Channel Aware Adversarial Attacks are Not Robust

Sujata Sinha and Alkan Soysal

Wireless@VT, Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA

***Abstract*—Adversarial Machine Learning (AML) has shown significant success when applied to deep learning models across various domains. This paper explores channel-aware adversarial attacks on DNN-based modulation classification models within wireless environments. Our investigation focuses on the robustness of these attacks with respect to channel distribution and path-loss parameters. We examine two scenarios: one in which the attacker has instantaneous channel knowledge and another in which the attacker relies on statistical channel data. In both cases, we study channels subject to Rayleigh fading alone, Rayleigh fading combined with shadowing, and Rayleigh fading combined with both shadowing and path loss. Our findings reveal that the distance between the attacker and the legitimate receiver largely dictates the success of an AML attack. Without precise distance estimation, adversarial attacks are likely to fail.**

I. Introduction

Machine learning (ML) techniques find wide applicability in data-rich domains such as natural language processing, computer vision, and speech recognition. Specifically, in the wireless domain, where data is high dimensional, ML models actively learn data representations and automate signal detection [1], waveform design [2], and radio signal classification [3], [4].

However, ML models are susceptible to adversarial attacks. While a key strength of ML techniques is their ability to adapt to new data, adversaries can discover blind spots, thereby subverting these advantages. Evasion of a trained target classifier occurs when an adversary exploits the system vulnerabilities and carefully crafts adversarial perturbations that fool the target system [5], [6].

Vulnerabilities to adversarial attacks extend to any system that employs a deep neural network (DNN). For example, in the wireless domain, autoencoder-based communication systems and signal detection in OFDM are susceptible to such attacks [7]. It is also common to leverage AI techniques to disrupt radio access for 5G and 6G systems [8]–[10] and cooperative spectrum sensing [11]. For dynamic channel access agents, [12] presents jamming attacks using generative adversarial networks for modulation classifiers. Previous works extensively explore the threats and impacts of adversarial attacks on the DNN-based modulation recognition in various environments [7], [13]–[16].

In this paper, we consider channel-aware adversarial attacks on automatic modulation classification models that employ DNN-based classifiers to assign high-dimensional spectrum data to one of the several mutually exclusive modulation constellations. A wireless adversary can exploit the properties of DNNs to fool the receiver into making incorrect class predictions. Analogous to computer vision literature, where the perturbations are imperceptible to a human observer, the interference designed for a legitimate receiver in the wireless domain should be of the same power (or below) as the noise level [17]. However, differences exist between adversarial attacks crafted for computer vision and those for wireless communications. An adversary who sends over-the-air perturbations to a target receiver cannot manipulate data directly at the input to the classifier. Perturbations crafted by the adversary undergo phase and amplitude changes as they traverse the communication channel to the legitimate receiver. Consequently, the received perturbations may not meet the minimum power requirements or may lose features necessary to fool the target receiver. Accurately estimating the channel parameters and crafting perturbations that align with the domain constraints of a wireless communication system are inherently challenging.

Channel-aware attacks aim to account for realistic channel effects, ensuring that a legitimate receiver is still fooled even after attributes such as amplitude or phase of a transmitted perturbation signal change during air travel. The authors in [18] introduce distinct strategies for crafting these perturbations, tailoring each to varying levels of uncertainty in the channel between the adversary and receiver, the data inputs received by the target system, and the level of system knowledge the adversary can access. In a related work [19], the authors examine multiple concurrent perturbations sent over different channels to a wireless receiver, successfully leading the target modulation classifiers to make erroneous decisions.

In this work, we investigate the assumptions based on the relationship between the efficacy of "channel-aware" adversarial attacks and the degree of domain knowledge available to the adversary. We assume the adversary uses either instantaneous or statistical knowledge about the channel between themselves and the target receiver to craft perturbations for different channel models. For the adversary-receiver channel, we consider three models: (i) Rayleigh fading only, (ii) Rayleigh fading with shadowing, and (iii) Rayleigh fading with shadowing and path loss. We identify whether such attacks are robust across various propagation environments and at different adversary-receiver distances. Our findings report that the reliability of adversarial attacks increases with the accuracy of estimated adversary-channel parameters. For a more practical assumption of statistical channel knowledge, we conclude that channel-aware adversarial attacks are not robust, demonstrating significant vulnerabilities when channel parameters vary.

