Multi-Objective GAN-Based Adversarial Attack Technique for Modulation Classifiers

Paulo Freitas de Araujo-Filho, Georges Kaddoum, *Senior Member, IEEE*,
Mohamed Naili, Emmanuel Thepie Fapi, and Zhongwen Zhu, *Senior Member, IEEE*

Abstract—**Deep learning is increasingly being used for many tasks in wireless communications, such as modulation classification. However, it has been shown to be vulnerable to adversarial attacks, which introduce specially crafted imperceptible perturbations, inducing models to make mistakes. This letter proposes an input-agnostic adversarial attack technique that is based on generative adversarial networks (GANs) and multi-task loss. Our results show that our technique reduces the accuracy of a modulation classifier more than a jamming attack and other adversarial attack techniques. Furthermore, it generates adversarial samples at least 335 times faster than the other techniques evaluated, which raises serious concerns about using deep learning-based modulation classifiers.**

Index Terms—**Adversarial Attacks, Wireless Security, Modulation Classification, Deep Learning, Generative Adversarial Networks.**

I. Introduction

Due to its success in the most diverse fields, deep learning has been increasingly investigated and adopted in wireless communications. It has been recently used for channel encoding and decoding [1], resource allocation [2], [3], and automatic modulation classification (AMC) [4], [5]. More specifically, deep learning-based modulation classifiers have been replacing traditional AMC techniques because they achieve better classification performance without requiring manual feature engineering [6]–[8].

However, deep learning models have been shown to be vulnerable to adversarial attacks, which puts into question the security and reliability of wireless communication systems that rely on such models [6], [9]–[12]. Adversarial attacks introduce specially crafted imperceptible perturbations that cause wrong classification results. Thus, they can force a deep learning-based modulation classifier on a receiver to misidentify the modulation mode used so that a signal is not correctly demodulated and the communication compromised.

Adversarial attacks can be classified as white or black-box attacks, depending on the knowledge they require from their target models. White-box attacks require a complete knowledge of the classifier's model, such as training data, architecture, learning algorithms, and hyper-parameters [13]. Black-box attacks, on the other hand, assume a more feasible scenario in which the attacker has access to only the model's output [13]. Furthermore, the authors of [14] define three more restrictive and realistic black-box threat models: query-limited, partial-information, and decision-based. The query-limited scenario considers that attackers have access to only a limited number of the model's outputs. The partial-information scenario considers that attackers have access to only the probabilities of some of the model's classes. Finally, the decision-based scenario considers that attackers have access to only the model's decision, i.e., the class to which it assigns a given data sample.

Although existing adversarial attacks pose risks to the use of deep learning in wireless communications, they require a complete knowledge about the target model [7], [15] or take too long to craft adversarial perturbations [11], [16], [17]. In this letter, we propose a novel input-agnostic decision-based adversarial attack technique that reduces the accuracy of modulation classifiers more and crafts perturbations significantly faster than existing techniques. Our technique is necessary for assessing the risks of using deep learning-based AMC in the more realistic scenario of decision-based black-box attacks. Moreover, it can significantly contribute to developing classifiers that are robust against adversarial attacks. The main contributions of our work are as follows: First, we combine generative adversarial networks (GANs) [18] and multi-task loss [19] to generate adversarial samples, by simultaneously optimizing their ability to cause wrong classifications and not being perceived. Second, we reduce the accuracy of modulation classifiers more and craft adversarial samples in a shorter time than existing techniques while following the decision-based black-box scenario. Third, we propose an input-agnostic adversarial attack technique that does not depend on the original samples to craft perturbations. It allows adversarial perturbations to be prepared in advance, further reducing the time for executing the adversarial attack. Finally, our work verifies that modulation classifiers are at an increased risk and urgently need to be enhanced against adversarial attacks.

II. Related Works

Although adversarial attacks were initially explored in computer vision applications, they have recently been investigated for wireless communication applications, such as AMC. The authors of [7] and [15] evaluate the robustness of a modulation

This work was funded and supported by Ericsson - Global Artificial Intelligence Accelerator (GAIA) in Montreal, Mitacs Accelerate Fellowship, the Tier 2 Canada Research Chair on the Next Generations of Wireless IoT Networks, Fonds de recherche du Québec B2X Scholarship, and CAPES.

