ADVERSARIAL ATTACKS ON DEEP LEARNING BASED MMWAVE BEAM PREDICTION IN 5G AND BEYOND

B. Kim[†], Y. Sagduyu[∗], T. Erpek[∗] and S. Ulukus[†]

[†] Dept. of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA
[∗]Intelligent Automation, Inc., Rockville, MD, USA

ABSTRACT

Deep learning provides powerful means to learn from spectrum data and solve complex tasks in 5G and beyond such as beam selection for initial access (IA) in mmWave communications. To establish the IA between the base station (e.g., gNodeB) and user equipment (UE) for directional transmissions, a deep neural network (DNN) can predict the beam that is best slanted to each UE by using the received signal strengths (RSSs) from a subset of possible narrow beams. While improving the latency and reliability of beam selection compared to the conventional IA that sweeps all beams, the DNN itself is susceptible to adversarial attacks. We present an adversarial attack by generating adversarial perturbations to manipulate the over-the-air captured RSSs as the input to the DNN. This attack reduces the IA performance significantly and fools the DNN into choosing the beams with small RSSs compared to jamming attacks with Gaussian or uniform noise.

1. INTRODUCTION

Due to the algorithmic and computational advances in deep learning (DL), applications from different domains, such as computer vision and speech recognition, have been empowered by deep neural networks (DNNs) in solving complex problems by effectively learning from rich data representations. Similarly, DL has been applied for various wireless communication tasks, such as waveform design, signal classification, spectrum sensing, and interference management [1] by capturing the inherent complex characteristics of the spectrum data.

One particular application of DL is in the domain of initial access (IA), where user equipments (UEs) need to establish their initial connection to an access point or base station when they attempt to join the communication network for the first time [2]. As 5G and beyond communication systems rely on millimeter wave (mmWave) and higher frequency bands to sustain high data rates over large available bandwidths, transmissions become more directional using narrow beams. This paradigm makes the beam alignment in the IA process more difficult as many narrow beams need to be swept to find the most suitable beam for each UE [2].

In the IA, the transmitter such as the 5G base station (gNodeB) sequentially transmits pilot signals over different narrow beams. The UE calculates the received signal strength (RSS) for each beam, determines the beam that provides the highest RSS, and informs the gNodeB of this beam selection. Since the time for the IA is limited, it is essential to reduce the number of beams swept as checking each beam consumes time and delays the UE to gain access to time-sensitive services such as ultra-reliable low-latency communications (URLLC) in 5G. To overcome this issue, a DL-based approach has been proposed in [3, 4] to reduce the number of beams that need to be swept before making the IA decision, compared to the conventional beam sweeping approach that needs to sweep all possible beams. In this DL-based approach, the UE predicts the best beam from a large set of narrow beams by using only the RSSs for a subset of these possible beams.

DNNs are highly susceptible to carefully crafted adversarial perturbations that induce incorrect output or misclassification, as first observed in computer vision applications [5]. Wireless medium is shared and open to jamming attacks that can also be launched via adversarial machine learning (AML) to target the underlying DNNs. There are various attacks built upon AML including exploratory (inference) attacks [6, 7], adversarial (evasion) attacks [8–17], poisoning (causative) attacks [18, 19], membership inference attacks [20], Trojan attacks [21], and spoofing attacks [22]. These attacks are harder to detect than conventional jamming attacks [23, 24].

In this paper, we consider an adversarial attack that aims to manipulate the input to a DNN in test time. Most of the applications of adversarial attacks to the wireless domain have focused on wireless signal classifiers such as modulation classifiers [8–16] and spectrum sensing classifiers [17–19]. There have been efforts to extend adversarial attacks to

communication problems such as autoencoder-based end-to-end communications [25] and power control for MIMO communications [26].