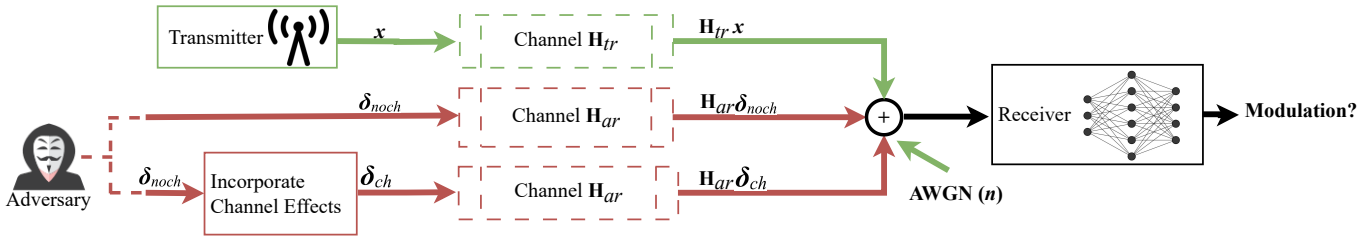

Fig. 1: Communication channel between a legitimate transmitter/receiver pair and an adversary. The adversary crafts adversarial perturbation without considering realistic channel effects $\boldsymbol{\delta}_{noch}$ or incorporates realistic channel effects into $\boldsymbol{\delta}_{noch}$. To incorporate realistic channel effects, we employ the MRPP technique [18].

II. System Model

We consider a wireless communication system comprising a legitimate transmitter/receiver pair and an adversary, as depicted in Fig. 1. The receiver employs a pre-trained DNN-based modulation classifier on the received signal to predict the modulation constellation used by the legitimate transmitter. The adversary transmits over-the-air perturbations to the target receiver with the aim of generating erroneous classification outcomes.

Let $\mathbf{x} \in \mathcal{X} \subset \mathbb{C}^p$ denote the in-phase and quadrature time samples of the waveform transmitted by the legitimate user, where p is the number of complex-valued samples. In the absence of an adversary, the legitimate receiver observes

$$\mathbf{y} = \mathbf{H}_{tr}\mathbf{x} + \mathbf{n} \tag{1}$$

where $\mathbf{H}_{tr} = \text{diag}\{h_{tr_1}, \ldots, h_{tr_p}\} \in \mathbb{C}^{p\times p}$ is the diagonal channel matrix between the transmitter and receiver, and $\mathbf{n} \in \mathbb{C}^p$ is the complex Gaussian noise. The classifier $f(\mathbf{y};\boldsymbol{\theta})$, parameterized by $\boldsymbol{\theta}$, categorizes the received signal $\mathbf{y}$ into the appropriate modulation constellation. Specifically, for each input $\mathbf{y}$, the classifier f assigns a label $\hat{l}(\mathbf{y},\boldsymbol{\theta}) = \arg\max_k f_k(\mathbf{y},\boldsymbol{\theta})$, where $f_k(\mathbf{y},\boldsymbol{\theta})$ represents the output of the classifier corresponding to the kth modulation type.

The presence of an adversary during the testing phase introduces a perturbation $\boldsymbol{\delta} \in \mathbb{C}^p$ at the legitimate receiver. The adversarial attack aims to fool the target DNN-based classifier, modifying the received signal as

$$\mathbf{y}_{adv} = \mathbf{H}_{tr}\mathbf{x} + \mathbf{n} + \mathbf{H}_{ar}\boldsymbol{\delta} \tag{2}$$

$$= \mathbf{y} + \mathbf{H}_{ar}\boldsymbol{\delta} \tag{3}$$

where $\mathbf{H}_{ar} = \text{diag}\{h_{ar_1}, \ldots, h_{ar_p}\} \in \mathbb{C}^{p\times p}$ is the diagonal channel matrix between the attacker and receiver.

Our work investigates the robustness of perturbations received at the classifier f through three different adversary-receiver channel models. The first channel model introduces random fading according to the Rayleigh distribution, denoted as $\mathbf{H}_{ray}$. Consequently, in this channel model, $\mathbf{H}_{ar} = \mathbf{H}_{ray}$. The second channel model considers both fast (Rayleigh) fading and slow (lognormal) fading effects, represented by $\mathbf{H}_{ar} = \sqrt{\psi}\mathbf{H}_{ray}$, where ψ is lognormally distributed. The third channel model incorporates propagation path loss, characterized by the path loss exponent (γ) and distance (d), in addition to Rayleigh fading with shadowing. Here, we have $\mathbf{H}_{ar} = \sqrt{K(\frac{d_0}{d})^{\gamma}\psi}\mathbf{H}_{ray}$[1].

In addition to these three channel models between the adversary and the receiver, we also consider two scenarios concerning the adversary's knowledge of these channel models. In the first, we assume the adversary knows the instantaneous channel state information from the adversary to the receiver. In the second, the adversary knows only the statistical properties of the channel. It is important to note that in all cases, the training data and target model f with parameters $\boldsymbol{\theta}$ are assumed to be known.