P. F. de Araujo-Filho and G. Kaddoum are with the Electrical Engineering Department, École de Technologie Supérieure (ÉTS), University of Quebec, Montreal, Canada (email: paulo.freitas-de-araujo-filho.1@ens.etsmtl.ca, georges.kaddoum@etsmtl.ca).

P. F. de Araujo-Filho is with the Centro de Informática, Universidade Federal de Pernambuco (UFPE), Recife, Brazil (email: pfaf@cin.ufpe.br).

M. Naili, E. T. Fapi, and Z. Zhu are with Ericsson GAIA Montreal, Canada (email: mohamed.naili@ericsson.com, emmanuel.thepie.fapi@ericsson.com, zhongwen.zhu@ericsson.com).

classifier against four white-box adversarial attack techniques: fast gradient sign method (FGSM), projected gradient descent (PGD), basic iterative method (BIM), and momentum iterative method (MIM). The works show that the classifier's accuracy is significantly compromised. However, they do not measure the extent of the perturbation or the time it takes to craft adversarial samples. The work in [10] extends the white-box techniques FGSM, momentum iterative fast gradient sign method (MI-FGSM), and PGD to a power allocation application. It shows that adversarial attacks also pose a significant risk to regression-based applications, such as power allocation.

Several other works focus on black-box attacks, as they are more realistic for not requiring complete knowledge about the model [13]. The authors of [16] propose a boundary attack technique that requires access to only the classifier's decision. It relies on a probabilistic distribution to iteratively craft adversarial samples and reduce their distance to the original sample. Although it compromises the accuracy of classifiers, it takes more than a minute to craft a single adversarial sample. The authors of [17] propose an iterative algorithm to produce universal perturbations and show that state-of-the-art image classification neural networks are highly vulnerable. However, it takes more than 20 seconds to craft each adversarial sample. The authors of [11] propose an algorithm to craft adversarial attacks that is shown to require significantly less power than conventional jamming attacks to compromise the performance of a modulation classifier. Although the algorithm reduces the craft time of adversarial perturbations, it still requires hundreds of milliseconds to craft each adversarial sample.

III. Adversarial Attacks Formulation

Although deep learning models may be trained with a large amount of data, it is impractical to train them to cover all possible input feature vectors. As a result, the decision boundary found by a trained model may differ from the real one. The discrepancy creates room for a trained model to make mistakes [7]. Adversarial attacks craft perturbations to corrupt data samples so that they fall within that discrepancy area and are misclassified by a trained model. However, this is not a trivial task as the perturbations must be large enough to cause misclassifications but small enough to not be perceived. Therefore, given a sample x, the goal of an adversarial attacker is to find a perturbation δ and construct an adversarial sample $x_{adv} = x + \delta$ while satisfying

$$\min ||x_{adv} - x|| < \rho \tag{1}$$

and

$$f(x_{adv}) \neq f(x), \tag{2}$$

where $||\cdot||$ represents a chosen distance metric, ρ is the maximum imperceptible perturbation according to that metric, and f is the trained classifier target of the attack.

IV. Proposed Adversarial Attack Technique

In our work, we consider that our proposed adversarial attack technique is deployed as a malicious software on software-defined wireless receivers, an essential piece of modern wireless communication and 5/6G. Although injecting such malicious software is out of the scope of our work, it may be done by infecting software-defined radios with malware [20]. The malware can send samples to the receiver's modulation classifier and has access to its decisions. It intercepts incoming signals, craft perturbations δ, add the perturbations to original samples to form adversarial samples $x_{adv} = x+\delta$, and forward adversarial samples x_{adv} to the modulation classifier. Thus, the receiver's modulation classifier f identifies the modulation mode of x as $f(x_{adv})$. Since $f(x_{adv}) \neq f(x)$, the signal is not correctly demodulated, and the communication is compromised. Figure 1 shows our attack model. The analog-to-digital converter (ADC) forwards clean samples to the modulation classifier, but they are tampered by the adversarial attacker.

