The Best Defense Is a Good Offense: Adversarial Attacks to Avoid Modulation Detection

Muhammad Zaid Hameed, András György, and Deniz Gündüz, *Senior Member, IEEE*

***Abstract*— We consider a communication scenario, in which an intruder tries to determine the modulation scheme of the intercepted signal. Our aim is to minimize the accuracy of the intruder, while guaranteeing that the intended receiver can still recover the underlying message with the highest reliability. This is achieved by perturbing channel input symbols at the encoder, similarly to adversarial attacks against classifiers in machine learning. In image classification, the perturbation is limited to be imperceptible to a human observer, while in our case the perturbation is constrained so that the message can still be reliably decoded by the legitimate receiver, which is oblivious to the perturbation. Simulation results demonstrate the viability of our approach to make wireless communication secure against state-of-the-art intruders (using deep learning or decision trees) with minimal sacrifice in the communication performance. On the other hand, we also demonstrate that using diverse training data and curriculum learning can significantly boost the accuracy of the intruder.**

***Index Terms*— Secure communication, deep learning, adversarial attacks, modulation classification.**

I. Introduction

SECURING wireless communication links is as essential as increasing their efficiency and reliability, for military, commercial, as well as consumer communication systems. The standard approach to securing communications is to encrypt the transmitted data. However, encryption may not always provide full security (e.g., in case of side-channel attacks), or strong encryption may not be available due to complexity limitations (e.g., for IoT devices). To further improve security, encryption can be complemented with other techniques, preventing the adversary from even recovering the encrypted bits.

As outlined in [2], an adversary implements its attacks on a wireless communication link in four steps: 1) tunes into the frequency of the transmitted signal; 2) detects whether there is signal or not; 3) intercepts the signal by extracting its features; and 4) demodulates the signal by exploiting the extracted features, and obtains a binary stream of data. Preventing any of these steps can strengthen the security of the communication link. While encryption focuses on protecting the demodulated bit stream, physical layer security [3], [4] targets the fourth step by minimizing the mutual information available to the intruder. Recently, there has also been significant interest in preventing the second step through covert communications [5]. In this work, we instead focus on the third step, and aim at preventing the adversary from detecting the modulation scheme used for communications.

Modulation detection is the step between signal detection and demodulation in communication systems, and thus plays an important role in data transmission, as well as in detection and jamming of unwanted signals in military communications and other sensitive applications [6]. Recently, deep learning techniques have led to significant progress in modulation-detection accuracy: methods based on convolutional and other deep neural networks can detect the modulation scheme directly from raw time-domain samples [7]–[11], surpassing the accuracy of conventional modulation detectors based on likelihood function or feature-based representations (see [6] for a survey of these approaches).

Our aim in this article is to prevent an intruder that employs a state-of-the-art modulation detector from successfully identifying the modulation scheme being used. If the intruder is unable to identify the modulation scheme, it is unlikely to be able to decode the underlying information or employ modulation-dependent jamming attacks to prevent communication. To achieve this goal, we introduce modifications to the transmitted signal. The main challenge here is to guarantee that the intended receiver of the (modified) transmitted signal can still receive the underlying message reliably, while preventing the intruder from detecting the modulation scheme being used. Otherwise, reducing the accuracy of the modulation-detecting intruder would be trivial at the price of increasing – possibly by a lot – the bit-error rate (BER) of the intended receiver. We assume that the intended receiver is oblivious to the modifications employed by the transmitter, and therefore, the goal of the transmitter is to introduce as small modifications to the transmitted signal as possible that are sufficient to fool the intruder but not larger than the error-correction capabilities of the intended receiver.

Introducing small variations into the modulation scheme that can fool an intruder is similar to adversarial attacks

Manuscript received March 24, 2020; revised August 1, 2020; accepted August 26, 2020. Date of publication September 21, 2020; date of current version October 21, 2020. This work was supported in part by the Imperial College London President's Ph.D. Scholarship and in part by the European Research Council (ERC) through the Starting Grant BEACON under Grant 677854. This article was presented in part at the 7th IEEE Global Conference on Signal and Information Processing (GlobalSIP 2019). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Lifeng Lai. *(Corresponding author: András György.)*

Muhammad Zaid Hameed was with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. He is now with the Department of Computing, Imperial College London, London SW7 2AZ, U.K. (e-mail: muhammad.hameed13@imperial.ac.uk).

András György is with DeepMind, London N1C 4AG, U.K. (e-mail: agyorgy@google.com).

Deniz Gündüz is with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: d.gunduz@imperial.ac.uk).

Digital Object Identifier 10.1109/TIFS.2020.3025441

on classifiers, in particular, deep neural networks (DNNs) [12], [13]. In the literature, adversarial attacks are mostly considered in the area of image classification, where they pose security risks by exposing the vulnerabilities of classifiers against very small changes in the input that are imperceptible to humans but lead to incorrect decisions. In contrast, we exploit the same approach here to defend a communication link against an intruder that employs DNNs or other standard classification methods for interception.

In [14], an adversarial attack for a deep-learning-based modulation classifier has been proposed where the adversary assumes the availability of noisy symbols received at the modulation classifier for generating the adversarial attack, which makes it impractical and limited in scope. A similar method has been proposed recently in [15], where modifications are employed by the transmitter to evade a DNN-based jammer, and the receiver uses another DNN (an autoencoder) to preprocess the received signal and filter out the modifications. However, no analysis has been provided on the impact of this method on the BER of the intended receiver. In contrast to [15], we do not limit our approach to a DNN-based jammer and consider a receiver that is completely oblivious to the modifications in the transmitter. Moreover, we also consider the impact of the defensive perturbations on the BER at the legitimate receiver. The results of [15] has been extended in [16] to the detection of wireless communication protocols, and targeted adversarial attacks are also considered to generate the perturbations.[1]

A number of concurrent works have also appeared in the literature (parallel or following the original publication of our preprint on arXiv [18] and the conference version of our article [1]). Most similar to ours is [19], which proposes modifications in the transmitted signal using an adversarial residual network at the transmitter to evade the modulation detector at an intruder, while the legitimate receiver is able to decode the signal with small bit error rate. Compared to this article, we use different adversarial attack techniques, propose different ways of improving the modulation-detection accuracy of the intruder, and analyze the trade-off between the code rate and the BER for defensive perturbations and an improved intruder. Adversarial perturbations have also been applied to attack a legitimate receiver in [20], [21]. In these works, the signal at the receiver is perturbed by an over-the-air attack, i.e., by transmitting an adversarial signal, to make the modulation classifier at the legitimate receiver fail (in comparison, in our case the transmitter changes the signal to fool the intruder). In [20], modifications in the transmitted signals are also proposed to evade the modulation detector at the intruder and are evaluated in terms of the BER at the receiver, but the modifications in the transmitted signals are not optimized with respect to the BER and induce larger errors at the legitimate receiver. The over-the-air attack scenario has been considered in [21], and attack methods of various strength have been devised under more realistic assumptions about the capabilities of the attacker, in particular about its information on the signal received by the modulation classifier (fully known vs. its distribution being estimated based on samples available to the attacker) and on the channel noise from the attacker to the receiver (knowing the exact realization or just the noise distribution). While these attack methods share the underlying idea with our defensive perturbations, they face a much easier problem, as the attacks are not constrained by ensuring a low BER at a distinct receiver.

[1] Targeted adversarial attacks [17] aim to modify the data so that the attacked classifier predicts an incorrect class selected by the attacker.

While we consider adversarial attack methods that affect the behavior of trained classifiers (i.e., the modulation classifier of the intruder in our case) by perturbing their input data (these attacks are known as test-time or evasion attacks), another class of adversarial machine learning algorithms, called poisoning attacks, aim to compromise the training procedure of classifiers and other machine learning models by modifying their training data [22]. Poisoning attacks have been used in launching and avoiding jamming attacks in wireless communication [23]–[25]; however, since these methods address the training of the machine learning models employed by the jammer and the transmitter, they are orthogonal to our developments.

In summary, our main contributions are as follows:

- We propose a novel defense mechanism that modifies the channel input symbols at the transmitter in order to reduce the modulation-classification accuracy at the intruder while maintaining a low BER at the legitimate receiver.
- We provide a thorough experimental evaluation of the effect of these modifications on the BER of different modulation schemes.
- We demonstrate that by using training data obtained from different SNR values and employing curriculum learning, an intruder can learn a classifier that is much more robust against both the channel noise and the defensive perturbations, improving upon the state of the art in our experiments when no defense mechanism is applied.
- We also demonstrate that the usual trade-off between the code-rate and the BER is still achievable under our defensive modulation schemes, which introduce an unusual compound noise due to the combined effects of the introduced perturbations and the channel noise. More precisely, we show that by using a stronger error correction code, the BER at the legitimate receiver can be reduced while the intruder achieves the same or worse modulation-classification accuracy.

The rest of the article is organized as follows: The system model is described in Section II, followed by the description of our novel modulation perturbation methods in Section III. Experimental results are presented in Section IV, while conclusions are drawn and future work is discussed in Section V.

II. System Model

Consider a transmitter that maps a binary input sequence $\mathbf{w} \in \{0,1\}^m$ into a sequence of n complex channel input symbols, $\mathbf{x} \in \mathbb{C}^n$, employing forward error correction coding. The input data is first encoded by the channel encoder, and then modulated for transmission. Formally, the modulated signal $\mathbf{x}$ is obtained as $\mathbf{x} = M_s(\mathbf{w})$, where $s \in \mathcal{S}$ is the employed modulation scheme with $\mathcal{S}$ denoting the finite set of available modulation schemes, and for any s, $M_s : \{0,1\}^m \rightarrow \mathbb{C}^n$ denotes the whole encoder function with modulation s. We assume that M_s satisfies the power constraint $(1/n)\|\mathbf{x}\|_2^2 \leq 1$ for any input sequence $\mathbf{w}$. After encoding, signal $\mathbf{x}$ is sent over a noisy channel, assumed to be an additive white Gaussian noise (AWGN) channel: baseband signals $\mathbf{y}_1$ and $\mathbf{y}_2$, received by the intended receiver and the intruder, respectively, are given by

$$\mathbf{y}_i = M_s(\mathbf{w}) + \mathbf{z}_i = \mathbf{x} + \mathbf{z}_i, \quad i = 1, 2, \tag{1}$$

where $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{C}^n$ are independent channel noise (also independent of $\mathbf{x}$) with independent zero-mean complex Gaussian components with variance σ_1^2 and σ_2^2, respectively.

