2022

Security Hardening of Intelligent Reflecting Surfaces Against Adversarial Machine Learning Attacks

Ferhat Ozgur Catak

Murat Kuzlu
Old Dominion University, mkuzlu@odu.edu

Haolin Tang

Evren Catak

Yanxiao Zhao

Original Publication Citation

Catak, F. O., Kuzlu, M., Tang, H., Catak, E., & Zhao, Y. (2022). Security hardening of intelligent reflecting surfaces against adversarial machine learning attacks. *IEEE Access*, *10*, 100267-100275. https://doi.org/10.1109/ACCESS.2022.3206012

Received 19 August 2022, accepted 31 August 2022, date of publication 14 September 2022, date of current version 27 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3206012

RESEARCH ARTICLE

Security Hardening of Intelligent Reflecting Surfaces Against Adversarial Machine Learning Attacks

FERHAT OZGUR CATAK[1], (Member, IEEE), MURAT KUZLU[2], (Senior Member, IEEE), HAOLIN TANG[3], EVREN CATAK[4], (Member, IEEE), AND YANXIAO ZHAO[3], (Senior Member, IEEE)

[1]Department of Electrical Engineering and Computer Science, University of Stavanger, Rogaland, 4021 Stavanger, Norway
[2]Department of Engineering Technology, Old Dominion University, Norfolk, VA 23529, USA
[3]Department of Electrical and Computer Engineering, Virginia Commonwealth University, Richmond, VA 23284, USA
[4]4034 Stavanger, Norway

Corresponding author: Yanxiao Zhao (yzhao7@vcu.edu)

This work was supported in part by the Commonwealth Cyber Initiative (CCI) and Virginia Commonwealth University Presidential Research Qurest Fund (PeRQ). ICC is an Investment in the Advancement of Cyber Research and Development, Innovation, and Workforce Development in VIRGINIA. For more information about CCI, visit (www.cyberinitiative.org).

ABSTRACT Next-generation communication networks, also known as NextG or 5G and beyond, are the future data transmission systems that aim to connect a large amount of Internet of Things (IoT) devices, systems, applications, and consumers at high-speed data transmission and low latency. Fortunately, NextG networks can achieve these goals with advanced telecommunication, computing, and Artificial Intelligence (AI) technologies in the last decades and support a wide range of new applications. Among advanced technologies, AI has a significant and unique contribution to achieving these goals for beamforming, channel estimation, and Intelligent Reflecting Surfaces (IRS) applications of 5G and beyond networks. However, the security threats and mitigation for AI-powered applications in NextG networks have not been investigated deeply in academia and industry due to being new and more complicated. This paper focuses on an AI-powered IRS implementation in NextG networks along with its vulnerability against adversarial machine learning attacks. This paper also proposes the defensive distillation mitigation method to defend and improve the robustness of the AI-powered IRS model, i.e., reduce the vulnerability. The results indicate that the defensive distillation mitigation method can significantly improve the robustness of AI-powered models and their performance under an adversarial attack.

INDEX TERMS Security, next-generation networks, adversarial machine learning, model poisoning, intelligent reflecting surfaces.

I. INTRODUCTION

In recent years, next-generation networks, also called NextG or 5G and beyond, have been paying attention more in academia and industry along with high demand and new ways of communication need from consumers. According to the report released by the International Telecommunication Union (ITU), the mobile data traffic based on NextG will constantly increase each year and reach thousands of exabytes [1]. NextG networks aim to connect billions of devices, systems, and applications to meet high data rate and low latency requirements to support new applications, especially delay-sensitive services using the Internet, from digital twins, virtual reality, metaverse, industry 4.0, self-driving cars, online education, to eHealth services and many more [2]. Fortunately, NextG networks can meet these requirements and support these applications with advanced communication, computing, and Artificial Intelligence (AI) technologies. AI is an extraordinary contributor among them to innovative technologies in NextG networks [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Giambattista Gruosso.

Intelligent Reflecting Surfaces (IRS) is one of those innovative technologies, in addition to Massive Multiple-Input Multiple-Output (MIMO) and millimeter wave, to improve the performance of NextG wireless networks in terms of data rate and channel capacity. Recently, IRS has received extensive attention in the literature due to its powerful capability of reconfiguring wireless communication environments. IRS is typically composed of a large amount of low-cost passive reflecting elements [4]. By cooperatively tuning the phase shifts of all reflecting elements, the reflected signals can be constructively or destructively added to the receiver [5]. Consequently, wireless communication environments could be changed dynamically to enhance or degrade communication performance.

Inspired by the tremendous achievements of AI, AI-powered models have also been applied to IRS-driven wireless communication in NextG wireless networks to improve performance [4], [6], [7], [8]. However, the security threats (e.g., model poisoning or adversarial machine learning attacks) and mitigation methods (e.g., adversarial training or defensive distillation) have not been investigated in AI-powered applications of NextG networks due to being new, complicated, and multi-disciplinary topics (e.g., next-generation communications, cybersecurity, and AI) [9], [10].

To fill the gap, this paper will focus on AI-powered IRS applications in 5G and beyond networks, and their vulnerabilities, which have received limited attention. Vulnerabilities of an AI-powered model are one of the top security concerns and deserve a thorough investigation. For example, a trained AI model might be manipulated by adding noise to the data, i.e., targeted and non-targeted adversarial attacks. The adversarial attacks are generated by adding a perturbation to a legitimate data point, i.e., an adversarial example, to fool the AI-powered models.

The major contributions of this paper are summarized as follows:

- Evaluate the vulnerabilities of an AI-powered IRS model under widely used adversarial attacks, including Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), Projected Gradient Descent (PGD), and Momentum Iterative Method (MIM).
- Propose a defensive distillation mitigation method to train a more robust model to improve the robustness of the AI-powered IRS model.
- Conduct the comprehensive simulations to assess the robustness of the proposed AI-powered IRS system with undefended and defended models under the above-mentioned adversarial attacks.

The results indicate that AI-powered models used in NextG networks are vulnerable to adversarial attacks, while the models can be more secure against adversarial attacks through the proposed defensive distillation mitigation method. Note that the scope of this study is limited to one of 5G physical layer applications, i.e., AI-powered IRS, its vulnerability analysis under selected adversarial attacks, and the proposed defensive distillation mitigation method. Other attack types like the Carlini & Wagner (C&W) attack are compute-intensive and require more iterations than traditional methods. Our study uses a less compute-intensive and more efficient way to create adversarial examples.

