IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 February 2024, accepted 11 March 2024, date of publication 18 March 2024, date of current version 25 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3378263

RESEARCH ARTICLE

Security of AI-Driven Beam Selection for Distributed MIMO in an Adversarial Setting

ÖMER FARUK TUNA AND FEHMI EMRE KADAN, (Member, IEEE)
Ericsson Research, Ericsson Turkey, 34467 Istanbul, Turkey

Corresponding author: Ömer Faruk Tuna (omer.tuna@ericsson.com)

This work was supported in part by the Scientific and Technological Research Council of Turkey (TUBITAK) through the 1515 Frontier Research and Development Laboratories Support Program under Project 5169902, and in part by the European Union's Horizon Europe Research and Innovation Program and Smart Networks and Services Joint Undertaking (SNS JU) through AI-powered eVolution towards opEn and secuRe edGe architEctures (VERGE) Project under Grant 101096034.

ABSTRACT In distributed multiple-input multiple-output (D-MIMO) networks, beam selection is necessary to predict the best beam and radio units (RUs) to serve the users in an optimum way. Finding the best RU and beam requires measuring the downlink channel for all possible RU/beam pairs, which becomes a resource-heavy operation, especially at the millimeter Wave band. To overcome this problem, artificial intelligence (AI) solutions are investigated which aim to infer the best RU/beam from sounding the channel for a subset of RUs and beams. While fairly accurate AI models can be obtained, these models have some intrinsic vulnerabilities to adversarial attacks where carefully designed perturbations are applied to the input of the AI model. In this study, we consider four different adversarial attack methods that craft perturbations using gradients of the AI cost function under two different beam reporting scenarios considering sequential and one-shot reporting of reference signal received power values for all RUs and demonstrate their effectiveness over traditional methods by extensive simulations, showing the necessity of smart defense techniques. To this aim, we propose an effective mitigation solution based on scrambling of RUs against these kinds of adversarial attack threats and verify the efficacy of our solution via detailed simulations. The proposed defense method provides up to 10 dB better signal strengths at the user side by selecting more accurate RU/beam pairs under adversarial attacks.

INDEX TERMS Adversarial machine learning, beam selection, cell-free massive MIMO, deep learning, distributed MIMO, security, 6G.

I. INTRODUCTION

Distributed multiple-input multiple-output (D-MIMO), also known as cell-free massive MIMO, is a new network type considered for beyond 5G communication systems where many radio units (RUs) are geographically distributed in a region to increase the coverage and reliability. RUs are connected to a central processor (CP) with high baseband capability to optimize resource allocation and performance through coordination. Previous research studies [1] show the benefits of D-MIMO compared to traditional uncoordinated small-cells, which are used in all previous generations, and collocated massive MIMO which is currently used in 5G systems. D-MIMO is well-suited for high-frequency bands with challenging radio propagation, thanks to its high line-of-sight probability and increased macro diversity [2].

Machine learning (ML) is a subcategory of artificial intelligence (AI) which helps us to abstract and extract knowledge from data and then apply that data algorithmically to make highly accurate predictions. One of the main reasons why machine learning has become such a popular topic in the last decade is the success of a specific category of algorithms known as deep learning that have been immensely successful in various domains such as computer vision and machine translation. Wireless communication has been

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

one of the recent domains that we started to use deep learning algorithms effectively. There are different cases where modern machine learning algorithms can help us, and one possible scenario is where we don't have a good physical model. One example for such a case is when we want to model the traffic pattern in a city where we don't have any useful tool to formulate the problem in the first place. In such a scenario, machine learning models could help us and based on seeing a lot of data, it can draw useful conclusions about the possible trajectories about how the people might move around the city. Another motivation for using machine learning in wireless communication is that there might be some problems for which we know the optimal solution, yet the optimal solution is computationally very expensive. In such a case, one might use machine learning techniques to find a good enough approximation to the original highly complex problem.

Wireless networks must perform demanding tasks such as beamforming [3], [4] with robustness against attackers [5], [6] in a dynamic spectrum environment influenced by traffic, channel, and interference impacts. Deep learning has become an indispensable tool for solving a wide range of variety of wireless communication problems [7]. In a D-MIMO network, there are lots of functions that are needed to be performed in order to orchestrate the network resources to optimize the network and maximize spectral efficiency of the served users. Most of these functions are planned to be handled by AI/ML due to various reasons such as reducing complexity and overhead. One possible scenario where we can use AI/ML is power allocation task where the aim is to efficiently allocate the power of RUs among users to optimize the system performance [8], [9]. The power control is applied by RUs to allocate their transmission power among user equipment (UE) allowing interference management and spectral/energy efficiency enhancement in D-MIMO. Although there is an exact analytical solution for this problem by using sequential second order cone programming, the computational complexity of the solution is extremely high [10] especially when the number of RUs and UEs in the network gets larger. Hence low complexity AI/ML based solutions are proposed to approximate the exact solution. Another scenario where we can use AI/ML techniques is beam selection task where the aim is to find the best beam(s) by sweeping a subset of all possible beams instead of a complete beam sweeping procedure to decrease beam training overhead and energy consumption [11]. We can also use AI to efficiently find user-centric RU sets to serve UEs [12]. Although there are many RUs in a D-MIMO network, using all RUs to serve a single UE might not be practical due to the high fronthaul load. A better approach is to find a suitable subset of RUs to serve each UE, allowing for control of fronthaul loads and providing better energy efficiency. The reason for using AI/ML technique in beam/RU selection task is that there is a complex dependence to the positions of the RUs and UEs, and hence developing an analytical solution is very challenging. In summary, there is a need for AI functionality at the CP to take care of many important tasks.

Although AI-empowered solutions can be efficiently used to solve many tasks in D-MIMO, adversarial attacks pose severe security concerns for AI-driven systems. Very small and undetectable changes in input data samples might be sufficient to trick the most advanced classifiers resulting in inaccurate predictions. To understand the effects of adversarial attacks on D-MIMO networks, performance of different attack methods should be evaluated and efficient defense methods should be developed to mitigate the potential disruptive effects. All potential scenarios and parameter sets should be carefully considered to effectively use developed techniques in practical networks.

A. RELATED WORK

AI/ML models have been discovered to be vulnerable to malicious attacks [13]. Very small and often undetectable changes in input data samples are enough to fool state-of-the-art classifiers in inference time and lead to incorrect predictions. In the past few years, extensive research studies have shown the vulnerability of AI-driven systems in different domains. Despite the distributed nature of the communication domain and the heterogeneity of the network, there is still a risk of adversarial attacks in the telecommunications domain. Previous research [14] indicates that adversarial attacks with optimized perturbations can target AI-driven tasks such as power control and degrade the performance of a D-MIMO network in terms of both spectral and energy efficiency. In [15], it is shown that adversarial attacks on power control in the training stage can ruin the performance of the AI model with supervised learning. In [16], the authors analyze the effects of over-the-air adversarial attacks on AI-driven beam selection in collocated massive MIMO systems where all beams are swept by a single base station. In [17], the authors investigate the effects of poisoning attacks implemented at the training stage for beam selection in collocated massive MIMO and propose a machine unlearning method to mitigate the effects of adversaries. Finally, [18] studies the effects of blackbox (without knowledge of the true AI model at the victim) adversarial attacks on beam selection for various use cases again for collocated massive MIMO. There are different defense solutions proposed in the literature for adversarial attacks. The most popular ones are adversarial training [19], [20], [21], [22], [23], ensemble adversarial training [24], defensive distillation [25], [26], [27], [28], squeezed models [29], and auto encoder-based input denoising [30], [31]. However, there is no existing adversarial defense mechanism that achieves both efficiency and effectiveness against adversarial samples [32].

