



Adversarial security mitigations of mmWave beamforming prediction models using defensive distillation and adversarial retraining

Murat Kuzlu¹ · Ferhat Ozgur Catak² · Umit Cali³ · Evren Catak⁴ · Ozgur Guler⁵

Published online: 29 November 2022

© The Author(s), under exclusive licence to Springer-Verlag GmbH, DE 2022

Abstract

The design of a security scheme for beamforming prediction is critical for next-generation wireless networks (5G, 6G, and beyond). However, there is no consensus about protecting beamforming prediction using deep learning algorithms in these networks. This paper presents the security vulnerabilities in deep learning for beamforming prediction using deep neural networks in 6G wireless networks, which treats the beamforming prediction as a multi-output regression problem. It is indicated that the initial DNN model is vulnerable to adversarial attacks, such as Fast Gradient Sign Method, Basic Iterative Method, Projected Gradient Descent, and Momentum Iterative Method, because the initial DNN model is sensitive to the perturbations of the adversarial samples of the training data. This study offers two mitigation methods, such as adversarial training and defensive distillation, for adversarial attacks against artificial intelligence-based models used in the millimeter-wave (mmWave) beamforming prediction. Furthermore, the proposed scheme can be used in situations where the data are corrupted due to the adversarial examples in the training data. Experimental results show that the proposed methods defend the DNN models against adversarial attacks in next-generation wireless networks.

Keywords Security · Next-generation networking · 6G mobile communication · Adversarial machine learning

1 Introduction

The first 5G standard was announced and approved by 3GPP in December 2017 [1]. The early standardization work on

5G is expected to provide a solid and stable foundation for the early adoption of 5G services. In addition, 5G will be essential for Internet of Things (IoT) applications and future mobile networks. There are many challenges in the design of 5G networks [2], including a security scheme for beamforming prediction. It is an essential part of wireless networks, studied in communication systems and signal processing. Designing and implementing beamforming algorithms in next-generation wireless networks is also crucial. In current wireless networks, deep learning (DL)-based beamforming prediction is vulnerable to adversarial machine learning attacks [3]. Therefore, designing a security scheme for beamforming prediction in 6G networks is critical.

6G is the latest wireless communication technology among cellular networks currently under development. In 6G solutions, artificial intelligence (AI)-based algorithms, especially DL, would be one of the main components of wireless communication systems [4] to improve the overall system performance. The existing solutions in 5G would be migrated to the AI domain, specifically into the DL area. Therefore, it is crucial to design secure DL solutions for the AI-based models in 6G wireless networks. The new attack surface, in addition to the existing 5G security problems, is DL security

✉ Ferhat Ozgur Catak
f.ozgur.catak@uis.no

Murat Kuzlu
mkuzlu@odu.edu

Umit Cali
umit.cali@ntnu.no

Evren Catak
evren.catak@ieee.org

Ozgur Guler
oguler@ekare.ai

¹ Batten College of Engineering & Technology, Old Dominion University, Norfolk, VA, USA

² Department of Electrical Engineering & Computer Science, University of Stavanger, Rogaland, Norway

³ Norwegian University of Science and Technology, Trondheim, Norway

⁴ Independent researcher, Stavanger, Norway

⁵ eKare, Inc Fairfax, Fairfax, VA, USA

vulnerabilities. Researchers and companies should mitigate their DL models' security problems before deploying them to production environments. They need to identify, document, and perform risk assessments for new types of security threats in the next-generation wireless communication systems.

5G networks were commercially launched in late 2018 [5]. After the first commercial launch of the 5G network, the planning of the next-generation networks, such as 6G, commenced providing communication services for future demands. The most important key for this next generation is the use of advanced communications and AI technologies [6]. In the literature, many studies focus on next-generation wireless networks (5G, 6G, and beyond) and the integration of current emerging AI tools into these networks [7–11]. Next-generation wireless networks have been considered as one of the most important drivers in the ability of current and future information age applications (i.e., virtual and augmented reality, remote surgery, holographic projection, metaverse, etc.) to meet the forecast requirements, such as ultra-broadband, ultra-reliable, low latency communication, massive access, and real-time services with low cost. The authors in [12] reviewed AI-based solutions in 6G networks to achieve these requirements and emphasized several solutions for ultra-broadband transmission (terahertz channel estimation and spectrum management), secure communication (authentication, access control, and attack detection), and ultra-reliability and low latency services (intelligent resource allocation). The study [13] investigated next-generation wireless networks in core services, key performance indices (KPIs), enabling technologies, architecture, challenges and possible solutions, opportunities, and future research trends. It also evaluated core services for 5G and 6G networks. Further, it indicated that several emerging technologies will play a key role in 6G networks, i.e., AI for improving the system performance, blockchain for managing the system security, and quantum computing for computing efficiency. The authors in [14] provided a comprehensive review of DL-based solutions focusing on emerging physical layer techniques, such as massive multiple-input multiple-output (MIMO), multi-carrier (MC) waveform, reconfigurable intelligent surface (RIS) communications, and security for 6G networks. It also indicated that AI will significantly contribute to improving next-generation networks' performance. The study [15] addressed the key role of next-generation networks for humans and systems and discussed how ML-based solutions will improve these networks in terms of performance, control, and security and solve problems in various network layers, i.e., the physical, medium access, and application. Many researchers of 6G networks have explored AI by adopting it as the top solution in many extremely complex scenarios. Yang et al. in [16] pre-

sented an AI-enabled intelligent 6G networks architecture, which can support several services, such as discovery, automatic network adjustment, smart service provisioning, and intelligent resource management. It also discusses AI-based methods and how to apply them to 6G networks by efficiently optimizing network performance, including intelligent spectrum management, mobile edge computing, mobility, and handover management.

Utilizing DL-based algorithms for the next-generation wireless network is a great opportunity to improve the overall system performance. However, it may lead to potential security problems, i.e., AI-model poisoning. While AI-based algorithms offer significant advantages for 6G networks, potential security issues related to AI-based models are typically overlooked. As such, the wireless research community should give particular attention to the security and privacy concerns regarding next-generation networks [17]. The authors in [18] provide an overview of 6G wireless networks in terms of the security and privacy challenges, promising security solutions and technologies, and 6G network specifications. The authors in [19] investigate the role of AI in IoT security for possible cyber attacks, emphasizing the model poisoning attack, i.e., where a machine learning model's training data are poisoned. The study [20] provides a comprehensive review of the opportunities and challenges in AI-based security and private provision, as well as proposes solutions for 6G and beyond networks.

In our recent works [3] and [21], we only investigated FGSM attacks, which can be mitigated using the adversarial training method. In this work, four different adversarial attacks (FGSM, BIM, PGD, and MIM) are investigated to build robust beamforming DL models using two mitigation methods. The DL-based beamforming prediction solutions provide satisfactory results; however, these solutions cannot work under an attack, such as adversarial machine learning attacks. This paper presents a DL security scheme for beamforming prediction using deep neural networks (DNNs) in next-generation wireless networks, which treats the beamforming prediction as a multi-output regression problem. The results showed that the proposed scheme is more secure against adversarial attacks because it is robust to the perturbations of the adversarial samples of the training data.

The rest of the paper is organized as follows: Section 2 describes two publicly available cyber-attack tools and the proposed framework. Section 3 covers the background information regarding adversarial machine learning and mitigation methods and system overview. Section 4 describes the experiments for three scenarios, respectively. Section 5 discusses the proposed scheme along with observations, and Sect. 6 concludes the paper.

2 Cybersecurity frameworks

Cybersecurity frameworks help enterprises manage potential cyber risks in a better way and decide future plans for any cyber threat detection or investigation of a security incident in the application and system development. Widely used cybersecurity frameworks are discussed along with the proposed framework below.

