiFiFi est un générateur de BDD virtuelle basé sur des modèles de transmission RF avec des dégradations. Dans ce papier, les dégradations modélisées sont les effets de l'oscillateur, à savoir le décalage de la fréquence porteuse (CFO), le déséquilibre IQ, le bruit de phase, ainsi que les effets de non-linéarité de l'amplificateur de puissance (PA). Différents niveaux de similarité de dégradation sont modélisés et représentés par un pourcentage noté $p\%$. Plus ce pourcentage est élevé, plus la différence entre les émetteurs est grande [4]. Ici, différents scénarios de transmission sont étudiés avec un intervalle de $p\%$, pour le déséquilibre IQ, le CFO, et le PA. Les différents modèles et valeurs utilisés pour l'expérience sont donnés dans [4] et accessibles sur github [2].

2.1.2 Expérience - simulation

Trois BDD virtuelles en OFDM sont créées avec différents niveaux de similarité d'empreinte RF $p = 10\%, 7\%$ et 5% .

Chaque BDD est composée de 5 transmetteurs ayant transmis en mode *Préambule*. Le pas d'apprentissage du réseau est fixé à 10^{-5} , le dropout dr à 0.25 et la taille de batch à 64.

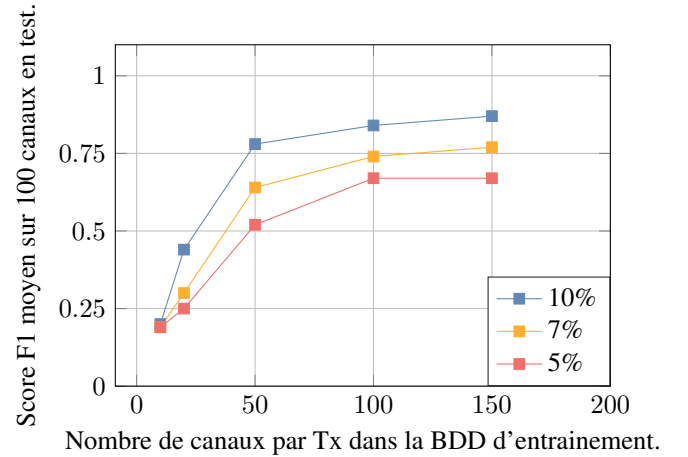


FIGURE 1 : Score-F1 moyen obtenu en test en fonction du nombre de canaux en entraînement.

La Figure 1 présente les scores F1 obtenus en test en fonction du nombre de canaux de propagation par transmetteur vus pendant l'entraînement. Le score F1 est obtenu en test avec d'autres réalisations de canal, et moyenné sur 5 simulations (données et entraînement). Les résultats révèlent qu'il faut plus de 100 canaux par émetteur, ce qui signifie qu'il faut capturer les signaux de chaque émetteur à 100 endroits différents. Ce processus est trop long et trop complexe pour de nombreuses applications d'identification d'empreinte RF.

Si changer de canal de propagation pour un émetteur ne prend que 5 secondes, Cela signifie qu'un nouvel enregistrement avec un canal de propagation différent est créé toutes les cinq secondes. La collecte de données suffisantes pour un seul émetteur nécessite donc plus de 500 secondes (environ 8 minutes). Et si le processus n'est pas automatisé et que le changement de contexte prend une minute entière, le temps nécessaire par émetteur dépasserait 1 heure et 30 minutes.

3 Diversifier les données pour garantir la robustesse et la résilience

3.1 Méthodologie

Lorsque le signal émis est toujours le même (préambule), avoir des canaux de propagation différents permet de créer de la diversité en réception comme présenté en partie supérieure de la Figure 2. Le réseau de neurones amené à classifier la donnée peut alors se concentrer sur l'empreinte RF. Dans le cas où il n'y a pas cette diversité, le canal de propagation peut devenir une signature saillante pour le réseau. C'est ici que se positionne notre contribution : nous proposons d'introduire de la diversité au niveau du signal émis, comme présenté sur la partie inférieure de la Figure 2. Au lieu d'émettre une séquence unique, nous émettons une longue séquence de symboles aléatoires, correspondant à un flux de bits aléatoires modulé en OFDM, appelé *Payload* ajoutant ainsi de la diversité. Cette BDD est ensuite utilisée pour entraîner le réseau à reconnaître l'empreinte RF dans un environnement de propagation bruyant