The intended receiver, upon receiving the sequence of noisy channel symbols $\mathbf{y_1}$, demodulates the received signal, and decodes the underlying message bits with the goal of minimizing the (expected) BER $\mathbb{E}[(\mathbf{w}, \mathbf{y_1})]$, where

$$e(\mathbf{w}, \mathbf{y}_1) \triangleq \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}\{w_i \neq \hat{w}_i\}, \tag{2}$$

$\hat{\mathbf{w}}$ is the decoded bit sequence from $\mathbf{y}_1$, and the expectation is over the uniformly random input bit sequence $\mathbf{w}$ and the noise sequence $\mathbf{z_1}$.[2]

The intruder aims to determine the modulation scheme employed by the transmitter based on its received noisy channel output $\mathbf{y}_2$. The transmitter, on the other hand, wants to communicate without its modulation scheme being correctly detected by the intruder, while keeping the BER at an acceptable level.

Formally, the aim of the intruder is to determine, for any sequence of channel output symbols $\mathbf{y}_2 \in \mathbb{C}^n$, the modulation method used by the transmitter. This leads to a *classification* problem where the label $s \in \mathcal{S}$ is the employed modulation scheme, and the input to the classifier is the received channel sequence $\mathbf{y}_2 \in \mathbb{C}^n$. We consider the case in which the intruder implements a score-based classifier, and assigns to $\mathbf{y}_2$ the label $\hat{s} = \mathrm{argmax}_{s' \in \mathcal{S}} f_{\boldsymbol{\theta}}(\mathbf{y}_2, s')$, where $f_{\boldsymbol{\theta}} : \mathbb{C}^n \times \mathcal{S} \rightarrow \mathbb{R}$ is a score function parametrized by $\boldsymbol{\theta} \in \mathbb{R}^d$, which assigns a score (pseudo-likelihood) to each possible class $s' \in \mathcal{S}$ for every $\mathbf{y}_2$, and finally selects the class with the largest score. With a slight abuse of notation, we denote the resulting class label by $\hat{s} = f_{\boldsymbol{\theta}}(\mathbf{y}_2)$. The goal of the intruder is to maximize the probability $\Pr(s = \hat{s})$ of correctly detecting the modulation scheme, which we will also refer to as the success probability of the intruder.[3] For state-of-the-art modulation detection schemes [7]–[11], $f_{\boldsymbol{\theta}}$ is a convolutional neural network classifier, $\boldsymbol{\theta}$ is the vector of the weights of the neural network, while the $f_{\boldsymbol{\theta}}(\mathbf{y}_2, s')$ are the so-called logit values for the class labels $s' \in \mathcal{S}$.

The performance of both the intended receiver, measured by the BER, and the intruder, measured by the detection accuracy, depend on the signal-to-noise ratio (SNR) of the corresponding channels, $\frac{1}{\sigma_1^2}$ and $\frac{1}{\sigma_2^2}$, respectively. We assume that these SNR values are known by the legitimate receiver and the intruder, which can employ a specific $f_{\boldsymbol{\theta}}$ for each SNR value. We will also assume that the intruder has access to training data at the SNR value $\frac{1}{\sigma_2^2}$ to train $f_{\boldsymbol{\theta}}$. This can be done offline as the intruder can generate as much training data as required at a specific SNR value.

[2] For any event E, $\mathbb{I}\{E\} = 1$ if E holds, and 0 otherwise. Furthermore, for any real or complex vector $\mathbf{v}$, v_i denotes its ith coordinate.

[3] Here we assume an underlying probabilistic model about how the the bit sequence $\mathbf{w}$ and modulation scheme is selected.

III. Modulation Perturbation to Avoid Detection

In this article we intend to modify the encoding processes M_s such that, given a modulation scheme $s \in \mathcal{S}$, the new encoding method M_s' ensures that the intruder's success probability gets smaller, while the BER of the receiver (using the same decoding procedure for M_s) does not increase substantially. Our solution is motivated by adversarial attacks for image classification, where it is possible to modify images such that the modification is imperceptible to a human observer, but it makes state-of-the-art image classifiers to err [12], [13]. Adversarial examples are particularly successful in fooling high-dimensional DNN classifiers. Applying the same idea to our problem, we aim to find defensive modulation schemes M_s' such that $M_s'(\mathbf{w}) \approx M_s(\mathbf{w})$, but the intruder misclassifies the new received signal $\mathbf{y}_2' = M_s'(\mathbf{w}) + \mathbf{z}_2$ with higher probability.

A. Adversarial Attack in an Idealized Scenario

Following directly the idea of adversarial attacks on image classifiers [13], an idealized yet impractical adversarial attack mechanism is proposed in [14] which modifies a correctly classified channel output sequence $\mathbf{y}_2$ (i.e., for which $s = f_{\boldsymbol{\theta}}(\mathbf{y}_2)$) with a perturbation $\boldsymbol{\delta} \in \mathbb{C}^n$ such that $f_{\boldsymbol{\theta}}(\mathbf{y}_2 + \boldsymbol{\delta}) \neq f_{\boldsymbol{\theta}}(\mathbf{y}_2)$, the true label, while imposing the restriction $\|\boldsymbol{\delta}\|_2 \leq \epsilon$ for some small positive constant ϵ. Thus, to mask the modulation scheme, the goal is to find, for each correctly classified $\mathbf{y}_2$ separately, a perturbation $\boldsymbol{\delta}$ that maximizes the zero-one loss:

$$\text{maximize} \;\; \mathbb{I}\{f_{\boldsymbol{\theta}}(\mathbf{y}_2 + \boldsymbol{\delta}) \neq s\} \;\text{such that}\; \|\boldsymbol{\delta}\|_2 \leq \epsilon, \tag{3}$$

where $s = f_{\boldsymbol{\theta}}(\mathbf{y}_2)$ is the true modulation label.

If the maximum is 1, such a $\boldsymbol{\delta}$ results in a successful adversarial perturbation and a successful adversarial example $\mathbf{y}_2 + \boldsymbol{\delta}$ (i.e., one for which the intruder makes a mistake). This approach, however, has two limitations. First of all, as opposed to image classifiers, we are not concerned with the visual similarity of the perturbed signal $\mathbf{y}_2 + \boldsymbol{\delta}$ to the original one, $\mathbf{y}_2$. The reason for bounding the perturbation $\boldsymbol{\delta}$ is instead to guarantee that the BER at the intended receiver is still limited. Moreover, in practice we do not have access to $\mathbf{y}_2$, as it does not only depend on $\mathbf{x}$, but also on the channel noise $\mathbf{z}_2$, which is not available at the transmitter. Therefore, the above mechanism, analyzed in [14], is an *oracle* scheme working under some idealized assumptions, and we use it only as a baseline.

It remains to give an algorithm that finds an adversarial perturbation $\boldsymbol{\delta}$ solving problem (3). However, we note that the target function $\mathbb{I}\{f_{\boldsymbol{\theta}}(\mathbf{y}_2 + \boldsymbol{\delta}) \neq f_{\boldsymbol{\theta}}(\mathbf{y}_2)\}$ is binary, and so no gradient-based search is directly possible. To alleviate this, usually a surrogate loss function $L(\boldsymbol{\theta}, \mathbf{y}_2, s)$ to the zero-one loss is used (which is often also used in training the classifier $f_{\boldsymbol{\theta}}$), which is amenable to gradient-based (first-order) optimization. For classification problems, a standard choice is the cross-entropy loss defined as $L(\boldsymbol{\theta}, \mathbf{y}_2, s) = -\log(1 + e^{-f_{\boldsymbol{\theta}}(\mathbf{y}_2, s)})$, and one can search for adversarial perturbations by solving

$$\text{maximize} \quad L(\boldsymbol{\theta}, \mathbf{y}_2 + \boldsymbol{\delta}, s) \text{ such that } \|\boldsymbol{\delta}\|_2 \le \epsilon. \tag{4}$$

Different methods are used in the literature to solve (4) approximately [13], [17], [26], [27]. In this article we use the state-of-the-art projected (normalized) gradient descent (PGD) attack [28] to generate adversarial examples, which is an iterative method: starting from $\mathbf{y}^0 = \mathbf{y}_2$, at each iteration t it calculates

$$\mathbf{y}^t = \Pi_{\mathcal{B}_\epsilon(\mathbf{y}_2)}\big(\mathbf{y}^{t-1} + \beta \ \text{sign}(\nabla_{\mathbf{y}} L(\boldsymbol{\theta}, \mathbf{y}^{t-1}, s))\big), \tag{5}$$

where $\beta > 0$ denotes the step size, sign denotes the sign operation, and $\Pi_{\mathcal{B}_\epsilon(\mathbf{y}_2)}$ denotes the Euclidean projection operator to the L_2-ball $\mathcal{B}_\epsilon(\mathbf{y}_2)$ of radius ϵ centered at $\mathbf{y}_2$, while ∇ denotes the gradient. The attack is typically run for a specified number of steps, which depends on the computational resources; in practice $\mathbf{y}^t$ is more likely to be a successful adversarial example for larger values of t. We will refer to this *idealized* modulation scheme as the *Oracle Scheme* (*Oracle*).

Note that this formulation assumes that we have access to the logit function $f_{\boldsymbol{\theta}}$ of the intruder; these methods are called *white-box* attacks. If $f_{\boldsymbol{\theta}}$ is not known, one can create adversarial examples against another classifier $f_{\boldsymbol{\theta}'}$, and hope that it will also work against the targeted model $f_{\boldsymbol{\theta}}$. Such methods are called *black-box* attacks, and are surprisingly successful against image classifiers [29]. We will also consider black-box attacks against intruders in our experimental evaluations.