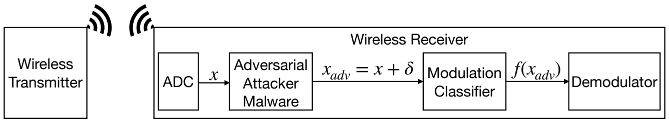

Fig. 1. Our attack model considers the adversarial attacker as malicious software on the wireless receiver

We propose a novel multi-objective adversarial attack technique by combining a GAN and multi-task loss. GANs estimate generative models by simultaneously training two competing neural networks: generator and discriminator [21]. The generator learns the probabilistic distribution of training data, and the discriminator learns how to distinguish between real data and data produced by the generator. We train a GAN so that its generator produces adversarial perturbations $\delta = G(z)$ from random latent vectors z and its discriminator learns to distinguish between clean samples x and adversarial samples $x_{adv} = x + G(z)$. We adopt the Wasserstein GAN (WGAN), which minimizes the Wasserstein distance between two probability distributions. It is easier to train than the original GAN, and does not suffer from the gradient vanishing problem [22], [23]. Although other GAN formulations, such as WGAN Gradient Penalty (WGAN-GP) [24], try to overcome WGAN's difficulty in enforcing the Lipschitz constant, the work in [25] shows that WGAN-GP does not necessarily outperform WGAN. In future work, we will evaluate our technique with other GAN formulations, such as WGAN-GP.

The WGAN discriminator estimates the Wasserstein distance by maximizing the difference between average critic score on real and fake samples. Besides, since we want the generator to produce perturbations rather than adversarial samples, fake samples are designated as $x + G(z)$ instead of $G(z)$. Thus, we minimize the discriminator loss given by $L_D = D(x + G(z)) - D(x)$. On the other hand, the WGAN generator has the opposite goal of maximizing the average critic score on fake samples. Hence, we minimize the generator loss given by $L_G = -D(x+G(z))$. However, such a L_G only accounts for minimizing the difference between x and x_{adv}, which corresponds to the condition of equation (1). It does not consider the condition of equation (2), which is to ensure that x and x_{adv} are assigned to different classes.

To ensure that our GAN considers the conditions of both equation (1) and equation (2), we modify the generator's loss to simultaneously optimize two objective functions that are given by L_{G1} and L_{G2}. L_{G1} represents the task of minimizing the difference between x and x_{adv} and is given by the original generator loss, hence $L_{G1} = -D(x + G(z))$. L_{G2} represents the task of ensuring that x and x_{adv} are assigned to different classes. It is given by the cross entropy loss between the class f assigns to x_{adv} and the label of x, hence $L_{G2} = CE(f(x + G(z)), y)$, where CE stands for the cross entropy loss largely adopted in classification problems and y is the label of x. During training, our technique leverages its access to the classifier's decisions to simultaneously optimize its ability to cause wrong classifications and not being perceived.

While most works that simultaneously learn multiple tasks manually tune a weighted sum of losses, we leverage the multi-task loss proposed in [19]. That work uses aleatoric uncertainty, which is a quantity that stays constant for all input data and varies between different tasks, to simultaneously optimize any two losses by optimally balancing their contributions as

$$L = \frac{1}{2\sigma_1^2} L_1 + \frac{1}{2\sigma_2^2} L_2 + \log \sigma_1 \sigma_2, \tag{3}$$

where L_1 and L_2 are any two losses, and σ_1 and σ_2 are learnable weights automatically tuned when training a neural network. Thus, while we train the GAN discriminator with

$$L_D = D(x + G(z)) - D(x), \tag{4}$$

we combine L_{G1} and L_{G2} with equation (3), where $L_1 = L_{G1}$ and $L_2 = L_{G2}$, so that our generator loss becomes

$$L_G = \frac{-D(x + G(z))}{2\sigma_1^2} + \frac{CE(f(x + G(z)), y)}{2\sigma_2^2} + \log \sigma_1 \sigma_2. \tag{5}$$

Figure 2 shows the training model, and Algorithm 1 shows the execution steps of our proposed adversarial attack technique.

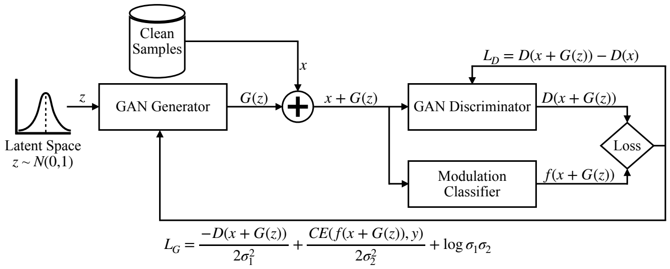

Fig. 2. Our proposed training model

Algorithm 1: Proposed Adversarial Attack Technique

```
1: Train a GAN according to equations (4) and (5)
2: for Each incoming sample $x$ do
3:     Compute $G(z)$
4:     Construct the adversarial sample $x_{adv} = x + G(z)$
5: end for
```