B. MOTIVATION

Although there are existing adversarial attack and defense frameworks for wireless networks using collocated MIMO, it is not straightforward to extend the related methods to

D-MIMO systems due to their special physical and transport layer architecture, and specific areas such as beam/RU selection. Existing studies for beam/RU selection task in D-MIMO only covers developing AI techniques without any attack/defense consideration. To the best of our knowledge, there is not any related prior work investigating attack and defense techniques for beam/RU selection in D-MIMO networks.

C. CONTRIBUTIONS

In this study, we focus on the AI-driven beam selection task in D-MIMO for the initial access stage where a small number of beams are transmitted to find the best possible beam with the help of AI. We consider a scenario where malicious software infects a UE and reads/modifies measurements of the UE modem to perturb the measurement reports to be sent back to the network. The contributions of the paper is listed below:

- We first investigate the efficiency of various adversarial attack methods on the beam/RU selection performance of the D-MIMO network. We propose four different attack approaches for crafting adversarial perturbations, two of which are effective even only partial channel knowledge is available at the attacker side. We consider both the whitebox setting where the attacker has full access to the original AI model and the blackbox setting where the attacker trains a surrogate model to imitate the functionality of the original AI model. The former approach is considered to see the performance upper bound whereas the latter one shows more practical results. We empirically show that malicious software in a compromised UE can easily target the beam selection model in the network and result in dramatic performance decreases, even under the most constrained circumstances.
- Then, we propose a defense mechanism relying on a scrambling operation prior to feeding the input to the AI model to mitigate those adversarial attacks. We perform detailed simulations for various attack types and UE reporting scenarios to show the effectiveness of the proposed technique. We also provide a practical solution to use different scramblers in different D-MIMO sites in a secure way.

The paper is structured as follows: Section II goes over the system model together with the details of AI-driven beam selection task in D-MIMO, some of the widely known adversarial attack types, available defense techniques in the literature and the details of our proposed defense solution. In Section III, we provide the results of our comprehensive simulations for different adversarial attack scenarios, and we conclude the paper in Section IV.

II. SYSTEM MODEL

We consider a D-MIMO network with P RUs each with B antennas and Q UEs each with A antennas, and we assume that all UEs are served by all (or some subset of) RUs that are connected to a CP via fronthaul links. Fig. 1 shows an example D-MIMO network.

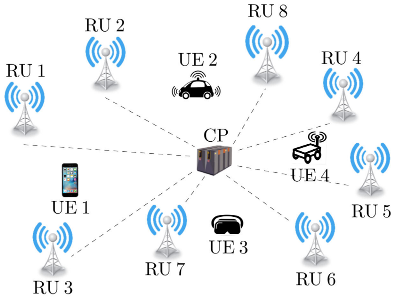


FIGURE 1. A typical D-MIMO network with 8 RUs and 4 UEs.

In a D-MIMO network, as in current 5G networks, RUs should transmit some broadcast signals (such as synchronization signal blocks, SSBs) for initial access of UEs to the system. However, as there are many RUs each with some number of beams to be transmitted, the total number of SSBs will be very high. This causes beam sweeping overhead to be large which is not desired as it will increase the latency of UEs to connect to the system. Besides, this potentially will cause a huge energy consumption due to a high number of transmissions and measurements. An example of a full beam sweeping approach is given in Fig. 2, where there are P RUs each having B SSB beams, resulting in PB beams in total. Herein, all beams from all RUs scan the coverage area sequentially, where only one beam is active in a given time interval.

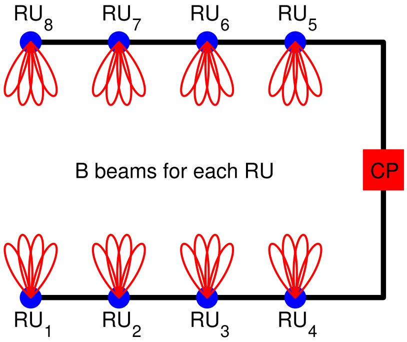


FIGURE 2. Sequential full beam sweeping in a D-MIMO network. In this example, there are $P = 8$ RUs each with $B = 4$ beams making 32 beams in total.

As the common information is transmitted by SSBs, all UEs inside the coverage region receive the same data and hence we will only focus on a single UE. The received signal $\mathbf{y}_{p,b} \in \mathbb{C}^{A\times 1}$ by the UE from the b-th beam of the p-th RU can be written as

$$\mathbf{y}_{p,b} = \mathbf{H}_p \mathbf{w}_{p,b} s + \mathbf{n}_{p,b}, \quad p = 1, 2, \ldots, P, \quad b = 1, 2, \ldots, B, \tag{1}$$

where $\mathbf{H}_p \in \mathbb{C}^{A\times B}$ is the channel matrix and $\mathbf{w}_{p,b} \in \mathbb{C}^{B\times 1}$ is the corresponding beam vector, s is the unit-norm scalar reference symbol (primary/secondary synchronization sequence, PSS/SSS, or demodulation reference symbol, DMRS) in the SSB signal which is the same for all RUs and beams, and $\mathbf{n}_{p,b} \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2\mathbf{I})$ is the internal noise of the UE receiver. The UE uses the reference symbol s to measure the reference signal received power (RSRP) values for each beam transmitted by RUs using a combiner[1] $\mathbf{u} \in \mathbb{C}^{A\times 1}$

$$\mathbf{u}^H\mathbf{y}_{p,b} = \mathbf{u}^H\mathbf{H}_p\mathbf{w}_{p,b}s + \mathbf{u}^H\mathbf{n}_{p,b}, \quad \forall p, b. \tag{2}$$

The UE can measure the RSRP information by estimating the coefficient $c_{p,b} = \mathbf{u}^H\mathbf{H}_p\mathbf{w}_{p,b}$ using the equation (2) as

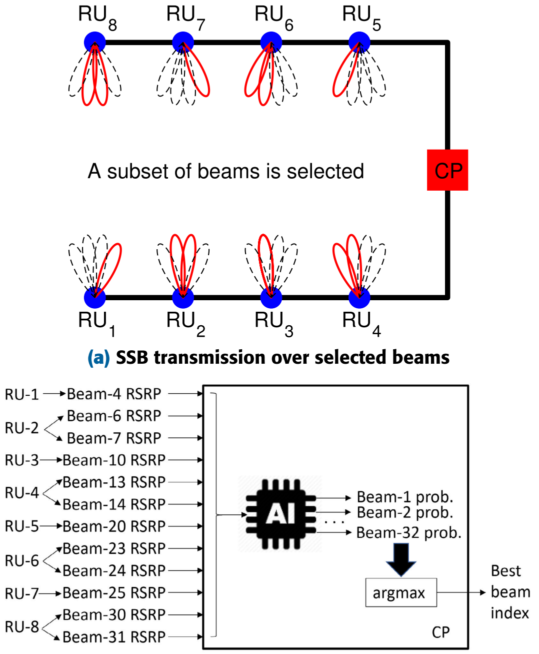
$$\widehat{c}_{p,b} = \mathbf{u}^H\mathbf{y}_{p,b}s^* = \mathbf{u}^H\mathbf{H}_p\mathbf{w}_{p,b} + \mathbf{u}^H\mathbf{n}_{p,b}s^*, \quad \forall p, b, \tag{3}$$

where $\widehat{c}_{p,b}$ is the least-squares estimation of $c_{p,b}$ and s^* is the complex conjugate of s. The coefficient $\widehat{c}_{p,b}$ involves both RSRP (derived from $|\widehat{c}_{p,b}|^2$) and delay (derived from $\angle\widehat{c}_{p,b}$) information of the augmented channel. Due to the transformed noise term $\mathbf{u}^H\mathbf{n}_{p,b}s^*$ in (3), the estimate $\widehat{c}_{p,b}$ includes an additive Gaussian noise with variance $\sigma_c^2 = \sigma_n^2\mathbf{u}^H\mathbf{u}$.