B. Adversarial Attack Through Channel Input Modification

As mentioned before, the *Oracle* scheme is infeasible in practice as the transmitter can only modify the channel input $\mathbf{x} = M_s(\mathbf{w})$ but not $\mathbf{y}_2$ directly. Thus, the new modulation scheme is defined as

$$M_s'(\mathbf{w}) = \alpha(M_s(\mathbf{w}) + \boldsymbol{\delta}), \tag{6}$$

where we will consider different choices for $\boldsymbol{\delta} \in \mathbb{C}^n$, and the multiplier $\alpha = \sqrt{n}/\|M_s(\mathbf{w}) + \boldsymbol{\delta}\|_2$ is used to ensure that the new channel input $\bar{\mathbf{x}} = M_s'(\mathbf{w})$ satisfies the average power constraint $(1/n)\|\bar{\mathbf{x}}\|_2^2 \le 1$. The signals received at the receiver and at the intruder are $\bar{\mathbf{y}}_1 = \bar{\mathbf{x}} + \mathbf{z}_1$ and $\bar{\mathbf{y}}_2 = \bar{\mathbf{x}} + \mathbf{z}_2$, respectively. The difficulty in this scenario is that the effect of any carefully designed perturbation $\boldsymbol{\delta}$ may be (and, in fact, is in practice) at least partially masked by the channel noise. Furthermore, since now the perturbed signal is transmitted at the actual SNR of the channel, the effective SNR of the system is decreased, as the transmitted signal already includes the perturbation $\boldsymbol{\delta}$, which can be treated as noise from the intended receiver's point of view.

Our first and simplest method to find a perturbation $\boldsymbol{\delta}$ disregards the effects of the channel noise and the resulting BER at the receiver. In this method, called the *Perturbation-Based Defensive Modulation Scheme (PDMS)*, we aim to solve the optimization problem (4) with $\mathbf{x}$ in place of $\mathbf{y}_2$, via (5) initialized at $\mathbf{y}^0 = \mathbf{x}$ and with projection to $\mathcal{B}_\epsilon(\mathbf{x})$ (for a specified number of iterations t and perturbation size ϵ).

C. BER-Aware Adversarial Attack

Next, we consider methods that also take into account the BER, $e(\bar{\mathbf{y}}_1, \mathbf{w})$ at the receiver (see Eq. 2): that is, instead of enforcing the perturbation $\boldsymbol{\delta}$ to be small and hoping for only a slight increase in the BER, we optimize also for the latter. There is an inherent trade-off between these two targets: a larger $\boldsymbol{\delta}$ results in a bigger reduction in the detection accuracy of the intruder, but will also increase the BER at the receiver. We consider two methods to handle this trade-off:

In the first one, called *BER-Aware Defensive Modulation Scheme (BDMS)*; we consider a (signed) linear combination of our two target functions in order to balance the above two effects,

$$L_\lambda(\boldsymbol{\theta}, \bar{\mathbf{x}}, s, \mathbf{z}_1, \mathbf{z}_2) = L(\boldsymbol{\theta}, \bar{\mathbf{x}} + \mathbf{z}_2, \boldsymbol{\delta}) - \lambda e(\bar{\mathbf{x}} + \mathbf{z}_1, \mathbf{w})$$

for some $\lambda > 0$, where $\bar{\mathbf{y}}_i = \bar{\mathbf{x}} + \mathbf{z}_i$, $i = 1, 2$, and aim to find a perturbation $\boldsymbol{\delta}$ or, equivalently, a modulated signal $\bar{\mathbf{x}} = \mathbf{x} + \boldsymbol{\delta}$ that maximizes the expectation

$$\mathbb{E}_{\mathbf{z}_1, \mathbf{z}_2}[L_\lambda(\boldsymbol{\theta}, \mathbf{x}, s, \mathbf{z}_1, \mathbf{z}_2)] \tag{7}$$

with respect to the channel noise $\mathbf{z}_1, \mathbf{z}_2$. Here we can use stochastic gradient ascent[4] to compute an approximate local optimum, but in practice we find that enforcing $\boldsymbol{\delta}$ to be small during iterations improves the performance; hence, we use a stochastic version of PGD optimization (5): starting at $\mathbf{x}^0 = \mathbf{x}$, our candidate for $\bar{\mathbf{x}}$ is iteratively updated as

$$\mathbf{x}^t = \Pi_{\mathcal{B}_\epsilon(\mathbf{x})}\big(\mathbf{x}^{t-1} + \beta \cdot \text{sign}(\nabla_{\mathbf{x}} L(\boldsymbol{\theta}, \mathbf{x}^{t-1}, s, \mathbf{z}_1^t, \mathbf{z}_2^t))\big),$$

where $\mathbf{z}_i^t$ are independent copies of $\mathbf{z}_i$, respectively, for $i = 1, 2$, and $t = 1, 2, \ldots$. Although $\mathbb{E}_{\mathbf{z}_1}[e(\bar{\mathbf{x}} + \mathbf{z}_1, \mathbf{w})]$ is differentiable, $e(\mathbf{y}, \mathbf{w})$ for a given fixed value of $\mathbf{y}$ is not (since it takes values from the finite set $\{0, 1/n, \ldots, 1\}$). Similarly to [30], we approximate the gradient of the expected error using simultaneous perturbation stochastic approximation (SPSA) [31] as

$$\hat{\nabla}_{\mathbf{y}}\, e(\mathbf{y}, \mathbf{w}) \triangleq \frac{1}{K}\sum_{k=1}^{K} \frac{e(\mathbf{y} + \eta\mathbf{r}_k, \mathbf{w}) - e(\mathbf{y} - \eta\mathbf{r}_k, \mathbf{w})}{2\eta}\mathbf{r}_k^\top, \tag{8}$$

where $\mathbf{r}_1, \ldots, \mathbf{r}_K$ are random vectors selected independently and uniformly from $\{-1, 1\}^n$ (the notation $\hat{\nabla}$ is used to indicate that this is not a real gradient).

In the alternative *BER-Aware Orthogonal Defensive Modulation Scheme (BODMS)*, instead of maximizing the combined target (7), we try to maximize the cross-entropy loss $L(\boldsymbol{\theta}, \bar{\mathbf{y}}_2, s)$ while not increasing (substantially) the BER $e(\bar{\mathbf{y}}_1, \mathbf{w})$. In order to do so, we maximize $L(\boldsymbol{\theta}, \bar{\mathbf{y}}_2, s)$ using stochastic PGD (again, in every step we choose independent noise realizations), but we restrict the steps in the directions

[4] However, similarly to the literature on adversarial attack methods, we often call these methods gradient *descent* instead of ascent.

where the BER does not change. Thus, in every step we update $\mathbf{x}^{t-1}$ in a direction *orthogonal* to the gradient of the BER defined as

$$\nabla_o L(\boldsymbol{\theta}, \mathbf{x}^{t-1} + \mathbf{z}_2^t, s) \triangleq \nabla_{\mathbf{x}} L(\boldsymbol{\theta}, \mathbf{x}^{t-1} + \mathbf{z}_2^t, s) - \langle \nabla_{\mathbf{x}} L(\boldsymbol{\theta}, \mathbf{x}^{t-1} + \mathbf{z}_2^t, s), \mathbf{d}_e \rangle \mathbf{d}_e$$

where $\mathbf{d}_e = \hat{\nabla}_{\mathbf{x}} e(\mathbf{x}^{t-1} + \mathbf{z}_1^t, \mathbf{w}) / \|\hat{\nabla}_{\mathbf{x}} e(\mathbf{x}^{t-1} + \mathbf{z}_1^t, \mathbf{w})\|_2$ is the (approximate) gradient direction of the BER (computed, e.g., using SPSA as in Eq. 8).

IV. Experimental Evaluation

In this section, we test and compare the performance of the proposed methods through numerical simulations. We assume that the binary source data is generated independently and uniformly at random, and is encoded using a rate 2/3 convolutional code before modulation. Eight standard baseband modulation schemes are considered: GFSK, CPFSK, PSK8, BPSK, QPSK, PAM4, QAM16, QAM64. A square-root raised cosine filter is used for pulse shaping of the modulated data with a filter span of 10, roll-off factor of 0.25 and upsampling factor of 8 samples per symbol and the modulated data is sent over an AWGN channel with SNR varying between −20 dB and 20 dB. We consider identical SNRs during both the training of the intruder and at test time. After hard decision demodulation, the receiver uses Viterbi decoding to estimate the original source data.

We follow the setup of [8] for modulation detection: The intruder has to estimate the modulation scheme after receiving 128 complex I/Q (in-phase /quadrature) channel symbols; this is because we assume that the modulation detection is only the first step for the intruder, which then uses this information for either trying to decode the message or to interfere with its transmission. Therefore, the modulation detection should be completed based on a short sequence of channel symbols. As the classifier, we first consider the deep convolutional neural network architecture of [8] for the intruder (given in Table I-a), which operates on the aforementioned 256-dimensional data. We train this network for 100 epochs with a batch size of 100 samples and use the Adam optimizer [32] with a learning rate of 0.001.

For each modulation scheme, we generate data resulting in approximately 245000 I/Q channel symbols (note that for different modulation schemes this corresponds to different number of data bits), split into blocks of 128 I/Q symbols ($n = 128$), as explained above. The last 300 blocks for each modulation scheme are reserved for testing the performance (tests are repeated 20 times), while we train a separate classifier for each SNR value based on the above data. As shown in Fig. 1 (see the curve with label *NoPerturb*), for high SNR values the accuracy of the modulation classification is close to 90%. As expected, the classification accuracy degrades as the SNR decreases (as the noise masks the signal), but even at −10 dB, the intruder can achieve a 40% detection accuracy (as opposed to the 12.5% accuracy a completely random detector would achieve).

In the experiments, we compare this performance with the following defensive modulation schemes and baselines:

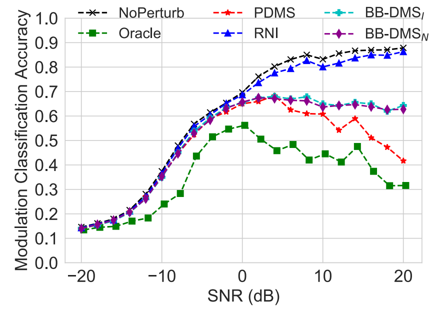


Fig. 1. Modulation-classification accuracy of the intruder as a function of SNR for different defensive modulation schemes.