V. Methodology and Experimental Evaluation

We use the RADIOML 2016.10A dataset and VT-CNN2 modulation classifier designed by DeepSiG and publicly available in [4], [26] to evaluate our proposed adversarial attack technique. The dataset is constructed by modulating and exposing signals to an additive white Gaussian noise (AWGN) channel that includes sampling rate offset, random process of center frequency offset, multipath, and fading effects, as described in [4], [26]. Since our technique crafts adversarial samples on receivers, it is not subject to channel effects. In future work, we will consider them to enhance our proposed technique so that it sends adversarial samples over the air.

After modulation and channel modeling, the signals are normalized and packaged into 220,000 samples of in-phase and quadrature components with length 128, each associated with a modulation scheme and a signal-to-noise ratio (SNR). SNR is a measure of a signal's strength. It is the ratio between the power of the signal and of the background noise, i.e., $SNR_{[dB]} = 10 \log(\frac{P_{signal}}{P_{noise}})$, where P is the signal power. Eleven different modulation schemes (eight digital and three analog) are possible: 8PSK, BPSK, QPSK, QAM16, QAM64, CPFSK, GFSK, PAM4, WBFM, AM-DSB, and AM-SSB. Twenty different SNRs, ranging from -20 dB to 18 dB in steps of 2 dB, are possible. Twenty percent of the samples are reserved as a testing set to measure the VT-CNN2 modulation classifier's accuracy on clean and adversarial samples.

The VT-CNN2 modulation classifier relies on deep convolutional neural networks and classifies samples among the eleven modulation schemes in the dataset. Figure 3 shows VT-CNN2's architecture. Although the softmax layer gives the probability of membership for each class, we consider the classifier's output to be only its final decision, i.e., the modulation class that has the highest probability. Thus, $f(x+G(z))$ is the predicted label of one of the modulation schemes considered.

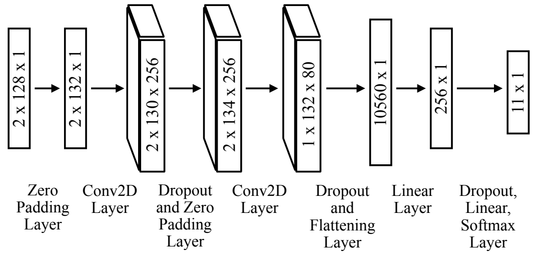

Fig. 3. VT-CNN2 neural network architecture

Finally, Figures 4 and 5 show the GAN's generator and discriminator architectures. They were optimized using the Optuna framework [27], which automatically searches for the optimal hyper-parameters, and the early stopping mechanism to avoid overfitting. Table I shows the hyper-parameter values used in the GAN after tuning. All experiments were conducted using an AMD Ryzen Threadripper 1920X 12-core 2.2GHz processor with 64GB of RAM and an NVIDIA GeForce RTX 2080 in a Pytorch environment.

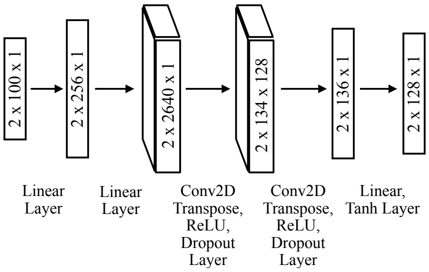

Fig. 4. GAN generator architecture

VI. Results and Discussion

As previously mentioned, the goal of adversarial attacks is to introduce imperceptible perturbations capable of re-