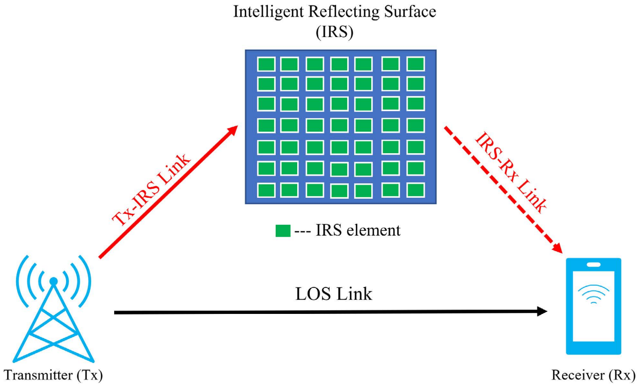


FIGURE 1. **A typical IRS-assisted wireless communication system.**

The remainder of the paper is organized as follows: Section II provides the background information about the IRS and common adversarial attacks. Section III presents the system overview, including the AI model and defense distillation for mitigation. Section IV shows experimental results, and Section V discusses the results along with observations. Section VI concludes the paper.

II. PRELIMINARIES

This section provides background information and related works, including IRS and popular adversarial attacks using FGSM, BIM, PGD, and MIM.

A. INTELLIGENT REFLECTING SURFACES (IRS)

IRS is commonly proposed to improve wireless communication quality in various applications. Consider a typical IRS-aided wireless communication system as depicted in Figure 1. The IRS is deployed to enhance the communication performance between a transmitter and a receiver. The receiver gets the Line of Sight (LOS) signal through the LOS link as well as constructive reflected signals from IRS through the IRS-Rx link at the same time such that the communication performance between the transmitter and receiver could be significantly improved.

In the literature, there are several studies on IRS and security concerns [4], [7], [8]. AI-powered models, e.g., neural networks, have been integrated into IRS-aided systems to improve wireless communication performance. Authors in [11] propose the concept of Intelligent Spectrum Learning (ISL) to optimize IRS to tackle the interfering signals by dynamically controlling the IRS elements. The ISL first employs a well-trained convolutional neural network to realize a multi-class classification for the incident signals, and then the IRS elements can be turned on/off depending on the class of that signal by using an IRS binary control. Moreover, a dynamic "think-and-decide" function allows the reflection of incident signals to be blocked or passed based on the state of the IRS element block. Therefore, the Signal-to-Interference-plus-Noise Ratio (SINR) of the overall system can be improved. The study [12] presents a novel deep learning-based channel tracking algorithm in IRS-assisted UAV communication systems. The algorithm first designs a deep neural network with off-line training for signal denoising, and then a stacked bi-directional long short-term memory is developed to track the time-varying channel. Simulations demonstrate that this algorithm improves channel tracking performance while requiring fewer overheads for pilots than the benchmark algorithm. An IRS architecture is deployed to prevent the communications of multiple legitimate users from eavesdropping in the presence of multiple eavesdroppers [13]. They propose an approach that uses deep reinforcement learning to determine the optimal beamforming policy since the system is highly dynamic and complex.

It is challenging to acquire channel knowledge to estimate the Tx-IRS and IRS-Rx channel link in an IRS-assisted system since all the reflecting elements are expected to be nearly passive. Authors in [14] propose a new IRS architecture where all elements are passive except for a few active sensing elements and adopted a deep learning technique to assist the IRS in addressing this problem. Specifically, the transmitter and receiver first transmit two orthogonal uplink pilots to the active elements of IRS, and the active elements estimate the sampled channel vectors to construct the multipath signature as the environment descriptors. Motivated by recent advances in deep learning, this paper then proposes to train a neural network to observe the environment descriptors to predict the achievable rate with each IRS interaction vector. Based on the predictions, the IRS interaction vector corresponding to the highest predicted achievable rate will be used to reflect the transmitted data from the transmitter to the receiver. In our paper, we refer to the model above as the AI-powered IRS model and will investigate and examine the vulnerability of this model and apply the defensive distillation mitigation method.

B. ADVERSARIAL ATTACKS

Machine Learning (ML)-based models are trained to automatically learn the underlying patterns and correlations in data using algorithms. Once an ML-based model is trained, it can be used to predict the patterns in new data. The accuracy of the trained model is essential to achieving a high performance, which can also be called a generalization. However, the trained model can be manipulated by targeted and non-targeted adversarial ML attacks to fool the models. There are various kinds of adversarial ML attacks, such as evasion attacks, data poisoning attacks, and model inversion attacks.

Liu *et al.* [15] conducted a comprehensive survey on adversarial ML for wireless and mobile systems. Adversarial ML approaches can be used to generate and detect adversarial samples, which are samples that have been specifically designed to deceive a machine learning model. These samples can fool a model into misclassifying an input and can be used to exploit certain blind spots in image classifiers. The article reviews the state-of-the-art adversarial ML approaches to generating and detecting adversarial samples. It provides detailed discussions highlighting the open issues and challenges these approaches face.

An evasion attack aims to cause the ML-based models to misclassify the adversarial examples as legitimate data points, i.e., targeted and non-targeted evasion attacks. Targeted attacks aim to force the models to classify the adversarial example as a specific target class. Non-targeted attacks aim to push the models to classify the adversarial example as any class other than the ground truth. Data poisoning aims to generate malicious data points to train the ML-based models to find the desired output. It can be applied to the training data, which causes the ML-based models to produce the desired outcome. Model inversion aims to generate new data points close to the original data points to find the sensitive information of the specific data points.

These adversarial attack types are given as follows.

1) FAST GRADIENT SIGN METHOD (FGSM)

FGSM is one of the most popular and straightforward approaches to constructing adversarial examples. It is called one-step gradient-based attack. It is used to compute the gradient of the loss function with respect to the input, $\mathbf{x}$, and then the attacker creates the adversarial example by adding the sign of the gradient to the input data. It was first introduced by Goodfellow *et al.* [16]. The gradient sign is computed using the backpropagation algorithm. The steps are summarized as follows:

- Compute the gradient of loss function, $\nabla_{\mathbf{x}}\ell(\mathbf{x}, \mathbf{y})$
- Add the gradient to the input data, $\mathbf{x}_{adv} = \mathbf{x} + \epsilon \times sign(\nabla_{\mathbf{x}}\ell)$,

where ϵ is the budget. FGSM attack has been used in [17] to attack models.