After measuring the RSRP values, the UE sends these values back to the network using uplink control channels so that the CP can collect the RSRP values from RUs and select the best RU and beam pair to serve the corresponding UE.

A. AI-DRIVEN BEAM SELECTION

To find the best beam ($\mathbf{w}_{p,b}$ vector) to be used to transmit data in the downlink, the p-th RU should know the augmented channel $\mathbf{u}^H\mathbf{H}_p$ whose dimension is $1 \times B$. This can be done by transmitting B different beams with independent $\mathbf{w}_{p,b}$ vectors and getting $\widehat{c}_{p,b}$ feedback for each of those beams. Considering all RUs, this process requires PB different beams to be transmitted resulting in a large overhead and high energy consumption. To mitigate this problem, an AI-driven beam selection process is proposed in the literature [16], [17], [18]. In this approach, a subset of all possible beams is selected and only the selected beams are sequentially sent to the UEs. The UEs measure the corresponding RSRP values of each selected beam and send the measurements back to the network in the uplink. An AI model $f(\cdot, \boldsymbol{\theta})$ with model weights $\boldsymbol{\theta}$, whose output is a $PB \times 1$ vector of probabilities of each beam, is trained and used in the network to find the best beam using the measured RSRP values. The AI solution is proven to be effective in finding the best beam index even if it is among the ones which are not selected to send SSBs. In Fig. 3, an example block diagram of this solution is given. In this example, there are $P = 8$ RUs each with 4 beams. To decrease the total number of beams transmitted, the network chooses only 12 of beams indicated in Fig. 3a. Beams are numbered from 1 to 32 in the counterclockwise order from RU 1 to RU 8. The UE measures the corresponding RSRP values and reports the measurements back to the network. The RSRP measurements are collected by the CP and fed to an AI model to generate the probability of each beam being the best one. The final beam decision is made by selecting the beam index with the highest probability.



(a) SSB transmission over selected beams

(b) AI execution with reported RSRPs

FIGURE 3. In this example, there are $M = 12$ beams (indicated by red solid lines in (a) out of total $PB = 32$ chosen to transmit SSBs to UEs. The trained AI model is used to predict the best beam index using the measured RSRPs for the selected M beams transmitted from $P = 8$ RUs.

B. ADVERSARIAL ATTACKS

The goal of adversarial attacks in classification tasks is to craft a perturbation $\boldsymbol{\delta}$ under given constraints (such as $|[\boldsymbol{\delta}]_i| < \epsilon, \ \forall i$) which yields to an incorrect prediction as $y_{\text{adv}} = \operatorname{argmax} f(\mathbf{x} + \boldsymbol{\delta}, \boldsymbol{\theta})$ which differs from a prediction on a clean sample $y = \operatorname{argmax} f(\mathbf{x}, \boldsymbol{\theta})$. The success criteria of the attack might change depending on the type of task. The attack can be considered successful if the model predicts a class other than the actual class. For an AI-driven beam selection task, the objective of the attacker can be to force the target model to predict a beam index that is different than the best beam, which falls under the category of untargeted attacks. However, the motivation of the attacker might also be to fool the model to predict as worst option possible to lower the network performance, which can be considered as a targeted attack. Besides, in a practical scenario as in most of the wireless tasks, the attacker will not be able to have complete knowledge of the target AI model, including its architecture and weights, and therefore the attacker needs to craft perturbations in a blackbox manner by training and using a surrogate AI model that imitates the original one.

One of the most popular adversarial attacks in the literature is the Fast Gradient-Sign Method (FGSM) [19] which utilizes the derivative of the model's loss function with respect to the input sample to decide in which direction the feature values of the input vector should be altered to minimize the objective loss function of the model. As soon as this direction is found, it perturbs all features simultaneously in the opposite direction to maximize the loss. Then, Kurakin et al. [33] suggested a small yet effective improvement to the FGSM,

[1] We assume that the combiner is the same for all RUs and SSB beams.

known as the Basic Iterative Method (BIM). In this attack method, instead of taking one step of size ϵ in the gradient sign's direction, the attacker takes several but smaller steps α and uses the given ϵ value to clip the result. Crafting perturbations under ℓ_∞ norm for BIM attack is given by (4).

$$\begin{aligned} \mathbf{x}_1^* &= \mathbf{0} \quad \text{and for all } j = 1, 2, \ldots, j_{\max} \\ \mathbf{x}_{j+1}^* &= \text{Clip}_\epsilon \left\{ \mathbf{x}_j^* + \alpha \cdot \text{sign}\left(\nabla_{\mathbf{x}} \left(L(\mathbf{x}_j^*, \boldsymbol{\theta}, y_{\text{true}}) \right) \right) \right\} \\ y_{\text{true}} &= \operatorname{argmax} f(\mathbf{x}, \boldsymbol{\theta}) \end{aligned} \tag{4}$$

where $\mathbf{x}_j^*$ is the crafted perturbation at j-th iteration, $\boldsymbol{\theta}$ represents the parameters of the AI model, L is the loss function, $\nabla_{\mathbf{x}}$ shows the derivative with respect to the input vector $\mathbf{x}$, ϵ is a tunable parameter, limiting the maximum level of perturbation for ℓ_∞ norm, α is the step size, y_{true} is the actual class which can be estimated by running the AI model and finding the most probable class and $\text{Clip}_\epsilon\{\cdot\}$ is the clipping operator that clips entries of the argument larger than ϵ to ϵ and smaller than $-\epsilon$ to $-\epsilon$. The final perturbation vector can be found after $j_{\max}$ iterations or at the stage where the rate of change of $\mathbf{x}_j^*$ becomes small enough.

We can easily modify (4) to produce a targeted variant of BIM. At each intermediate step, we can try to minimize the loss with respect to the target class while at the same time maximizing the loss with respect to the original class as in (5).

$$\begin{aligned} \mathbf{x}_1^* &= \mathbf{0} \quad \text{and for all } j = 1, 2, \ldots, j_{\max} \\ \mathbf{x}_{j+1}^* &= \text{Clip}_\epsilon \Big\{ \mathbf{x}_j + \alpha \cdot \text{sign}\Big(\nabla_{\mathbf{x}} \big(L(\mathbf{x}_j{}^*, \boldsymbol{\theta}, y_{\text{true}}) \\ &\qquad - L(\mathbf{x}_j^*, \boldsymbol{\theta}, y_{\text{target}}) \big) \Big) \Big\} \\ y_{\text{true}} &= \operatorname{argmax} f(\mathbf{x}, \boldsymbol{\theta}), \quad y_{\text{target}} = \operatorname{argmin} f(\mathbf{x}, \boldsymbol{\theta}), \end{aligned} \tag{5}$$

where y_{target} is the targeted class which can be estimated by running the AI model and finding the least probable class. It should be noted that perturbations with the above attack algorithms are specific to a particular sample. So, another research direction is to find perturbations that when added to "any" input sample can fool the target model. For this, the universal adversarial perturbation (UAP) method [34] is proposed for cases where complete input knowledge is not available. As a final remark, in case the original AI model $f(\cdot, \boldsymbol{\theta})$ is not available at the attacker side, a well-trained surrogate model $\tilde{f}(\cdot, \boldsymbol{\theta})$ can be used to make the calculations in (4) and (5).

C. ADVERSARIAL ATTACK SCENARIOS IN D-MIMO

Current UE modems are vulnerable to malicious software [35]. A highly capable malware can settle down in a UE and read/modify RSRP values measured by the UE modem. The malware can train a local AI model that imitates the actual AI model used in the CP to carefully craft adversarial perturbations. In Fig. 4, a typical attack scenario is shown.

The malware infecting the UE can read the measured SSB RSRP values from the UE modem in a period to collect sufficient RSRP data and it can learn the labels (the best beam index) by observing the RSRP values for data channels. In this way, it can train a surrogate model that imitates the original one.