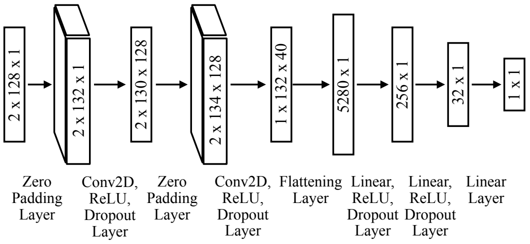

Fig. 5. GAN discriminator architecture

TABLE I
HYPER-PARAMETERS VALUES

Hyper-Parameter	Value
Optimizer	Adam
Generator Learning Rate	0.00049
Discriminator Learning Rate	0.00055
Batch Size	128
Latent Dimension	100
Dropout Rate	0.10

ducing the accuracy of a modulation classifier. Therefore, we evaluated our proposed attack technique by measuring the VT-CNN2's accuracy on clean and adversarial samples, and the perturbation-to-noise ratio (PNR). PNR measures the ratio between the perturbation and noise power levels so that $PNR_{[dB]} = 10\log(\frac{P_{perturbation}}{P_{noise}})$, where P is the signal power. The larger the PNR, the larger the perturbation is in comparison to the noise, becoming more distinguishable and more likely to be detected. Perturbations are considered imperceptible when they are in the same order as or below the noise level, i.e., PNR $<$ 0 dB.

Figure 6 shows the VT-CNN2's accuracy versus PNR for SNRs of 10, 0, and -10 dB. Without attacks, the classifier achieves different accuracy depending on the SNR because larger noises make it harder for the classifier to achieve correct results. Under our proposed adversarial attack, the classifier's accuracy is significantly reduced in all cases. At 0 dB PNR, our technique reduces the accuracy by 37% for 10 dB SNR, 56% for 0 dB SNR, and 7% for -10 dB SNR. Our technique reduces the accuracy more for 0 dB than for 10 dB SNR because, for signals with the same strength, larger SNRs mean lower noise levels so that it is more challenging to produce imperceptible perturbations that still significantly compromise the accuracy. However, although the noise at -10 dB SNR is the highest, allowing our technique to produce larger perturbations, the accuracy reduction is not as significant as at 0 dB SNR or 10 dB SNR. If $f(x + G(z))$ in equation (5) gives too many wrong results regardless of the adversarial perturbation $G(z)$, it is harder for our technique to find what perturbation would reduce the classifier's accuracy the most. Thus, the fact that our technique relies on the classifier's decisions to train the GAN diminishes its capacity to produce wrong classifications when the classifier's accuracy is low. Since the classifier's accuracy is around only 22% at -10 dB SNR, the adversarial perturbations that our proposed technique crafts are less effective. Nevertheless, our proposed adversarial attack technique still significantly reduces the classifier's accuracy.

We further examine the influence of perturbations on signal waveforms. We verify that the signal waveform after perturbation (adversarial sample) is consistent with the original waveform (clean sample), i.e., amplitude, frequency, and phase do not significantly change. Thus, while our technique's perturbations mislead the classifier, they are not easily recognized by human eyes. Figure 7 illustrates the time domain waveform of an 8PSK signal before and after the perturbation is introduced. Similar results were achieved for the other modulation schemes considered, such that clean and adversarial samples always have very similar waveforms without significant changes in their amplitude, frequency, and phase.

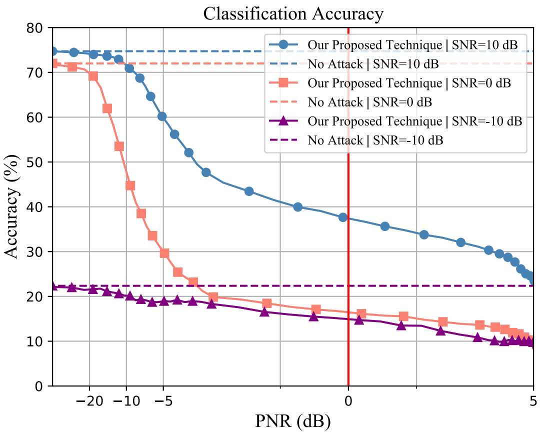

Fig. 6. Modulation classifier's accuracy versus PNR with and without our proposed adversarial attack technique

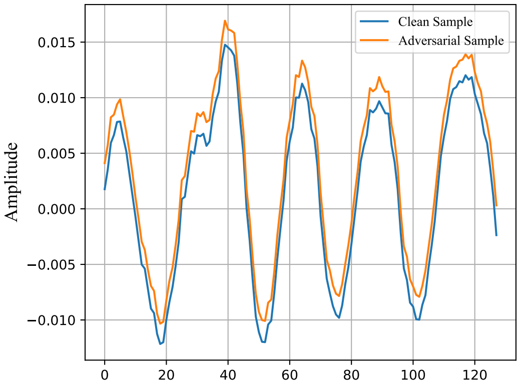

Fig. 7. Waveform comparison of a 8PSK signal with SNR=10 dB before (clean sample) and after (adversarial sample) our proposed adversarial attack