Our goal in this paper is to investigate the vulnerability of a DNN that is used for mmWave beam prediction as part of the IA process in 5G and beyond communications. We first generate a non-targeted attack using the fast gradient method (FGM) [27] to cause any misclassification at the DNN classifier at the receiver (independent of the wrong labels for beams). Then, we introduce the k-worst beam attack that not only causes a misclassification but also enforces the DNN classifier to select one of the k worst beams to further reduce the IA performance. We compare the non-targeted FGM attack and k-worst beam attack with benchmark attacks with Gaussian and uniform noise added across RSSs from input beams. We show that the beam prediction of the IA process is highly vulnerable to adversarial attacks that can significantly reduce the beam prediction accuracy. The effect of this attack translates to notable reduction in communication performance in terms of data rate since the UEs connect to the gNodeB using beams that are not well aligned and the corresponding RSSs drop significantly.

The rest of the paper is organized as follows. Section 2 provides the system model. Section 3 describes the adversarial attacks considered in this paper, namely, the non-targeted attack and the k-worst attack. Section 4 presents the performance results. Section 5 concludes the paper.

2. SYSTEM MODEL

We consider a mmWave network that consists of a directional transmitter (e.g., the gNodeB in 5G), an omnidirectional receiver (e.g., a UE) during the IA, and an adversary. A pre-trained DL-based classifier is applied to the RSS values measured at the receiver to select the best beam without a need to sweep all beams, as proposed in [3, 4]. Concurrently, the adversary attempts to cause misclassification at the receiver by launching an over-the-air adversarial attack.

In the conventional IA process with exhaustive beam sweeping, the transmitter transmits pilot signals over all narrow beams and the receiver uses the RSS values from all $\mathcal{N}$ possible beams where $N = |\mathcal{N}|$ to select the best beam. The transmitter transmits the pilot signal over each beam at a separate time slot. DL reduces the number of beams swept such that the DNN classifier at the receiver only uses the RSS values from $\mathcal{M}$ subset of beams, where $\mathcal{M} \subseteq \mathcal{N}$ and $M = |\mathcal{M}|$, to select the best beam. For that purpose, the transmitter transmits the pilot signals over a smaller set of beams, each at a separate time slot. Overall, the total time needed for the IA is reduced. In the DL-based approach, the DNN classifier at the receiver is denoted by $f(.; \boldsymbol{\theta}) : \mathcal{X}_M \to \mathbb{R}^N$, where $\mathcal{X}_M \subset \mathbb{R}^M$ is the input to the DNN which corresponds to the RSS values of M beams and

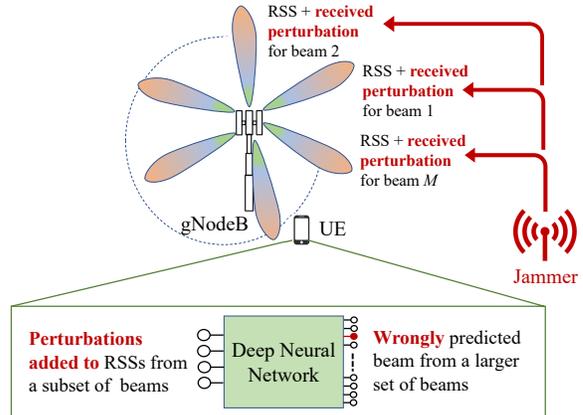**Fig. 1**. Adversarial attack on DL-based beam prediction.

$\boldsymbol{\theta}$ is the set of the receiver's DNN parameters. The classifier f assigns the best beam $\hat{l}(\boldsymbol{x}_M, \boldsymbol{\theta}) = \arg\max_j f_j(\boldsymbol{x}_M, \boldsymbol{\theta})$ to every input $\boldsymbol{x}_M \in \mathcal{X}_M$. In this formulation, $f_j(\boldsymbol{x}_M, \boldsymbol{\theta})$ is the output of classifier f corresponding to the jth beam.