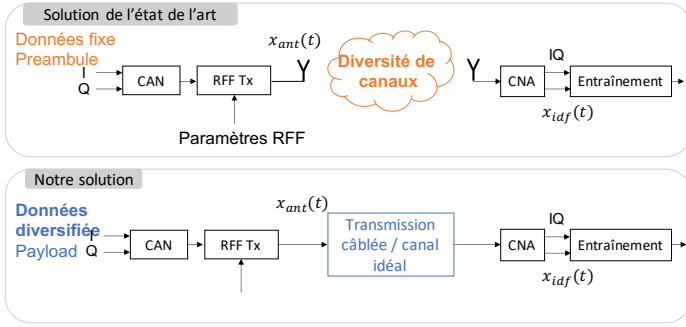


FIGURE 2 : Ajout de diversité dans les données transmises.

et diversifié. Une première étude est réalisée à l'aide de BDD virtuelle créée par RiFiFi en fonction des paramètres décrits précédemment, avec cette fois un mode de transmission *Payload*.

3.2 Preuve de concept en simulation

Trois grandes BDD avec des similarités de 10%, 7% et 5% ont été créées, chacune constituée de symboles OFDM aléatoires formant 180 000 séquences de 256 échantillons IQ. Le réseau est entraîné sans diversité de canaux (scénario S_1), avec les mêmes paramètres d'apprentissage que précédemment. Après l'entraînement, le réseau est évalué avec et sans canaux de propagation (S_2 et S_1). La Table 1 présente les scores F1 pour les deux scénarios. Nous avons choisi de créer 180 000 signaux d'entraînement par émetteur afin d'obtenir la même diversité qu'avec 1 800 signaux et 100 canaux par signal, comme dans le cas de l'augmentation traditionnelle des données. Les résultats montrent qu'il faut effectivement un très grand ensemble de données pour effectuer l'identification. Cependant, un seul enregistrement par émetteur est suffisant, ce qui simplifie l'enregistrement des données. En utilisant cette approche, l'enregistrement pour un émetteur prend environ 30 secondes, soit une réduction d'au moins 16 fois (en supposant une durée d'enregistrement optimiste : toutes les 5 secondes) par rapport à la méthode décrite précédemment.

TABLE 1 : Score F1 en fonction de la similarité (Payload).

Similarité p%	$N_{sig.}$	Test S_1	Test S_2
10%	180000	77%	59%
7%	180000	60%	47%
5%	180000	45%	39%

4 Expérimentations sur des données réelles

4.1 Base de données

Pour valider les résultats nous créons un dispositif expérimental avec six radio logicielles (SDR), cinq sont utilisées comme émetteurs et une, toujours la même, est utilisée comme récepteur, une X310. La Table 2 fournit des détails sur chaque émetteur : nom, références et l'étiquette qui lui a été attribuée dans la BDD. Les émetteurs sont tous différents, à l'exception des deux dispositifs E310. Cependant, le GPS de Tx3 a été

désactivé, ce qui se traduit par un CFO moins stable. La puissance d'émission de chaque SDR a été ajustée pour garantir des niveaux de puissance reçus identiques, ce qui évite au réseau de se fier aux écarts de puissance, qui peuvent varier en fonction des conditions du canal. Afin d'évaluer la capacité du réseau à détecter les émetteurs dans des conditions réelles, nous créons différents ensembles de données. Pour chaque scénario de transmission, les deux modes différents sont pris en compte : *Preamble* et *Payload*.

TABLE 2 : Émetteurs de la BDD expérimentale.

Label	Émetteur	Constructeur	Carte RF
Tx1	Blade RF	Nuand	AD9361
Tx2	ADALM Pluto	Analog device	AD9363
Tx3	E310 (no gps)	ETTUS	AD9361
Tx4	E310	ETTUS	AD9361
Tx5	X310	ETTUS	UBX-160

Dans le mode *Preamble*, tous les émetteurs envoient la même séquence de 30 symboles OFDM pendant 8 secondes. Dans le mode *Payload*, chaque émetteur envoie des symboles OFDM aléatoires pendant une durée de 30 secondes. Cinq scénarios de transmission sont ensuite définis :

- S_1, S_{1bis} : Transmission câblée, l'émetteur est allumé juste avant la transmission. BDD d'entraînement.
- S_2 : Identique à S_1 , mais les émetteurs sont allumés plusieurs heures avant la transmission. Les composants des émetteurs sont considérés comme chauds, ce qui peut avoir une incidence sur l'empreinte RF (ex : CFO).
- S_3 : Transmission « Over-The-Air » (OTA) dans une chambre anéchoïque, tous les émetteurs étant placés au même endroit.
- S_4 : Transmission OTA réalisée dans un bureau, avec des émetteurs placés à des endroits différents.