TABLE I

Neural Network Architectures: (a) the Architecture of [8] Used for Modulation Classification; (b) the Modified Architecture Used in Our Black-Box Modulation Scheme *BB-DMS$_N$*

Layer	Output dimensions
Input	128 x 2
Convolution (128 filters, size 8 x 2) + ReLU	121 x 128
Max Pooling (size 2, strides 2)	60 x 128
Convolution (64 filters, size 16 x1) + ReLU	45 x 64
Max Pooling (size 2, strides 2)	22 x 64
Flatten	1408
Dense + ReLU	128
Dense + ReLU	64
Dense + ReLU	32
Output: Dense + softmax	8

(a)

Layer	Output dimensions
Input	128 x 2
Convolution (64 filters, size 4 x 2) + ReLU	125 x 64
Max Pooling (size 2, strides 2)	62 x 64
Convolution (32 filters, size 8 x1) + ReLU	55 x 32
Max Pooling (size 2, strides 2)	27 x 32
Convolution (32 filters, size 8 x1) + ReLU	20 x 32
Max Pooling (size 2, strides 2)	10 x 32
Flatten	320
Dense + ReLU	100
Output: Dense + softmax	8

(b)

- Our three defensive modulation schemes, *PDMS*, *BDMS*, and *BODMS*, as well as the *Oracle* scheme as a baseline;
- Adding uniform random noise of L_2-norm ϵ to a block, called *random noise insertion* (*RNI*), which is then normalized for the power constraints;
- Black-box defensive modulation schemes that do not use the classifier of the intruder, but calculate *PDMS* against different classifiers. We consider two variants: *Black-box DMS-identical* (*BB-DMS$_I$*) uses a classifier that has the same architecture as that of the intruder's, but is trained separately (assuming no channel noise). Alternatively, *BB-DMS-non-identical* (*BB-DMS$_N$*) employs a classifier with a different architecture than the intruder's, and is also trained assuming no channel noise – its architecture is shown in Table I-b.

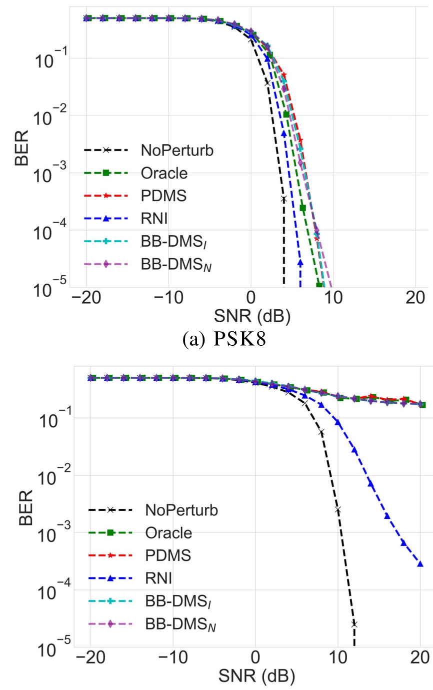
All the above schemes, except for *RNI*, are implemented using the projected (normalized) gradient descent (PGD) [28] method from the CleverHans Library [33], with 20 iterations,

$\beta = 0.2$ and $\epsilon = 3$. $\epsilon = 3$ results in significant reduction in modulation-classification accuracy without incurring too large BER at the intended receiver and has been determined by running experiments over different values of ϵ. *RNI* uses the same ϵ. Note that a perturbation of this size accounts for about 7% of the total energy of a block (which is 128 due to our normalization to the energy constraint). *Oracle* serves as an upper bound on the achievable defensive performance given the parameters, while the role of *RNI* is to analyze the effect of carefully crafted perturbations instead of selecting them randomly. *BB-DMS$_I$*, and *BB-DMS$_N$* explore the more practical situation in which the exact classifier of the intruder is not known, but its training method and/or a similar classifier is available to the transmitter.

A. Defensive Modulation Schemes With Norm-Bounded Perturbations

We first consider defensive modulation schemes with a bound on the L_2 norm of the applied perturbation. Fig. 1 shows the modulation-classification accuracy for several methods. It can be seen that adding random noise (*RNI*) helps very little compared to no defense at all (*NoPerturb*). The basic defense mechanism *PDMS* and its black-box versions *BB-DMS$_I$* and *BB-DMS$_N$* become effective from about -5 dB SNR, and, as expected, *PDMS* outperforms *BB-DMS$_I$* and *BB-DMS$_N$*. For smaller SNR values the classification accuracy is relatively small (the channel noise already makes classification hard), and only the *Oracle* defense gives noticeable improvement. As expected, the performance of *PDMS* gets closer to its lower bound, *Oracle*, as the SNR increases (note that the two methods coincide at the limit of infinite SNR). The similar performance of *BB-DMS$_I$*, *BB-DMS$_N$*, and *PDMS* for medium SNR values shows a similar transferability of adversarial attacks in our situation as was observed in other machine learning problems, such as in image classification [29], [34], although this effect deteriorates quickly as the SNR increases and *PDMS* becomes more effective. Note that the two black-box schemes, *BB-DMS$_I$* and *BB-DMS$_N$*, perform very similarly. In a practical scenario, the transmitter may not know the exact architecture of the intruder's classifier. Nevertheless, adversarial attacks designed against one classifier are generally effective against another classifier [29], and the results in Fig. 1 confirm this observation in our scenario as well, and demonstrate that black-box defenses are possible in general. Observe that the classification accuracy of *PDMS* increases up to 0 dB SNR, when the channel noise during both the training phase and test phase is higher than the defensive perturbation and thus, channel noise is the main cause of the performance limitation of the intruder, while the accuracy decreases for higher SNR when the defensive perturbation is larger compared to the channel noise and the defense mechanisms start working.

Table II shows the modulation-classification accuracy for the individual modulation schemes at channel SNR of 20 dB. It can be seen that a defensive perturbation of the same norm ϵ affects different modulation schemes differently, where CPFSK and BPSK appear to be the most robust against



(a) PSK8

(b) QAM64

Fig. 2. BER vs SNR for PSK8 and QAM64 modulated signal for different defensive modulation schemes.

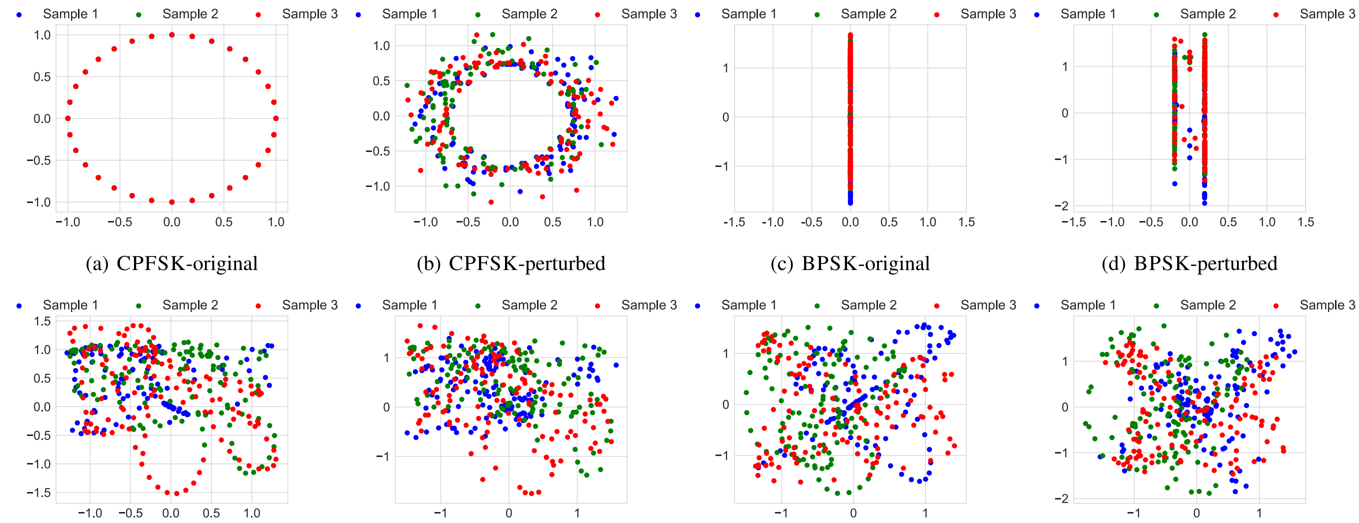
TABLE II

CLASSIFICATION ACCURACY OF *PDMS* FOR DIFFERENT MODULATION SCHEMES WITH $\epsilon = 3$ AT SNR 20 dB

Modulation	*NoPerturb*	*PDMS*
GFSK	1.0	0.02
CPFSK	1.0	0.963
PSK8	0.986	0.0167
BPSK	1.0	0.76
QPSK	0.996	0.07
PAM4	1.0	0.096
QAM16	0.48	0.376
QAM64	0.526	0.376

defensive perturbations. Note that QAM16 and QAM64 are very difficult to classify even without any perturbations, which is in line with the observation made in [8]. Modulated signals without any perturbation and *PDMS*-modulated signals are presented in Fig. 3, which shows that, even after perturbation, CPFSK retains the modulated signal constellation and the perturbed BPSK signals are still different from the output of any other modulation scheme. On the other hand, it becomes difficult to distinguish QAM16 and QAM64 signals.

The reduced classification accuracy of the intruder for *PDMS*, *BB-DMS$_I$*, and *BB-DMS$_N$* is achieved at the cost of an increased BER at the legitimate receiver. To illustrate this effect, Fig. 2 shows the BER for PSK8 and QAM64; the other modulation schemes, except for QAM16, show similar relative behavior to PSK8, but with the error dropping sharply for medium SNR values, with a few dB difference among different modulation schemes (up to about 5 dB for PSK8). On the other hand, the price of using any defense mechanism on QAM64 is severe, resulting in a significantly higher



(a) CPFSK-original (b) CPFSK-perturbed (c) BPSK-original (d) BPSK-perturbed

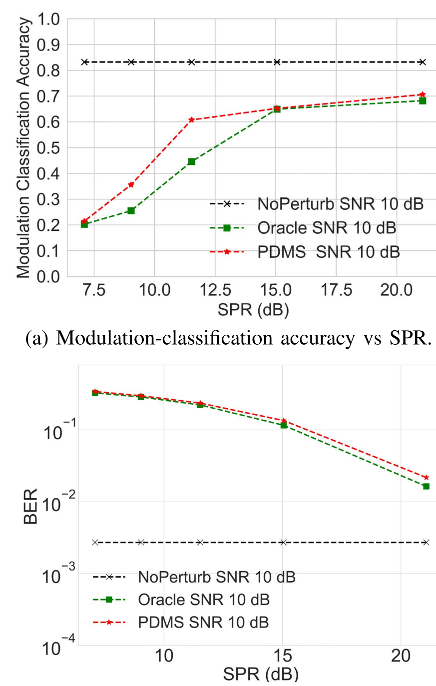
(e) QAM16-original (f) QAM16-perturbed (g) QAM64-original (h) QAM64-perturbed

Fig. 3. Original constellation points and the perturbed channel input symbols with *PDMS* for CPFSK, BPSK, QAM16 and QAM64 for the first three inputs samples (3×128 channel symbols) at the modulation classifier.