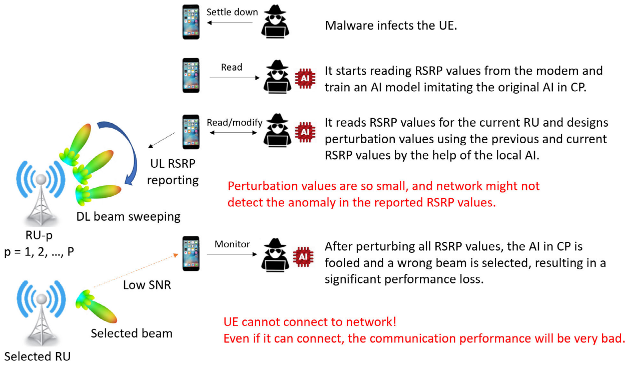

Settle down
Malware infects the UE.
Read
AI
It starts reading RSRP values from the modem and train an AI model imitating the original AI in CP.
Read/modify
AI
It reads RSRP values for the current RU and designs perturbation values using the previous and current RSRP values by the help of the local AI.
UL RSRP reporting
DL beam sweeping
RU-p
p = 1, 2, ..., P
Perturbation values are so small, and network might not detect the anomaly in the reported RSRP values.
Monitor
AI
After perturbing all RSRP values, the AI in CP is fooled and a wrong beam is selected, resulting in a significant performance loss.
Low SNR
Selected beam
Selected RU
UE cannot connect to network!
Even if it can connect, the communication performance will be very bad.

FIGURE 4. A block diagram of a typical adversarial attack scenario where the attacker aims to fool the target AI model.

We consider two different attack scenarios under two different settings:

Scenario 1: In this scenario, the RSRP reports are sent in uplink so that in each report only the RSRP values corresponding to the beams of the current RU are included. This is compatible with what is in the current 5G NR implementation. Due to this assumption, the attacker crafts the adversarial samples with missing information (to craft adversarial samples for the p-th RU, only the RSRP values for $1, 2, \ldots, p$-th RU beams can be used.) As the best perturbation calculation requires all RSRP information for all RUs one at a time, we assume that the attacker uses the UAP technique, allowing to craft perturbations with limited knowledge.

Scenario 2: In this scenario, all the RSRP reports are sent in one shot for all RSRP information of all RUs. In this case, since the attacker will know all the parts of the AI model's input, there won't be any need to apply UAP and the attacker can use existing attack algorithms, such as BIM.

Setting 1, Whitebox: In the whitebox setting, we assume that the attacker perfectly knows the original AI model implemented at the CP. This might not be realistic scenario but it can be used to observe the theoretical performance limit.

Setting 2, Blackbox: In the blackbox setting, we assume that the attacker has no access to the actual AI model in the CP. So, the attacker trains a separate surrogate model imitating the actual AI model's functionality, and crafts the perturbations accordingly. The training is performed as shown in Fig. 4.

To make the attack practical, we assume the following:

1) The attacker designs a perturbation so that for each RSRP value, the added perturbation has a magnitude of at most ϵ dB, where ϵ is a small quantity. (It should be smaller than typical large-scale channel variations, such as shadowing standard deviation.) This is required for attackers not to be detected by the network.
2) Due to uplink channel effects, the measured RSRP values might not be perfectly delivered to the network. We add some noise to the RSRP values (in both

clean and perturbed cases) to model this effect at both training and inference stages.

D. PROPOSED ATTACK ALGORITHMS

We consider two different attack algorithms for each of the two scenarios to see their effects on the beam selection performance of the network. The attacker uses the original $f(\cdot, \boldsymbol{\theta})$ or a surrogate $\widetilde{f}(\cdot, \boldsymbol{\theta})$ AI model according to the whitebox/blackbox setting. All algorithms use the RSRP measurements $r_{p,b} = |\widehat{c}_{p,b}|^2$ and the given perturbation amount ϵ to derive perturbation values $\delta_{p,b}$ satisfying $|\delta_{p,b}| \leq \epsilon, \quad \forall p, b$. For all methods, we use the cross entropy loss function as the loss function L.

1) UAP-BASED ATTACK FOR SCENARIO 1

For Scenario 1, only some portion of the input of AI is known at any RSRP measurement stage at the UE side. We assume that the RU, beam pairs used to transmit SSBs form the set $\mathcal{B}$, where $|\mathcal{B}| = M$, and after the p_0-th RU SSB transmission stage, the UE and hence the attacker, has only the knowledge of RSRP values for $r_{p,b}, \quad (p, b) \in \mathcal{B}, \ p \leq p_0$. A UAP-based algorithm is proposed to craft perturbation values by filling out the unknown input values (for $p > p_0$) by drawing some typical input vectors $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N$ from the attacker's set of collected data. (Here as shown in Fig. 4, the attacker monitors the RSRP measurements done by the UE modem and collects RSRP data for each transmitted SSB beam. The UAP method draws N measurement vectors $r_1, r_2, \ldots, r_N$ (each with dimension $M \times 1$) from the stored data set and updates the known part (for RUs with $p \leq p_0$) of each vector by their known values. By this way, the UAP method creates N different potential AI inputs whose known parts are the same, but the unknown parts include different measurements corresponding to random UE locations drawn from the past measurements.) The final perturbations are calculated using principal component analysis (PCA) that finds the principal singular vector of the matrix of possible perturbation vectors. The detailed algorithm is presented in Algorithm 1.

Algorithm 1 generates perturbation values for the p_0-th RU after it sends its SSB beams to the UE. We run Algorithm 1 for each $p_0 = 1, 2, \ldots, P$ to design all perturbation values sequentially. In Stage 6, as $-\mathbf{v}_1$ is also a valid right singular vector as $\mathbf{v}_1$ corresponding to the same singular value of $\mathbf{R}$, we check both $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ to find the best perturbation vector.

2) GENERATIVE AI (GAI)-BASED ATTACK FOR SCENARIO 1

This method is also proposed for Scenario 1. In this case, the attacker trains different AI models to predict the RSRP values for $p > p_0$ using the RSRP measurements obtained for $p \leq p_0$. A different predictive AI model $g_{p_0}(\cdot, \boldsymbol{\theta})$ is trained for each p_0 value (considering whitebox or blackbox setting) as shown in Fig. 5.

After estimating the unknown RSRP values, the BIM method is used to design the perturbation vector. The detailed algorithm is presented in Algorithm 2.