2) BASIC ITERATIVE METHOD (BIM)

BIM is one of the most popular attacks called an iterative gradient-based attack. This attack is derived from the FGSM attack. It is used to compute the gradient of the loss function with respect to the input, $\mathbf{x}$, and then the attacker creates the adversarial example by adding the sign of the gradient to the input data. The gradient sign is computed using the backpropagation algorithm. The steps are summarized as follows:

- Initialize the adversarial example as $\mathbf{x}_{adv} = \mathbf{x}$
- Iterate i times, where $i = 0, 1, 2, 3, \ldots, N$
  - Compute the gradient of loss function, $\nabla_{\mathbf{x}}\ell(\mathbf{x}_{adv}, \mathbf{y})$
  - Add the gradient to the input data, $\mathbf{x}_{adv} = \mathbf{x}_{adv} + \epsilon \times sign(\nabla_{\mathbf{x}}\ell)$,

where ϵ is the budget, and N is the number of iterations. The BIM attack has been used in [17] to attack models.

3) PROJECTED GRADIENT DESCENT (PGD)

PGD is one of the most popular and powerful attacks [18]. It is used to compute the gradient of the loss function with respect to the input, $\mathbf{x}$, and then the attacker creates the adversarial example by adding the sign of the gradient to the input data. The gradient sign is computed using the backpropagation algorithm. The steps are summarized as follows:

- Initialize the adversarial example as $\mathbf{x}_{adv} = \mathbf{x}$
- Iterate i times, where $i = 0, 1, 2, 3, \ldots, N$
  - Compute the gradient of loss function, $\nabla_{\mathbf{x}}\ell(\mathbf{x}_{adv}, \mathbf{y})$
  - Add random noise to the gradient, $\hat{\nabla}_{\mathbf{x}}\ell(\mathbf{x}_{adv}, \mathbf{y}) = \nabla_{\mathbf{x}}\ell(\mathbf{x}_{adv}, \mathbf{y}) + \mathcal{U}(\epsilon)$
  - Add the gradient to the input data, $\mathbf{x}_{adv} = \mathbf{x}_{adv} + \alpha \times sign(\hat{\nabla}_{\mathbf{x}}\ell)$,

where ϵ is the budget, N is the number of iterations, and α is the step size. PGD can generate stronger attacks than FGSM and BIM.

4) MOMENTUM ITERATIVE METHOD (MIM)

MIM is a variant of the BIM adversarial attack, introducing momentum and integrating it into iterative attacks [19]. It is used to compute the gradient of the loss function with respect to the input, $\mathbf{x}$, and then the attacker creates the adversarial example by adding the sign of the gradient to the input data. The gradient sign is computed using the backpropagation algorithm. The steps are summarized as follows:

- Initialize the adversarial example $\mathbf{x}_{adv} = \mathbf{x}$ and the momentum, $\mu = 0$
- Iterate i times, where $i = 0, 1, 2, 3, \ldots, N$
  - Compute the gradient of loss function, $\nabla_{\mathbf{x}}\ell(\mathbf{x}_{adv}, \mathbf{y})$
  - Update the momentum, $\mu = \mu + \frac{\eta}{\epsilon} \times \nabla_{\mathbf{x}}\ell(\mathbf{x}_{adv}, \mathbf{y})$
  - Add random noise to the gradient, $\hat{\nabla}_{\mathbf{x}}\ell(\mathbf{x}_{adv}, \mathbf{y}) = \nabla_{\mathbf{x}}\ell(\mathbf{x}_{adv}, \mathbf{y}) + \mathcal{U}(\epsilon)$
  - Add the gradient to the input data, $\mathbf{x}_{adv} = \mathbf{x}_{adv} + \alpha \times sign(\hat{\nabla}_{\mathbf{x}}\ell)$,

where ϵ is the budget, N is the number of iterations, η is the momentum rate, and α is the step size.

Note that there are many types of adversarial attacks and defenses. The existing defenses and adversarial attacks for images can be applied to attack and defend on intelligent reflecting surfaces and other fields [20], [21], [22], [23]. The cleverly-designed adversarial examples can fool the deep neural networks with high success rates on the test images. The adversarial examples can also be transferred from one model to another model. In our experiments, we generated the adversarial inputs with untargeted attacks.

III. SYSTEM OVERVIEW

This section presents the overall system model for the proposed AI-powered IRS system, as illustrated in Figure 2. According to the figure, it is assumed that data collected from User Equipments (UEs) is provided to the IRS prediction model. The undefended model covers only conventional training of deep neural networks, while the defended model covers the defensive distillation-based training method. The defensive distillation method covers the teacher and student models. The teacher model is typically a large deep neural network, while the student model is usually a small and shallow neural network. In the figure, the training of the prediction model (i.e., student model) is protected against adversarial ML attacks in base stations. Adversarial attacks are applied to models, i.e., undefended and defended models, to evaluate the models' robustness under any attacks.

A. DEEP NEURAL NETWORKS

As we briefly discussed in Section II-A, a neural network is designed for mapping the observed environment descriptors to the predicted achievable rate in the AI-powered IRS model. This subsection introduces the neural network architecture and training details below.

- **Neural Network Architecture:** The input of the neural network model is defined as a stack of the environment descriptors (i.e., uplink pilot signals) received from both transmitter and receiver. Since the training process is designed to build a function mapping descriptors to reflection vectors, the output target of the neural network is to be a set of predictions on the achievable rates of every possible reflection beamforming vector. The neural network is built as a Multi-Layer Perceptron (MLP) network, which is well-demonstrated as an effective universal approximator. The MLP is adopted to establish the connection between the environment descriptors and the predicted achievable rates using reflection beamforming vectors, as shown in Figure 3. The MLP is composed of four fully connected layers. ReLU activation function is adopted, and a dropout layer is added after the activation function for every layer except for the last layer. The MLP consists of the following dimensions: M(Input), $[M, 2M]$(Layer1), $[2M, 4M]$(Layer2), $[4M, 4M]$(Layer3), $[4M, M]$(Layer4), where M is the number of the antenna elements on IRS.
- **Training Details:** The training dataset has 54300 data samples since the candidate receiver locations contain 54300 points as discussed in III-C. The dataset is split into two sets, namely a training set and a testing set with 85% and 15% of the points, respectively. To measure the quality of the predictions and make the predicted achievable rates close to the real achievable rates in the dataset, we define the loss function with Mean-Squared-Error (MSE) between them. In the training process, the batch size is set to 500 samples, and the training epochs is set to 20. The dropout rate is set to 50%, and a L_2 regularization term with the factor of 10^{-4} is added to the loss function. The learning rate decreases by 50% every 3 epochs starting at 0.1 with Stochastic Gradient Descent (SGD) optimizer.

