SafeAMC: Adversarial training for robust modulation classification models

Javier Maroto
Signal Processing Laboratory (LTS4)
EPFL, Switzerland

Gérôme Bovet
armasuisse Science&Technology
Cyber-Defence Campus, Switzerland

Pascal Frossard
Signal Processing Laboratory (LTS4)
EPFL, Switzerland

***Abstract*—In communication systems, there are many tasks, like modulation classification, for which Deep Neural Networks (DNNs) have obtained promising performance. However, these models have been shown to be susceptible to adversarial perturbations, namely imperceptible additive noise crafted to induce misclassification. This raises questions about the security but also about the general trust in model predictions. We propose to use adversarial training, which consists of fine-tuning the model with adversarial perturbations, to increase the robustness of automatic modulation classification (AMC) models. We show that current state-of-the-art models can effectively benefit from adversarial training, which mitigates the robustness issues for some families of modulations. We use adversarial perturbations to visualize the learned features, and we found that the signal symbols are shifted towards the nearest classes in constellation space, like maximum likelihood methods when adversarial training is enabled. This confirms that robust models are not only more secure, but also more interpretable, building their decisions on signal statistics that are actually relevant to modulation classification.**

***Index Terms*—Modulation classification, robustness, adversarial training, deep learning, security**

I. Introduction

Communication systems are important for both civil and military applications. Deep learning [1] has proved its usefulness across multiple fields of research in the last decade and it presents numerous advantages that are attractive for wireless communication systems. Compared with the previous state-of-the-art approaches, which are mainly based on feature extraction from the signals [2], Deep Neural Networks (DNNs) are capable of end-to-end learning and their performance scales with high quantities of data. DNNs have been successful in multiple tasks like wireless resource allocation [3] anomaly detection [4], or automatic modulation classification (AMC) [5], which is the main focus of this work. AMC estimates modulation schemes from the received data and has multiple applications ranging from detecting daily radio stations and managing spectrum resources, to eavesdropping and interfering with radio communications.

Recent studies have highlighted security issues in DNNs models [6]–[11]. Specifically, those models have been shown to be vulnerable to adversarial examples, which are carefully crafted but almost imperceptible perturbations, namely adversarial perturbations, that are added to a real data sample. For AMC, a perturbation can be crafted by solving the following optimization:

$$\delta_i^* = \arg\max_{\delta_i} \mathcal{L}_{y_i,\theta}(x_i + \delta_i) \quad \text{s.t.} \quad \|\delta_i\|_\infty \leq \varepsilon \tag{1}$$

where δ_i^* is the adversarial perturbation that is added to the clean signal x_i, $\mathcal{L}_{y_i,\theta}$ is the model loss function that depends on y_i, the signal modulation, and θ, the model parameters, and ε is a fixed value that constrains the norm of the perturbation to be small and imperceptible.

This security issue combined with the black-box nature of DNNs raises a critical question: can we actually trust the predictions of neural networks? The fact that a negligible change in the input alters the prediction, implies that DNNs base their decisions on features that do not seem to be all aligned with the target task. Understanding the reasons for such vulnerabilities and making systems more robust form an active line of research [12], [13].

In this work, we propose a new framework, SafeAMC, that addresses these trust and security issues. It provides a way to reduce the AMC model susceptibility against adversarial perturbations and properly measures both the model robustness and how secure it would be in practice, in a real communication system. Our main contributions are the followings:

- We are the first to differentiate between practical security and robustness in AMC, using specific properties of communication systems;
- We propose adversarial training to defend against adversarial examples in AMC, and show that it increases the robustness and security of state-of-the-art AMC models using popular modulation classification datasets;
- Using feature analysis on the constellation diagram space, we show that robust models learn better features. They are correlated with the optimal ones computed by Bayes-optimal maximum likelihood methods;
- Instead of using a fixed constraint for adversarial perturbations, we promote constraining the perturbation with respect to the signal energy, which is a better measure of adversarial perceptibility in AMC.

In Section II we discuss some related works to AMC and adversarial perturbations. Then, in Section III, we introduce our adversarial learning framework with two dedicated use cases, namely robustness and security against adversarial attacks. We provide results of the experiments realized along

This work has been sponsored by armasuisse Science and Technology with the project ROBIN (project code Aramis 047-22).

with our newly defined framework in Section IV. The feature importance of standardly and adversarially trained models is presented in Section V. We finally conclude and provide future work directions in Section VI.