Moreover, we compare our results to those of a jamming attack, which adds Gaussian noise to signals, and two other adversarial attack techniques: those proposed in [17] and [11]. Figure 8 shows the VT-CNN2's accuracy on clean samples and adversarial samples produced by the jamming attack and the three adversarial attack techniques evaluated for SNR=10 dB. Perturbations introduced by adversarial attacks are specially crafted to reduce the classifier's accuracy the most while not being perceived. Thus, our technique and the techniques from [17] and [11] are significantly more harmful than attacks that introduce random noises, such as the jamming attack. Moreover, our proposed attack technique is the one that reduces the accuracy the most.

Finally, we evaluate how long it takes for each technique to craft adversarial samples. Table II shows the mean execution time for crafting adversarial samples. Our proposed technique achieves significantly shorter times than the other two tech-

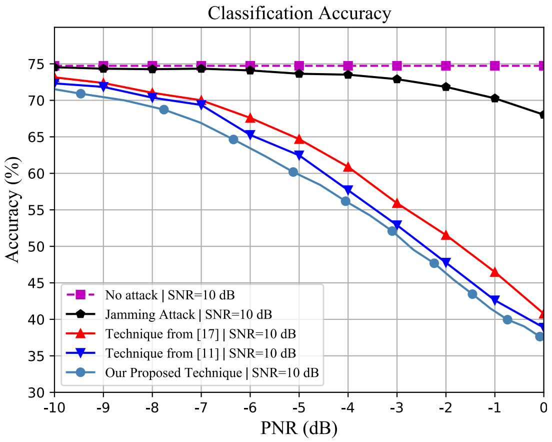

Fig. 8. Modulation classifier's accuracy versus PNR without and subject to different adversarial attack techniques

niques by crafting adversarial samples in less than 0.7 ms. Thus, it is more than 335 times faster than the second-fastest attack technique. Techniques that take too long to craft perturbations might be too late so that the signals they aim to perturb have already been correctly demodulated. Thus, such a time reduction is essential to compromise fast modulation classifiers and is a great advantage of our technique. Moreover, since our technique is input-agnostic, it can prepare perturbations in advance and just add them to incoming signals. Therefore, our proposed technique represents a severe risk to using deep learning-based modulation classifiers.

TABLE II
MEAN EXECUTION TIME FOR CRAFTING ADVERSARIAL SAMPLES

Adversarial Attack Technique	**Mean Execution Time per Sample**
Technique from [17]	20189 ms
Technique from [11]	234 ms
Our Proposed Technique	0.6980 ms

VII. CONCLUSION

In this letter, we verified that deep learning is exposed to security risks that must be considered despite its advantages. Our results showed that it is possible to quickly craft small imperceptible perturbations that completely compromise modulation classifiers' accuracy and hence wireless receivers' performance. Therefore, it is urgently necessary to enhance deep learning-based modulation classifiers' robustness against adversarial attacks. As future work, we will evaluate the use of other GAN formulations, such as WGAN-GP, modify our attack model to consider adversarial attacks transmitted over the air, and investigate adversarial attack defense strategies.

REFERENCES

[1] F. Liang, C. Shen, and F. Wu, "An Iterative BP-CNN Architecture for Channel Decoding," *IEEE J. of Sel. Topics in Signal Process.*, vol. 12, no. 1, pp. 144–159, 2018.

[2] L. Sanguinetti, A. Zappone, and M. Debbah, "Deep Learning Power Allocation in Massive MIMO," in *2018 52nd Asilomar Conf. on Signals, Syst., and Comput.*, 2018, pp. 1257–1261.

[3] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu, and N. D. Sidiropoulos, "Learning to optimize: Training deep neural networks for wireless resource management," in *2017 IEEE 18th Int. Workshop on Signal Process. Advances in Wireless Commun. (SPAWC)*, 2017, pp. 1–6.

[4] T. J. O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in *Int. Conf. on Eng. Appl. of Neural Networks*. Springer, 2016, pp. 213–226.

[5] T. J. O'Shea, T. Roy, and T. C. Clancy, "Over-the-Air Deep Learning Based Radio Signal Classification," *IEEE J. of Sel. Topics in Signal Process.*, vol. 12, no. 1, pp. 168–179, 2018.