As shown in Fig. 1, the adversary generates an adversarial attack by jamming the spectrum with adversarial perturbation, $\boldsymbol{\delta} = [\delta_1, \delta_2, \cdots, \delta_M]$, under some suitable power constraint P_{max} with respect to the original RSS values, $\boldsymbol{x} \in \mathbb{R}^N$. The adversarial perturbation $\boldsymbol{\delta}$ is determined by solving the following optimization problem:

$$\arg\min_{\boldsymbol{\delta}} \quad \sum_{i=1}^{M} \delta_i$$

$$s.t. \quad \hat{l}(\boldsymbol{x}_M, \boldsymbol{\theta}) \neq \hat{l}(\boldsymbol{x}_M + \boldsymbol{\delta}, \boldsymbol{\theta}), \quad \sum_{i=1}^{M} \delta_i \leq P_{max}. \quad (1)$$

Note that $\boldsymbol{x}_M \in \mathbb{R}^M$ is the M elements from $\boldsymbol{x}$ which correspond to the RSS values at the receiver for all beams.

Solving (1) is hard due to the inherent structure of the DNN. Thus, different methods have been proposed to approximate the adversarial attack such as FGM. FGM is a computationally efficient way of creating adversarial attacks by linearizing the loss function of the DNN. We denote the loss function of the DNN classifier by $L(\boldsymbol{\delta}, \boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{y} \in \{0, 1\}^N$ is the one-hot encoded class vector. Then, FGM linearizes this loss function in a neighborhood of $\boldsymbol{x}$ and uses this linearized function for optimization.

We assume that the adversary knows the architecture ($\boldsymbol{\theta}$ and $L(\cdot)$) of the DNN classifier at the receiver, as well as the exact RSS values for all N beams that are received at the receiver. Later in the paper, we relax the assumption about knowing all N RSS values to knowing only M RSS values, and compare the results under these two assumptions. In the next section, we describe the two adversarial attacks considered in this paper, namely, the non-targeted attack and the k-worst attack.

Algorithm 1: Non-targeted FGM attack

Inputs: RSS values $\boldsymbol{x}_M$, true label $\boldsymbol{y}^{true}$, desired accuracy ε_{acc}, power constraint P_{max}, and model of the classifier

Initialize: $\varepsilon \leftarrow 0, \varepsilon_{max} \leftarrow P_{max}, \varepsilon_{min} \leftarrow 0$

$\boldsymbol{\delta}_{norm} = \frac{\nabla_{\boldsymbol{x}_M} L(\boldsymbol{\theta}, \boldsymbol{x}_M, \boldsymbol{y}^{true})}{(||\nabla_{\boldsymbol{x}_M} L(\boldsymbol{\theta}, \boldsymbol{x}_M, \boldsymbol{y}^{true})||_1)}$

if $\hat{l}(\boldsymbol{x}_M, \boldsymbol{\theta}) == \boldsymbol{y}^{true}$ **then**
   while $\varepsilon_{max} - \varepsilon_{min} > \varepsilon_{acc}$ **do**
      $\varepsilon_{avg} \leftarrow (\varepsilon_{max} + \varepsilon_{min})/2$
      $\boldsymbol{x}_{adv} \leftarrow \boldsymbol{x}_M + \varepsilon_{avg}\boldsymbol{\delta}_{norm}$
      if $\hat{l}_m(\boldsymbol{x}_{adv}) == \boldsymbol{y}^{true}$ **then** $\varepsilon_{min} \leftarrow \varepsilon_{avg}$
      else $\varepsilon_{max} \leftarrow \varepsilon_{avg}$
   end
end

$\varepsilon = \varepsilon_{max}, \boldsymbol{\delta}^* = \varepsilon\boldsymbol{\delta}_{norm}$