Algorithm 1 UAP-Based Attack for Scenario 1

```
Input: $r_{p,b}, \ (p, b) \in \mathcal{B}, \ p \leq p_0, \ \epsilon$
Data: $\{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N\}$ and $g(\cdot, \boldsymbol{\theta}) = f(\cdot, \boldsymbol{\theta})$ or $\widetilde{f}(\cdot, \boldsymbol{\theta})$
Output: $\delta_{p,b}, \ (p, b) \in \mathcal{B}, \ p = p_0$
1 Form a column vector $\mathbf{r}$ by augmenting the input $\mathbf{r}_{p,b}$ RSRP values.
2 Define a matrix $\widehat{\mathbf{R}} \in \mathbb{C}^{N \times M}$ using $\mathbf{r}$ and the vectors
    $\{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N\}$ taken from the data set. Firstly, initialize the
    matrix $\widehat{\mathbf{R}}$ as $\widehat{\mathbf{R}} = [\mathbf{r}_1 \ \mathbf{r}_2 \ \cdots \mathbf{r}_N]^T$. Then update $\widehat{\mathbf{R}}$ using the known
    $\mathbf{r}_{p,b}$ values, i.e., $[\widehat{\mathbf{R}}]_{i,j} = [\mathbf{r}]_j$ for all known term indices $j$ and for
    all $i = 1, 2, \ldots, N$.
3 For each row $\widehat{\mathbf{r}}_i^T$ of $\widehat{\mathbf{R}}$, run the AI model $g(\cdot, \boldsymbol{\theta})$ and find an estimate
    of the best beam index (true class) by $y_{\text{true},i} = \text{argmax } g(\widehat{\mathbf{r}}_i^T, \boldsymbol{\theta})$.
4 For each row $\widehat{\mathbf{r}}_i^T$ of $\widehat{\mathbf{R}}$ and the estimated true class $y_{\text{true},i}$, apply the
    BIM in (4) to generate the matrix
    $\boldsymbol{\Delta} \in \mathbb{C}^{N \times M} = [\boldsymbol{\rho}_{\widehat{\mathbf{r}}_1}, \boldsymbol{\rho}_{\widehat{\mathbf{r}}_2}, \ldots, \boldsymbol{\rho}_{\widehat{\mathbf{r}}_N}]^T$ of perturbation vectors.
5 Compute the principal right singular vector $\mathbf{v}_1$ of $\mathbf{R}$ as
    $\boldsymbol{\Delta} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^H$, and $\mathbf{v}_1$ is the first column of $\mathbf{V}$.
6 Compute two perturbations $\boldsymbol{\delta}_1 = \epsilon \cdot \text{sign}(\mathbf{v}_1), \ \boldsymbol{\delta}_2 = -\boldsymbol{\delta}_1$ and
    calculate total losses corresponding to the perturbed inputs, i.e.,
    $L_u = \sum_{i=1}^{N} L(\widehat{\mathbf{r}}_i + \boldsymbol{\delta}_u, \boldsymbol{\theta}, y_{\text{true},i})$ for $u = 1, 2$. Find the index
    $u_0 \in \{1, 2\}$ such that $u_0 = \underset{u \in \{1,2\}}{\text{argmax}} \ L_u$.
7 $\boldsymbol{\delta} = \boldsymbol{\delta}_{u_0}$.
8 Choose the elements of $\boldsymbol{\delta}$ corresponding to the beams of $p_0$-th RU
    to determine $\delta_{p,b}$ values for beam pairs with $(p, b) \in \mathcal{B}, \ p = p_0$.
9 return $\delta_{p,b}, \ (p, b) \in \mathcal{B}, \ p = p_0$
```

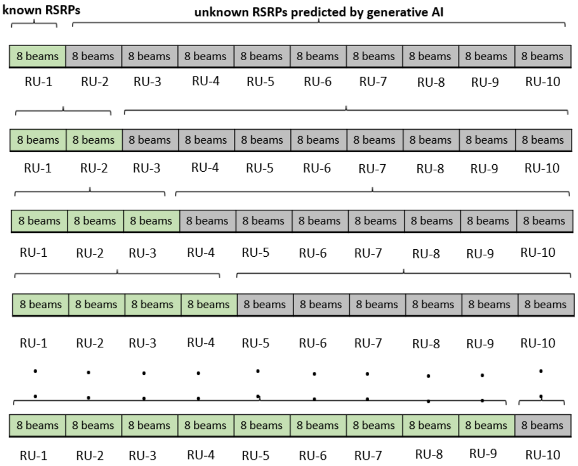


FIGURE 5. Predicting missing RSRP values via generative AI models. In this example, $B = 8$ beams are transmitted from each of $P = 10$ RUs.

Algorithm 2 GAI-Based Attack for Scenario 1

```
Input: $r_{p,b}, \ (p, b) \in \mathcal{B}, \ p \leq p_0, \ \epsilon$
Data: $g_{p_0}(\cdot, \boldsymbol{\theta})$ and $g(\cdot, \boldsymbol{\theta}) = f(\cdot, \boldsymbol{\theta})$ or $\widetilde{f}(\cdot, \boldsymbol{\theta})$
Output: $\delta_{p,b}, \ (p, b) \in \mathcal{B}, \ p = p_0$
1 Using the input RSRP values and the predictive AI model $g_{p_0}(\cdot, \boldsymbol{\theta})$
    estimate the unknown RSRPs and generate the full RSRP vector
    $\mathbf{r} \in \mathbb{C}^{M \times 1}$.
2 Run the AI model $g(\cdot, \boldsymbol{\theta})$ and find an estimate of the best beam
    index (true class) by $y_{\text{true}} = \text{argmax } g(\widehat{\mathbf{r}}, \boldsymbol{\theta})$.
3 Using the vector $\mathbf{r}$ and the estimated class $y_{\text{true}}$, apply the BIM
    method in (4) to design the perturbation vector $\boldsymbol{\delta}$.
4 Choose the elements of $\boldsymbol{\delta}$ corresponding to the beams of $p_0$-th RU
    to determine $\delta_{p,b}$ values for beam pairs with $(p, b) \in \mathcal{B}, \ p = p_0$.
5 return $\delta_{p,b}, \ (p, b) \in \mathcal{B}, \ p = p_0$
```

As in Algorithm 1, the attacker iteratively runs Algorithm 2 for each $p_0 = 1, 2, \ldots, P$ to design all perturbation values.

3) ATTACKS FOR SCENARIO 2

For Scenario 2 where all RSRP measurements are available at one-shot at the UE and attacker side, the attacker can directly design the perturbation values using the BIM as in (4) or targeted BIM (T-BIM) as in (5).

E. PROBLEMS WITH EXISTING DEFENSE SOLUTIONS

Adversarial training is one of the most widely used defense technique [19], [20] against adversarial attacks. The aim of adversarial training is to augment training data with previously crafted adversarial samples and then use this augmented data for training in order to increase robustness. However, there are important drawbacks of adversarial training some of which are listed below:

- First of all, adversarial training procedure necessitates applying known adversarial attacks to whole training data and this is a computationally expensive process. Even so, it still does not provide perfect robustness to adversarial attacks. This means that adversarially trained models might still suffer some degree of performance problems when facing adversarial samples [22].
- Secondly, there is a tradeoff between the performance and robustness when using adversarial training method [23]. It means that applying adversarial training has a negative impact on the natural (clean) performance of the model [21].
- It is known that adversarially trained models can be made robust to whitebox attacks if the perturbations computed during training closely maximize the model's loss. However, it is shown that adversarially trained models are still vulnerable to blackbox adversarial attacks which are the practical attack scenarios in wireless communications. As a solution to mitigate blackbox adversarial attacks, ensemble adversarial training is proposed [24]. This method augments model's training data with adversarial examples crafted on other static pre-trained models. By doing this, it decouples adversarial example generation from the learned model's parameters and enhances the diversity of perturbations observed during training. However, ensemble adversarial training is an extremely expensive process, as one needs to train tens of different models using adversarial training and still not provide a perfect performance.

Another commonly used technique is defensive distillation method but the main problem with this method is that it does not work for blackbox attacks [28]. Other techniques such as squeezed models [29], and auto encoder-based input denoising [30], [31] cannot be directly used in attack scenarios considered in this study.

As a result, we need to find an alternative defense solution that is both effective and practical while not suffering from any of the aforementioned shortcomings.

F. PROPOSED DEFENSE SOLUTION

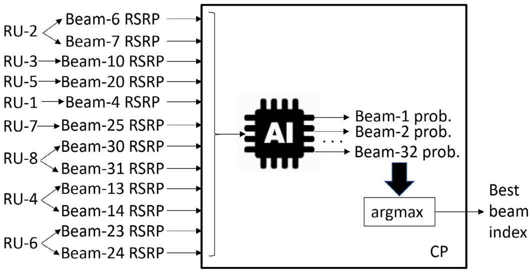


FIGURE 6. An example AI input ordering after scrambling. Here the new order of RUs is 2, 3, 5, 1, 7, 8, 4, 6.