Overall, we explore six scenarios. Each scenario combines a different set of channel models and assumptions about the adversary's knowledge. Notably, the last scenario assumes that the parameters of path loss are unknown even when the adversary knows the statistical properties of Rayleigh and lognormal fading. Although very practical, such a case has yet to be considered in prior literature. Prior work assumed that path-loss parameters are also known when channel distribution is assumed to be known to the attacker.

III. Channel-aware Adversarial Attacks

In this paper, an adversary in a wireless setting crafts perturbations during the training phase and deploys these attacks against the legitimate receiver in the testing phase. During training, the adversary leverages the known target model, f, classifier parameters, $\boldsymbol{\theta}$, the training dataset, and specific domain knowledge, such as one of the six scenarios outlined in Section II.

The primary objective of this paper is to explore the impacts of different channel models and the assumptions about the adversary's available knowledge. Instead of introducing a new method of attack, we utilize the Maximum Received Perturbation Power (MRPP) attack from [18] designed to account for the effects of various channel conditions. Moreover, we examine a *no-channel* attack where $\mathbf{H_{ar}} = \mathbf{I}$, signifying the absence of a wireless channel between the attacker and receiver. Although not "channel-aware", this scenario, which is first proposed in [17], will serve as a basis for designing perturbations in the presence of a wireless channel and is expected to represent an upper performance bound for the

[1]It should be noted that the standard path loss formula [20] calculates the power loss at the receiver relative to the transmit power, necessitating the square of the channel matrix's Euclidean norm for computation.

attacker. Lastly, we consider a *with-channel* attack, where the adversary deploys the *no-channel* attack as-is, deliberately neglecting the impact of any wireless channel. This serves as a worst-case scenario, providing a lower bound on the efficacy of channel-aware attacks. Fig. 1 illustrates these three cases. In the following, we define these attacks in more detail.

A. Fast Gradient Method

Consider a DNN classifier model, $f(\mathbf{y};\boldsymbol{\theta})$, with model input, $\mathbf{y}$. The adversary adds a perturbation, $\boldsymbol{\delta}$, to the model input so that the input becomes $\boldsymbol{y}_{\text{adv}} = \boldsymbol{y}+\boldsymbol{\delta}$. Denoting $\boldsymbol{l}_{\text{true}}$ to be the true class label of $\boldsymbol{y}$, the Fast Gradient Method (FGM) linearizes the loss function, $L(\boldsymbol{\theta}, \boldsymbol{y}_{\text{adv}}, \boldsymbol{l}_{\text{true}})$, using

$$L(\boldsymbol{\theta}, \boldsymbol{y}_{\text{adv}}, \boldsymbol{l}_{\text{true}}) \approx L(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{l}_{\text{true}}) + \boldsymbol{\delta}^T \nabla_x L(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{l}_{\text{true}}) \quad (4)$$

which is maximized by setting $\boldsymbol{\delta}^* = \epsilon \nabla_x L(\boldsymbol{\theta}, \mathbf{y}, \mathbf{y}_{true})$, where ϵ is the scaling factor to satisfy the power constraint p_{max} of the adversary.

B. No-channel Attacks

No-channel attack assumes the channel between the attacker and the receiver can be represented with $\mathbf{H}_{ar} = \mathbf{I}$, i.e., there are no fading or path-loss components. Since no channel is present, this attack is the same for both channel knowledge assumptions and all three channel models we consider.

Reference [17] designs a universal adversarial perturbation (UAP) on a sample set of N_s signal vectors from the training dataset. Let arbitrarily collected subset of training data be $\{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(N_s)}\}$, and their associated labels be $\{\mathbf{l}^{(1)}_{true}, \ldots, \mathbf{l}^{(N_s)}_{true}\}$. FGM is employed for each vector in the sample set, and adversarial perturbations, $\boldsymbol{\delta}^{(j)}_{noch}$, are crafted corresponding to each clean input instance $\mathbf{y}^{(j)}$.

$$\boldsymbol{\delta}^{(j)}_{noch} = \sqrt{p_{max}} \frac{\nabla_x L(\boldsymbol{\theta}, \mathbf{y}^{(j)}, \mathbf{l}^{(j)}_{true})}{\|\nabla_x L(\boldsymbol{\theta}, \mathbf{y}^{(j)}, \mathbf{l}^{(j)}_{true})\|_2}, \quad j = 1, \ldots, N_s \quad (5)$$

To obtain a UAP that reflects the common characteristics of N_s perturbations, we apply the principal component analysis (PCA) to N_s perturbations, $\boldsymbol{\delta}^{(j)}_{noch}$, given in (5). The first principal component is the desired perturbation vector, which is denoted as $\boldsymbol{\delta}_{noch}$.