2.1 Cybersecurity frameworks

ML Cyber Kill Chain: Lockheed Martin's *Cyber Kill Chain* is a methodology designed to help companies assess the risks they face and the potential impact on their organization.¹ The methodology breaks down the seven phases of a cyber-attack and the critical activities performed during each step. The seven phases are 1. *Reconnaissance* 2. *Weaponization* 3. *Delivery* 4. *Exploitation* 5. *Installation* 6. *Command and Control/Actuation*, and 7. *Actions on Objectives*. By assessing the activities during each phase of their organization's potential cyber-attack, users can understand the impact of a successful cyber-attack on their organization.

MITRE ATT&CK: It is a framework designed to enable analysts and defenders to identify the stages of an attack and construct and execute a response plan.² MITRE ATT&CK is a comprehensive catalog of attack techniques used by both state and non-state actors. It allows organizations to track a potential adversary's movement and understand their methods to gain access and move laterally across a network. It can be used to identify malicious attackers' activity and generate a more effective response strategy. This framework was designed to be used as a common language and modular so that organizations can determine which techniques they need to focus on.

MITRE Atlas: MITRE developed another framework for AI-based applications, namely MITRE Atlas (Adversarial Threat Landscape for Artificial-Intelligence Systems).³ It is a knowledge resource for AI systems that includes adversary tactics, methodologies, and case studies based on real-world demonstrations from security groups, the state-of-the-art from academic research. It is similar to the MITRE Att&ck framework.

2.2 Proposed framework

In this study, the Cyber Kill Chain and MITRE Atlas frameworks are matched to detect and fix the vulnerabilities of ML models, which will be the new component of potential

AI-based 6G networks. In this way, we aim to show both threats and protection methods. The DL-based beamforming prediction models are used for MIMO systems as the proof-of-concept study of new cyber threats for 6G networks. Figure 1 shows the cyber kill chain for AI-based applications with three (3) stages.

Stage-1 is the process of acquiring artifacts about the target AI model for beamforming prediction (i.e., reconnaissance) and building the adversarial machine learning generated craftily designed noise, i.e., weaponization. The adversary can collect the artifacts of AI models in 6G solutions (e.g., AI models in base stations), such as the weights and hyper-parameters used in the training process and datasets from publicly available resources. After this step, the adversary can replicate the AI model in its environment to build malicious inputs.

Stage-2 is the process of building the replicated model, finding the vulnerabilities, and building the craftily generating malicious pilot signals (i.e., inputs of the beam prediction model) into the target AI model (i.e., delivery).

Lastly, Stage-3 is the process of executing the target AI model with the malicious input signals (i.e., exploitation and installation), which can cause the AI model to produce erroneous results. The adversary can then use the malicious input signals to exploit the AI model and install the backdoor in the AI model (i.e., command and control). The adversary can use the backdoor to take control of the AI model and the target system (actions on objective).

Detailed information on the hostile tactics and methodologies parts of MITRE Atlas is given below, which will take place in the Cyber kill chain stages.

- i) In the reconnaissance phase, the adversary gathers information about the organization and its networks, systems, and employees. This information can be used to build a profile of the organization, the employees working there, and the organization's network and systems and to make a social engineering attack.
- ii) In the weaponization phase, the adversary uses the information gathered during the reconnaissance phase to develop the tools they need to launch an attack against the organization successfully.

The adversary can then focus on the delivery phase, using the same tools to deliver information or files to the organization's network. The adversary will use the information gathered during the reconnaissance phase to determine the best delivery mechanism to get the information, it wants to deliver to the organization's network. Once the adversary has delivered the information, they need to exploit a vulnerability in the organization's network. In this phase, it can use the information gathered during the reconnaissance phase to identify the software operated by the organization, operating

¹ <https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html>

² <https://attack.mitre.org>

³ <https://atlas.mitre.org>

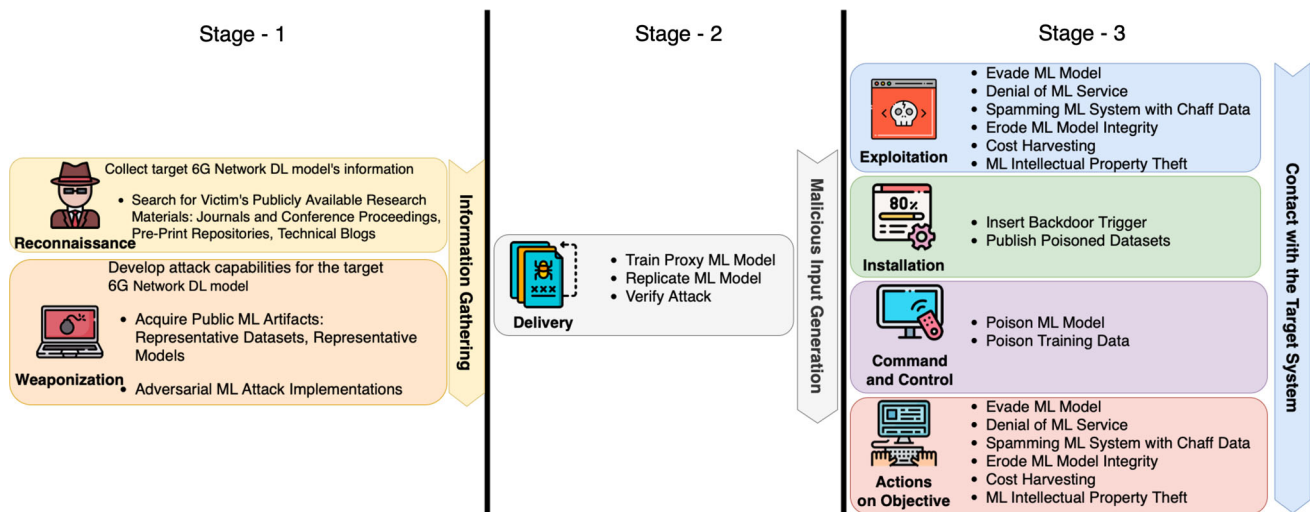


Fig. 1 Cyber kill chain for AI-based applications of 6G wireless communication networks

systems, and applications running on the organization's systems. During the exploitation phase, the adversary uses some of the information gathered in the reconnaissance phase to identify the best way to exploit the organization's network. They can use the reconnaissance phase information to identify the best software, operating systems, and applications to exploit the organization's network. When the adversary has exploited the organization's network, they have the ability to install malicious software on the organization's systems. This malicious software can then exploit the organization's network further or monitor the organization's network. During the command and control phase, the adversary can use the malicious software installed during the exploitation phase to install additional malicious software on the organization's systems. This malicious software can then be used to control the organization's systems. During the actions on objectives phase, the adversary can use the malicious software installed during the exploitation phase to access the organization's systems and steal information or interfere with the organization's network. After the adversary has completed all the steps of the cyber kill chain, they have been able to launch a successful cyberattack on the organization's network. The organization's ability to continue to operate its network can be affected by the adversary's activities during each phase of the cyber kill chain.

3 Background

In this section, a brief overview of the beamforming prediction, the existing adversarial machine learning attacks, such as Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), Projected Gradient Descent (PGD), and Momentum Iterative Method (MIM), along with the existing

solutions, i.e., adversarial training and defensive distillation, for the beamforming prediction in 6G wireless networks are presented. We also introduce the proposed scheme using deep neural networks to protect the beamforming prediction in 6G wireless networks.

3.1 Adversarial machine learning

In adversarial machine learning, the attacker tries to generate a perturbation to the adversarial examples, which would affect the prediction phase of the machine learning model [22]. The goal of the attacker is to manipulate the trained model output so that the attacker can benefit from the user's perspective. Adversarial machine learning attacks work well if the attacker has access to the training data. However, the proposed scheme is robust to the perturbations of the adversarial samples of the training data, which in turn makes the proposed scheme robust to adversarial machine learning attacks.