II. Related work

The task of recognizing modulations can be modeled mathematically using maximum likelihood. It computes the likelihood function of the received signal with all possible modulations and estimates the most likely modulation. While in theory this is Bayes-optimal, it is computationally expensive and requires prior knowledge of the channel characteristics, so sub-optimal approximations are used in practice [2], [14].

Deep learning [1] has been proposed as an alternative solution since it has linear computational complexity and can be trained end-to-end. It learns the channel conditions directly from the data instead of using priors, making it more flexible for AMC. On the one hand, inspired by successes in speech recognition, some works [15], [16] propose architectures based on long short-term memory networks (LSTMs) [17] for modulation classification. On the other hand, other works [8], [18], [19] employ convolutional neural networks (CNNs) [20], which have the advantage of inducing translation invariance on the input. The current state-of-the-art in AMC [5] uses a model based on the ResNet Deep Network architecture [21].

Despite their good performance in numerous application domains, DNNs are black-box models, giving less explainable predictions than simpler models, and they are vulnerable to adversarial perturbations [6], [7]. These perturbations have the property of influencing discriminative features of the model [12], and put into question the relevance of the features learned by the classifier [13]. Multiple "defenses" against adversarial perturbations have been proposed in different application domains. Some authors advocate for randomized smoothing [22], [23], while others propose some form of loss function regularization [24], [25]. Currently, the best defense generally relies on adversarial training [26], where the model is fined-tuned on adversarial perturbations to increase its robustness.

In the specific case of modulation classification, adversarial perturbations have also been shown to require much less power than additive white gaussian noise (AWGN) to fool the network [8], [10], [11]. Adversarial perturbations in this case are constrained relative to the signal power, using the signal-to-perturbation ratio (SPR) metric [8]. Some defenses have been proposed for AMC models. On the one hand, some methods preprocess the input, like gaussian smoothing [27], to reduce the effectiveness of the adversarial perturbations. On the other hand, some methods modify the deep network model itself by adding a pretrained autoencoder [28] or using an Assorted Deep Ensemble (ADE) [29], in which different model architectures and signal domains are combined. These defense methods can protect against adversarial examples, but they may be vulnerable if the attacker has knowledge of the full model. A concurrent work [30] solves this issue by using adversarial training, which our work extends by using the SPR metric to craft the attacks and evaluating on the state-of-the-art networks and benchmarking datasets.

In this work, we present SafeAMC, a new framework based on adversarial training to make DNNs less susceptible to adversarial perturbations. We show that it increases the robustness of these models, and makes them safer.

III. Framework

In this Section we present our framework that relies on adversarial training to improve DNNs robustness and evaluates how robust and secure it is against adversarial attacks.

A. Adversarial training

The objective of adversarial training is to find the model parameters that minimize the susceptibility of the model to adversarial perturbations. That is, adversarial training tries to solve the following min-max optimization problem over the N training samples:

$$\min_{\theta} \frac{1}{N} \sum_{i} \max_{\|\delta_i\|_\infty \leq \varepsilon} \mathcal{L}_{y_i,\theta}(x_i + \delta_i) \tag{2}$$

This formulation is a generalization of standard training, for which adversarial perturbations are not considered ($\delta_i = 0$).

In practice, the procedure to train a model adversarially is straightforward. First, we take a training sample and compute the adversarial perturbation using an optimization algorithm like FGSM or Projected Gradient Descent (PGD) [26]. Then, we compute the model classification loss on that perturbed sample. Next, we backpropagate this loss with respect to the model parameters. Finally, we update them to minimize the loss using gradient descent. This whole process is repeated for all training samples multiple times.

B. Robustness and security evaluation in AMC

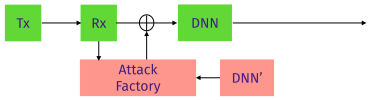
While robustness is derived from the model susceptibility to adversarial perturbations, in practice, when measuring how secure it is, we find unrealistic to assume the attacker has direct access to the model. This motivates us to use two different frameworks to measure robustness and security [11], which are shown in Figure 1. The *robustness framework* is similar to the traditional approach, adding the attack just before the model. We use it to measure the theoretical robustness of our model. The *security framework* measures how secure the model is against malicious adversarial attacks. We simulate a man-in-the-middle attack, where Gaussian noise is added to the adversarial example sent by the attacker.

Our framework, SafeAMC, trains the model using the robustness framework to learn features that are invariant to small perturbations according to (2). Then, it evaluates the model on both frameworks to measure the robustness and the security of the model.