C. With-channel Attacks

With-channel attack assumes a non-identity channel matrix, $\mathbf{H}_{ar}$, but does not account for the influence of the communication channel between the adversary and receiver. This attack uses the perturbation vector, $\boldsymbol{\delta}_{noch}$, of *no-channel* attack. However, the received perturbation at the target classifier becomes $\boldsymbol{\delta}_{rx} = \mathbf{H}_{ar}\boldsymbol{\delta}_{noch}$ resulting in a significantly different perturbation at the model input.

D. Channel-aware Attacks

The goal of channel-aware attacks is to craft a transmit perturbation, $\boldsymbol{\delta}_{tx}$, at the attacker so that the received perturbation at the classifier input is as close as possible to $\boldsymbol{\delta}_{noch}$ after traversing through the channel, i.e., $\boldsymbol{\delta}_{noch} \approx \mathbf{H}_{ar}\boldsymbol{\delta}_{tx}$. One such example of crafting perturbations that account for channel effects is the *MRPP attack* of [18]. MRPP attack utilizes the adversary-receiver channel and ensures that the direction and power of transmitted perturbations remain minimally affected. In an MRPP attack, the main idea is to invert the channel using the pseudo-inverse of the channel matrix. Since the amount of knowledge about the channel influences the design of the MRPP attack, the MRPP perturbation signals differ depending on whether instantaneous or statistical knowledge is available.

When the instantaneous channel between the attacker and receiver is available, the attacker designs the transmit perturbation as a function of the instantaneous channel and the no-channel perturbation using

$$\boldsymbol{\delta}^{inst}_{mrpp} = \frac{\mathbf{H}^H_{ar}\|\nabla_x L(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{l}_{true})\|_2}{\|\mathbf{H}^H_{ar}\nabla_x L(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{l}_{true})\|_2}\boldsymbol{\delta}_{noch} \quad (6)$$

After (6) goes through the channel, the received perturbation at the input of the target classification model becomes $\mathbf{H}_{ar}\boldsymbol{\delta}^{inst}_{mrpp}$ which is a close approximation of $\boldsymbol{\delta}_{noch}$.

Note that for instantaneous channel knowledge available to the adversary, we consider the adversary-receiver channel models mentioned earlier: (a) Rayleigh Fading ($\mathbf{H}_{ar} = \mathbf{H}_{ray}$), (b) Rayleigh Fading with lognormal shadowing ($\mathbf{H}_{ar} = \sqrt{\psi}\mathbf{H}_{ray}$), and (c) Rayleigh Fading with lognormal shadowing and path loss ($\mathbf{H}_{ar} = \sqrt{K(\frac{d}{d_0})^{\gamma}\psi}\mathbf{H}_{ray}$). Note that the path loss parameters K, d_0 and γ are assumed to be constants.

When only the distribution of the channel between the adversary and receiver is available, the attacker utilizes a sample set of channel realizations generated by known distributions. Following the same reasoning used for UAP design in Section III-B, the attacker generates N_s realizations of the channel $\{\mathbf{H}^{(1)}_{ar}, \ldots, \mathbf{H}^{(N_s)}_{ar}\}$. Each $\mathbf{H}^{(j)}_{ar}$ is used with perturbations $\boldsymbol{\delta}^{(j)}_{noch}$ in (5) to craft

$$\boldsymbol{\delta}^{(j)}_{mrpp} = \frac{{\mathbf{H}^{(j)}_{ar}}^H\|\nabla_x L(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{l}_{true})\|_2}{\|{\mathbf{H}^{(j)}_{ar}}^H\nabla_x L(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{l}_{true})\|_2}\boldsymbol{\delta}^{(j)}_{noch}, \quad j = 1, \ldots, N_s \quad (7)$$

Finally, PCA is applied to (7) to obtain $\boldsymbol{\delta}^{stat}_{mrpp}$. After going through the channel, the received perturbation at the input of the target classification model becomes $\mathbf{H}_{ar}\boldsymbol{\delta}^{stat}_{mrpp}$. Note that the received perturbation, when statistical channel knowledge is used, is not as close an approximation to $\boldsymbol{\delta}_{noch}$ as is the received perturbation when instantaneous channel knowledge is available.