[6] B. Flowers, R. M. Buehrer, and W. C. Headley, "Evaluating Adversarial Evasion Attacks in the Context of Wireless Communications," *IEEE Trans. on Inf. Forensics and Secur.*, vol. 15, pp. 1102–1113, 2020.

[7] Y. Lin, H. Zhao, X. Ma, Y. Tu, and M. Wang, "Adversarial Attacks in Modulation Recognition With Convolutional Neural Networks," *IEEE Trans. on Rel.*, vol. 70, no. 1, pp. 389–401, 2021.

[8] R. Sahay, C. G. Brinton, and D. J. Love, "A Deep Ensemble-based Wireless Receiver Architecture for Mitigating Adversarial Interference in Automatic Modulation Classification," *arXiv preprint arXiv:2104.03494*, 2021.

[9] Y. Lin, H. Zhao, Y. Tu, S. Mao, and Z. Dou, "Threats of Adversarial Attacks in DNN-Based Modulation Recognition," in *IEEE INFOCOM 2020 - IEEE Conf. on Comput. Commun.*, 2020, pp. 2469–2478.

[10] B. Manoj, M. Sadeghi, and E. G. Larsson, "Adversarial Attacks on Deep Learning Based Power Allocation in a Massive MIMO Network," *arXiv preprint arXiv:2101.12090*, 2021.

[11] M. Sadeghi and E. G. Larsson, "Adversarial Attacks on Deep-Learning Based Radio Signal Classification," *IEEE Wireless Commun. Lett.*, vol. 8, no. 1, pp. 213–216, 2019.

[12] O. Ibitoye, R. Abou-Khamis, A. Matrawy, and M. O. Shafiq, "The Threat of Adversarial Attacks on Machine Learning in Network Security–A Survey," *arXiv preprint arXiv:1911.02621*, 2019.

[13] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial Examples: Attacks and Defenses for Deep Learning," *IEEE Trans. on Neural Networks and Learn. Syst.*, vol. 30, no. 9, pp. 2805–2824, 2019.

[14] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *Int. Conf. on Mach. Learn.* PMLR, 2018, pp. 2137–2146.

[15] H. Zhao, Y. Lin, S. Gao, and S. Yu, "Evaluating and Improving Adversarial Attacks on DNN-Based Modulation Recognition," in *GLOBECOM 2020 - 2020 IEEE Global Commun. Conf.*, 2020, pp. 1–5.

[16] W. Brendel, J. Rauber, and M. Bethge, "Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models," *arXiv preprint arXiv:1712.04248*, 2017.

[17] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal Adversarial Perturbations," in *Proc. of the IEEE Conf. on Comput. Vision and Pattern Recognit. (CVPR)*, July 2017.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Inf. Process. Syst.*, vol. 27, 2014.

[19] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. of the IEEE Conf. on Comput. Vision and Pattern Recognit.*, 2018, pp. 7482–7491.

[20] K. Li, X. Yu, H. Zhang, L. Wu, X. Du, P. Ratazzi, and M. Guizani, "Security Mechanisms to Defend against New Attacks on Software-Defined Radio," in *2018 Int. Conf. on Comput., Netw. and Commun. (ICNC)*, 2018, pp. 537–541.

[21] P. Freitas de Araujo-Filho, G. Kaddoum, D. R. Campelo, A. Gondim Santos, D. Macêdo, and C. Zanchettin, "Intrusion Detection for Cyber–Physical Systems Using Generative Adversarial Networks in Fog Environment," *IEEE Internet of Things J.*, vol. 8, no. 8, pp. 6247–6256, 2021.

[22] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein Generative Adversarial Networks," in *Int. Conf. on Mach. Learn.* PMLR, 2017, pp. 214–223.

[23] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative Adversarial Networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, 2018.

[24] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved Training of Wasserstein GANs," *arXiv preprint arXiv:1704.00028*, 2017.

[25] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, "Are GANs Created Equal? A Large-Scale Study," *arXiv preprint arXiv:1711.10337*, 2017.

[26] T. J. O'Shea and N. West, "Radio Machine Learning Dataset Generation with GNU Radio," *Proc. of the 6th GNU Radio Conf.*, 2016.

[27] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proc. of the 25rd ACM SIGKDD Int. Conf. on Knowl. Discovery and Data Mining*, 2019.