To mitigate the adversarial effectiveness of the attacker that attacks over UEs by modifying RSRP measurement reports, we propose a defensive solution that significantly reduces the attack's success without degrading clean (natural) performance. Our solution is based on permuting the order of RUs during both AI model training and inference phases. Given its characteristics, we name this proposed technique as the scrambling. The success of our proposed solution lies in the effectiveness of hiding the actual model from potential intruders, thereby preventing anyone from training a surrogate model which can infer the exact input/output relationship. Since the scrambling operation is not known by the attacker, the original AI input ordering will be different than that of the surrogate AI model used by the attacker, resulting in degradation in the attack performance. In Fig. 6, we present an example scrambling operation where we use the same setting as in Fig. 3 with the new RU ordering 2, 3, 5, 1, 7, 8, 4, 6. Notice that a new AI should be trained together with each new scrambler to be used.

The following facts allow us to efficiently use scramblers where the last one describes a potential limitation:

Fact 1: The beam selection problem does not include a strong correlation between different RUs. Therefore, when the RU ordering is changed and a new AI is trained accordingly, we do not observe any significant performance change compared to the case with no reordering when there is no attack. Notice that a similar operation would not be possible in other domains like image, text, or audio processing as the scrambling of the input will completely change the semantics of the input and break the strong correlation between neighbor input points, thus causing a poor performance model.

Fact 2: The main motivation of an adversarial attack is to determine sensitive points of the AI function where the rate of change of the objective loss function with respect to the input is large. The attacker calculates the perturbation values to push the input towards those sensitive points to degrade the performance of the AI. The sensitive points depend on the input ordering and when a scrambler is applied, their locations also change. Therefore, without the knowledge of the new ordering, the attacker cannot determine new sensitive points in the AI function and hence the attack becomes ineffective.

Fact 3: The computational complexity of the scrambling operation is very low as it only permutes the ordering of RUs.

Fact 4: There are $P!$ different permutations of RUs and the probability of choosing the correct scrambler will be $1/P!$ for

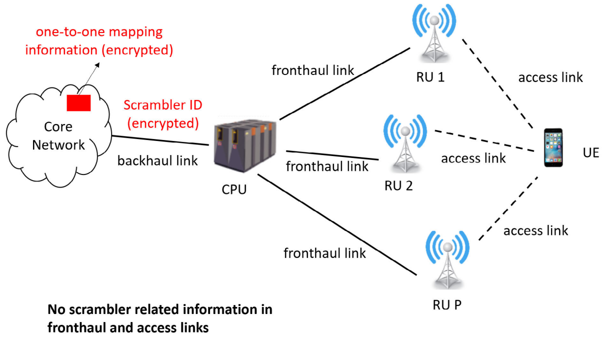


FIGURE 7. Protection of the scrambler information.

the attacker. For example, assuming $P \geq 10$, the probability will be less than one in a million, which is extremely low. To make sure to find the correct scrambler, the attacker should train $P!$ different surrogate AI models which are impractical.

Fact 5: If it is required to implement the proposed solution to previously deployed models that are already functioning in any D-MIMO site, then a retraining operation will be needed based on the assigned scrambling pattern.

For this method to be successful against adversarial attacks, the scrambling operation should be protected and not known/extracted by the attacker. For this purpose, we propose to use different scrambling patterns in each D-MIMO site (D-MIMO site refers to a separate D-MIMO network set up in a different geographic area. Each D-MIMO site consists of a set of RUs which are cooperatively serving multiple users whereas different D-MIMO sites do not coordinate.) to increase security. Furthermore, the scrambling information for each D-MIMO site can be generated at the core network, and the corresponding scrambler information can be transmitted to the related D-MIMO site via backhaul links with effective encryption techniques. This approach protects the scrambler information as no scrambler information is shared in fronthaul (CP-RU) and access (RU-UE) links.

To have a site-specific scrambler, a related function can be placed in the core network that generates different scramblers and uses D-MIMO site ID (which can be defined as similar to physical cell identifier in current 5G networks which is determined by the primary and the secondary synchronization signal indices) to map those scramblers to different sites. This function can first generate a codebook of scramblers and map each scrambler to a specific site in a one-to-one manner. Finally, the related scrambling operation, which can be seen as a permutation of $\{1, 2, \ldots, P\}$ and labeled by a scrambler ID, can be encrypted and sent to the related D-MIMO site via backhaul links. Here we use an ID for each permutation to decrease the communication overhead in the backhaul links. In Fig. 7, we present the block diagram of the solution proposed to protect the scrambler information. Here, the one-to-one mapper in the core network matches scrambler IDs with D-MIMO site IDs. The encrypted scrambler ID information is sent to related D-MIMO sites via backhaul links and a related AI model is trained to be used together with the related scrambler at the CP.

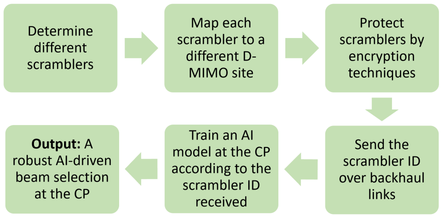

Determine different scramblers
Map each scrambler to a different D-MIMO site
Protect scramblers by encryption techniques
Send the scrambler ID over backhaul links
Train an AI model at the CP according to the scrambler ID received
Output: A robust AI-driven beam selection at the CP

FIGURE 8. Flowchart of the proposed defense solution.

In our proposed method, for each D-MIMO site, there needs to be one site-specific AI model for the beam selection task and that model needs to be trained by using site-specific data. A flowchart summarizing the main steps for our proposal is shown in Figure 8.

To summarize, we propose to use a scrambler before the AI model to shuffle the places of RUs as shown in Fig. 6 so that the attacker cannot design an effective perturbation to fool the AI model. When the scrambler operation is protected, it will not be possible for attackers to design effective adversarial perturbations eliminating the need to employ previously mentioned complex adversarial defense techniques.

Table 1 compares the proposed method with two benchmarks methods. As indicated by Table 1, the proposed method provides a robust solution with low beam sweeping overhead.

III. NUMERICAL RESULTS

To show the effects of the adversarial attack methods UAP and GAI for Scenario 1 and BIM and T-BIM for Scenario 2 under whitebox/blackbox settings, we perform various numerical simulations. As a baseline technique, we also consider white Gaussian noise (WGN) attack where the attacker injects a white Gaussian noise into the input with standard deviation ϵ dB. For all four adversarial attacks, the perturbation value for each RSRP value is in $[-\epsilon, \epsilon]$ dB. We only consider small enough ϵ values (compared to typical shadowing standard deviation values causing natural fluctuations in the RSRP values) for the attacker not to be able to detected by the network. We place $P = 10$ RUs each with $B = 32$ antennas in a region given in Fig. 9 and the single-antenna UE position is randomly chosen inside this region. We assume that each RU sends 8 SSB beams as shown in Fig. 9 to decrease the transmission overhead by 75 percent compared to the full transmission of $PB = 320$ beams. Here the selection of P, B and M (the number of total beams to be sent which is 80 here) and the directions of SSB beams in Fig. 9 are arbitrary. We select $M = 80$ to be significantly less than $PB = 320$ to focus on a case where AI can reduce the overhead significantly. One can also consider different values and beam directions to see the effect on performance.

All simulation parameters are given in Table 2. All RSRP data is generated in MATLAB using the simulation scenario given by Fig. 9 according to the channel models provided by 3GPP TR 38.901 [36].

We use the original and surrogate AI models shown in Fig. 10 for simulations. The surrogate model is chosen as a smaller deep neural network (DNN) model compared to

TABLE 1. Comparison with two benchmarks.

Method	Description	Overhead	Robustness
Sequential full beam sweeping	P RUs sequentially sweep B SSB beams and the best beam is selected according to the maximum RSRP. No AI is used.	PB beam transmission/reporting	Robust to attacks as all potential beams are sent and there is no AI-based RSRP prediction.
AI-based beam selection	Only $M < PB$ beams are sent and the best beam is selected using an AI method with M reported RSRP values.	$M < PB$ beam transmission/reporting	Sensitive to adversarial attacks with small crafted perturbations.
AI-based beam selection with scrambling	Only $M < PB$ beams are sent and the best beam is selected using an AI method with M reported RSRP values. A scrambler is used prior to AI operation.	$M < PB$ beam transmission/reporting	Robust to adversarial attacks when the attacker does not have the scrambling information.