Le réseau est entraîné à l'aide des données S_1 , et plusieurs tailles de BDD d'entraînement sont proposées. Pour le mode *Preamble*, la taille de la BDD varie de 9 000 à 45 000 signaux par émetteur, tandis que pour le mode *Payload*, elle varie de 9 000 à 180 000 signaux par émetteur. L'entraînement se termine lorsque le réseau atteint un score F1 de 98 % sur les données d'entraînement. Ensuite, chaque réseau est évalué sur l'ensemble de données de test de chaque scénario S_k chacun composé de 5 000 signaux par émetteur. Les paramètres d'apprentissage sont définis comme suit : le pas d'apprentissage, $\gamma = 10^{-4}$, $dr = 0, 5$ et 64 pour la taille de batch.

La Table 3 donne les scores F1 obtenus en mode *Preamble* et *Payload* pour chaque scénario et pour différentes tailles $N_{sig.}$ de BDD d'apprentissage. La valeur du score F1 est calculée en moyenne sur trois graines aléatoires du réseau. La colonne 1 montre les résultats obtenus en utilisant la partie test de S_1 , tandis que la colonne 1bis fournit les résultats d'un second enregistrement effectué plusieurs jours plus tard pour assurer la reproductibilité du cas idéal. Les résultats des colonnes 1 et 1bis montrent que l'augmentation du nombre de signaux dans l'ensemble de données d'entraînement améliore le score F1 en test pour les deux cas. La comparaison du score F1 obtenu avec les scénarios S_1/S_{1bis} par rapport à S_2 montre une baisse de la performance de classification. Cela suggère que la température de la SDR, qu'elle soit chaude ou froide, a un impact relatif

sur la performance. Toutefois, l'augmentation du nombre de signaux dans la BDD d'entraînement compense cet effet.

L'évaluation du réseau sur S_3 montre une performance médiocre en mode préambule, et la diminution de la performance est perceptible au fur et à mesure que la difficulté du scénario augmente S_4 . En mode préambule, l'augmentation de la taille de la BDD n'a aucun effet sur les résultats car les données manquent de diversité. Pour améliorer la résilience, il faudrait réaliser de plusieurs enregistrements avec différents environnements de propagation pour ajouter de la diversité.

Toutefois, comme la création de cette BDD prendrait beaucoup de temps, nous préférons examiner les résultats de l'augmentation basée sur les données obtenus grâce au mode *Payload*. Comparé au mode *Preamble*, le mode *Payload* nécessite un plus grand nombre de signaux pour atteindre la même précision dans S_1 , ce qui est cohérent avec les conclusions tirées des données RiFiFi. Les résultats obtenus dans les colonnes 1, 1bis et 2 sont très similaires à ceux obtenus en mode *Preamble*. Les scénarios 3 et 4 obtiennent de meilleurs résultats en mode *Payload* qu'en mode *Preamble*, et l'augmentation de la taille de la BDD améliore les performances. Cela indique que la diversité des données liée au mode *Payload* améliore la robustesse du réseau. Toutefois, lorsque l'environnement de propagation introduit de grandes distorsions, la qualité des performances n'est pas suffisante et un prétraitement pourrait améliorer la qualité du signal, et donc la précision de la classification.

La Table 4 montre la matrice de confusion obtenue en test sur S_4 en mode *Payload*. Les résultats montrent des différences de classification entre les émetteur probablement due à la qualité de ceux-ci. Par exemple, la SDR Pluto, Tx2, connue pour sa faible qualité, est plus facile à distinguer que la X310 Tx5 de plus haute gamme. Cela suggère que l'approche proposée dans cet article, pourrait être utile pour identifier différents types d'émetteurs présentant des caractéristiques RFF distinctes, comme c'est apparemment le cas pour la Pluto et la E310.

5 Conclusion

Cet article propose une approche moins complexe pour créer des BDD d'entraînement destinées à l'identification d'empreintes RF par DL. Notre solution met l'accent non plus sur la diversité des canaux, mais sur celle des données, afin d'améliorer la résilience du réseau face à des environnements de propagation variés. Notre solution est au moins 16 fois plus rapide que les méthodes d'augmentation des données de l'état de l'art. Elle permet de distinguer de manière fiable certains

TABLE 3 : Score F1 en fonction de $N_{sig.}$ dans la BDD d'entraînement et du scénario de test.