BER in the high SNR regime; QAM16 behaves similarly with somewhat smaller BER values. For the *Oracle* defensive scheme, we directly feed the perturbed signal to the decoder to calculate the BER, which is lower than the BER at the decoder when the *PDMS*, *BB-DMS*$_I$, and *BB-DMS*$_N$ defensive schemes are employed for PSK8, while these BERs are essentially the same for QAM64. We observe that *BB-DMS*$_I$ and *BB-DMS*$_N$ achieve very similar BER performances. Since the two black-box schemes perform similarly in terms of both the modulation classification accuracy at the intruder and the BER at the decoder, we consider only *BB-DMS*$_N$ in the remainder of the article, and represent it by *BB-DMS* for convenience.

This negative effect on the BER can be suppressed if the perturbation size is decreased, which, at the same time, results in increased detection accuracy. This is shown in Fig. 4 as a function of the signal-to-perturbation ratio $SPR \triangleq n/\|\boldsymbol{\delta}\|_2^2$ (recall $n = 128$, and $SPR \approx 11.5$dB corresponds to $\epsilon = 3$). In every case, *PDMS* trades off increased BER for reduced detection accuracy compared to the case when no defense mechanism is applied. Also, increasing the number of iterations used in the defensive schemes to compute the perturbations has limited impact on modulation-classification accuracy and BER as the total perturbation is limited to have L_2-norm ϵ.

Fig. 5 shows the trade-off between the average modulation-detection accuracy of the intruder and the BER for the individual modulation schemes for an intruder DNN trained at an SNR of 10 dB (i.e., the training samples are generated with this channel SNR) when the maximum perturbation norm ϵ of *PDMS* takes values in the range $[1, 6]$ (smaller ϵ values correspond to points with smaller BER and larger classification accuracy on each curve). It can be seen that an effective perturbation that results in a reduction in the modulation-classification accuracy also causes an increase in the BER. The trade-off between the two is different for different modulation schemes for the same perturbation constraint ϵ (note that the reported classification accuracy is an average computed over all modulation schemes). It can be seen that an increase in ϵ needed to reduce the average modulation-classification accuracy results in large BER for QAM16 and QAM64. Note that in our experiments BPSK, QPSK, GFSK and CPFSK have zero error rate for this ϵ range, hence they are not included in the figure.



(a) Modulation-classification accuracy vs SPR.

(b) BER vs SPR.

Fig. 4. Effect of signal-to-perturbation ratio (SPR) on the modulation-classification accuracy and the BER (QAM64).

B. BER-Aware Defense Schemes

A more systematic way of improving the BER is to use our BER-aware modulation schemes *BDMS* and *BODMS*.

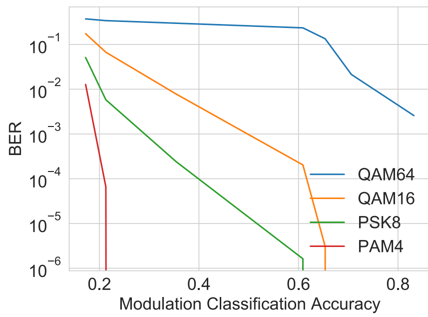


Fig. 5. Trade-off between the modulation-classification accuracy and the BER for *PDMS* with code rate $2/3$, where the L_2 norm of the perturbations is limited by $\epsilon \in [1, 6]$. The accuracy is averaged over all modulation schemes while the BER is shown for each modulation scheme separately. BPSK, QPSK, GFSK and CPFSK have zero error rate for these perturbations.

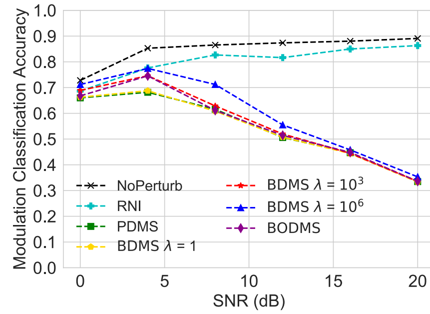


Fig. 6. Modulation-classification accuracy of BER-aware defense mechanisms ($\epsilon = 3$).

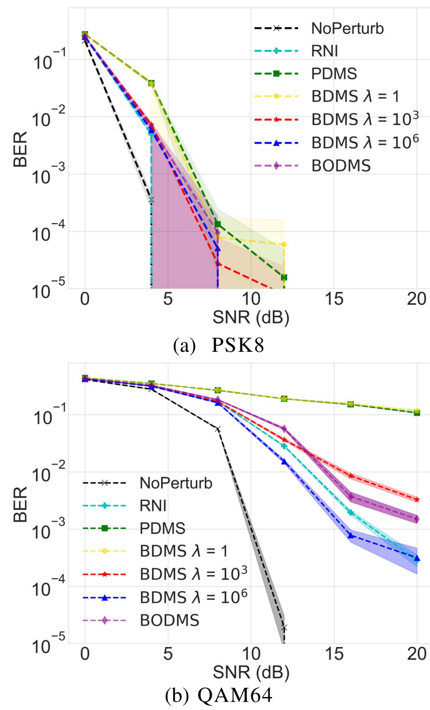


Fig. 7. BER for PSK8 and QAM64 for BER-aware modulation schemes ($\epsilon = 3$). In figure (a) the BER for *RNI* is 0 beyond 5 dB, and all BER values are 0 when the SNR is larger than 12 dB.

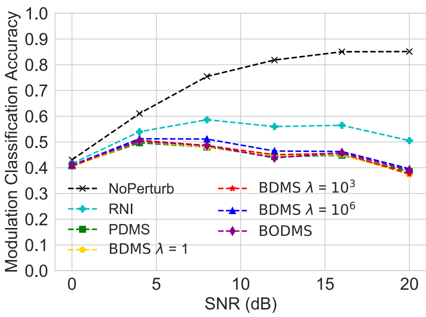
In the numerical experiments, due to the large computational overhead of calculating the SPSA gradient estimates in (8) (with $K = 400$), we only used 400 signal blocks to measure the test performance (instead of the $300 \times 20 = 6000$ blocks used previously). Also, to keep the required computation feasible, in (8) we used error rates calculated over 100 signal blocks (that is, over 12800 perturbed channel input symbols simultaneously). This approximation allowed us to run Viterbi decoding once for every hundred blocks, instead of running it from the beginning for every block, causing a substantial reduction in computational complexity. The approximate gradient of e computed this way was then used to calculate one step of the optimization (i.e., the next candidate perturbation) for each of the 100 blocks simultaneously. The drawback of this approximation is twofold: (i) instead of taking the gradient for a single perturbation, for each perturbation the error gradient is computed as an average coming from perturbing each of the 100 blocks simultaneously (this affects negatively the accuracy of the optimization); (ii) the applied method introduces delays in the transmission as it assumes that all signal blocks perturbed together are available at the transmitter at the same time (this gives some optimistic bias to the optimization compared to non-delayed real-time encoding). Nevertheless, we believe that the negative effects are stronger here, and the performance of our modulation schemes (*BDMS* and *BODMS*) could be improved if the BER of the individual signal blocks were used for gradient estimation in SPSA.

Fig. 6 and Fig. 7 show, respectively, the modulation-classification accuracy and the BER for *BDMS* and *BODMS*, also compared to *PDMS*, *RNI* and *NoPerturb*, against a DNN-based intruder, which is trained with channel input symbols at specific SNR values. The performance of *BDMS* is presented for three different values of λ, namely $1, 10^3, 10^6$. As before, the BER is shown for PSK8 and QAM64, as again QAM64 is the modulation scheme most affected by our perturbations, and except for QAM16 (which is similar to QAM64), and the error rate for the other modulation schemes is similar to (in fact smaller than) that of PSK8 and is very small under any defense mechanisms at high SNR values.

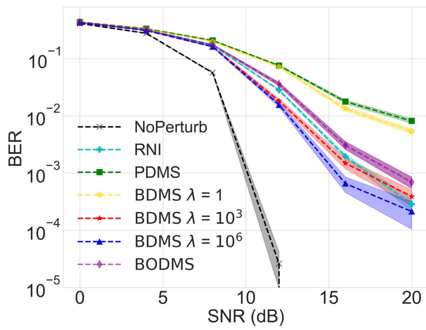
It can be seen that at high SNR (at least 12 dB), all defensive schemes achieve roughly the same classification accuracy, while *BODMS* and *BDMS* for large λ provide significant improvement in the BER (shown for PSK8 and QAM64). Note, however, that the errors are still significantly higher than for the standard QAM64 modulation with no perturbation.

For larger λ values, the BER of *BDMS* for QAM64 is smaller than or approximately the same as for *RNI*, which adds uniform random noise of the same perturbation size, while it significantly outperforms *RNI* in classification accuracy (for PSK8, both *RNI* and *BDMS* achieve low BER, although it can be much smaller for *RNI*). Note that *BODMS* approaches the performance of *BDMS* with a large λ ($10^3 - 10^6$), without the need to tune the hyperparameter λ, and these methods provide a good compromise between the effectiveness of the defense and the increase in the BER.

In addition to DNN-based detectors at the intruder, we also examine defense against one of the best standard modulation detection schemes in the literature, a multi-class decision



(a) Modulation-classification accuracy



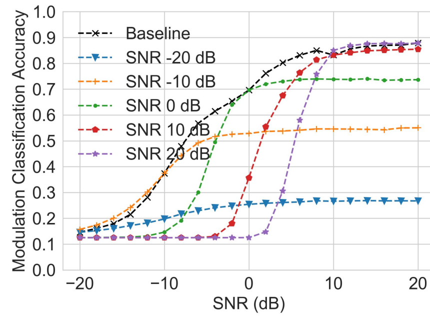
(b) BER

Fig. 8. Modulation-classification accuracy of tree-based intruder and BER (QAM64) for BER-aware modulation schemes ($\epsilon = 3$).