IV. Results

We conduct a range of experiments that aim to understand the importance of domain knowledge in crafting effective perturbations. We begin by investigating the relationship between the amount of channel knowledge available and the efficacy of the perturbation received at the target classifier. We study the influence of perturbations that traverse three distinct statistical channel models: (i) Rayleigh fading, (ii) Rayleigh fading with log-normal shadowing, and (iii) Rayleigh fading with shadowing and path loss. We particularly analyze the effect of path loss parameters, i.e., path loss exponents (γ)

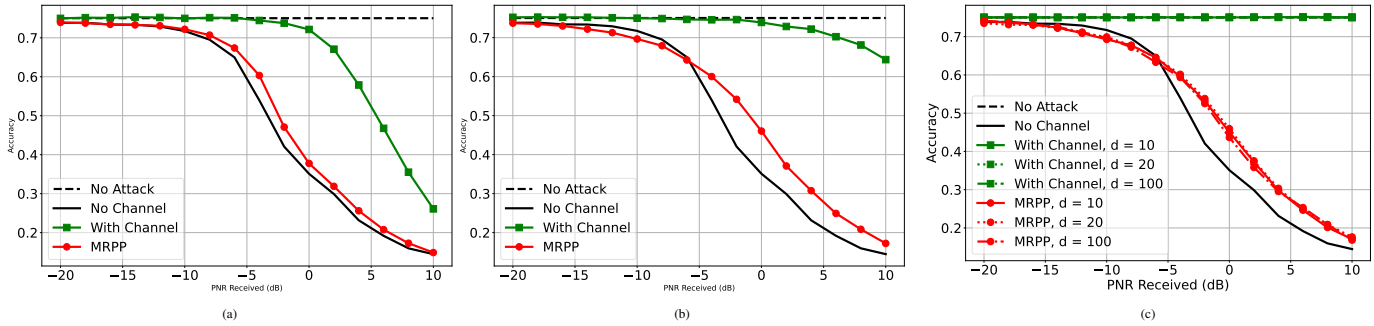

Fig. 2: The performance of adversarial attacks under instantaneous channel knowledge assumption: (a) Rayleigh fading only; (b) Rayleigh fading with shadowing; (b) Rayleigh fading with shadowing and path loss (variation in adversary-receiver distance).

and adversary-receiver distance (d) values, on the transmission power requirements and perturbation efficiency. We use $N_s = 40$ and set the channel parameters as $K = 1, d_0 = 1, \psi \sim$ Lognormal$(0, 8)$. Rayleigh channel consists of uncorrelated in-phase and quadrature components, which are Gaussian random variables ($\sim \mathcal{N}(0, 0.5)$).

A. Dataset, Receiver Model, and Evaluation Metrics

We employ GNU radio ML dataset RML2016.10a [21] where the channel between the legitimate transmitter and receiver, $\mathbf{H}_{tr}$, follows the GNU Radio Dynamic Channel Model hierarchical block [21]. This dataset contains $220,000$ complex-valued data points, each corresponding to a 128-dimensional in-phase/quadrature sample of the received waveform, $\mathbf{y}$, of a specific modulation constellation. In total, there are 11 modulation constellations: BPSK, QPSK, 8PSK, QAM16, QAM64, CPFSK, GFSK, PAM4, WBFM, AM-SSB, and AM-DSB. Each modulation type contains waveforms with a signal-to-noise ratio (SNR) ranging from -20 dB to 18 dB with a step size of 2 dB. We restrict our analysis to samples with SNR of 10 dB for our experiments. The target classifier at the legitimate receiver is the VTCNN2 classifier [22].

To measure the effectiveness of received adversarial perturbations, we adopt perturbation-to-noise ratio (PNR) as a metric [17], [18], which is given by:

$$\text{PNR (dB)} = \frac{\text{Received Perturbation Power } (\mathrm{P_{rx}})}{\text{Noise Power } (\mathrm{P_n})} \text{ (dB)} \quad (8)$$

Increasing the PNR introduces a higher power adversarial perturbation at the target receiver. A high power perturbation reduces the accuracy of the target model while concurrently distorting within the underlying signal ($\mathbf{y}$).

B. Robustness Against Available Domain Knowledge

In our numerical results, we report the accuracy of the target modulation classifier at the legitimate receiver. As a baseline, we use the attack in [17], which is designed for $\mathbf{H}_{ar} = \mathbf{I}$ using FGM (*no-channel* attack). For various channel models, we report the performances of the baseline attack from [17], the *with-channel* attack, and the "channel-aware" MRPP attack from [18]. For the channel models with path loss, we consider three path-loss exponents ($\gamma = \{1.7, 2.7, 4\}$) and three distances ($d = \{10, 20, 100\}$).

1) Instantaneous Channel Knowledge: We start our analysis with the case where instantaneous channel knowledge is available to the adversary. Fig. 2 depicts the accuracy of the target classifier under different channel model assumptions when the instantaneous channel is available to the attacker. Let us first focus on the MRPP attack. We observe from Fig. 2(a) through Fig. 2(c) that adversarial perturbations crafted with the exact channel information perform similarly across different channel models. Compared to the Rayleigh fading only case in Fig. 2(a), the performance of the MRPP attack under the combination of Rayleigh fading shadowing in Fig. 2(b) is better for low PNRs (< -5 dB) but worse in higher PNRs. On the other hand, when comparing Figs. 2(a) and (b) to Fig. 2(c), we see that adding a path loss component to the channel model has almost no effect on the MRPP attack when the channel is instantaneously known. This robustness to path loss is expected because the known instantaneous effects of the channel can be inverted. In Fig. 2(c), we report only the effect of distance. The effect of the path-loss exponent is very similar and therefore is omitted.