B. DEFENSIVE DISTILLATION

As mentioned previously, in this paper, we leverage the defensive distillation mitigation method to improve the robustness of our AI-powered IRS model. Defensive distillation is a method that applies defensive knowledge distillation to train a more robust model [24]. Knowledge distillation was

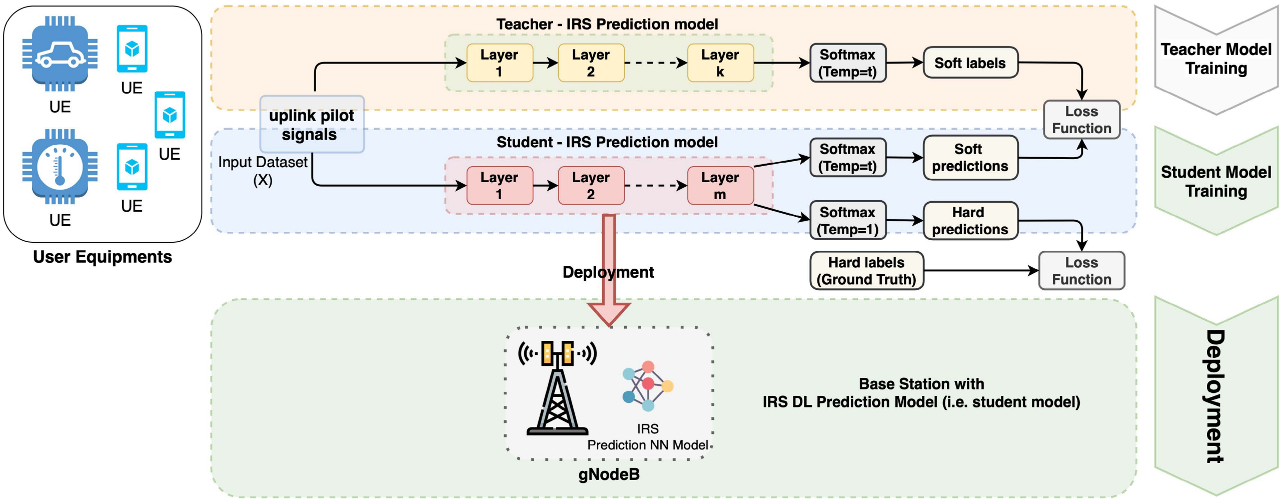


FIGURE 2. Overview of the proposed AI-powered IRS system architecture.

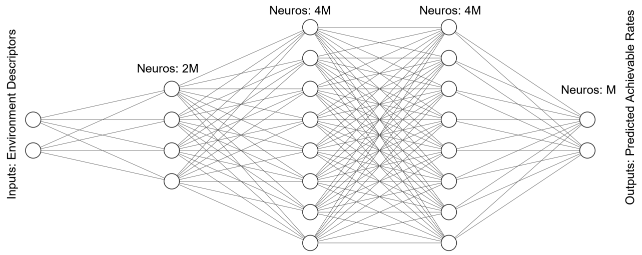

Inputs: Environment Descriptors
Neuros: 2M
Neuros: 4M
Neuros: 4M
Neuros: M
Outputs: Predicted Achievable Rates

FIGURE 3. The adopted neural network architecture is composed of four fully connected layers. The number of the neurons of the four layers is (2*M*, 4*M*, 4*M*, *M*), where *M* indicates the number of the antenna elements on IRS.

previously introduced by Hinton *et al.* [25] to compress the knowledge of a large, densely connected neural network (the teacher) into a smaller, sparsely connected neural network (the student). It has been shown that the student could achieve a similar performance as the teacher by mimicking the teacher's output, and the teacher would be used as a soft label to train the student. Furthermore, the student could be trained to be more resistant to adversarial attacks than the teacher by using the label of the teacher as the label of the student [26].

The architecture of the defensive distillation consists of the following steps:

- **Step 1:** Train a model with cross-entropy loss as the classification task's base model (teacher).
- **Step 2:** Train the same model (teacher) with defensive distillation loss (soft label + cross-entropy) to generate the respective soft label.
- **Step 3:** Train a model with the soft label generated in step 2 as the label (student) to obtain the robust model.

The defensive distillation loss function is defined as

$$\mathcal{L}_D(\theta) = (1-\lambda)\,\mathcal{L}_{CE}(\theta) + \lambda\mathcal{L}_{KL}\left(P_T(y|\theta), P_T(y)\right), \quad (1)$$

where $\mathcal{L}_{CE}(\theta)$ and $\mathcal{L}_{KL}\left(P_T(y|\theta), P_T(y)\right)$ denote the cross entropy and Kullback Leibler (KL) divergence losses, respectively. $P_T(y|\theta)$ is the output of the teacher model with

Algorithm 1 Training the Defensive Distillation

```
Input: Training data set $\mathcal{D}$, base model $M_T$, $\lambda$, $\alpha$, $\epsilon$, number of iterations $N$
Output: Defensive distillation model $M_D$
Train the base model $M_T$ by minimizing the cross entropy loss $\mathcal{L}_{CE}$ on $\mathcal{D}$
Initialize the defensive distillation model $M_D = M_T$
while $iter < N$ do
    Get a batch of samples $X$ and labels $Y$ from $\mathcal{D}$
    Calculate the cross entropy loss $\mathcal{L}_{CE}$ and KL divergence loss $\mathcal{L}_{KL}$ of $X$
    Calculate the defensive distillation loss $\mathcal{L}_D$ using Eq. 1

    Calculate the adversarial samples $X_{adv}$ by FGSM, BIM, MIM, and PGD with $\epsilon$
    Calculate the new loss $\mathcal{L}'_D$ with the adversarial samples $X_{adv}$
    Update the weights of the defensive distillation model $M_D$ by minimizing the new loss $\mathcal{L}'_D$
    $iter \leftarrow iter + 1$
end while
return $M_D$
```

parameters θ. $P_T(y)$ is the output of the soft label. λ is a trade-off parameter between cross entropy and KL divergence losses. Algorithm 1 shows the pseudocode.