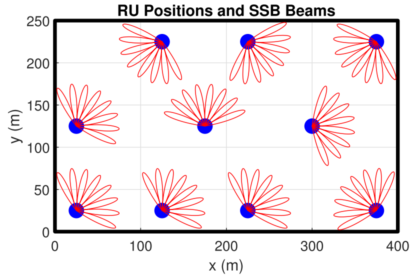


FIGURE 9. RU positions and SSB beams used in simulations.

TABLE 2. Simulation parameters.

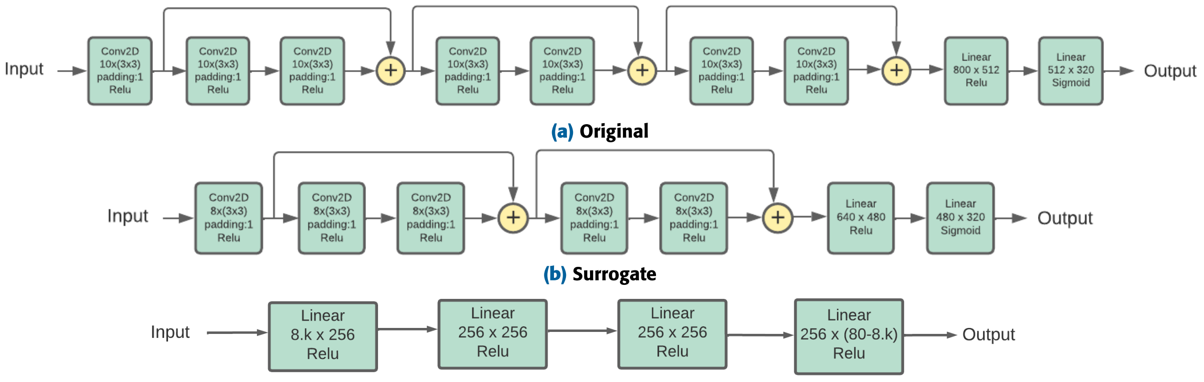
Parameter	Model and/or Value
Carrier frequency, bandwidth	39 GHz, 100 MHz
Path-loss and shadowing	According to 3GPP TR 38.901 [36]
Average RU transmit power	35 dBm
UE receiver noise power	$\sigma_n^2 = -84$ dBm
(P, B, M, A)	$(10, 32, 80, 1)$
N in UAP	256
$(\alpha, j_{\max})$ in BIM and T-BIM	$(0.1\epsilon, 20)$
RU positions and SSB beams	According to Fig. 9
Infected UE position	Random in the area given by Fig. 9

the original model to make the attack practical. We use 1.7×10^6 training data, and 1.7×10^5 test data to test the performance of models. We train the models for 30 epochs with an initial learning rate (LR) of 0.001 and batch size of 128 using Adam optimizer and decrease LR to half after each 10 epoch. Both models are built based on multiple successive residual blocks (each consisting of 2 convolutional layers) followed by 2 dense layers. Input and output channels for the convolutional layers are set to 10 and 8 for the original and surrogate model, and 3×3 kernel is used for both models. Final model performances of the two models on test data are 99.73% and 99.62%, respectively. The model training took around 110 minutes in a local machine with Intel Core i5 processor (2.4 GHz base frequency, 8 MB cache and 4 cores). The input to the models is 80 RSRP values and the output is a vector with size 320, showing the probability of each beam being the best one. Notice that there are $PB = 320$ beams in total whereas the network only sends 80 of them to decrease the overhead and energy consumption. In GAI method, to estimate the missing RSRP values using the previous measurements, we assume that the attacker uses 9 simple 4-layer DNNs as shown in Fig. 10c where the missing $80 - 8k$ RSRP values are estimated using the known $8k$ RSRP values for $k = 1, 2, \ldots, 9$. We only consider ϵ values less than shadowing standard deviation of the channel as larger ϵ values might cause attacker to be detected.

A. EFFECTIVENESS OF THE ADVERSARIAL ATTACKS

For comparison, we evaluate the RSRP error which is the RSRP difference between the actual best beam and the beam inferred by the AI method. Even a few dB RSRP error is important as it directly lowers the signal-to-noise-ratio by that much at the UE side due to a wrong beam selection. In Fig. 11, we observe the RSRP error statistics for $\epsilon = 3, 6$ dB under the whitebox/blackbox setting. Here, the x-axis shows the percentiles of the cumulative distribution function of the RSRP errors. We first observe that RSRP errors are larger in the whitebox setting for each method compared to the blackbox setting. This is expected since the perturbations crafted using the original model can maximize the same original model's loss function better than the surrogate model. Another direct consequence is that all four adversarial attack methods outperform WGN attacks in all cases. This result proves the effectiveness of the adversarial attacks over known techniques. A final observation is that in all cases, the best method among BIM and T-BIM is better than the best one of UAP and GAI. This is because Scenario 2 allows more complete input data information for the attacker to efficiently design perturbation compared to Scenario 1.

In Fig. 12, we observe the 50-th and 90-th percentile RSRP errors for various ϵ values under the whitebox/blackbox setting. Similar results obtained from Fig. 11 are also valid here. We observe that UAP outperforms GAI in the blackbox setting for all ϵ values whereas the situation depends on the ϵ value in the whitebox setting. When we compare BIM and T-BIM, we conclude that T-BIM is better for large ϵ values whereas BIM is more useful when ϵ is small. According to the scenario (1 or 2), setting (whitebox/blackbox), and the perturbation amount (ϵ) it is possible to determine the best adversarial attack method (UAP, GAI, BIM, T-BIM) using the findings of this study. As a final observation, in blackbox setting, all methods are ineffective at small ϵ values as shown in Fig. 12d, which is because of limited knowledge about the AI model and small input variations.