Unlike the MRPP attack, the performance of the with-channel attack depends on path loss parameters. The channel weakens the transmitted perturbation as the distance between the attacker and receiver increases. The with-channel attack uses the same transmit power for all channel realizations, resulting in very low received perturbation power. On the other hand, the MRPP attack overcomes this limitation by increasing its transmit power, ensuring that the received perturbation power remains the same, regardless of the distance.

2) Statistical Channel Knowledge: Next, we move to the case where statistical channel knowledge is available at the receiver. Similar to Fig. 2 in the instantaneous channel case, Fig. 3 depicts the accuracy of the target classifier under different channel model assumptions when statistical channel knowledge is available to the attacker. Recall from Section II that statistical knowledge refers to the parameters of the channel distributions. Under the statistical channel knowledge assumption, the adversary does not know about the deterministic path loss parameters. For our experiments, we assume

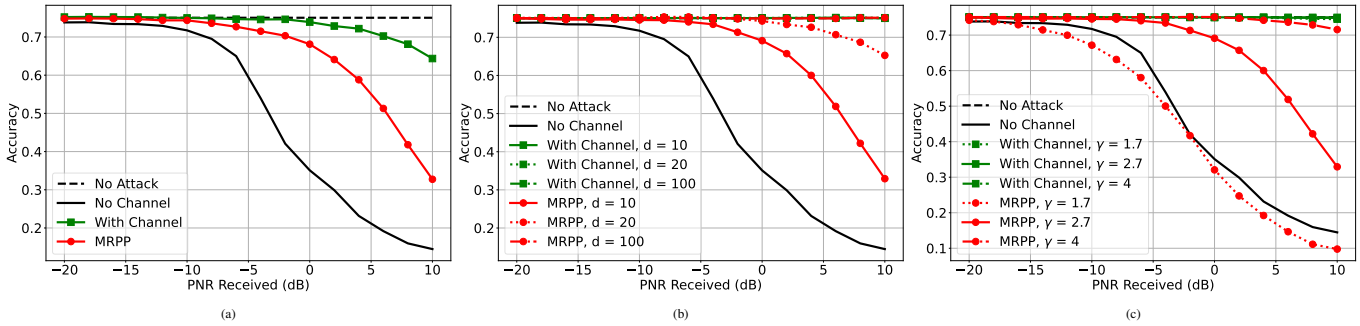

Fig. 3: The performance of adversarial attacks under statistical channel knowledge assumption: (a) Rayleigh fading with shadowing; (b) Rayleigh fading with shadowing and path loss (variation in adversary-receiver distance); (c) Rayleigh fading with shadowing and path loss (variation in path loss exponent).

that the adversary has estimated the distance and path-loss exponent as $d = 10$ and $\gamma = 2.7$, respectively. However, it is important to note that the actual distance and path-loss exponent values used in the testing phase vary.

Since the instantaneous channel is not known, the adversary can no longer invert the actual realization of the channel for the MRPP attack. This results in losing the advantage of a *channel-aware* attack. The performance of the MRPP attack (indicated by red lines with circle markers), as seen in Fig. 3(a) and 3(b) is much worse than that in Fig. 2(b) and Fig.2(c), respectively. This contrasting performance underlines the importance of having instantaneous channel knowledge at the adversary, even within the same channel model. While MRPP with instantaneous channel knowledge inverts the channel effects, statistical channel knowledge penalizes the adversary by using the same transmission power across all channel parameter variations.

Finally, we observe from Fig. 3(b) and 3(c) that the existence of path loss drastically changes the efficacy of the MRPP attack. Specifically, for lower PNRs values (PNR ≤ 0 dB), when the path loss parameters γ or d are underestimated ($\gamma = 4, d = \{20, 100\}$), the MRPP attack fails to fool the target modulation classifier.

C. Path Loss Exponent Uncertainty

In this section, we further analyze the importance of the path loss exponent (γ) on the efficacy of the MRPP attack by focusing on PNR $= 0$ dB. For a given γ and for both instantaneous and statistical channel knowledge assumptions, we derive the transmit perturbation signals as explained in Sec.III. Then, we plot both the accuracy of the target modulation classifier and the required transmit perturbation power as functions of γ.