The DNN model's input is the pilot signal received at BSs with omni or quasi-omni beam patterns, and the output is the beamforming vectors. However, the transmitted pilot signal is distorted due to the various elements of the propagation, i.e., reflection, diffraction, and scattering. This creates an RF signature of the environment when the pilot signal is received at BSs. The RF signature and the pilot signals are needed to learn to predict the beamforming directions when these pilot signals are received collectively at the many BSs. Taking DL-based beamforming prediction model, here, we use $h(\mathbf{x}, \omega) : \mathbb{C}^k \mapsto \mathbb{R}^m$ to denote that an uplink pilot signals, $\mathbf{x} \in \mathbb{C}^k$, to beamforming vectors, $\mathbf{y} \in \mathbb{R}^m$ where ω shows the parameters of the prediction model, h . Given the budget ϵ (i.e., the norm vector of the noise), the attacker tries to find a noise vector $\sigma \in \mathbb{C}^k$ to maximize the loss function ℓ output

[23]. The attacker uses the lowest possible budget to corrupt the inputs, aiming to increase the distance (i.e., MSE) between the model's prediction and the real beam vector. Therefore σ is calculated as:

$$\sigma^* = \arg \max_{|\sigma|_p \leq \epsilon} \ell(\omega, \mathbf{x} + \sigma, \mathbf{y}) \quad (1)$$

where $\mathbf{y} \in \mathbb{R}^m$ is the label (i.e., beamforming vectors), and p is the norm value and it can be 0, 1, 2, ∞ .

Methods for constructing adversarial examples can be categorized into groups: gradient-based and content-based attacks, respectively [24]. In this study, gradient-based attacks were chosen as adversarial attacks because of their simplicity and variety. These attacks use the gradient of the loss function to generate adversarial examples, i.e., incorrectly labeled. These adversarial attack types are given as follows.

(i) **Fast Gradient Sign Method (FGSM):** FGSM is a simple, fast, and single-step attack type that can quickly generate adversarial examples. It was first introduced by Goodfellow et al. in 2014 [25]. The gradient sign is computed using backpropagation and is quite fast. In this method, the noise, i.e., different from random noise, is added to data in the same direction (+/-) along with the loss function. The noise is adjusted by epsilon (ϵ), which is a small number controlling the size of an adversarial attack. We can summarize the FGSM using the following equation:

$$\mathbf{x}^{adv} = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_x \ell(\omega, \mathbf{x}, y)) \quad (2)$$

(ii) **Basic Iterative Method (BIM):** BIM is an extension of the FGSM single-step attack. In this method, adversarial examples are updated by iterating many times. However, this increases the computing cost and complexity. Unlike FGSM, BIM manipulates the selected input with a smaller step size, and each value is calculated as in (ϵ), the neighborhood of the original input [26]. It takes an iterative approach by applying FGSM multiple times to a small step size α instead of taking one large step, i.e., ϵ/α . We can summarize the BIM using the following equations:

$$\begin{aligned} \mathbf{x}_0^{adv} &= \mathbf{x}, \\ \mathbf{x}_{N+1}^{adv} &= \text{Clip}_{\mathbf{x}, \epsilon} \{ \mathbf{x}_N^{adv} + \epsilon \cdot \text{sign}(\nabla_x \ell(\omega, \mathbf{x}_N^{adv}, y)) \} \end{aligned} \quad (3)$$

(iii) **Projected Gradient Descent (PGD):** The PGD attacks are similar to FGSM and BIM attack types. However, it has a different method to generate adversarial examples. It initials the search for the adversarial example at random points in a suitable region, then runs several iterations to

find an adversarial example with the greatest loss, but the size of the perturbation is smaller than a specified amount referred to as epsilon, ϵ [27]. PGD can generate stronger attacks than FGSM and BIM.

(iv) **Momentum Iterative Method (MIM):** MIM is a variant of the BIM adversarial attack, introducing momentum and integrating it into iterative attacks. It improves the convergence of BIM to stabilize the direction of the gradient at each step [28]. The step size of the ϵ also determines the attack level of MIM as an attack parameter.

Figure 2 shows a typical adversarial machine learning-based malicious input generation process.

3.2 Mitigation methods

The DL-based beamforming prediction is vulnerable to adversarial machine learning attacks in wireless networks. Adversarial training and defensive distillation are two existing mitigation methods for adversarial machine learning attacks that mitigate wireless communication networks.

3.2.1 Adversarial training

The first mitigation method is iterative adversarial training. In this approach, the DNN model is trained with the regular training data, and then the DNN model is trained with the adversarial examples using the correct labels. The DNN model is trained multiple times with regular and adversarial examples. The iterative adversarial training attempts to minimize the adversarial samples' effect on the training process. However, iterative adversarial training is not efficient in practice. To obtain a robust model, the victim model must be trained with all attack types and different parameters. Therefore, the training period of the model can be quite long.

Algorithm 1 shows the pseudo-code of adversarial training.

Algorithm 1 Iterative adversarial training-based mitigation

Input h : vulnerable model, Ω : attacks, Π : epsilon values, \mathbf{x}_{train} : training data, \mathbf{y}_{train} training data output, \mathbf{x}_{test} : test data, \mathbf{y}_{test} : test data output
Output \hat{h} : robust model

```

1: for  $\epsilon \in \Pi$  do                                ▷ For each epsilon budget
2:   for attack  $\in \Omega$  do                          ▷ For each epsilon budget
3:      $\mathbf{x}^{adv} \leftarrow \text{attack}(\mathbf{x}_{train}, \epsilon)$   ▷ Generate malicious inputs with
      attack and  $\epsilon$  budget.
4:      $\mathbf{x}^{adv\_train} \leftarrow \mathbf{x} \cup \mathbf{x}^{adv}$         ▷ Merge newly created malicious
      inputs  $\mathbf{x}^{adv}$  and  $\mathbf{x}_{train}$ 
5:      $h.fit(\mathbf{x}^{adv\_train}, \mathbf{y}_{train})$           ▷ Re-train the model  $h$  with new
      training data
6:   end for
7: end for
```

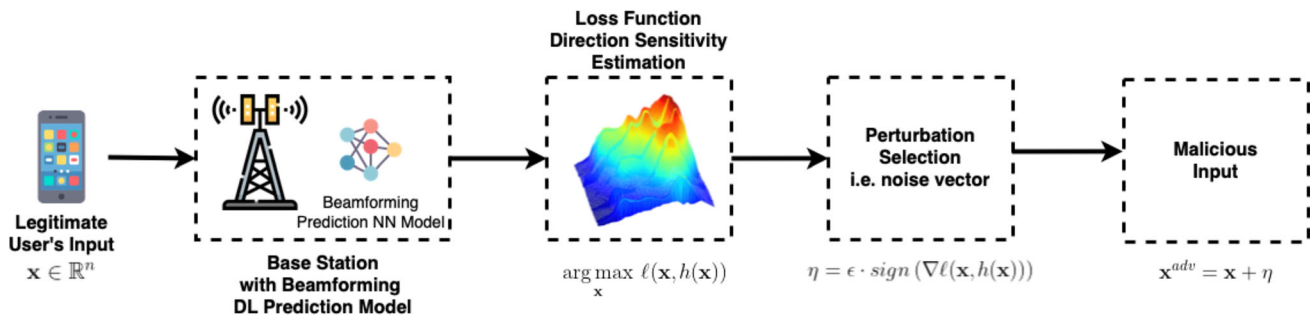


Fig. 2 Typical adversarial machine learning-based malicious input generation

3.2.2 Defensive distillation

Knowledge distillation was previously introduced by Hinton et al. in [29] to compress a large model into a smaller one. Papernot et al. in [30] proposed this technique for the adversarial machine learning defense against attacks. The defensive distillation mitigation method includes larger teacher and compressed student models. The first step is to train the teacher model with a high-temperature (T) parameter to soften the softmax probability outputs of the DNN model. Equation (4) shows the modified softmax activation function as follows:

$$p_i = \frac{\exp(\frac{z_i}{T})}{\sum_j \exp(\frac{z_j}{T})} \quad (4)$$

where p_i is the probability of i^{th} class, and z_i are the logits. The second step is to use the previously trained teacher model to obtain the soft labels of the training data. In this step, the teacher model predicts each of the samples in the training data using the same temperature (T) value, and the predictions are the labels (i.e., soft labels) for the training data to train the student model. The student model is trained with the soft labels acquired from the teacher model, again with a high T value in the softmax. After the student model's training phase, the T parameter is set to 1 during the prediction time of the student model. Figure 3 shows the overall steps for this technique.