C. DATASET PREPARATION

To examine the performance of the AI-powered IRS model, a publicly available ray-tracing-based DeepMIMO dataset [27] is adopted to generate the training dataset. DeepMIMO dataset is a parameterized dataset designed for constructing the MIMO channels based on ray-tracing data obtained from the accurate ray-tracing scenario simulation. Similar to the simulation setup in [14], the outdoor ray-tracing scenario 'O1' is selected as shown in Figure 4. Base

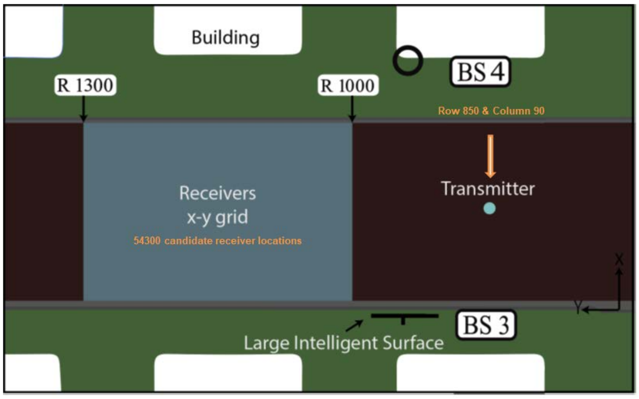


FIGURE 4. The adopted ray-tracing scenario where the large intelligent surface (i.e., IRS) is deployed to reflect the signal from the fixed transmitter to the candidate receivers.

Station 3 (BS 3) is set as an IRS, which is equipped with an UPA (Uniform Planar Array) with 32 × 32 (M = 1024) or 64 × 64 (M = 4096) antennas at the mmWave 28GHz setup. The transmitter is fixed in row R850 and column 90, and the candidate receiver locations are in the uniform x-y grid from row R1000 to R1300 (i.e., 54300 points). Both the transmitter and receiver are assumed to have a single antenna. The antenna elements have a gain of 3dBi and a transmit power of 35dBm. Table 1 summarizes the adopted parameters in the DeepMIMO dataset. The generated DeepMIMO dataset includes the channel vectors between the IRS and the transmitter/receiver of the specified subcarriers for all candidate user locations in the x-y grid. With these channel vectors and given the randomly selected active elements, we can construct the sampled active channel vectors between the active elements of IRS and the transmitter/receiver. Note that the channel vectors depend on the various elements of the surrounding environment [14]. Therefore, the sampled active channel vectors (i.e., environment descriptors) can be used to describe the wireless environment and fed into the deep neural networks described earlier.

D. PERFORMANCE METRIC

This study evaluates the AI-powered IRS model through the Mean Squared Error (MSE) performance metric. MSE scores are utilized to analyze the model vulnerabilities under undefended and defended conditions. The equation regarding the MSE score is given below.

$$MSE = \frac{\sum (Y_t - \hat{Y}_t)^2}{n} \tag{2}$$

where:

- Y_t: The actual t$^{\text{th}}$ instance,
- $\hat{Y}_t$: The forecasted t$^{\text{th}}$ instance,
- n: The total number of instance

MSE score measures the average squared difference between the actual and predicted values. A high MSE score represents a high prediction error.

TABLE 1. The adopted DeepMIMO dataset parameters.

DeepMIMO Dataset Parameter	Value
Frequency band	28GHz
Active BSs	3
Number of Antennas	$(M_x, M_y, M_z) \in \{(1, 32, 32); (1, 64, 64)\}$
Active users (receivers)	From row R1000 to R1300
Active transmitter	row R850 column 90
System bandwidth	100MHz
Number of OFDM subcarriers	512
OFDM sampling factor	1
OFDM limit	64
Number of channel paths	1
Antenna spacing	0.5λ

IV. EXPERIMENTAL RESULTS

This section analyses the results obtained from the experiments related to AI-powered IRS models against adversarial machine learning attacks. Results are represented in three ways: (1) bar plots showing the impact of each adversarial machine learning attack on the performance of undefended and defended models, i.e., MSE, (2) histogram plots showing the MSE metric values for each attack of defended and undefended models, and (3) the table showing the prediction performance results of defended and undefended models for each adversarial attack. Figure 5-6 show the bar plots, while Figure 7-10 show the histogram plots. Table 2 shows the prediction performance results of the defended and undefended AI-powered IRS models against the attacks.

The trained AI-powered IRS model is implemented using Python 3.7.13 and the TensorFlow 2.8.2 framework running on Google Colab Tesla T4 GPU with 16GB of memory. Adversarial inputs are generated using Cleverhans 4.0.0. library.

The adversarial attack on AI-powered models has become more popular with various attack methods. This study uses FGSM, MIM, BIM, and PGD methods to generate adversarial examples. The performance of each model is evaluated through the MSE metric.

Figure 5 shows MSE values for the selected attack methods under attack powers from $\epsilon = 0.01$ to $\epsilon = 0.8$. MSE values look similar for MIM, BIM, and PGD methods, i.e., around 0.09, for all attack powers. On the other hand, MSE values increase along with a higher attack power ($\epsilon > 0.5$) for BIM attacks and go from 0.009 to 0.0128. The results also indicate that AI-powered models are dramatically vulnerable to adversarial attacks. The mitigation methods have been widely used to increase the AI-powered model's robustness against adversarial attacks. In this study, the defensive distillation method is applied in the model to reduce the vulnerability against adversarial attacks. The performance of the AI-powered model is evaluated in terms of MSE after applying the mitigation method. Figure 6 shows the models' performance, i.e., MSE values, against adversarial attacks from $\epsilon =$ 0.01 to $\epsilon = 0.8$ after applying the selected mitigation method. The figure shows that the AI-powered model is still sensitive to adversarial attacks. However, the model's robustness is better against adversarial attacks. According to the figure, the model can resist any attack under low attack power ($\epsilon < 0.3$).