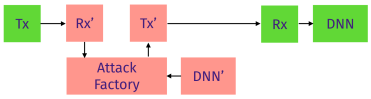
IV. Performance analysis

A. Adversarial attacks fool towards similar modulations

We examine how the model predictions change between normal and adversarial settings with and without using SafeAMC.



(a) Robustness framework



(b) Security framework

Fig. 1: Frameworks used to measure robustness and security. The green and red blocks illustrate the communication system and the attacker, respectively. Tx and Rx are the transmitter and receiver, while DNN is the AMC model. We analyze the scenario where the attacker surrogate model (DNN') is identical to our AMC model (white-box attack).

For the sake of clarity, we will only measure the robustness in this experiment. For the dataset, we use the RML2016.10a dataset [31], which has 220000 IQ signals of 128 time samples each, ranging from -20dB to 18dB signal-to-noise ratio (SNR), with 11 possible modulations to predict. We use a 70-30% train-test split, using 5% of the training as validation for hyperparameter tuning. For the model, we use the VT_CNN2_BF architecture [10]. All adversarial examples are crafted with 20 dB SPR, since it affects significantly the performance of the model without compromising the inperceptibility of the perturbations [11]. We use l_∞ PGD with 7 iterations (PGD-7) of step size 0.36 (relative to the l_∞ ball radius) and PGD-20 of step size 0.125 for the adversarial examples crafted during training and testing, respectively.

On the one hand, when SafeAMC is not used (Figure 2), the model is consistently fooled towards modulation classes with similar constellation diagrams (e.g. BPSK towards PAM4) when adversarial attacks are used. On the other hand, using SafeAMC (Figure 3) greatly improves the model robustness against adversarial perturbations at the cost of a small performance penalty on the original data. We observe that, for QAM16/QAM64, 20 dB SPR is too strong and can change the underlying modulation of the signal. Thus, for those modulations, the best defense can at most obtain random chance adversarial performance.

B. Robustness and security evaluation with SafeAMC

For our next experiment, we do a full evaluation of the robustness and security when using SafeAMC with two state-of-the-art DNN models, the VT_CNN2_BF and ResNet models. For the data, we use the more difficult RML2018.01a dataset [5], which contains more than 2.5 million IQ signals of 1024 time samples each, ranging from -20dB to 30dB SNR, with 24 possible modulations. We only use 1 million signals as training set as proposed in [5], and split the rest in 5% for validation and 95% for testing. The adversarial settings are the same as the previous experiment, but we also consider training and testing with 15 dB and 25 dB SPR adversarial perturbations.

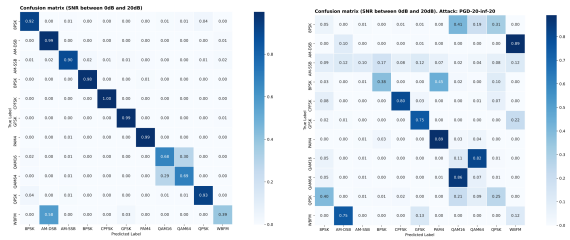


Fig. 2: Confusion matrices of the VT_CNN2_BF model before (left) and after (right) adding PGD-20 l_∞ adversarial perturbations on the RML2016.10a dataset. Results for IQ signals with SNR $\geq$ 0dB.

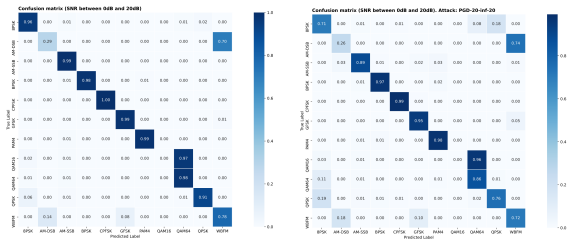


Fig. 3: Confusion matrices of the VT_CNN2_BF model trained with SafeAMC before (left) and after (right) adding PGD-20 l_∞ adversarial perturbations on the RML2016.10a dataset. Results for IQ signals with SNR $\geq$ 0dB.

For the security framework, we add 20 dB SNR AWGN to the signal transmitted by the attacker to model the channel noise. The comparison of the accuracy of the AMC models on all the training schemes, different testing scenarios, and different SafeAMC frameworks is shown in Table I.