In the figure, the training of the beamforming prediction model (i.e., student model) is protected from adversarial machine learning attacks. The teacher model is trained as the first step, the student model is trained with the predictions made by the teacher model, and the real labels with the student model's predictions are used as the loss function inputs as the second step. In this way, the knowledge of the teacher model is compressed and transferred to the student model. The student model is deployed to the base stations in the last stage.

This technique significantly reduces the effects of gradient-based untargeted attacks. This is because defensive distillation

has the effect of lowering the gradients down to zero, the usage of the standard objective function is no longer practical.

3.3 Dataset description and scenarios

The generic DL dataset generation framework for massive MIMO (DeepMIMO) and millimeter-wave channels is used in experiments [31]. This framework consists of two parts: (i) creating the DeepMIMO channels based on accurate ray-tracing data obtained from the Wireless InSite simulator, developed by Remcom [32] for mmWave and massive MIMO models, and (ii) configuring a generic (parametrized) system and channel parameters to generate DeepMIMO dataset for the different applications. The ray-tracing simulation is used to generate channels based on geometry-based characteristics. They include primarily (1) the correlation between the channels at different locations and (2) the dependence on the environment geometry/materials. The generic (parametrized) dataset allows researchers to tune several parameters, such as the number of BSs, users, antennas, channel paths, system bandwidth, and subcarriers.

In this study, the DeepMIMO dataset is described for three ray-tracing scenarios, i.e., O1_60 (outdoor - 60 GHz), I1_2p5 (indoor - 2.5 GHz), and I3_60 (indoor - 60 GHz). A short description of each original scenario is given as follows:

- (i) O1_60 is an outdoor scenario of two streets and one intersection, which includes 18 base stations supporting more than a million users. Its operating frequency is 60 GHz [33].
- (ii) I1_2p5 is an indoor distributed massive MIMO scenario of a 10x10x5 (m) room with two conference tables, which includes 64 distributed antennas in the ceiling at 2.5 m height. It can support more than 150 thousand users and, its operating frequency is 2.5 GHz [34].
- (iii) I3_60 is an indoor conference room scenario, i.e., 10x11x3 (m) conference room with its hallways, which includes two access points inside the conference room at

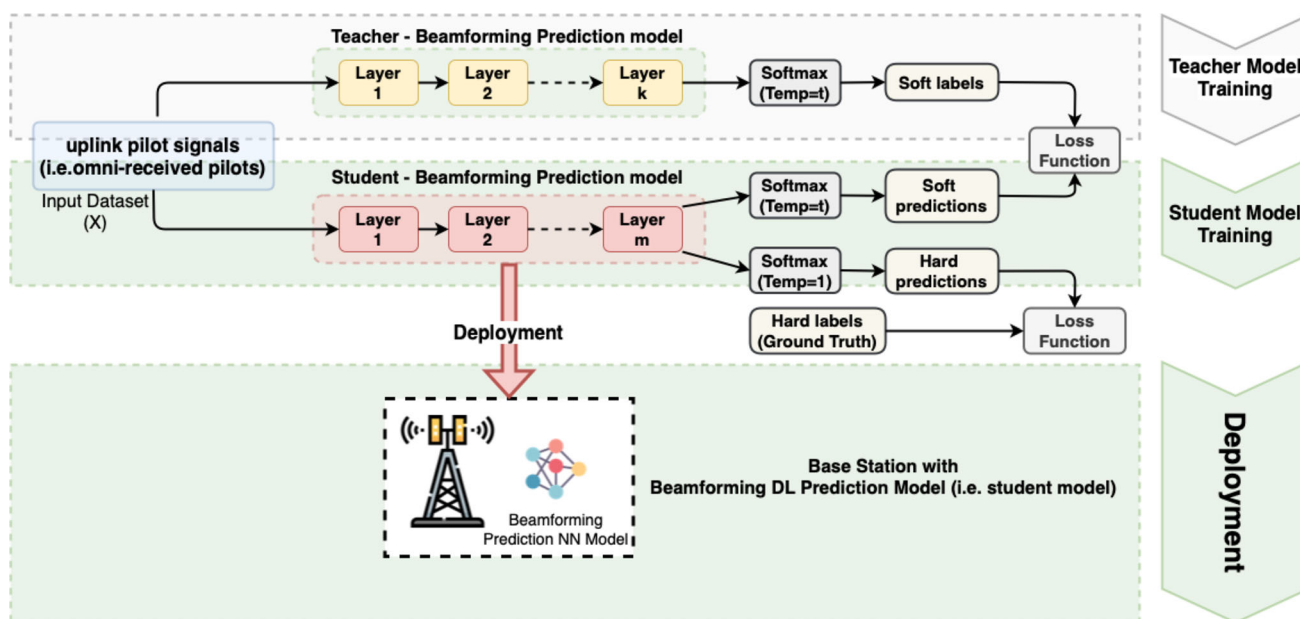


Fig. 3 Defensive Distillation

Table 1 DeepMIMO dataset parameters for each scenario

	O1_60	I1_2p5	I3_60
Number of active *BSs/APs	4	1	2
Number of active users	54300	12060	7260
Number of BS antennas	256	64	32
System bandwidth	0.5 GHz	0.02GHz	0.5 GHz
Number of subcarriers	1024	64	32

*BS: Base Station, AP: Access Point

2 m height. It can support more than 118 thousand users, and its operating frequency is 60 GHz [35].

These scenarios are revised in terms of the number of BSs/APs, and active users. The revised DeepMIMO dataset parameters are given for each scenario in Table 1.

3.4 System overview

This section gives a high-level system overview of the proposed security scheme for beamforming prediction in 6G wireless networks. The proposed security scheme is a two-phase approach: (1) adversarial training and (2) defensive distillation. In the adversarial training phase, the proposed scheme uses a modified version of the adversarial training algorithm proposed in [3]. The adversarial training algorithm is used to train the deep learning models to defend against

adversarial attacks. The complex number system is used in digital wireless communication, especially in the modulation and demodulation of wireless signals. However, adversarial machine learning attacks try to penetrate the decision boundaries of the victim DL models using real numbers, and the final malicious inputs are in the real number domain. To overcome this problem, the complex numbers are broken into their corresponding real and imaginary parts.

In the defensive distillation phase, the proposed scheme uses a modified version of the defensive distillation algorithm proposed in [4]. The defensive distillation algorithm is used to improve the deep learning models against adversarial attacks. The proposed security scheme is implemented in the mmWave beamforming prediction in 6G wireless networks. Figure 4 shows the system overview.

4 Experimental results

4.1 Research questions

- RQ1: Can we generate malicious inputs for beamforming vector prediction models using FGSM [36], PGD, BIM, and MIM attacks in the complex domain?
- RQ2: Is there any correlation between noise vector norm value (i.e., epsilon) and prediction performance with the MSE metric?