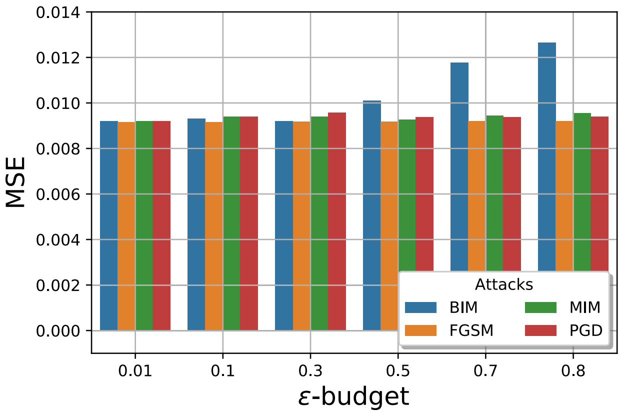


FIGURE 5. MSE values of the undefended models for each adversarial machine learning attack under different attack powers (ϵ).

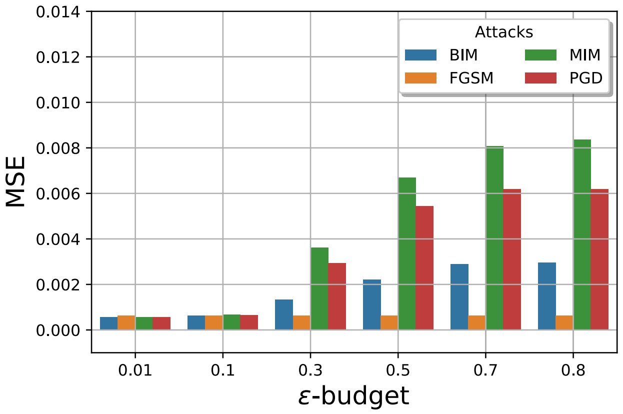
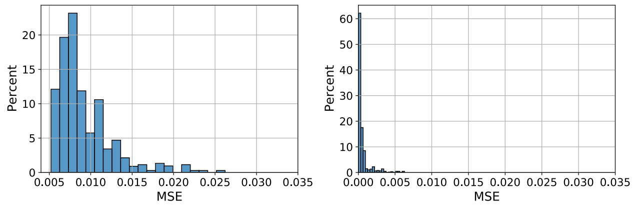


FIGURE 6. MSE values of the defended models for each adversarial machine learning attack under different attack powers (ϵ).

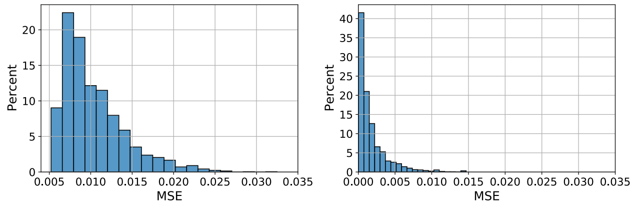
The MSE values increase along with a high attack power ($\epsilon > 0.3$) as expected. However, the impact of the mitigation method on the performance is not the same for all attacks. For example, the MSE values can go up to 0.006 and 0.008 under the PGD and MIM attack, respectively, while only going up to 0.003 under the BIM attack with a very high attack power ($\epsilon = 0.8$). It is very interesting that there is no impact on the attack power under the FGSM attack if the mitigation method is applied to the model. The results also indicate that the defensive distillation method significantly contributes to the model's robustness against adversarial attacks.

The histogram plots investigate the distribution of MSE values for undefended and defended models under adversarial attacks. In Figure 7-10, (a) represents the undefended models, while (b) represents defended models for each attack, i.e., FGSM, BIM, MIM, and PGD, respectively. According to the results, the undefended models, i.e., (a), represent a little right-skewed distribution, which has a peak to the left of the distribution and data values that taper off to the right. MSE values vary from 0.005 to 0.025 for all attack types, and around 50% percent of MSE values are between 0.006 and 0.009. It is compatible with Figure 5-6. On the other hand, it is difficult to define the histogram plots for defended models, i.e., (b). According to the results, Figure 7, 8, 10 (b) represent a little right-skewed distribution like the undefended model ones, while Figure 9 (b) does not represent any distribution. The most MSE values are clustered around 0.0, i.e., 30% - 60%. It means the AI-powered model can correctly predict the target values. It is also clear that the percent of the high MSE values (< 0.015) is much lower than the undefended model. The defended models are more effective against FGSM and BIM attacks, as shown in Figure 7 and 8. It is obvious that the mitigation methods can dramatically improve the model robustness under FGSM attacks, i.e., 90% of MSE values are less than 0.005. On the other hand, the defended models are not successful against MIM and PGD attacks compared to FGSM and BIM, as shown in Figure 9 and 10. Although low MSE values, i.e., < 0.005, are clustered around 50%, the MSE values still go up to 0.015 for MIM and PGD attacks.



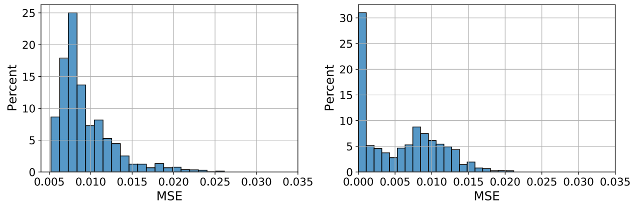
(a) Undefended **(b)** Defensive distillation

FIGURE 7. Distribution of MSE values for undefended and defended models under the FGSM attack.



(a) Undefended **(b)** Defensive distillation

FIGURE 8. Distribution of MSE values for undefended and defended models under BIM attack.



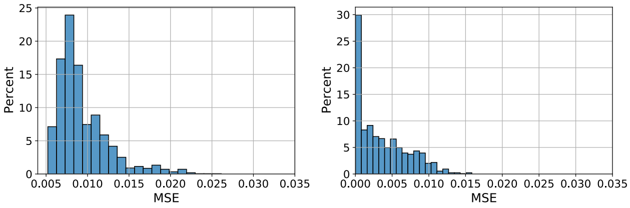
(a) Undefended **(b)** Defensive distillation

FIGURE 9. Distribution of MSE values for undefended and defended models under MIM attack.

Table 2 shows the impact of a specific ϵ value on the MSE performance metrics of the AI-powered IRS model for each adversarial attack in detail. The value of ϵ ranges from 0.01 to 0.8. The higher the value of ϵ means, the more powerful attack on the AI-powered model is expected. Except for BIM, the MSE values are usually around 0.0092-0.0095 for undefended models under any attack power and type. It reaches up to 0.012 under a high attack power

TABLE 2. Prediction performance results in terms of the MSE metric.