Algorithm 2: k-worst beam attack

Inputs: RSS values $\boldsymbol{x}_M$, true label $\boldsymbol{y}^{true}$, k worst beam indices, desired accuracy ε_{acc}, power constraint P_{max}, and model of the classifier

if $\hat{l}(\boldsymbol{x}_M, \boldsymbol{\theta}) == \boldsymbol{y}^{true}$ **then**
   for i in k worst beam indices **do**
      Initialize: $\varepsilon \leftarrow 0, \varepsilon_{max} \leftarrow P_{max}, \varepsilon_{min} \leftarrow 0, \boldsymbol{y}^{target} \leftarrow i$
      $\boldsymbol{\delta}_{norm} = \frac{\nabla_{\boldsymbol{x}_M} L(\boldsymbol{\theta}, \boldsymbol{x}_M, \boldsymbol{y}^{target})}{(||\nabla_{\boldsymbol{x}_M} L(\boldsymbol{\theta}, \boldsymbol{x}_M, \boldsymbol{y}^{target})||_1)}$
      while $\varepsilon_{max} - \varepsilon_{min} > \varepsilon_{acc}$ **do**
         $\varepsilon_{avg} \leftarrow (\varepsilon_{max} + \varepsilon_{min})/2$
         $\boldsymbol{x}_{adv} \leftarrow \boldsymbol{x}_M - \varepsilon_{avg}\boldsymbol{\delta}_{norm}$
         if $\hat{l}_m(\boldsymbol{x}_{adv}) == \boldsymbol{y}^{true}$ **then**
            $\varepsilon_{min} \leftarrow \varepsilon_{avg}$
         else $\varepsilon_{max} \leftarrow \varepsilon_{avg}$
      end
      $\boldsymbol{x}_{adv} \leftarrow \boldsymbol{x}_M - \varepsilon_{avg}\boldsymbol{\delta}_{norm}$
      if $\hat{l}_m(\boldsymbol{x}_{adv}) == i$ **then** break
   end
end

$\boldsymbol{\delta}^* = -\varepsilon_{avg}\boldsymbol{\delta}_{norm}$

---

3. ADVERSARIAL ATTACKS ON BEAM PREDICTION

3.1. Non-targeted FGM Attack

First, we consider a non-targeted FGM attack where the adversary searches for a perturbation that causes any misclassification at the receiver's DNN classifier. The adversary designs a perturbation that maximizes the loss function $L(\boldsymbol{\delta}, \boldsymbol{x}_M, \boldsymbol{y}^{true})$, where $\boldsymbol{y}^{true}$ is the true label of $\boldsymbol{x}_M$. FGM is used to linearize the loss function as $L(\boldsymbol{\theta}, \boldsymbol{x}_M + \boldsymbol{\delta}, \boldsymbol{y}^{true}) \approx L(\boldsymbol{\theta}, \boldsymbol{x}_M, \boldsymbol{y}^{true}) + \boldsymbol{\delta}^T \nabla_{\boldsymbol{x}_M} L(\boldsymbol{\theta}, \boldsymbol{x}_M, \boldsymbol{y}^{true})$ that is maximized by setting $\boldsymbol{\delta} = \alpha \nabla_{\boldsymbol{x}_M} L(\boldsymbol{\theta}, \boldsymbol{x}_M, \boldsymbol{y}^{true})$, where α is a scaling factor to constrain the adversarial perturbation power to P_{max}. The details of selecting α and generating the non-targeted FGM attack are presented in Algorithm 1.

3.2. k-worst Beam Attack

Next, we present an adversarial attack where the adversary not only causes a misclassification at the receiver's DNN classifier but also tries to change the beam to one of the worst k beams. Unlike attacks on some signal classification tasks (such as in the computer vision or wireless domains), where changing the label from 'signal 1' to 'signal 2' and from 'signal 1' to 'signal 3' may have the same effect on the classifier meaning that both correspond to an error in classification, changing the label from the best beam to second worst beam and from the best beam to the worst beam have totally different effects on the communication performance (as the signal quality on a beam strongly affects the achieved rate following the IA). Thus, the adversary first tries to change the label to the worst beam and if the adversary is not able to make this change, the adversary tries to change the label to the second worst beam. The adversary continues to do so until it tries for the kth worst beam.