Fig. 4 shows that the MRPP attack designed for instantaneous channel knowledge (indicated by the green solid line with a square marker) remains robust against variations in γ. This robustness holds as long as sufficient transmit power (indicated by the green dashed line with a square marker) is available. Note that the MRPP attack requires a transmit power greater than 20 dB to invert the channel when $\gamma > 4$. On the other hand, the MRPP attack designed for statistical channel knowledge (indicated by the red solid line with a

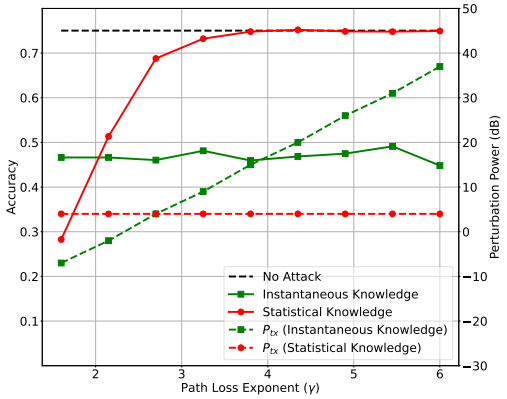

Fig. 4: Robustness of the MRPP attack with respect to γ.

TABLE I: Transmission power (dB) used by the adversary to send perturbations $\boldsymbol{\delta}_{tx}$ for distinct path loss exponent (γ) values.

Path Loss Exponent (γ)	**1.6**	**2.7**	**3.8**	**4.9**	**6**
Instantaneous Channel	-7 dB	4 dB	15 dB	26 dB	37 dB
Statistical Channel	4 dB	4 dB	4 dB	4 dB	4 dB

circle marker) fails to fool the target classifier as γ increases. Note that the transmit power (indicated by the red dashed line with a circle marker) of the MRPP attack remains constant with respect to γ since the instantaneous channel value is not available. Interestingly, when the actual exponent is small, $\gamma \leq 2.2$, the MRPP attack designed for statistical channel knowledge overestimates the required transmission power, making the attack successful but vulnerable to detection.

Table I reports the transmission powers for a few examples of path loss exponent values. We conclude two main points. First, for an adversary with statistical channel knowledge, the accuracy in estimating path loss parameter γ significantly impacts the efficacy of MRPP attacks. Second, the crafted perturbations under these conditions are not robust against variation in path loss exponent.

D. Adversary-Receiver Distance Uncertainty

Lastly, we study the influence of adversary-receiver distance (d) on the success of MRPP attacks for a received PNR of 0 dB. Similar to Section IV-C, we present the accuracy

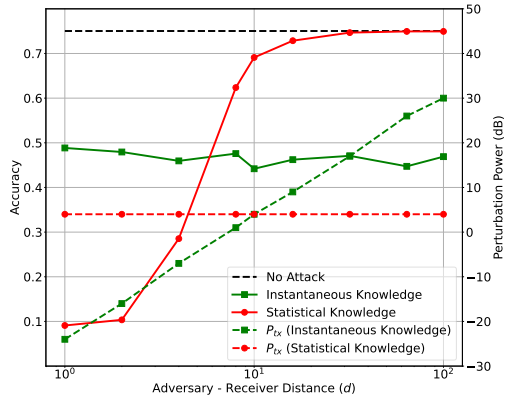

Fig. 5: Robustness of the MRPP attack with respect to d.

TABLE II: Transmission power (dB) used by the adversary to send perturbations $\boldsymbol{\delta}_{tx}$ for distinct adversary-receiver distance variations (d) values.

Distance (d)	**1**	**4**	**10**	**64**	**100**
Instantaneous Channel	-24 dB	-7 dB	4 dB	26 dB	30 dB
Statistical Channel	4 dB	4 dB	4 dB	4 dB	4 dB

of the target modulation classifier and the required transmit perturbation power as functions of d. Fig. 5 demonstrates that the performance of the MRPP attack varies with distance in a manner analogous to its variation with the path-loss exponent. Table II reports the transmission perturbation powers for a few distance values. We find that the MRPP attack is effective when the attacker has access to instantaneous knowledge. In contrast, the MRPP attack is not robust when only statistical channel knowledge is available. An overestimation of the path loss parameter, d, can enhance the effectiveness of MRPP attacks when $d < 10$. However, this overestimation also compromises the covertness of such attacks.

V. Conclusion

This study has undertaken an in-depth analysis of the influence of domain knowledge, specifically channel knowledge, on the effectiveness of adversarial perturbations targeting wireless communication systems. We conducted experiments to investigate the role of various channel models that incorporate Rayleigh fading, shadowing, and path loss. Our findings highlight the varying performance of adversarial attacks depending on the type of channel knowledge available to the adversary.