The results of this more extensive analysis show some clear patterns. First, the added AWGN of the security framework helps reduce the model fooling rate, showing that attacking the model is less effective in practice. This difference is more significant for standardly trained models since adversarially trained models rely on signal features that are less susceptible to adversarial perturbations and, by extension, to small AWGN noise. Second, the higher the perturbation strength used for testing, the higher the fooling rate. This is intuitive since the attacker is less constrained. Third, while adversarial training makes the model less susceptible to attacks, using stronger perturbations during training does not guarantee more robustness. Our ResNet model results exemplify this last point and show that, when using SafeAMC, the perturbation strength should be treated as a hyperparameter and tuned accordingly.

V. Feature Analysis

We now analyze the features learned by a VT_CNN2_BF neural network and show that, when trained with SafeAMC, they are more interpretable, since they correlate with the signal statistics used by the maximum likelihood (ML) classifier.

Since adversarial perturbations' objective is to change the model prediction, they essentially remove features from the correct class and add features from other classes to fool the

Model	Train attack	Accuracy	Robustness Framework Accuracy			Security Framework Accuracy		
			25dB SPR	20dB SPR	15dB SPR	25dB SPR	20dB SPR	15dB SPR
VT_CNN2_BF	None	**45.2%**	9.9%	5.1%	2.0%	13.1%	6.9%	3.3%
	20dB SPR	37.2%	29.2%	25.1%	13.7%	30.3%	26.8%	15.9%
	15dB SPR	32.4%	**30.4%**	**27.8%**	**21.4%**	**30.8%**	**28.8%**	**22.7%**
ResNet	None	**60.8%**	24.6%	17.7%	10.6%	**34.1%**	25.1%	15.4%
	20dB SPR	36.1%	**34.2%**	**32.5%**	**28.6%**	34.0%	**32.8%**	**29.3%**
	15dB SPR	31.2%	30.5%	30.0%	28.1%	30.1%	29.6%	28.1%

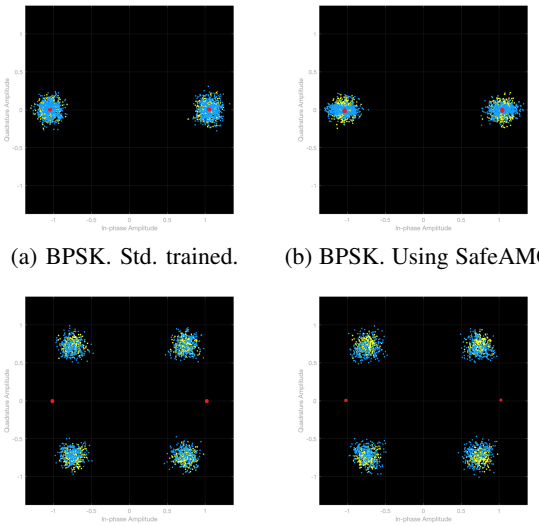
TABLE I: Robustness and security tested on different levels of attack strength for the VT_CNN2_BF and ResNet models on the RML2018.01a dataset. The train and test attacks were crafted using l_∞ PGD-7 and l_∞ PGD-20, respectively.

network. Thus, adversarial perturbations can help us understand what features the model has learned, an approach that is supported by several works in other application domains like computer vision [12], [32], [33]. The advantage AMC has over other application domains is that we can compare the model features with the ideal ones, which can be obtained by adversarially perturbing the Bayes-optimal ML method.

However, IQ signals are very high-dimensional, which makes visualizing the comparison untractable. Luckily, they can be divided into smaller equally long subunits, the signal symbols. The receiver takes each of these symbols and converts them into a pair of values (in-phase and quadrature components), that can be displayed as points in a two-dimensional plane, the constellation diagram [34]. For a given modulation of 2^N number of states, only 2^N different constellation pairs can be sent (the modulation states), each encoding N bits of information. In practice, due to communication noise, the position of the received constellation pairs is slightly shifted. Assuming this noise is AWGN, to recognize the original modulation, ML compares every received constellation point with the possible states defined by each modulation and determines which modulation is the most probable. This probability will be higher if each received point is very close to the modulation states. Thus, its adversarial perturbation will shift each constellation point closer to other modulation states.