Here, we consider the targeted FGM attack where the adversary aims to fool the DNN to a target label and tries to minimize the loss function $L(\boldsymbol{\delta}, \boldsymbol{x}_M, \boldsymbol{y}^{target})$, where $\boldsymbol{y}^{target}$ is one of the k worst beams. FGM is used to linearize the loss function as $L(\boldsymbol{\theta}, \boldsymbol{x}_M + \boldsymbol{\delta}, \boldsymbol{y}^{target}) \approx L(\boldsymbol{\theta}, \boldsymbol{x}_M, \boldsymbol{y}^{target}) + \boldsymbol{\delta}^T \nabla_{\boldsymbol{x}_M} L(\boldsymbol{\theta}, \boldsymbol{x}_M, \boldsymbol{y}^{target})$ that is minimized by setting $\boldsymbol{\delta} = -\alpha \nabla_{\boldsymbol{x}_M} L(\boldsymbol{\theta}, \boldsymbol{x}_M, \boldsymbol{y}^{target})$, where α is a scaling factor to constrain the adversarial perturbation power to P_{max}. First, we assume that the adversary knows the order of beam indices based on the real RSS values. Then, we relax this assumption by getting the order from the DNN's output. Since the DNN is only trained to find the best beam, there exists a discrepancy between the real order of beams and the order of beams obtained from the DNN output. The details of the algorithm are presented in Algorithm 2.

4. PERFORMANCE EVALUATION

This section describes the DNN classifier used at the receiver and the dataset that is used to train it, and then compares the performances of the two attacks that are introduced in this paper with two benchmark jamming attacks that inject Gaussian or uniform noise on different beams.

4.1. Deep Learning Framework for IA

We use the DNN structure of seven layers including the input and the output layers where five hidden layers have 32, 64, 126, 64, 32 neurons respectively. The input layer has

$M = 12$ neurons and the output layer is a Softmax layer that has $N = 24$ neurons where each neuron represents a likelihood score for each beam. Each dense layer uses RELU as an activation function and the output is batch normalized before sending it to the next layer. Cross-entropy is used as the loss function. Adam optimizer is used with initial and final learning rates 10^{-3} and 0.1, respectively. We follow the simulation setup of [3] and apply the attacks in a two-dimensional mmWave network where line-of-sight (LoS) mmWave channels with pathloss and shadowing effects are considered. A 10×10 antenna array is used at the transmitter to generate a beam width of about $15°$ using the standard planar array formulation. The location of the transmitter is fixed at $(0, 0)$ and the receiver's x and y positions are each uniformly randomly distributed between -25 m and 25 m. This bounds the simulation cell to an area of 50×50 m. The transmit power is set to 20 dBm. A total of 10^6 receiver positions are generated, which act as data samples. The adversary is also uniformly distributed in this area and its perturbations are reflected on the data samples.

4.2. Attack Performance Results

We compare the two attack schemes that we have described earlier with the Gaussian attack and the uniform attack that generate perturbations with Gaussian and equal power distribution for M beams, respectively. In the simulations, we use the perturbation-to-signal ratio (PSR) metric that shows the relative perturbation power with respect to the received signal power (namely, the RSS at the receiver). As the PSR increases, the attack becomes more likely to be detected.

Fig. 2 presents the accuracy of the classifier at the receiver under the non-targeted FGM attack and compares it with the Gaussian attack and the uniform attack. The non-targeted FGM attack significantly impacts the accuracy of the classifier for beam selection even for signal strength fluctuations that cannot be resolved with typical hardware. The Gaussian attack and the uniform attack do not perform well compared to the non-targeted attack in [-40dB,-20dB] region, while they are comparable in the high PSR region.