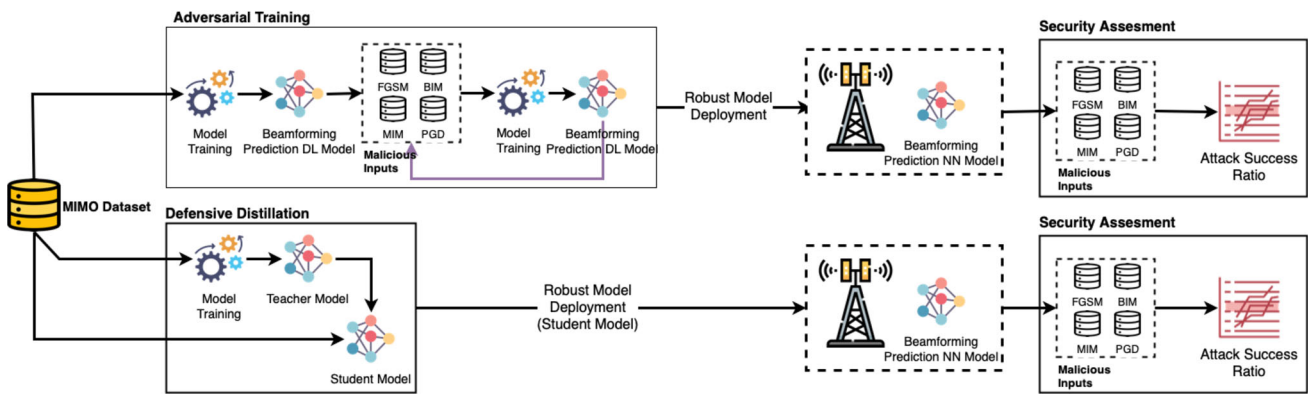


Fig. 4 System overview of the proposed security scheme for beamforming prediction in 6G wireless networks

- RQ3: What are the adversarial training and defensive distillation-based mitigation methods' protection performance metric results with different epsilon values?

4.2 RQ1 results

To answer this research question, first, we train a beamforming vector prediction model on a large number of simulated data to generate realistic malicious inputs. We then apply the attack algorithms FGSM, PGD, BIM, and MIM to generate malicious inputs and demonstrate that it is possible to generate malicious inputs for beamforming vector prediction models using the proposed attacks. Furthermore, we also examine the possibility of using the attacks to generate malicious inputs for other machine learning models. The complete paper demonstrates the feasibility of using the attacks for generating malicious inputs for beamforming vector prediction models. However, the research question of whether or not it is possible to generate malicious inputs for other machine learning models using the attacks is still open.

Figure 5 shows the prediction performance of the beamforming vector prediction models when the malicious inputs are generated using different attack algorithms in a simulation study.

The figure shows that all the attack algorithms can generate malicious inputs for the beamforming vector prediction model. As we can see from the figure, the BIM attack has the highest prediction error rate (i.e., attack success ratio). The PGD attack has the second-highest prediction error rate. The prediction error rate of the MIM attack has the third-highest prediction error rate. The FGSM attack has the lowest. The figures show that the beamforming vector prediction models are more sensitive to the BIM attack, whereas the FGSM attack is less sensitive to the predictions.

Concluding Remarks for RQ1: The attackers can generate malicious inputs for the beamforming vector prediction model. From the attacker's perspective, the most successful attack is BIM.

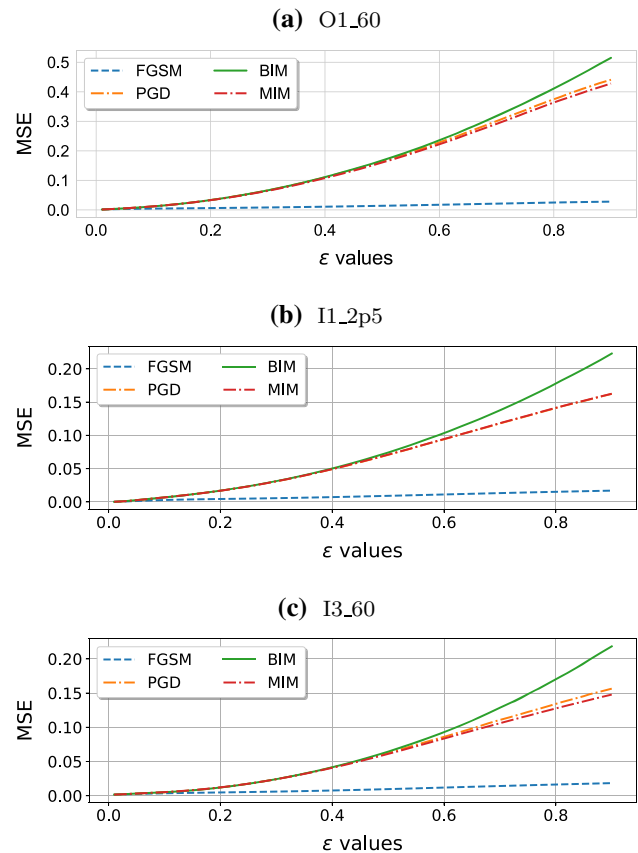


Fig. 5 Prediction performance of each model using the MSE metric. The x-axis shows the ϵ budget, and the y-axis shows the MSE value of the model

4.3 RQ2 results

We examined the correlation between noise vector norm value (i.e., epsilon) and prediction performance with MSE metric in a simulation study. The results show that the prediction performance with the MSE metric is strongly correlated with the noise vector norm value when noise is added to the input features and the target feature.

Table 2 Pearson correlation coefficients of the scenarios and attacks

	FGSM	PGD	BIM	MIM
O1_60	0.9910	0.9799	0.9718	0.9802
I1_2P5	0.9958	0.9917	0.9761	0.9920
I3_60	0.9925	0.9744	0.9539	0.9763

This simulation study examines the correlation between noise vector norm value (i.e., epsilon) and prediction performance with the MSE metric. The results show that the prediction performance with the MSE metric is strongly correlated with the noise vector norm value (i.e., epsilon) when noise is added to the input and target features.

Table 2 shows the Pearson correlation coefficients of the relation between epsilon budget and MSE value. Pearson correlation is a statistical measure of the linear correlation between two variables. It is a measure of the extent to which two variables vary together. A correlation of 1.0 means that the two variables vary completely; a correlation of 0 means that the two variables vary independently. In the case of Table 2, the correlation coefficient of the relation between epsilon budgets and MSE value is around 0.99. This means that the prediction performance with the MSE metric is strongly correlated with the noise vector norm value (i.e., epsilon) when noise is added to the input and target features.

Figures 6-8 show the MSE distributions of each input instance in the malicious inputs generated with different attack algorithms. The figures represent the distribution of the MSE obtained from the malicious inputs generated by adding noise to the input features and the target feature. They indicate that the MSE distribution of the malicious inputs generated by adding noise to the input features is not uniform. In the case of the MIM attack, the results show that the MSE distribution of the malicious inputs generated by adding noise to the input features, and the target feature has a minimal variance. The MIM attack adds noise to the input and target features. The BIM attack adds noise to the input features. Thus, the MSE distribution of the malicious inputs generated by adding noise to the input features, and the target feature has a more significant variance than the MIM attack. The PGD attack adds Gaussian noise to the input features. Thus, the MSE distribution of the malicious inputs generated by adding noise to the input features has a more significant variance than the FGSM attack. The FGSM attack adds Gaussian noise to the input features. Thus, the MSE distribution of the malicious inputs generated by adding noise to the input features has a more significant variance than the PGD attack.