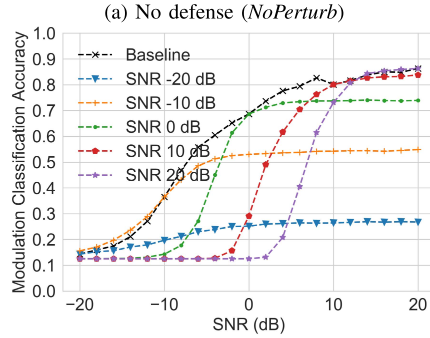
tree trained with expert features obtained from [35], [36]. Fig. 8 shows the modulation-classification accuracy and the BER achieved by employing various defense mechanisms against this intruder. It can be seen that the BER achieved against the tree-based classifier is approximately the same as the one achieved against the DNN-based classifier with *BDMS* and *BODMS*, while the accuracy of the DNN-based classifier is consistently higher, except for some high SNR values, when they are approximately the same. This demonstrates that our observations and conclusions also apply to intruders employing other types of detection mechanisms.

C. Robustness of the Intruder's Classifier

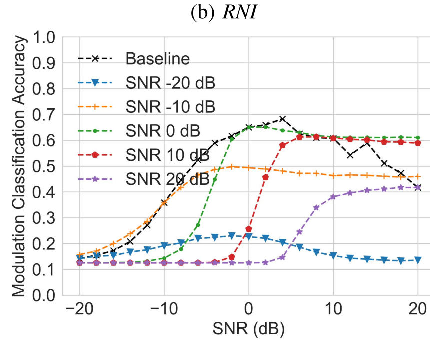
In the previous sections we assumed that the intruder knows the SNR of its received signals perfectly and trains its classifier for this SNR value. Although this may not be possible in practice due to estimation errors or variations in channel quality, assuming more accurate information at the intruder should allow us to design stronger defense mechanisms. In this subsection, we study the robustness of the intruder's detection network against errors in its SNR estimate; that is, we study its modulation-detection accuracy when it is trained for a specific channel SNR, but tested at different SNR values. We show in Fig. 9 the results for three cases: (a) when no defense mechanism is applied (i.e., *NoPerturb*); (b) when uniform noise is added (*RNI*); and (c) when our perturbation-based defense *PDMS* is applied. In each figure, we plot the detection accuracy with respect to the test channel SNR when the intruder is trained at five different SNR values. *Baseline* represents the case in which the test channel SNR matches the training SNR.



(a) No defense (*NoPerturb*)



(b) *RNI*



(c) *PDMS*

Fig. 9. Modulation-classification accuracy of the intruder as a function of the test channel SNR for total perturbation $\epsilon = 3.0$.

We can observe in Fig. 9a that the intruder network trained at channel SNR -20 dB is unable to learn any effective classifier for higher SNR values. As the channel SNR at the time of training increases, its performance improves for a larger range of test SNR values as evident from the plots for SNR -10 dB and 0 dB, but, as one would expect, the accuracy achieved is below the peak accuracy values in the *Baseline* curve. On the other hand, networks trained with high SNR values of 10 dB and 20 dB achieve higher accuracy, close to peak accuracy values in the *Baseline* curve, but tend to breakdown when SNR goes below a certain value (2 dB and 6 dB for intruder networks trained at 10 dB and 20 dB, respectively). It is due to the fact the DNNs learn the classifier function from the training data, and for those trained at high channel SNR, signals with higher noise may lie across decision boundaries learned from less noisy training data, and are wrongly classified.

Note that perturbations in Figs. 9b and 9c are generated with total L_2-norm $\epsilon = 3.0$ for each trained network and at each SNR value. It can be seen from Fig. 9b that adding random perturbations does not reduce the modulation-classification accuracy, yielding similar performance to the case when no defense mechanism is applied (Fig. 9a).

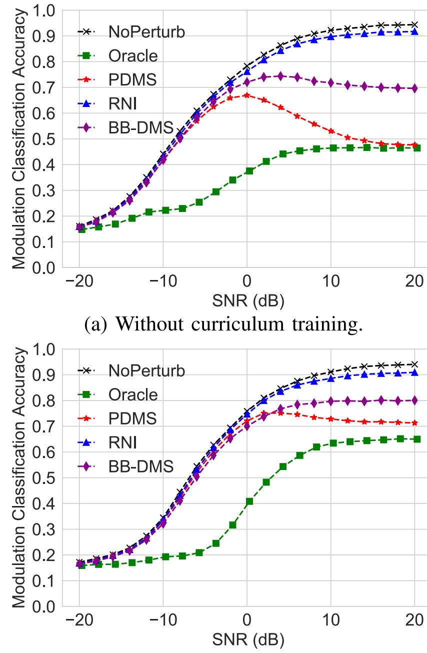
When *PDMS* is employed, if the network is trained for a low SNR value, then test data with lower noise level (higher SNR) will lie at a larger distance from the decision boundaries (learned from noisy data), since the decision boundaries are already accounting for a very high noise level. Therefore, in this case the total perturbation ϵ may not be enough to move the signal to the wrong side of a learned decision boundary of the intruder, resulting in a higher accuracy. On the other hand, when the network is trained for a higher SNR value than the test channel, there is not much variation in the training data due to the absence of noise, and an attacks with even limited perturbation is enough to move the data point to the other side of the learned decision boundary, changing the class label.

In case of *PDMS*, the intruder networks trained at low channel SNR values of 0 dB and 10 dB are more robust against *PDMS* as the decision regions learned by the intruder NN account for larger channel noise and the perturbation norm ϵ is too small compared to this noise at smaller test SNR values to move a perturbed signal over a decision boundary. Once the defensive perturbation ϵ becomes comparable in magnitude to the test data channel SNR then both intruder networks show similar performance for test SNR ($\geq$ 5 dB). In the case of an intruder network trained at SNR 20 dB, perturbation ϵ is large compared to channel noise for higher test SNR values, and thus, results in low modulation-detection accuracy. Also, since the signal is perturbed before transmission, these defensive perturbations are partially masked by the channel noise. This effect of the channel noise is prominent in accuracy curves, though *PDMS* perturbations significantly reduce the detection accuracy as evident in Fig. 9c.

D. Improving Intruder's Performance by Diversifying the Training Data

In this section, we consider the scenario when the intruder has training data available at different SNR values; more precisely, we use 21 different SNR levels uniformly spaced between -20 dB to 20 dB, leading to a total of 21×12966 samples. We consider two different training strategies: (i) randomly shuffle the training data of all channel SNR values to train the intruder's DNN; and (ii) curriculum learning [37], where the training data is arranged in descending order of their SNR values, and the training is started with samples of training data from SNR 20 dB, gradually adding samples with lower SNR values.

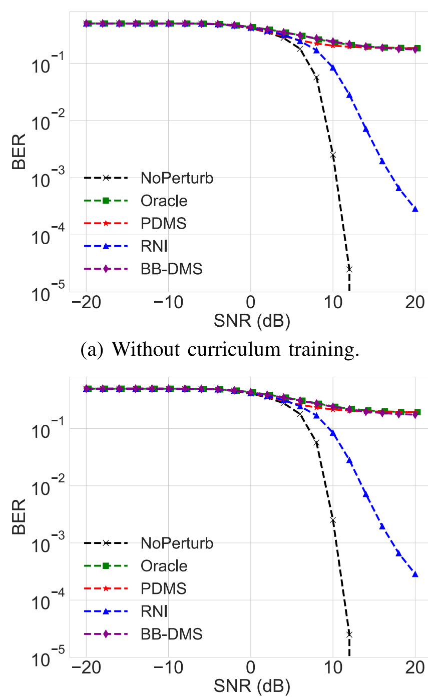
Fig. 10a shows that an intruder network trained with data from all SNR values achieves a higher modulation-classification accuracy for *NoPerturb* and against all defensive modulation strategies compared to the case when only samples from the same SNR values were used (cf. Fig. 1); this is most likely due to the approximately 20-fold increase in the number of training samples used. On the other hand, we can see in Fig. 10b that curriculum training achieves even higher robustness against all the defensive modulation schemes, and even the idealized defensive modulation scheme *Oracle* can be detected with more than 60% accuracy. This is



(a) Without curriculum training.

(b) With curriculum training.

Fig. 10. Modulation-classification accuracy for an intruder trained with/without curriculum training for a complete dataset of channel SNR values ranging from -20 dB to 20 dB ($\epsilon = 3$).

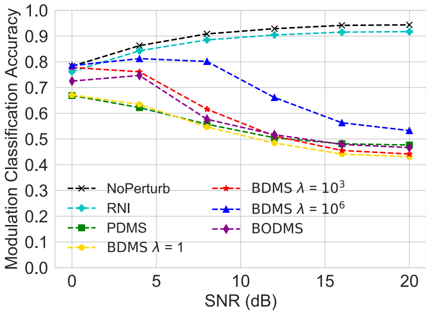


(a) Without curriculum training.

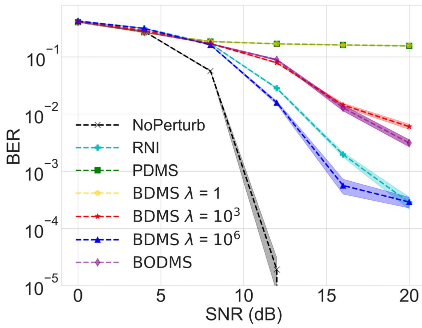
(b) With curriculum training.

Fig. 11. BER for QAM64 for intruder trained with/without curriculum training with complete dataset of channel SNR values ranging from -20 dB to 20 dB ($\epsilon = 3$).

because, in curriculum training, the neural network gradually learns, starting from easier concepts to more complex ones (more noisy channels in our case) and generalizes better to unseen data including those generated by defensive modulation schemes. In both cases, the improvement in detection accuracy is more for higher SNR values. Fig. 11 shows the BER for

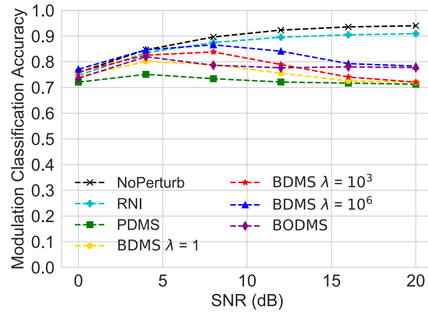


(a) Modulation-classification accuracy

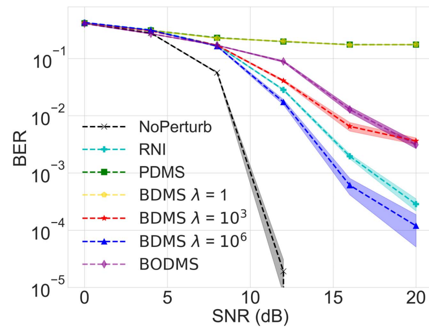


(b) BER for QAM64

Fig. 12. DNN classifier trained with training data of channel SNRs −20 dB to 20 dB without any curriculum ($\epsilon = 3$).