		ϵ **values**					
		0.01	**0.1**	**0.3**	**0.5**	**0.7**	**0.8**
FGSM	Undef.	0.009161	0.009162	0.009169	0.009177	0.009187	0.009193
	Distil.	0.000632	0.000631	0.00063	0.000628	0.000628	0.000629
BIM	Undef.	0.009205	0.009308	0.009191	0.010089	0.011761	0.012642
	Distil.	0.000555	0.000625	0.001321	0.002208	0.002895	0.002957
MIM	Undef.	0.009206	0.009402	0.009402	0.009269	0.009438	0.009539
	Distil.	0.00057	0.000663	0.003606	0.006696	0.008069	0.00836
PGD	Undef.	0.009206	0.009398	0.009582	0.009382	0.00937	0.009389
	Distil.	0.000555	0.00065	0.00294	0.005439	0.00618	0.006191



(a) Undefended

(b) Defensive distillation

FIGURE 10. Distribution of MSE values for undefended and defended models under PGD attack.

(BIM and $\epsilon = 0.8$). However, MSE values dramatically go down, e.g., from 0.0091/0.0092 to 0.0005/0.0006 for FGSM/BIM/MIM/PGD, once the mitigation method is applied. It is clear that the mitigation method significantly affects the robustness of the model, but not for all types of attacks. For example, MSE values are the same for the defended model under the FGSM attack at all attack powers. The mitigation method can handle FGSM-type attacks because of its simplicity. However, MSE values increase for the defended model under the other type of attacks at a high attack power level. For example, MSE values go from 0.0005 to 0.002, 0.0005 to 0.008, and 0.0005 to 0.006 for BIM, MIM, and PGD attacks, respectively. MSE values are the highest under the MIM attack (0.008 at 0.8 of the attack power). The MIM is the most effective adversarial attack type among the selected attacks.

V. DISCUSSION

This study investigates AI-powered IRS models in NextG networks and their vulnerabilities against adversarial attacks and the contribution of mitigation methods to the model robustness. The models' vulnerabilities are studied for various adversarial attacks, i.e., FGSM, BIM, MIM, and PGD, as well as the mitigation method, i.e., defensive distillation. The results show that AI-powered IRS models are vulnerable to adversarial attacks. On the other hand, the mitigation methods can significantly improve the model robustness under adversarial attacks. According to the results, adversarial attacks on AI-powered IRS models and the use of the proposed mitigation method can be summarized as:

Observation 1: AI-powered IRS models are vulnerable to adversarial attacks, especially BIM with a high attack power ($\epsilon > 0.5$).

Observation 2: There is no significant impact of the attack power (ϵ) on some adversarial attacks, i.e., FGSM.

Observation 3: The defensive distillation mitigation method significantly increases the model robustness, especially under FGSM and BIM attacks.

Observation 4: The MSE values histogram usually represents a smaller right-skewed distribution, especially for the undefended models.

Observation 5: Around 50% percent of MSE values are between 0.006 and 0.009 for the undefended models.

Observation 6: The most MSE values are clustered around 0.0, i.e., 30% - 60% for the defended model.

Observation 7: The most effective adversarial attack types are BIM and MIM for undefended and defended models, respectively.

VI. CONCLUSION AND FUTURE WORK

The next generation networks, i.e., NextG or 5G and beyond, have dramatically enhanced along with advanced communication, computing, and AI technologies in the last decade. AI is the most important contributor to NextG networks' improvement in terms of performance. This paper investigates the vulnerability of AI-powered IRS models against adversarial attacks (i.e., FGSM, BIM, PGD, and MIM) and the impact of the proposed mitigation method, i.e., defensive distillation, on the improvement of models' robustness in NextG networks. The results indicate that the AI-powered NextG networks are vulnerable to adversarial attacks. On the other hand, mitigation methods can make the models more robust against adversarial attacks. According to the overall results, the most effective adversarial attack types are BIM and MIM for undefended and defended models, respectively. The proposed mitigation method can provide better results for the attacks, including FGSM, BIM, MIM, and PGD, in terms of increasing the model robustness and reducing the vulnerability.

In future work, the authors will focus on automatic modulation classification using an AI-powered model in NextG networks and its vulnerability under adversarial attacks.

REFERENCES

[1] *IMT Traffic Estimates for the Years 2020 to 2030*, I. T. Union, Geneva, Switzerland, 2015.

[2] D. Jiang and G. Liu, "An overview of 5G requirements," *5G Mobile Commun.*, pp. 3–26, Oct. 2017.

[3] K. Koufos, K. Haloui, M. Dianati, M. Higgins, J. Elmirghani, M. Imran, and R. Tafazolli, "Trends in intelligent communication systems: Review of standards, major research projects, and identification of research gaps," *J. Sensor Actuator Netw.*, vol. 10, no. 4, p. 60, 2021.

[4] S. Sarp, H. Tang, and Y. Zhao, ''Use of intelligent reflecting surfaces for and against wireless communication security,'' in *Proc. IEEE 4th 5G World Forum (5GWF)*, Oct. 2021, pp. 374–377.

[5] Q. Wu and R. Zhang, ''Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network,'' *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 106–112, Nov. 2019.

[6] Q. Pan, J. Wu, X. Zheng, J. Li, S. Li, and A. V. Vasilakos, ''Leveraging AI and intelligent reflecting surface for energy-efficient communication in 6G IoT,'' 2020, *arXiv:2012.14716*.

[7] S. A. Abdel Hakeem, H. H. Hussein, and H. Kim, ''Security requirements and challenges of 6g technologies and applications,'' *Sensors*, vol. 22, no. 5, p. 1969, 2022.

[8] Y. Shi, Y. E. Sagduyu, and T. Erpek, ''Federated learning for distributed spectrum sensing in NextG communication networks,'' 2022, *arXiv:2204.03027*.

[9] F. O. Catak, M. Kuzlu, E. Catak, U. Cali, and D. Unal, ''Security concerns on machine learning solutions for 6G networks in mmWave beam prediction,'' *Phys. Commun.*, vol. 52, Jun. 2022, Art. no. 101626.