We created a custom simulated dataset (CRML2018) to be able to easily compare the adversarial perturbations generated with ML to the ones generated by a VT_CNN2_BF model trained with and without our SafeAMC framework. The dataset has 10000 signals of 1024 time samples for each of the 16 digital modulations used in the RML2018.01a dataset. For simplicity, we consider a 20 dB SNR AWGN channel. We use a 70-30% train-test split, using 5% of the training as validation. We generated FGSM adversarial examples of 20 dB SPR with the objective of maximizing the BPSK modulation probability. Figure 4 shows the constellation diagrams for two different signals and models, where the possible states of the BPSK modulation are in red, the received symbols in yellow and the VT_CNN2_BF adversarially perturbed symbols in blue. For the rows, we show a BPSK and a QPSK signal from the test set, and for the columns, we used a model trained with and without using SafeAMC.

The constellation diagrams show that SafeAMC helps the model learn features that are more aligned with the Bayes-optimal model, because the adversarial perturbations also shift



(a) BPSK. Std. trained. (b) BPSK. Using SafeAMC

(c) QPSK. Std. trained. (d) QPSK. Using SafeAMC

Fig. 4: Superposed constellation diagrams of both original (yellow) and adversarially perturbed (blue) BPSK and QPSK signals, for differently trained VT_CNN2_BF models. Despite the adversarial perturbation maximizing the BPSK class prediction, only when using the model trained with SafeAMC we observe it shifts the symbols closer to the BPSK possible states (red points). Thus, using SafeAMC helps the model learn better features, that are also used by maximum likelihood.

the symbols towards the BPSK possible states. For the generated BPSK signal, the perturbation of the standardly trained model shifts the symbols seemingly in random directions, not reducing the energy of the AWGN corruption but increasing the model confidence on the BPSK prediction. In contrast, the perturbation of the adversarially trained model is more confident after reducing the variance of the quadrature component. However, one would expect that in-phase component variance should be reduced too. We believe this is not the case because of the class unbalance in the dataset. In our dataset only the BPSK and PAM4 signals have almost no quadrature component, making it much more important than the in-phase component for prediction. Finally, for the generated QPSK signal, when comparing the two perturbations, we can observe that the shift towards BPSK is bigger in the adversarially trained model. However, the adversarial model is not fooled into predicting it as BPSK, while the standard model is fooled

by the perturbation with a smaller shift.

To sum up, SafeAMC does not only make the model less susceptible to adversarial attacks, but it also learns features that are better aligned with the task. This can be especially useful when we want to deploy our model on unknown communication channel conditions, where there is a distribution shift between the training and test data. We expect robust features to transfer better on varying channel conditions, thus reducing the decrease in performance.

VI. Conclusion and Future work

In this work, we propose adversarial training to make AMC models more robust against adversarial perturbations. We show that it greatly increases the robustness of the models and makes them less susceptible to adversarial attacks in real scenarios. Furthermore, we show that robust models learn better features for modulation classification. That is, their features correlate more with the features used by the Bayes-optimal ML model. Thus, showing that robustness may be the key to ensuring that DNNs learn features that are well-aligned with the task, and possibly transfer better between different channel conditions.

In the future, our analysis could be expanded beyond small l_∞ norm adversarial perturbations to other types of corruptions that are common in AMC and could have non-uniform constraints (e.g., Rayleigh or Rician channel corruptions). Moreover, to avoid changing the true class of high-order modulations when using adversarial perturbations, a per-class SPR based on the possible modulations could be defined to improve our robustness and security metrics.

References

[1] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio, *Deep learning*, vol. 1, MIT press Cambridge, 2016.

[2] Octavia A Dobre, Ali Abdi, Yeheskel Bar-Ness, and Wei Su, "Survey of automatic modulation classification techniques: classical approaches and new trends," *IET communications*, vol. 1, no. 2, pp. 137–156, 2007.

[3] Haoran Sun, Xiangyi Chen, Qingjiang Shi, Mingyi Hong, Xiao Fu, and Nikos D Sidiropoulos, "Learning to optimize: Training deep neural networks for wireless resource management," in *SPAWC*. IEEE, 2017, pp. 1–6.

[4] Raghavendra Chalapathy and Sanjay Chawla, "Deep learning for anomaly detection: A survey," *arXiv preprint arXiv:1901.03407*, 2019.

[5] Timothy James O'Shea, Tamoghna Roy, and T Charles Clancy, "Over-the-air deep learning based radio signal classification," *JSTSP*, vol. 12, no. 1, pp. 168–179, 2018.

[6] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[7] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard, "Universal adversarial perturbations," in *CVPR*, 2017, pp. 1765–1773.

[8] M. Sadeghi and E. G. Larsson, "Adversarial attacks on deep-learning based radio signal classification," *IEEE Wireless Commun. Letters*, vol. 8, no. 1, pp. 213–216, Feb 2019.