(a) Modulation-classification accuracy



(b) BER for QAM64

Fig. 13. DNN classifier trained with data of channel SNRs −20 dB to 20 dB with curriculum learning ($\epsilon = 3$).

QAM64 when defensive perturbations are used against these intruder modulation classifiers trained over the whole range of SNR values without and with curriculum training, respectively. The achieved BERs are similar to those achieved when the intruder classifiers are trained for a particular SNR in Fig. 2. This shows that the comparison of the detection accuracy discussed above is fair (i.e., the improved detection accuracy is not because the applied defensive perturbations are smaller).

Next, we consider the performance of BER-aware defensive modulation schemes when the intruder classifiers are trained with complete training data of all channel SNR values. The results without any curriculum learning are shown in Fig. 12 for the same DNN-based classifier. It can be seen that the modulation-classification accuracy is quite high, around 95%, when no defense mechanism is employed (*NoPerturb*), and over 90% when only noise is added (*RNI*). We can also observe that, compared to results in Fig. 10a, *BDMS* is less successful against this model for large λ (10^6); on the other hand, the BER is significantly improved, as demonstrated by comparing Fig. 11a and Fig. 12b. There is also a significant improvement in detection accuracy for essentially the same BER compared to the case when only training data for the same SNR value is used (cf. Fig. 6 and Fig. 7).

On the other hand, using this larger set of training data yields no significant improvement in the performance of the tree-based classifier, and the results are very similar to those reported in Fig. 8 (hence, they are omitted).

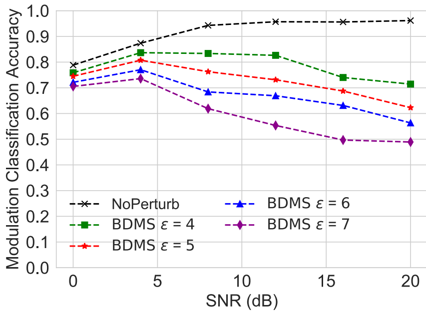
When the DNN-based classifier is trained using the complete dataset with curriculum learning, a significantly higher modulation-classification accuracy can be achieved against all defensive modulation schemes, as shown in Fig. 13. Compared to the non-curriculum learning results in Fig. 12, we can see that the improved detection accuracy also results in a smaller BER. This suggests that, for a fair comparison between the two approaches, we can increase the attack strength in the case of curriculum learning until we achieve similar BER values as in Fig. 12.

To this end, we increase the norm of perturbations for the *BDMS* scheme against the DNN-based intruder network trained with curriculum learning. Note that to make the defense mechanisms work, we need to increase the value of λ, and we have found that (the surprisingly large) $\lambda = 10^{20}$ works well in our experiments. The results are shown in Fig. 14. It can be seen that defensive perturbations with larger norms decrease the modulation-detection accuracy of the intruder, but they also result in significantly higher BER despite the very large λ value.

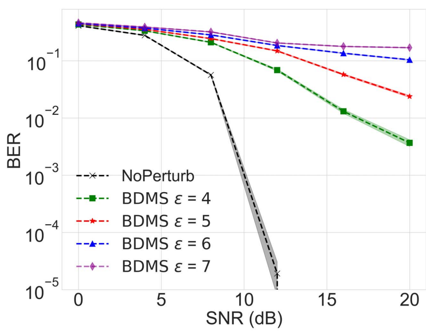
The results in this section showed that using more and diverse data and curriculum training can significantly improve the performance of the intruder and its robustness against various defense mechanisms. While designing better defense mechanisms against these intruders is an interesting and challenging future research direction, one method that can be employed directly at the transmitter is to reduce the code rate, which allows employing stronger attacks at the transmitter. This is explored in the next section.

E. The Effect of the Code Rate

Error correction codes have been traditionally designed and tested against independent Gaussian noise, and it is not clear how they perform in the presence of the adversarial perturbations we introduce, which are statistically very different from the channel noise. In the experiments below we show that the

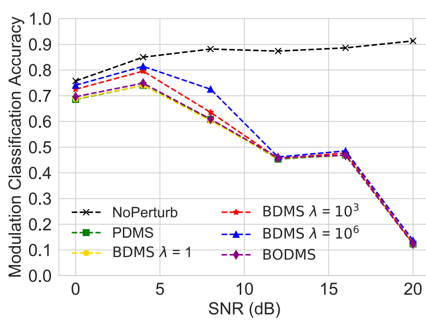


(a) Modulation-classification accuracy

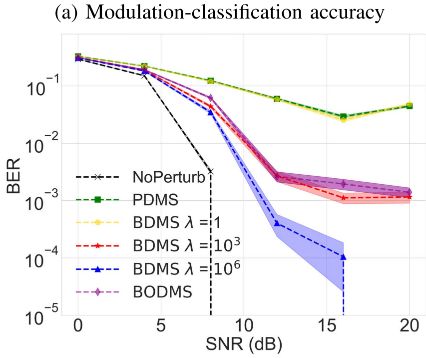


(b) BER for QAM64

Fig. 14. Modulation-classification accuracy and BER (QAM64) for an intruder trained with a dataset of channel SNR ranging from -20 dB to 20 dB with curriculum learning (code rate $= 2/3$, BDMS with $\lambda = 10^{20}$).

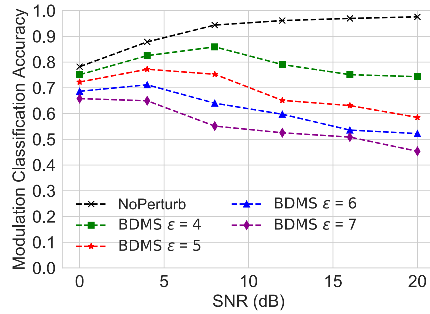


(a) Modulation-classification accuracy

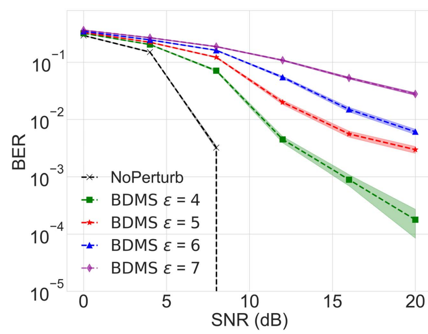


(b) BER for QAM64

Fig. 15. Modulation-classification accuracy of DNN-based intruder and bit error rate (QAM64) for code rate 1/2 for BER-aware modulation schemes ($\epsilon = 3$).



(a) Modulation-classification accuracy



(b) BER for QAM64

Fig. 16. Modulation-classification accuracy and BER (QAM64) for an intruder trained with a dataset of channel SNR ranging from -20 dB to 20 dB with curriculum learning (code rate $= 1/2$, BDMS with $\lambda = 10^{20}$).

conventional trade-off between the code rate and the BER still applies and can be exploited to achieve the desired BER level while keeping the adversary's accuracy low.

In our previous experiments we considered a fixed code rate of $2/3$. To illustrate the effect of the code rate, we evaluate the performance of our BER-aware defense schemes (with $\epsilon = 3$) for a channel code of rate $1/2$ against the usual DNN-based intruder trained for a specific SNR. The results, shown in Fig. 15 demonstrate that both the BER and the detection accuracy can be substantially reduced compared to the case when the code rate is $2/3$ (see Fig. 6 and Fig. 7 for comparison). For example, even for QAM64, *BDMS* (with $\lambda = 10^6$) achieves zero BER for high SNR values (at least 16 dB).

The very small BERs (obtained in the previous experiment) allow the application of more aggressive defensive perturbations when the intruder employs a stronger classifier. Accordingly, we evaluate the *BDMS* defensive scheme (with a large $\lambda = 10^{20}$) for different perturbation norms against a DNN-based intruder trained with curriculum learning over a range of SNR values (the setup is the same as for Fig. 13 except for the code rate). The results, shown in Fig. 16, demonstrate that, compared to Fig. 14, using a lower code rate of $1/2$, the modulation-classification accuracy of the intruder trained with curriculum learning can be reduced without incurring a large BER at the legitimate receiver.

V. Conclusion and Future Work

We proposed a novel approach to secure wireless communication by preventing an intruder from detecting the modulation scheme employed, which is typically the first step of a more advanced attack. In the proposed scheme, the I/Q symbols of the modulated waveform at the transmitter are perturbed using an adversarial perturbation derived against the modulation classifier of the intruder. The perturbation is designed using PGD, whose goal is to identify a perturbation with a limited norm that is sufficient to fool the intruder's classifier. More advanced methods are also proposed, whose goal is also to keep small the BER caused by the perturbation at the legitimate receiver. Experimental results verify the viability of our approach by showing that our methods are able to substantially reduce the modulation-classification accuracy of the intruder with minimal sacrifice in the communication performance. We have also shown that the intruder can improve its detection accuracy significantly by training with a dataset of samples taken from a range of SNR values, especially when curriculum learning is also employed. This provides robustness against channel noise as well as potential defense mechanisms against the intruder, and has led to improvements upon state-of-the-art modulation detectors in our experiments. Finally, we have shown that a better trade-off between the intruder's detection accuracy and the BER at the legitimate receiver can be achieved by sacrificing the communication rate.

An immediate challenge in the implementation of the proposed defense mechanism in practice is the computation of the proposed perturbations, which may introduce some delay. While this can be done in an offline fashion and tabularized for small n, some delay may be unavoidable for large n values, hence efficient methods to calculate the perturbations are of natural interest.

Utilizing the rapid advances in the field of adversarial machine learning, our defense methods can certainly be improved in the future by applying more advanced as well as more universal (e.g., black-box) adversarial attack methods. Another interesting avenue for future research is to develop sophisticated defensive perturbations that can exploit different channel characteristics both at the intruder and legitimate receiver. On the other end of the problem, one can develop better training strategies for the intruder that can achieve more robust performance against these defense mechanisms, for example, by applying adversarial training methods [28].

References

[1] M. Z. Hameed, A. György, and D. Gündüz, "Communication without interception: Defense against modulation detection," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2019, pp. 1–5.