In Fig. 3, we investigate the performance of the k-worst beam attack, where k is set as 4, 8 and 12, and compare it with the Gaussian attack and the uniform attack. Note that the accuracy definition in Fig. 3 is different from Fig. 2, where the accuracy is defined as the percentage that the label obtained from the DNN classifier is in the $N - k$ best beams since the attack is successful only if the k-worst beam attack fools the DNN classifier into choosing one of the worst k beams. As k increases for the k-worst beam attack, the DNN classifier accuracy decreases meaning that it is harder to enforce the beam selection to the worst beams. Also, the k-worst beam attack using the real order of RSSs outperforms the k-worst beam attack using the DNN order as there is a discrepancy between the real order and
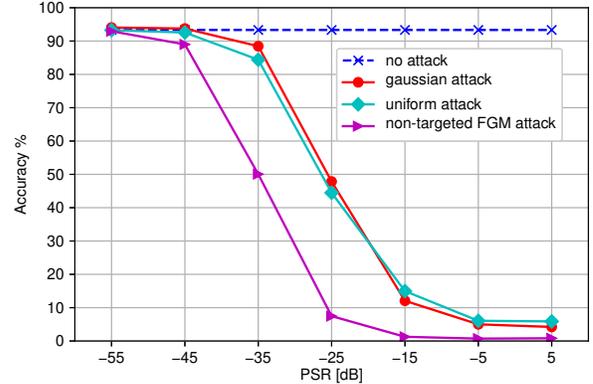

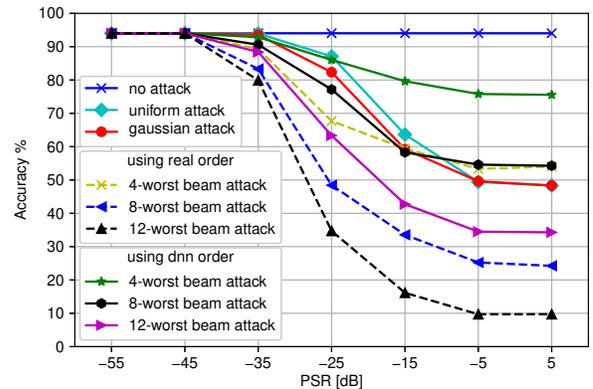**Fig. 2**. Classifier accuracy under the non-targeted attack.

Fig. 3. Classifier accuracy under the k-worst beam attack.

the DNN-predicted order of RSSs. Furthermore, the Gaussian attack and the uniform attack both saturate around 50% meaning that under both attacks the beams are misclassified to the best group or the worst group 50% of the time.

5. CONCLUSION

We presented an adversarial attack to fool the beam selection process of the IA based on a DNN classifier that uses only a subset of beams to predict the beam that is best oriented to the receiver. We investigated two different attacks, namely, the non-targeted FGM attack that only aims to fool the DNN classifier with misclassification to any other beam label, and the k-worst beam attack that not only fools the DNN classifier but also enforces the label that is chosen at the DNN to be in one of the k-worst beams. We showed that the adversarial attack can significantly decrease the accuracy of the DNN and fool the DNN into selecting the worst beam. We conclude that as DL finds applications for beam prediction in mmWave communication for 5G and beyond, the IA becomes vulnerable to adversarial attacks that can significantly reduce the beam selection performance.

6. REFERENCES

[1] T. Erpek, T. O'Shea, Y. E. Sagduyu, Y. Shi, and T. C. Clancy, "Deep learning for wireless communications," in *Development and Analysis of Deep Learning Architectures*. Springer, Cham, 2020.

[2] A. Alkhateeb, Y. H. Nam, M. S. Rahman, J. Zhang, and R. W. Heath, "Initial beam association in millimeter wave cellular systems: Analysis and design insights," *IEEE Trans. on Wireless Commun.*, 2017.

[3] T. S. Cousik, V. K. Shah, T. Erpek, Y. E. Sagduyu, and J. H. Reed, "Deep learning for fast and reliable initial access in AI-Driven 6G mmwave networks," *arXiv preprint arXiv:2101.01847*, 2021.