[9] Yun Lin, Haojun Zhao, Ya Tu, Shiwen Mao, and Zheng Dou, "Threats of adversarial attacks in dnn-based modulation recognition," in *INFOCOM*. IEEE, 2020, pp. 2469–2478.

[10] Bryse Flowers, R Michael Buehrer, and William C Headley, "Evaluating adversarial evasion attacks in the context of wireless communications," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1102–1113, 2019.

[11] Javier Maroto, Gérôme Bovet, and Pascal Frossard, "On the benefits of robust models in modulation recognition," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications III*. SPIE, 2021, vol. 11746, p. 1174611.

[12] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry, "Adversarial robustness as a prior for learned representations," *arXiv preprint arXiv:1906.00945*, 2019.

[13] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry, "Adversarial examples are not bugs, they are features," in *NeurIPS*, 2019, pp. 125–136.

[14] Fahed Hameed, Octavia A Dobre, and Dimitrie C Popescu, "On the likelihood-based approach to modulation classification," *IEEE Transactions on Wireless Communications*, vol. 8, no. 12, pp. 5884–5892, 2009.

[15] Sreeraj Rajendran, Wannes Meert, Domenico Giustiniano, Vincent Lenders, and Sofie Pollin, "Deep learning models for wireless signal classification with distributed low-cost spectrum sensors," *TCCN*, vol. 4, no. 3, pp. 433–445, Sep 2018.

[16] Youwei Guo, Hongyu Jiang, Jing Wu, and Jie Zhou, "Open set modulation recognition based on dual-channel lstm model," *arXiv preprint arXiv:2002.12037*, 2020.

[17] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] Timothy J O'Shea, Johnathan Corgan, and T Charles Clancy, "Convolutional radio modulation recognition networks," in *EANN*. Springer, 2016, pp. 213–226.

[19] N. E. West and T. O'Shea, "Deep architectures for modulation recognition," in *DySPAN*, Mar 2017, p. 1–6.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. of the ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[21] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[22] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter, "Certified adversarial robustness via randomized smoothing," in *ICML*. PMLR, 2019, pp. 1310–1320.

[23] Hadi Salman, Greg Yang, Jerry Li, Pengchuan Zhang, Huan Zhang, Ilya Razenshteyn, and Sebastien Bubeck, "Provably robust deep learning via adversarially trained smoothed classifiers," *arXiv preprint arXiv:1906.04584*, 2019.

[24] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard, "Robustness via curvature regularization, and vice versa," in *CVPR*, 2019, pp. 9078–9086.

[25] Gauri Jagatap, Ameya Joshi, Animesh Basak Chowdhury, Siddharth Garg, and Chinmay Hegde, "Adversarially robust learning via entropic regularization," *arXiv preprint arXiv:2008.12338*, 2020.

[26] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[27] Brian Kim, Yalin E Sagduyu, Kemal Davaslioglu, Tugba Erpek, and Sennur Ulukus, "Channel-aware adversarial attacks against deep learning-based wireless signal classifiers," *arXiv preprint arXiv:2005.05321*, 2020.

[28] Silvija Kokalj-Filipovic, Rob Miller, Nicholas Chang, and Chi Leung Lau, "Mitigation of adversarial examples in rf deep classifiers utilizing autoencoder pre-training," in *ICMCIS*. IEEE, 2019, pp. 1–6.

[29] Rajeev Sahay, Christopher G Brinton, and David J Love, "A deep ensemble-based wireless receiver architecture for mitigating adversarial interference in automatic modulation classification," *arXiv preprint arXiv:2104.03494*, 2021.

[30] Rajeev Sahay, David J Love, and Christopher G Brinton, "Robust automatic modulation classification in the presence of adversarial attacks," in *2021 55th Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 2021, pp. 1–6.

[31] Timothy J. O'Shea and Nathan West, "Radio machine learning dataset generation with gnu radio," *Proceedings of the GNU Radio Conference*, vol. 1, no. 11, Sep 2016.

[32] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry, "Adversarial examples are not bugs, they are features," *arXiv preprint arXiv:1905.02175*, 2019.

[33] Guillermo Ortiz-Jimenez, Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard, "Hold me tight! influence of discriminative features on deep network boundaries," *arXiv preprint arXiv:2002.06349*, 2020.

[34] S Andrew, "Tanenbaum computer networks," *Computer Networks, Englewood Cliffs*, pp. 141–148, 1996.