When instantaneous channel knowledge is available, MRPP attacks proved to be robust across different channel models and parameters, effectively inverting the channel effects. In contrast, these attacks were less effective under the constraint of statistical channel knowledge, particularly when path loss parameters were not accurately estimated. We demonstrated that the path-loss exponent (γ) and the adversary-receiver distance (d) significantly impact the efficacy of MRPP attacks, especially under statistical channel knowledge conditions.

Our analysis indicates the critical importance of channel knowledge in devising effective and efficient adversarial perturbations. This study has implications for the design of more resilient wireless communication systems, illuminating the need for countermeasures that take into account various channel conditions.

Acknowledgment

The authors acknowledge Advanced Research Computing at Virginia Tech for providing computational resources and technical support that have contributed to the results reported within this paper. URL: https://arc.vt.edu/

References

[1] H. Ye, G. Y. Li, and B.-H. Juang, "Power of deep learning for channel estimation and signal detection in OFDM systems," *IEEE Wireless Commun. Letters*, vol. 7, no. 1, pp. 114–117, 2017.

[2] T. Erpek, T. J. O'Shea, Y. E. Sagduyu, Y. Shi, and T. C. Clancy, "Deep learning for wireless communications," *Development and Analysis of Deep Learning Architectures*, pp. 223–266, 2020.

[3] T. J. O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in *EANN*, 2016.

[4] T. J. O'Shea, T. Roy, and T. C. Clancy, "Over-the-air deep learning based radio signal classification," *IEEE J. of Selected Topics in Signal Processing*, vol. 12, no. 1, pp. 168–179, 2018.

[5] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *ECML PKDD*, Sep. 2013.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *CoRR*, 2014.

[7] A. Bahramali, M. Nasr, A. Houmansadr, D. Goeckel, and D. Towsley, "Robust adversarial attacks against DNN-based wireless communication systems," in *SIGSAC*, 2021.

[8] Y. Shi, Y. E. Sagduyu, T. Erpek, and M. C. Gursoy, "How to attack and defend Next G radio access network slicing with reinforcement learning," *IEEE Open J. of Vehicular Technology*, vol. 4, pp. 181–192, 2022.

[9] Y. Shi and Y. E. Sagduyu, "Adversarial machine learning for flooding attacks on 5G radio access network slicing," in *ICC Wkshps*, 2021.

[10] Y. Siriwardhana, P. Porambage, M. Liyanage, and M. Ylianttila, "AI and 6G security: Opportunities and Challenges," in *EuCNC Summit*, 2021.

[11] Z. Luo, S. Zhao, Z. Lu, J. Xu, and Y. E. Sagduyu, "When attackers meet AI: Learning-empowered attacks in cooperative spectrum sensing," *IEEE Trans. on Mobile Computing*, vol. 21, no. 5, pp. 1892–1908, 2020.

[12] T. Erpek, Y. E. Sagduyu, and Y. Shi, "Deep learning for launching and mitigating wireless jamming attacks," *IEEE Trans. on Cognitive Communications and Networking*, vol. 5, no. 1, pp. 2–14, 2018.

[13] B. Flowers, R. M. Buehrer, and W. C. Headley, "Evaluating adversarial evasion attacks in the context of wireless communications," *IEEE Trans. on Inf. Forensics and Security*, vol. 15, pp. 1102–1113, 2019.

[14] Y. Lin, H. Zhao, Y. Tu, S. Mao, and Z. Dou, "Threats of Adversarial Attacks in DNN-based modulation recognition," in *IEEE INFOCOM*, 2020.

[15] M. Z. Hameed, A. György, and D. Gündüz, "The best defense is a good offense: Adversarial attacks to avoid modulation detection," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1074–1087, 2021.

[16] B. Kim, Y. Sagduyu, T. Erpek, and S. Ulukus, "Adversarial attacks on deep learning based mmWave beam prediction in 5G and beyond," in *IEEE Statistical Signal Processing Workshop*, 2021.

[17] M. Sadeghi and E. G. Larsson, "Adversarial attacks on deep-learning based radio signal classification," *IEEE Wireless Commun. Letters*, 2018.

[18] B. Kim, Y. E. Sagduyu, K. Davaslioglu, T. Erpek, and S. Ulukus, "Channel-aware adversarial attacks against deep learning-based wireless signal classifiers," *IEEE Trans. on Wireless Commun.*, vol. 21, no. 6, pp. 3868–3880, 2021.

[19] B. Kim, Y. E. Sagduyu, T. Erpek, K. Davaslioglu, and S. Ulukus, "Adversarial attacks with multiple antennas against deep learning-based modulation classifiers," in *IEEE Globecom Wkshps*, 2020.

[20] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

[21] T. J. O'Shea and N. West, "Radio machine learning dataset generation with GNU radio," in *Proceedings of the GNU Radio Conference*, 2016.

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