[4] T. S. Cousik, V. K. Shah, J. H. Reed, T. Erpek, and Y. E. Sagduyu, "Fast initial access with deep learning for beam prediction in 5G mmwave networks," *arXiv preprint arXiv:2006.12653*, 2020.

[5] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv: 1312.6199*, 2013.

[6] Y. Shi, Y. E. Sagduyu, T. Erpek, K. Davaslioglu, Z. Lu, and J. H. Li, "Adversarial deep learning for cognitive radio security: Jamming attack and defense strategies," in *IEEE ICC*, 2018.

[7] T. Erpek, Y. E. Sagduyu, and Y. Shi, "Deep learning for launching and mitigating wireless jamming attacks," *IEEE Trans. on Cogn. Commun. Netw.*, 2019.

[8] M. Sadeghi and E. G. Larsson, "Adversarial attacks on deep-learning based radio signal classification," *IEEE Commun. Lett.*, 2019.

[9] S. Kokalj-Filipovic and R. Miller, "Targeted adversarial examples against RF deep classifiers," in *ACM WiseML*, 2019.

[10] S. Bair, M. Delvecchio, B. Flowers, A. J. Michaels, and W. C. Headley, "On the limitations of targeted adversarial evasion attacks against deep learning enabled modulation recognition," in *ACM WiseML*, 2019.

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

[12] M. Z. Hameed, A. György, and D. Gündüz, "The best defense is a good offense: Adversarial attacks to avoid modulation detection," *IEEE Trans. on Inf. Forensics and Security*, 2021.

[13] 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," in *CISS*, 2020.

[14] ——, "Channel-aware adversarial attacks against deep learning-based wireless signal classifiers," *arXiv preprint arXiv:2005.05321*, 2020.

[15] ——, "How to make 5G communications "invisible" adversarial machine learning for wireless privacy," in *Asilomar Conf. Sig., Sys., and Comp*, 2020.

[16] ——, "Adversarial attacks with multiple antennas against deep learning-based modulation classifiers," in *IEEE Globecom*, 2020.

[17] Y. E. Sagduyu, Y. Shi, and T. Erpek, "IoT network security from the perspective of adversarial deep learning," in *IEEE SECON*, 2019.

[18] Y. Shi, T. Erpek, Y. E. Sagduyu, and J. Li, "Spectrum data poisoning with adversarial deep learning," in *IEEE MILCOM*, 2018.

[19] Y. E. Sagduyu, T. Erpek, and Y. Shi, "Adversarial deep learning for over-the-air spectrum poisoning attacks," *IEEE Trans. Mobile Comput.*, 2019.

[20] Y. Shi, K. Davaslioglu, and Y. E. Sagduyu, "Over-the-air membership inference attacks as privacy threats for deep learning-based wireless signal classifiers," in *ACM WiseML*, 2020.

[21] K. Davaslioglu and Y. E. Sagduyu, "Trojan attacks on wireless signal classification with adversarial machine learning," in *IEEE DySPAN*, 2019.

[22] Y. Shi, K. Davaslioglu, and Y. E. Sagduyu, "Generative adversarial network for wireless signal spoofing," in *ACM WiseML*, 2019.

[23] Y. E. Sagduyu, R. Berry, and A. Ephremides, "Jamming games in wireless networks with incomplete information," *IEEE Communications Magazine*, 2008.

[24] Y. E. Sagduyu, R. A. Berry, and A. Ephremides, "Wireless jamming attacks under dynamic traffic uncertainty," in *WiOpt*, 2010.

[25] M. Sadeghi and E. G. Larsson, "Physical adversarial attacks against end-to-end autoencoder communication systems," *IEEE Commun. Lett.*, 2019.

[26] B. Manoj, M. Sadeghi, and E. G. Larsson, "Adversarial attacks on deep learning based power allocation in a massive MIMO network," *IEEE ICC*, 2021.

[27] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *ICLR*, 2017.