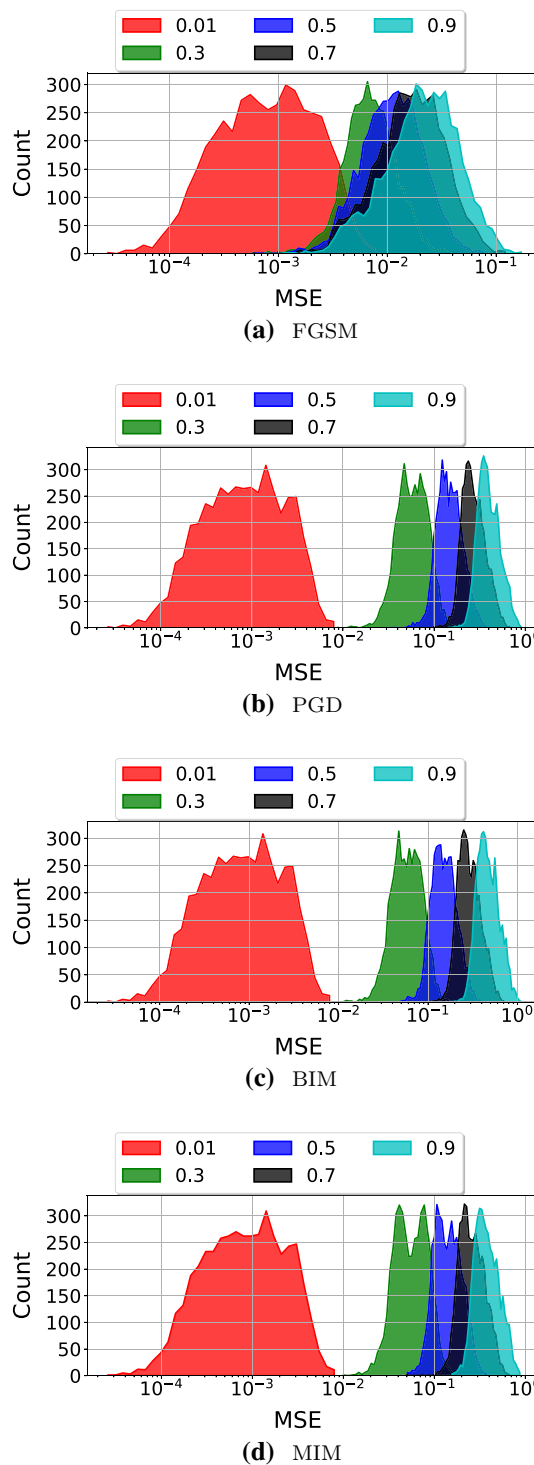


Fig. 6 O1_60: MSE distributions of the malicious inputs with five different epsilon values ($\epsilon \in \{0.01, 0.3, 0.5, 0.7, 0.9\}$)

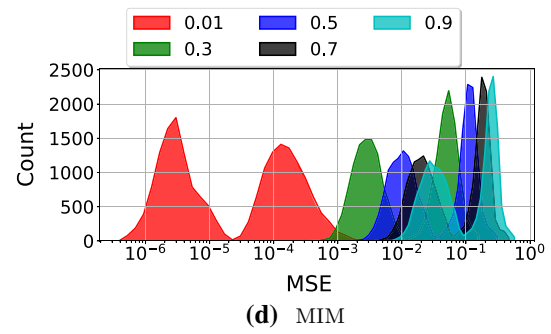
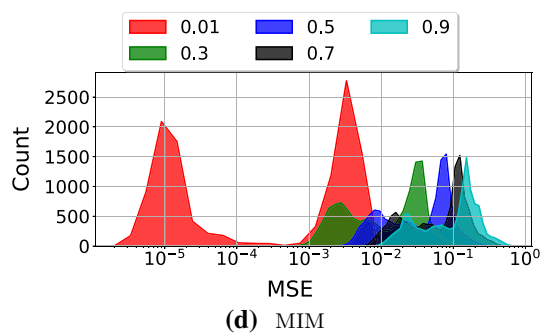
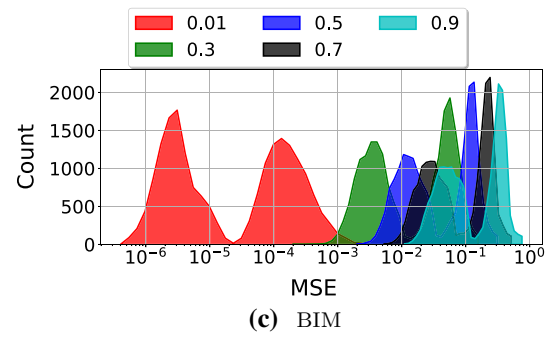
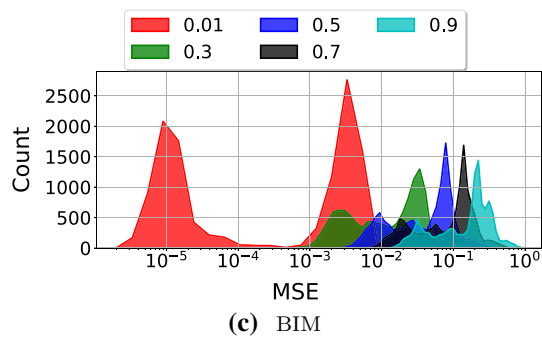
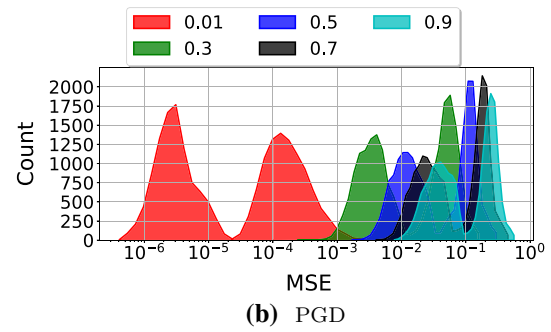
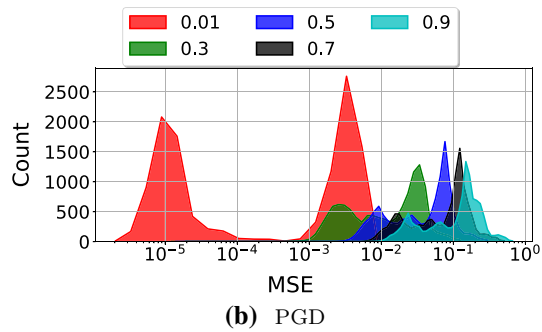
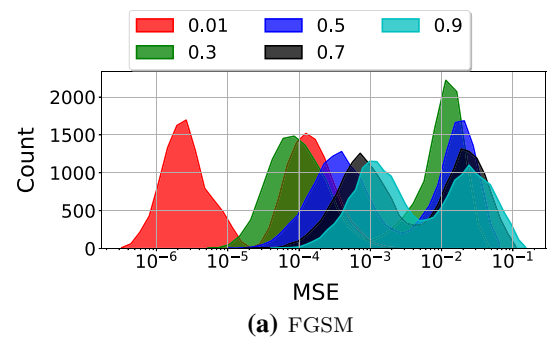
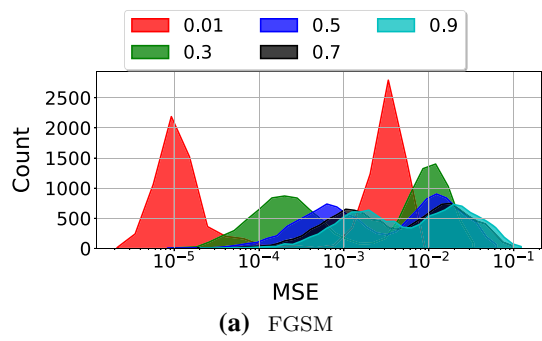


Fig. 7 I3_60: MSE distributions of the malicious inputs with five different epsilon values ($\epsilon \in \{0.01, 0.3, 0.5, 0.7, 0.9\}$)

Fig. 8 I1_2p5: MSE distributions of the malicious inputs with five different epsilon values ($\epsilon \in \{0.01, 0.3, 0.5, 0.7, 0.9\}$)

Concluding Remarks for RQ2: There is a strong negative correlation between ϵ and the DL model’s prediction performance. The confidence interval value of the correlations (i.e., p -value) is 0. The p -value is the probability connected to the likelihood of acquiring the correlation result.

4.4 RQ3 results

Figure 9 summarizes the experiment results for the adversarial training and defensive distillation mitigation methods for different epsilon values ($\epsilon : \{0.01, 0.03, 0.05, 0.08, 0.10\}$). Except for the I1_2p5 scenario, the MSE values of the model, which has been made robust by the defensive distillation mitigation method, are lower in the other two scenarios (i.e., the prediction performance is higher). In the I1_2p5 scenario, the adversarial training method is more successful in protecting against attacks with low ϵ values (i.e., $\epsilon < 0.08$). In contrast, in cases where the epsilon value is 0.08 or higher, the defensive distillation mitigation method creates a more successful defense.

Table 3 shows the experiment results for each scenario, attack, and mitigation method.

Concluding Remarks for RQ3: The defensive distillation mitigation method is more resilient against higher-order attacks, which are more difficult to detect.