[2] G. E. Prescott, "Performance metrics for low probability of intercept communication systems," Telecommun. Inf. Sci. Lab., Univ. Kansas, Lawrence, KS, USA, Tech. Rep. AFOSR-91-0018, Oct. 1993.

[3] A. D. Wyner, "The wire-tap channel," *Bell Syst. Tech. J.*, vol. 54, no. 8, pp. 1355–1387, Oct. 1975.

[4] D. Gündüz, D. R. Brown, and H. V. Poor, "Secret communication with feedback," in *Proc. Int. Symp. Inf. Theory Its Appl.*, Dec. 2008, pp. 1–6.

[5] B. A. Bash, D. Goeckel, and D. Towsley, "Square root law for communication with low probability of detection on AWGN channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2012, pp. 448–452.

[6] O. A. Dobre, A. Abdi, Y. Bar-Ness, and W. Su, "Survey of automatic modulation classification techniques: Classical approaches and new trends," *IET Commun.*, vol. 1, no. 2, pp. 137–156, Apr. 2007.

[7] G. J. Mendis, J. Wei, and A. Madanayake, "Deep learning-based automated modulation classification for cognitive radio," in *Proc. IEEE Int. Conf. Commun. Syst. (ICCS)*, Dec. 2016, pp. 1–6.

[8] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Trans. Cognit. Commun. Netw.*, vol. 3, no. 4, pp. 563–575, Dec. 2017.

[9] N. E. West and T. O'Shea, "Deep architectures for modulation recognition," in *Proc. IEEE Int. Symp. Dyn. Spectr. Access Netw. (DySPAN)*, Mar. 2017, pp. 1–6.

[10] B. Kim, J. Kim, H. Chae, D. Yoon, and J. W. Choi, "Deep neural network-based automatic modulation classification technique," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2016, pp. 579–582.

[11] X. Liu, D. Yang, and A. E. Gamal, "Deep neural network architectures for modulation classification," in *Proc. 51st Asilomar Conf. Signals, Syst., Comput.*, Oct. 2017, pp. 915–919.

[12] J. Bruna *et al.*, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 2–11.

[13] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–11.

[14] 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, Feb. 2019.

[15] S. Kokalj-Filipovic, R. Miller, N. Chang, and C. L. Lau, "Mitigation of adversarial examples in RF deep classifiers utilizing AutoEncoder pre-training," in *Proc. Int. Conf. Mil. Commun. Inf. Syst. (ICMCIS)*, May 2019, pp. 1–6.

[16] S. Kokalj-Filipovic, R. Miller, and J. Morman, "Targeted adversarial examples against RF deep classifiers," in *Proc. ACM Workshop Wireless Secur. Mach. Learn. WiseML*, 2019, pp. 6–11.

[17] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57.

[18] M. Z. Hameed, A. György, and D. Gündüz, "Communication without interception: Defense against deep-learning-based modulation detection," 2019, *arXiv:1902.10674*. [Online]. Available: https://arxiv.org/abs/1902.10674

[19] B. Flowers, R. M. Buehrer, and W. C. Headley, "Communications aware adversarial residual networks for over the air evasion attacks," in *Proc. MILCOM IEEE Mil. Commun. Conf. (MILCOM)*, Nov. 2019, pp. 133–140.

[20] B. Flowers, R. M. Buehrer, and W. C. Headley, "Evaluating adversarial evasion attacks in the context of wireless communications," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1102–1113, 2020.

[21] B. Kim, Y. E. Sagduyu, K. Davaslioglu, T. Erpek, and S. Ulukus, "Over-the-air adversarial attacks on deep learning based modulation classifier over wireless channels," 2020, *arXiv:2002.02400*. [Online]. Available: http://arxiv.org/abs/2002.02400

[22] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," *Mach. Learn.*, vol. 81, no. 2, pp. 121–148, 2010.

[23] Y. E. Sagduyu, Y. Shi, and T. Erpek, "IoT network security from the perspective of adversarial deep learning," in *Proc. 16th Annu. IEEE Int. Conf. Sens., Commun., Netw. (SECON)*, Jun. 2019, pp. 1–9.

[24] Y. Shi, T. Erpek, Y. E. Sagduyu, and J. H. Li, "Spectrum data poisoning with adversarial deep learning," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, Oct. 2018, pp. 407–412.

[25] T. Erpek, Y. E. Sagduyu, and Y. Shi, "Deep learning for launching and mitigating wireless jamming attacks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 5, no. 1, pp. 2–14, Mar. 2019.

[26] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–17.

[27] P.-Y. Chen, Y. Sharma, H. Zhang, J. Yi, and C.-J. Hsieh, "Ead: Elastic-Net attacks to deep neural networks via adversarial examples," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[28] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–10.

[29] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proc. ACM Asia Conf. Comput. Commun. Secur.*, Apr. 2017, pp. 506–519.

[30] J. Uesato, B. O'Donoghue, P. Kohli, and A. Oord, "Adversarial risk and the dangers of evaluating against weak attacks," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5025–5034.

[31] J. C. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Trans. Autom. Control*, vol. 37, no. 3, pp. 332–341, Mar. 1992.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[33] N. Papernot *et al.*, "Technical report on the CleverHans v2.1.0 adversarial examples library," 2016, *arXiv:1610.00768*. [Online]. Available: http://arxiv.org/abs/1610.00768

[34] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: From phenomena to black-box attacks using adversarial samples," 2016, *arXiv:1605.07277*. [Online]. Available: http://arxiv.org/abs/1605.07277

[35] A.-V. Rosti, "Statistical methods in modulation classification," M.Sc. thesis, Tampere Univ. Technol., Tampere, Finland, Jun. 1999.

[36] A. Abdelmutalab, K. Assaleh, and M. El-Tarhuni, “Automatic modulation classification based on high order cumulants and hierarchical polynomial classifiers,” *Phys. Commun.*, vol. 21, pp. 10–18, Dec. 2016.
[37] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, “Curriculum learning,” in *Proc. 26th Int. Conf. Mach. Learn. (ICML)*, 2009, pp. 41–48.

Muhammad Zaid Hameed received the B.Sc. degree in electrical engineering from the University of Engineering and Technology Lahore, Pakistan, in 2010, and the M.Sc. degree in communications and signal processing and the Ph.D. degree in electrical engineering from Imperial College London, London, U.K., in 2014 and 2020, respectively. He is currently a Post-Doctoral Researcher with the Resilient Information Systems Security Research Group, Imperial College London. His research interests include machine learning, federated learning, deep learning, reinforcement learning, and wireless communications. He received the Best Paper Award at the 7th IEEE Global Conference on Signal and Information Processing (GlobalSIP 2019).

András György received the M.Sc. (Eng.) degree (Hons.) in technical informatics from the Technical University of Budapest, in 1999, the M.Sc. (Eng.) degree in mathematics and engineering from Queen’s University, Kingston, ON, Canada, in 2001, and the Ph.D. degree in technical informatics from the Budapest University of Technology and Economics in 2003.

He was a Visiting Research Scholar with the Department of Electrical and Computer Engineering, University of California, San Diego, USA, in the spring of 1998. From 2002 to 2011, he was with the Computer and Automation Research Institute of the Hungarian Academy of Sciences, where he has been a Senior Researcher and the Head of the Machine Learning Research Group since 2006. From 2003 to 2004, he was also a NATO Science Fellow with the Department of Mathematics and Statistics, Queen’s University. He also held a part-time research position at GusGus Capital Llc., Budapest, Hungary, from 2006 to 2011. From 2012 to 2015, he was a Researcher with the Department of Computing Science, University of Alberta, Edmonton, AB, Canada. From 2015 to 2019, he was a Senior Lecturer with the Department of Electrical and Electronic Engineering, Imperial College London, London, U.K. Since 2018, he has been a Research Scientist with Deepmind, London, U.K. His research interests include machine learning, statistical learning theory, online learning, adaptive systems, optimization, and information theory.

Dr. György is an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and regularly serves as a senior program committee member or area chair of leading conferences in machine learning and information theory. He received the Best Paper Award at the 7th IEEE Global Conference on Signal and Information Processing (GlobalSIP 2019), the Gyula Farkas prize of the János Bolyai Mathematical Society in 2001, and the Academic Golden Ring of the President of the Hungarian Republic in 2003.

Deniz Gündüz (Senior Member, IEEE) received the B.S. degree in electrical and electronics engineering from METU, Turkey, in 2002, and the M.S. and Ph.D. degrees in electrical engineering from the NYU Tandon School of Engineering (formerly Polytechnic University) in 2004 and 2007, respectively.

After his PhD, he served as a Post-Doctoral Research Associate with Princeton University and a Consulting Assistant Professor with Stanford University. He served as a Research Associate with CTTC, Barcelona, Spain. In 2012, he joined the Electrical and Electronic Engineering Department of Imperial College London, U.K., as a Lecturer. He was promoted to Reader in 2016 and to Professor in 2020. He serves as the Deputy Head of the Intelligent Systems and Networks Group and leads the Information Processing and Communications Laboratory (IPC-Lab). He is also an Associate Researcher with the University of Modena and Reggio Emilia and held visiting positions with the University of Padova from 2018 to 2020 and Princeton University from 2009 to 2012. His research interests include the areas of communications and information theory, machine learning, and privacy. He was a recipient of the IEEE Communications Society-Communication Theory Technical Committee (CTTC) Early Achievement Award in 2017, the Starting Grant of the European Research Council (ERC) in 2016, the IEEE Communications Society Best Young Researcher Award for the Europe, Middle East, and Africa Region in 2014, the Best Paper Award at the 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP), the 2016 IEEE Wireless Communications and Networking Conference (WCNC), the Best Student Paper Awards at the 2018 IEEE Wireless Communications and Networking Conference (WCNC), and the 2007 IEEE International Symposium on Information Theory (ISIT). He served as a Symposium Co-Chair for the 2020 IEEE International Conference on Communications and a General Co-Chair for the 2019 London Symposium on Information Theory and 2016 IEEE Information Theory Workshop. He is an Area Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS and the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (JSAC). He also serves as an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He served as an Editor for the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING from 2016 to 2020 and the IEEE TRANSACTIONS ON COMMUNICATIONS from 2013 to 2018. He is a Distinguished Lecturer for the IEEE Information Theory Society from 2020 to 2022.