[10] F. O. Catak, M. Kuzlu, E. Catak, U. Cali, and O. Guler, ''Defensive distillation based adversarial attacks mitigation method for channel estimation using deep learning models in next-generation wireless networks,'' 2022, *arXiv:2208.10279*.

[11] B. Yang, X. Cao, C. Huang, C. Yuen, L. Qian, and M. D. Renzo, ''Intelligent spectrum learning for wireless networks with reconfigurable intelligent surfaces,'' *IEEE Trans. Veh. Technol.*, vol. 70, no. 4, pp. 3920–3925, Apr. 2021.

[12] J. Yu, X. Liu, Y. Gao, C. Zhang, and W. Zhang, ''Deep learning for channel tracking in IRS-assisted UAV communication systems,'' *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7711–7722, Sep. 2022.

[13] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, ''Deep reinforcement learning based intelligent reflecting surface for secure wireless communications,'' *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 375–388, Jan. 2021.

[14] A. Taha, M. Alrabeiah, and A. Alkhateeb, ''Enabling large intelligent surfaces with compressive sensing and deep learning,'' *IEEE Access*, vol. 9, pp. 44304–44321, 2021.

[15] J. Liu, M. Nogueira, J. Fernandes, and B. Kantarci, ''Adversarial machine learning: A multilayer review of the state-of-the-art and challenges for wireless and mobile systems,'' *IEEE Commun. Surveys Tuts.*, vol. 24, no. 1, pp. 123–159, 1st Quart., 2022.

[16] F. Michels, T. Uelwer, E. Upschulte, and S. Harmeling, ''On the vulnerability of capsule networks to adversarial attacks,'' 2019, *arXiv:1906.03612*.

[17] O. F. Tuna, F. O. Catak, and M. T. Eskil, ''Exploiting epistemic uncertainty of the deep learning models to generate adversarial samples,'' 2021, *arXiv:2102.04150*.

[18] Y. Jiang, G. Yin, Y. Yuan, and Q. Da, ''Project gradient descent adversarial attack against multisource remote sensing image scene classification,'' *Secur. Commun. Netw.*, vol. 2021, Jun. 2021, Art. no. 6663028.

[19] I. Fostiropoulos, B. Shbita, and M. Marmarelis, ''Robust defense against Lp-norm-based attacks by learning robust representations,'' Tech. Rep. [Online]. Available: https://par.nsf.gov/servlets/purl/10207644

[20] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, ''Adversarial attacks on medical machine learning,'' *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019, doi: 10.1126/science.aaw4399.

[21] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, ''Adversarial attacks on deep-learning models in natural language processing: A survey,'' *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 3, pp. 1–41, Apr. 2020, doi: 10.1145/3374217.

[22] P. Zelasko, S. Joshi, Y. Shao, J. Villalba, J. Trmal, N. Dehak, and S. Khudanpur, ''Adversarial attacks and defenses for speech recognition systems,'' 2021, *arXiv:2103.17122*.

[23] J. Zhang and C. Li, ''Adversarial examples: Opportunities and challenges,'' *IEEE Trans. neural Netw. Learn. Syst.*, vol. 31, no. 7, pp. 2578–2593, Sep. 2019.

[24] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, ''Distillation as a defense to adversarial perturbations against deep neural networks,'' in *Proc. IEEE Symp. Secur. Privacy (SP)*, 2016, pp. 582–597, doi: 10.1109/SP.2016.4.

[25] G. Hinton, O. Vinyals, and J. Dean, ''Distilling the knowledge in a neural network,'' 2015, *arXiv:1503.02531*. [Online]. Available: https://arxiv.org/abs/1503.02531

[26] X. Chen, Z. Shi, X. Qiu, and X. Huang, ''Adversarial multi-criteria learning for Chinese word segmentation,'' in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 1193–1203. [Online]. Available: http://aclweb.org/anthology/P/P17/P17-1110.pdf

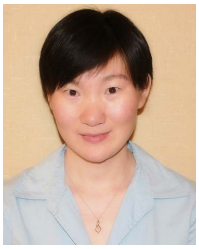
[27] A. Alkhateeb, ''DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications,'' 2019, *arXiv:1902.06435*.

FERHAT OZGUR CATAK (Member, IEEE) received the B.Sc. degree in electrical/electronic engineering, in 2002, and the Ph.D. degree in informatics, in 2014. He is currently an Associate Professor with the University of Stavanger, Norway. Previously, he worked at TUBITAK, Turkey, NTNU, and the Simula Research Laboratory, Norway. His research interests include cyber security, malware analysis, secure multi-party computation, and privacy methods.

MURAT KUZLU (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electronics and telecommunications engineering from Kocaeli University, Turkey, in 2001, 2004, and 2010, respectively. He joined the Department of Engineering Technology, Old Dominion University (ODU), in 2018, as an Assistant Professor. He worked as a Senior Researcher at Scientific and Technological Research Council of Turkey (TUBITAK), from 2006 to 2011. Before joining ODU, he was a Research Assistant Professor at the Advanced Research Institute, Virginia Tech. His research interests include cyber-physical systems, smart cities, smart grids, artificial intelligence, and next-generation wireless networks.

HAOLIN TANG received the B.S. degree in computer science and technology from Yunnan Normal University, China, in 2018. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Virginia Commonwealth University. His research interests include next-generation wireless communication, cyber security, computer vision, and machine learning.

EVREN CATAK (Member, IEEE) received the B.Sc. degree in electrical and electronics engineering from Eskisehir Osmangazi University, Turkey, in 2002, the M.Sc. degree in electronics engineering from Kadir Has University, Istanbul, Turkey, in 2012, and the Ph.D. degree in communication engineering from Yildiz Technical University, Istanbul, in 2017. She was a Postdoctoral Fellow at the Norwegian University of Science and Technology. Her research interests include the physical layer design of emerging communication systems, communication theory, signal processing, and wireless communications.



YANXIAO ZHAO (Senior Member, IEEE) received the Ph.D. degree from the Department of Electrical and Computer Engineering, Old Dominion University, in 2012. She is currently an Associate Professor with the Electrical and Computer Engineering Department, Virginia Commonwealth University (VCU). Her research interests include, but not limited to machine learning, cyber security, wireless networks, and the Internet of Things (IoT). Her research has been supported by different agencies, including NSF, NASA, Air Force, and Commonwealth Cyber Initiative.