5 Discussion

In this study, a comprehensive analysis of the mmWave beamforming prediction model’s vulnerabilities and mitigations has been provided. The model’s vulnerabilities are investigated for various adversarial attacks, i.e., FGSM, BIM, PGD, and MIM, while the mitigations for adversarial training and defensive distillation are explored. The results show that mmWave beamforming prediction models provide a satisfactory performance without any adversarial attacks. On the other hand, the models are very sensitive to adversarial attacks, especially BIM. For example, as shown in Figure 5, the MSE value can rise to 0.5 (for O1_60 scenario) and 0.20 (for I1_2p5 and I3_60 scenarios) under a heavy adversarial attack, i.e., $\epsilon = 0.9$. According to Fig. 6–8, the MSE distribution of the model performance is not uniform under the adversarial attack. Those attacks add noise to the input and/or target features. Figure 9 demonstrates the adversarial training and defensive distillation-based mitigation methods resulting in mmWave beamforming prediction. The defensive distillation mitigation method provides a better performance against higher-order attacks.

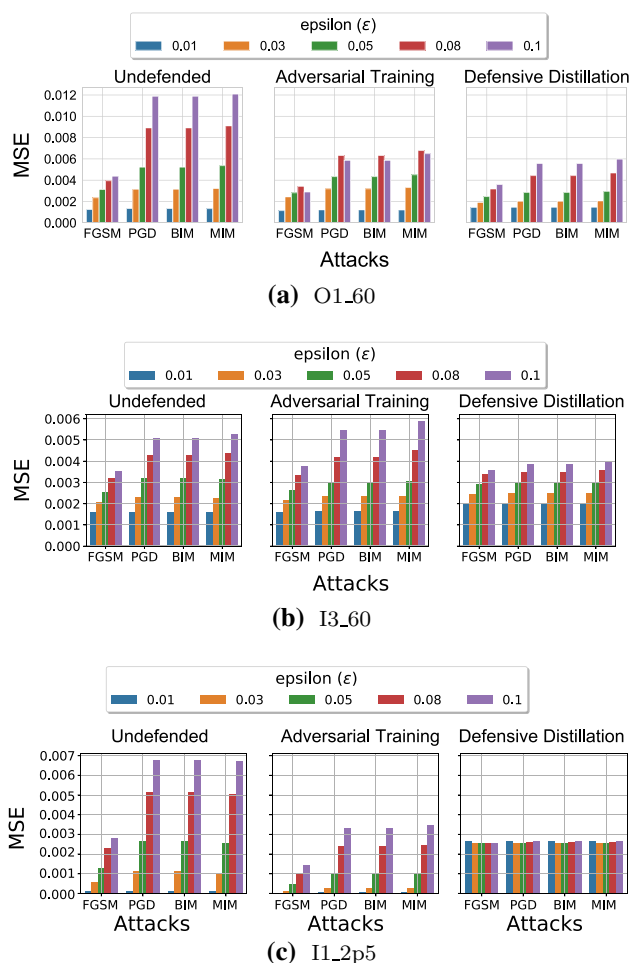


Fig. 9 Adversarial training and defensive distillation-based mitigation methods results

Observations derived from the results of adversarial attacks on mmWave beamforming prediction models and the use of mitigation methods can be summarized as:

Observation 1: The mmWave beamforming prediction models are vulnerable to adversarial attacks.

Observation 2: BIM is the most successful attack among those selected.

Observation 3: There is a strong negative correlation between attack power ϵ and the performance of DL-based mmWave beamforming prediction models.

Observation 4: The defensive distillation mitigation method is more resilient against higher-order attacks.

6 Conclusion and future work

This paper presents a DL security scheme for RF beamforming prediction models’ vulnerabilities and their mitigation techniques by satisfying the following research questions: (1) Can we generate malicious inputs for beamforming vec-

Table 3 Summary of the experiment results for each scenario, attack, and mitigation method

Sc.	Attack	Epsilon	Undefended	Adversarial Training	Defensive Distillation
O1_60	BIM	0.03	0.003160	0.003194	0.002033
		0.05	0.005232	0.004333	0.002864
		0.08	0.008918	0.006315	0.004427
		0.10	0.011888	0.005867	0.005574
	FGSM	0.03	0.002352	0.002440	0.001918
		0.05	0.003137	0.002819	0.002457
		0.08	0.003950	0.003431	0.003183
		0.10	0.004380	0.002900	0.003591
	MIM	0.03	0.003226	0.003303	0.002062
		0.05	0.005363	0.004510	0.002963
		0.08	0.009107	0.006778	0.004686
		0.10	0.012092	0.006488	0.005984
	PGD	0.03	0.003160	0.003194	0.002033
		0.05	0.005232	0.004333	0.002864
		0.08	0.008918	0.006315	0.004426
		0.10	0.011887	0.005868	0.005576
I3_60	BIM	0.03	0.002282	0.002345	0.002491
		0.05	0.003171	0.002995	0.002991
		0.08	0.004285	0.004185	0.003489
		0.10	0.005079	0.005460	0.003841
	FGSM	0.03	0.002055	0.002149	0.002446
		0.05	0.002551	0.002617	0.002898
		0.08	0.003173	0.003318	0.003361
		0.10	0.003508	0.003782	0.003578
	MIM	0.03	0.002270	0.002347	0.002496
		0.05	0.003161	0.003063	0.003006
		0.08	0.004354	0.004522	0.003566
		0.10	0.005245	0.005896	0.003998
	PGD	0.03	0.002282	0.002345	0.002491
		0.05	0.003171	0.002995	0.002991
		0.08	0.004285	0.004184	0.003489
		0.10	0.005079	0.005461	0.003841
I2_2p5	BIM	0.03	0.001113	0.000268	0.002583
		0.05	0.002655	0.001029	0.002548
		0.08	0.005145	0.002399	0.002586
		0.10	0.006786	0.003317	0.002649
	FGSM	0.03	0.000601	0.000138	0.002582
		0.05	0.001314	0.000479	0.002541
		0.08	0.002288	0.000989	0.002551
		0.10	0.002818	0.001428	0.002583
	MIM	0.03	0.001048	0.000252	0.002583
		0.05	0.002540	0.001003	0.002549
		0.08	0.005054	0.002457	0.002593
		0.10	0.006742	0.003453	0.002663

Table 3 continued

Sc.	Attack	Epsilon	Undefended	Adversarial Training	Defensive Distillation
	PGD	0.03	0.001113	0.000268	0.002583
		0.05	0.002655	0.001029	0.002548
		0.08	0.005144	0.002399	0.002586
		0.10	0.006785	0.003317	0.002649

tor prediction models using FGSM, PGD, BIM, and MIM attacks in the complex domain?; (2) Is there any correlation between noise vector norm value (i.e., epsilon) and prediction performance with MSE metric?; and (3) What are the adversarial training-based mitigation methods' protection performance metric results with different epsilon values? To investigate these questions, the experiments were performed with the selected DeepMIMO scenarios, i.e., *O1_60*, *I1_2p5*, and *I3_60 ray-tracing*. The results confirm that the original DL-based beamforming model is significantly vulnerable to FGSM, PGD, BIM, and MIM attacks, especially BIM. The MSE value increases in all three scenarios under a heavy BIM adversarial attack ($\epsilon=0.9$), i.e., 0.5 (for *O1_60* scenario) and 0.20 (for *I1_2p5* and *I3_60* scenarios). There is a high negative correlation between attack power (ϵ) and the performance of models, i.e., a high ϵ increases as the model's performance dramatically decreases. On the other hand, the results show that the proposed mitigation methods, i.e., the iterative adversarial training and defensive distillation approach, successfully increase the RF beamforming prediction performance and create more accurate predictions. The results prove that the proposed framework can enhance the performance of the DL-based beamforming model. In future work, the research team plans to investigate other AI-based solutions used in physical and media access layers of next-generation networks, i.e., channel coding, synchronization, positioning, channel estimations, symbol detection, resource allocation, and scheduling, and their cybersecurity risks.