$N_{sig.}$	Mode Préambule				
	1	1bis	2	3	4
9,000	76.7%	89.3%	74.5%	43.7%	27.0%
27,000	97.9%	97.2%	90.0%	34.5%	19.8%
45,000	97.8%	97.3%	89.3%	41.5%	24.9%
$N_{sig.}$	Mode Payload				
	1	1 bis	2	3	4
9,000	42.4%	47.8%	56.9%	37.8%	22.3%
45,000	91.1%	89%	88.7%	57.6%	48.4%
180,000	96.4%	94.5%	95.2%	60.46%	50.4%

TABLE 4 : Matrice de confusion obtenue sur S_4 (Payload)

Guess True	Tx ₁	Tx ₂	Tx ₃	Tx ₄	Tx ₅
TxTrue ₁	0.0	46.3	0.1	11.1	42.5
TxTrue ₂	1.9	90.8	0.2	6.8	0.2
TxTrue ₃	5.0	0.0	79.8	15.1	0.1
TxTrue ₄	0.7	0.0	1.4	61.8	36.1
TxTrue ₅	36.5	0.0	6.7	35.7	21.1

émetteurs, ce qui la rend particulièrement utile dans les applications où une reconnaissance précise n'est pas essentielle et où l'identification du type d'émetteur suffit. Dans le cadre de travaux futurs, nous pourrions étudier des méthodes plus efficaces pour générer des données Payload. Cela permettra de garantir que les données reflètent correctement les caractéristiques de l'émetteur et que le réseau peut être entraîné efficacement.

Remerciements

Ce travail a été financé par l'intermédiaire d'une bourse CREACH Lab et par le projet ANR-22-CE25-0007-01 (ANR RedInBlack).

Références

- [1] Amani AL-SHAWABKA et AL. : Exposing the Fingerprint : Dissecting the Impact of the Wireless Channel on Radio Fingerprinting. *In Proc. IEEE Conference on Computer Communications (INFOCOM)*, pages 646–655, Toronto, ON, Canada, 2020.
- [2] Alice CHILLET : *RiFiFi*.
- [3] Alice CHILLET, Baptiste BOYER, Robin GERZAGUET, Karol DESNOS et Matthieu GAUTIER : Tangled program graph for radio-frequency fingerprint identification. *In Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2023.
- [4] Alice CHILLET, Robin GERZAGUET, Karol DESNOS, Matthieu GAUTIER, Elena Simona LOHAN, Erwan NOGUES et Mikko VALKAMA : Understanding radio frequency fingerprint identification with RiFiFi virtual databases. *IEEE Open Journal of the Communications Society*, 2024.
- [5] Xinghao GUO, Zhen ZHANG et Jie CHANG : Survey of Mobile Device Authentication Methods Based on RF Fingerprint. *In Proc. IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–6, Paris, France, avril 2019.
- [6] Samer HANNA, Samurddhi KARUNARATNE et Danijela CABRIC : Wisig : A large-scale WiFi signal dataset for receiver and channel agnostic RF fingerprinting. *IEEE Access*, 10:22808–22818, 2022.
- [7] Cyrille MORIN, Leonardo S CARDOSO, Jakob HOYDIS, Jean-Marie GORCE et Thibaud VIAL : Transmitter classification with supervised deep learning. *In Proc. 14th EAI International Conference on Cognitive Radio-Oriented Wireless Networks (CrownCom)*, pages 73–86, 2019.
- [8] Kunal SANKHE, Mauro BELGIOVINE, Fan ZHOU, Luca ANGIOLONI, Frank RESTUCCIA, Salvatore D'ORO, Tommaso MELODIA, Stratis IOANNIDIS et Kaushik CHOWDHURY : No Radio Left Behind : Radio Fingerprinting Through Deep Learning of Physical-Layer Hardware Impairments. *IEEE Transactions on Cognitive Communications and Networking*, 6(1):165–178, 2020.
- [9] Guanxiong SHEN, Junqing ZHANG, Alan MARSHALL, Mikko VALKAMA et Joseph CAVALLARO : Radio Frequency Fingerprint Identification for Security in Low-Cost IoT Devices. *arXiv :2111.14275 [eess]*, 2021.
- [10] Nasim SOLTANI, Kunal SANKHE, Jennifer DY, Stratis IOANNIDIS et Kaushik CHOWDHURY : More Is Better : Data Augmentation for Channel-Resilient RF Fingerprinting. *IEEE Communications Magazine*, 58(10):66–72, 2020.