Acknowledgements This work was supported in part by the Commonwealth Cyber Initiative, an investment in the advancement of cyber R&D, innovation, and workforce development in Virginia. For more information about CCI, visit cyberinitiative.org

Data Availability Dataset used in the manuscript can be found at: <https://deepmimo.net>

Declarations

Conflict of Interest The authors have no conflicts of interest to declare. All co-authors have seen and agreed with the contents of the manuscript. We certify that the submission is original work and is not under review at any other publication.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent Not applicable.

References

- Lichtman, M., Rao, R., Marojevic, V., Reed, J., Jover, R.P.: in 2018 IEEE international conference on communications workshops (ICC workshops) (2018), pp. 1–6. <https://doi.org/10.1109/ICCW.2018.8403769>
- Catak, E., Durak-Ata, L.: Computers & electrical engineering **61**, 184 (2017). <https://doi.org/10.1016/j.compeleceng.2016.11.039>. <https://www.sciencedirect.com/science/article/pii/S0045790616309648>
- Catak, F.O., Kuzlu, M., Catak, E., Cali, U., Unal, D.: Security concerns on machine learning solutions for 6G networks in mmWave beam prediction. Phys. Commun. (2022). <https://doi.org/10.1016/j.phycom.2022.101626>
- Zheng, Z., Wang, L., Zhu, F., Liu, L.: Potential technologies and applications based on deep learning in the 6G networks. Comput. Electric. Eng. **95**, 107373 (2021)
- Liu, G., Huang, Y., Wang, F., Liu, J., Wang, Q.: 5G features from operation perspective and fundamental performance validation by field trial. China Commun. **15**(11), 33 (2018)
- De Alwis, C., Kalla, A., Pham, Q.V., Kumar, P., Dev, K., Hwang, W.J., Liyanage, M.: Survey on 6G frontiers: trends, applications, requirements, technologies and future research. IEEE Open J. Commun. Soc. **2**, 836 (2021)
- Zhang, Z., Xiao, Y., Ma, Z., Xiao, M., Ding, Z., Lei, X., Karagiannis, G.K., Fan, P.: 6G wireless networks: Vision, requirements, architecture, and key technologies. IEEE Vehic. Technol. Magazine **14**(3), 28 (2019)
- Giordani, M., Polese, M., Mezzavilla, M., Rangan, S., Zorzi, M.: Toward 6G networks: Use cases and technologies. IEEE Commun. Magazine **58**(3), 55 (2020)
- Saad, W., Bennis, M., Chen, M.: A vision of 6G wireless systems: Applications, trends, technologies, and open research problems. IEEE network **34**(3), 134 (2019)
- Khan, L.U., Yaqoob, I., Imran, M., Han, Z., Hong, C.S.: Perceptual enhancement of low light images based on two-step noise suppression. IEEE Access **8**, 147029 (2020). <https://doi.org/10.1109/ACCESS.2020.3015289>
- Sheth, K., Patel, K., Shah, H., Tanwar, S., Gupta, R., Kumar, N.: A taxonomy of AI techniques for 6G communication networks. Comput. Commun. **161**, 279 (2020)
- Du, J., Jiang, C., Wang, J., Ren, Y., Debbah, M.: Machine learning for 6G wireless networks: Carrying forward enhanced bandwidth, massive access, and ultrareliable/low-latency service. IEEE Vehic. Technol. Magazine **15**(4), 122 (2020). <https://doi.org/10.1109/MVT.2020.3019650>
- Gui, G., Liu, M., Tang, F., Kato, N., Adachi, F.: 6G: Opening new horizons for integration of comfort, security, and intelligence. IEEE Wire. Commun. **27**(5), 126 (2020). <https://doi.org/10.1109/MWC.001.1900516>

14. Ozpoyraz, B., Dogukan, A.T., Gevez, Y., Altun, U., Basar, E.: Deep learning-aided 6G wireless networks: A comprehensive survey of revolutionary phy architectures (2022)
15. Ali, S., Saad, W., Rajatheva, N., Chang, K., Steinbach, D., Sliwa, B., Wietfeld, C., Mei, K., Shiri, H., Zepernick, H.J., Chu, T.M.C., Ahmad, I., Huusko, J., Suutala, J., Bhadauria, S., Bhatia, V., Mitra, R., Amuru, S., Abbas, R., Shao, B., Capobianco, M., Yu, G., Claes, M., Karvonen, T., Chen, M., Girnyk, M., Malik, H.: 6G white Paper on Machine Learning in Wireless Communication Networks (2020)
16. Yang, H., Alphones, A., Xiong, Z., Niyato, D., Zhao, J., Wu, K.: Artificial-intelligence-enabled intelligent 6G networks. *IEEE Network* **34**(6), 272 (2020). <https://doi.org/10.1109/MNET.011.2000195>
17. Dang, S., Amin, O., Shihada, B., Alouini, M.S.: What should 6G be? *Nat. Electron.* **3**(1), 20 (2020)
18. Porambage, P., Gür, G., Osorio, D.P.M., Liyanage, M., Ylianttila, M.: in Proc. IEEE Joint Eur. Conf. Netw. Commun.(EuCNC) 6G Summit (2021), pp. 1–6
19. Kuzlu, M., Fair, C., Guler, O.: Role of artificial intelligence in the internet of things (IoT) cybersecurity. *Disc. Int. Things* **1**(1), 1 (2021)
20. Siriwardhana, Y., Porambage, P., Liyanage, M., Ylianttila, M.: in Proc. IEEE Joint Eur. Conf. Netw. Commun.(EuCNC) 6G Summit (2021), pp. 1–6
21. Catak, E., Catak, F.O., Moldsvor, A.: in 2021 IEEE International black sea conference on communications and networking (BlackSeaCom) (2021), pp. 1–6. <https://doi.org/10.1109/BlackSeaCom52164.2021.9527756>
22. Tuna, O., Faruk, C., Catak, F., Ozgur, E., Eskin, M., Taner, M.: arXiv e-prints [arXiv:2102.04150](https://arxiv.org/abs/2102.04150) (2021)
23. Bai, T., Luo, J., Zhao, J., Wen, B., Wang, Q.: arXiv e-prints [arXiv:2102.01356](https://arxiv.org/abs/2102.01356) (2021)
24. Vardhan, R.: An ensemble approach for explanation-based adversarial detection. Ph.D. thesis (2021)
25. Michels, F., Uelwer, T., Upschulte, E., Harmeling, S.: arXiv preprint [arXiv:1906.03612](https://arxiv.org/abs/1906.03612) (2019)
26. Lin, Y., Zhao, H., Ma, X., Tu, Y., Wang, M.: Adversarial attacks in modulation recognition with convolutional neural networks. *IEEE Trans. Reliabil.* **70**(1), 389 (2021). <https://doi.org/10.1109/TR.2020.3032744>
27. Jiang, Y., Yin, G., Yuan, Y., Da, Q.: Project gradient descent adversarial attack against multisource remote sensing image scene classification. *Sec. Commun. Net.* **2021** (2021)
28. Fostiropoulos, I., Shbita, B., Marmarelis, M.:
29. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network (2015)
30. Papernot, N., McDaniel, P., Wu, X., Jha, S., Swami, A.: Distillation as a defense to adversarial perturbations against deep neural networks (2016)
31. Alkhateeb, A.: arXiv preprint [arXiv:1902.06435](https://arxiv.org/abs/1902.06435) (2019)
32. Remcom, Wireless InSite. <http://www.remcom.com/wireless-insite>. Accessed: 2021-09-30
33. DeepMIMO, 'O1' scenario. <https://deepmimo.net/scenarios/o1-scenario/>. Accessed: 2021-09-30
34. DeepMIMO, 'I1' scenario. <https://deepmimo.net/scenarios/i1-scenario/>. Accessed: 2021-09-30
35. DeepMIMO, 'I3' scenario. <https://deepmimo.net/scenarios/i3-scenario/>. Accessed: 2021-09-30
36. Andriushchenko, M., Flammarion, N.: arXiv e-prints [arXiv:2007.02617](https://arxiv.org/abs/2007.02617) (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.