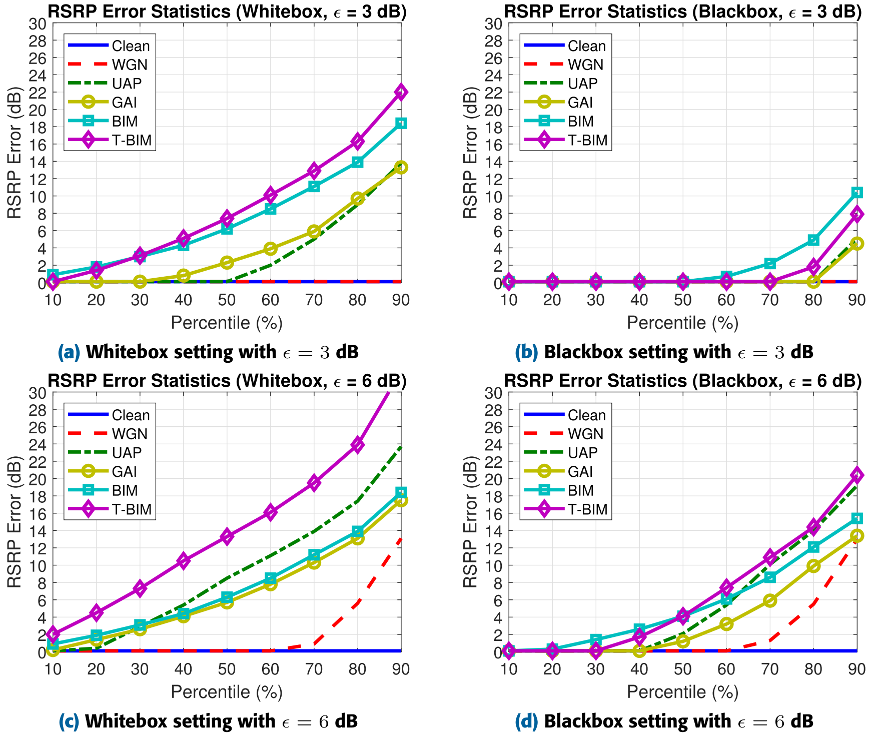

Input
Conv2D 10x(3x3) padding:1 Relu
Conv2D 10x(3x3) padding:1 Relu
Conv2D 10x(3x3) padding:1 Relu
+
Conv2D 10x(3x3) padding:1 Relu
Conv2D 10x(3x3) padding:1 Relu
+
Conv2D 10x(3x3) padding:1 Relu
Conv2D 10x(3x3) padding:1 Relu
+
Linear 800 x 512 Relu
Linear 512 x 320 Sigmoid
Output
Input
Conv2D 8x(3x3) padding:1 Relu
Conv2D 8x(3x3) padding:1 Relu
Conv2D 8x(3x3) padding:1 Relu
+
Conv2D 8x(3x3) padding:1 Relu
Conv2D 8x(3x3) padding:1 Relu
+
Linear 640 x 480 Relu
Linear 480 x 320 Sigmoid
Output
Input
Linear 8.k x 256 Relu
Linear 256 x 256 Relu
Linear 256 x 256 Relu
Linear 256 x (80-8.k) Relu
Output

(a) Original

(b) Surrogate

(c) Models in GAI method for estimation of missed RSRPs for $k = 1, 2, \ldots, 9$.

FIGURE 10. Original, surrogate and GAI AI models.



(a) Whitebox setting with $\epsilon = 3$ dB

(b) Blackbox setting with $\epsilon = 3$ dB

(c) Whitebox setting with $\epsilon = 6$ dB

(d) Blackbox setting with $\epsilon = 6$ dB

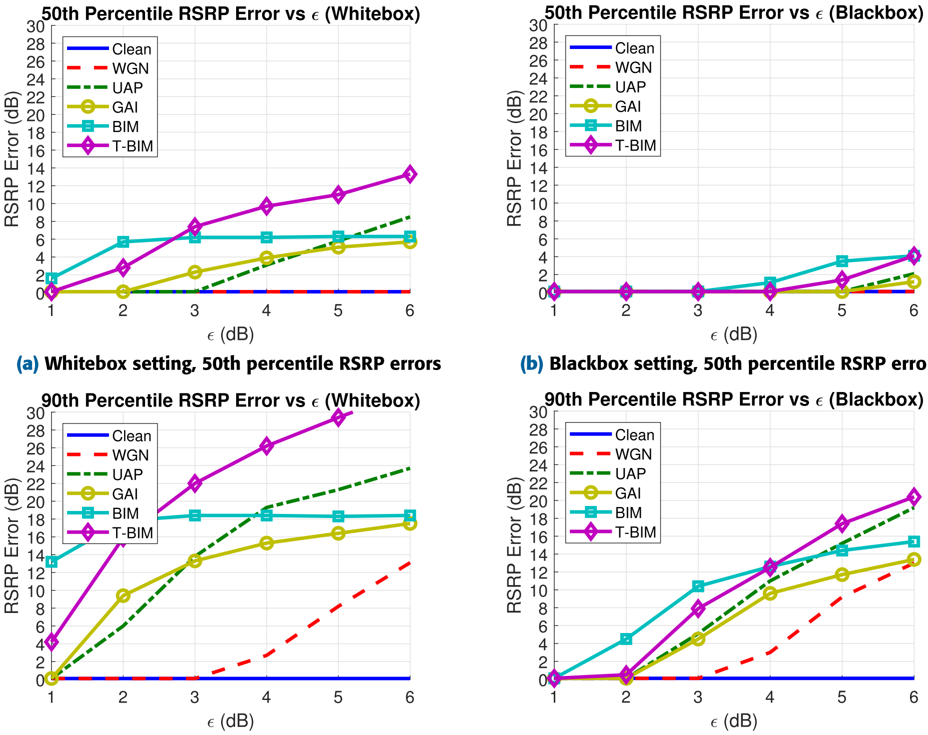
FIGURE 11. RSRP error statistics for $\epsilon = 3, 6$ dB.

Considering all results presented in Fig. 11 and 12, we conclude that adversarial attacks have substantial effects on network performance. Even in the most challenging scenario and setting from the attacker's perspective (Scenario 1, blackbox) the attacks can significantly degrade the beam selection performance. The results show the importance and necessity of a smart and effective defense technique against these threats.

B. PERFORMANCE OF THE PROPOSED DEFENSE METHOD

In this part, we present the results obtained with extensive simulations under the proposed defense method relying on a scrambling operation before the AI model. We use the same AI model at the CP and the same surrogate AI model at the attacker side. We only consider the blackbox setting as it is the practical one and investigate the same attack algorithms analyzed before under Scenario 1 and 2 accordingly. We assume that the new RU ordering after scrambling is $6, 10, 5, 1, 4, 9, 3, 8, 2, 7$, which is assumed to be unknown to the attacker. The attacker crafts the perturbation values as if the RU order is $1, 2, \ldots, 10$.

In Fig. 13, we present the RSRP error and RSRP gain statistics for $\epsilon = 3, 6$ dB with scrambling operation. Here RSRP errors (Fig. 13a, 13c) show the RSRP difference between the actual best beam and the beam selected by AI with scrambler, and RSRP gains (Fig. 13b, 13d) indicate the enhancement in the RSRP error compared to the no scrambler case. We first observe that for $\epsilon = 3$ dB, all



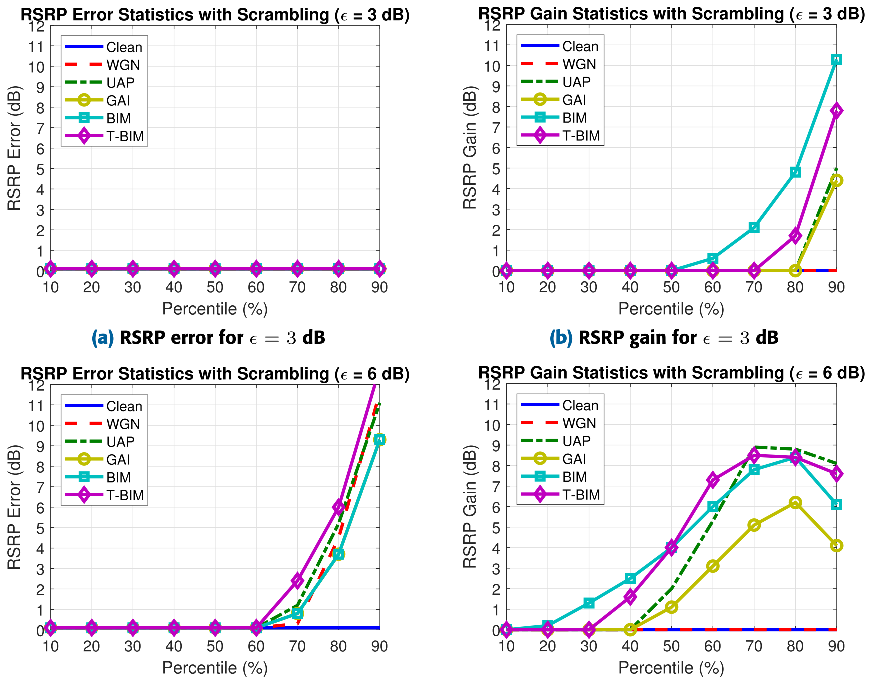
(a) **Whitebox setting, 50th percentile RSRP errors**

(b) **Blackbox setting, 50th percentile RSRP errors**

(c) **Whitebox setting, 90th percentile RSRP errors**

(d) **Blackbox setting, 90th percentile RSRP errors**

FIGURE 12. **RSRP errors for various ϵ values.**



(a) **RSRP error for $\epsilon = 3$ dB**

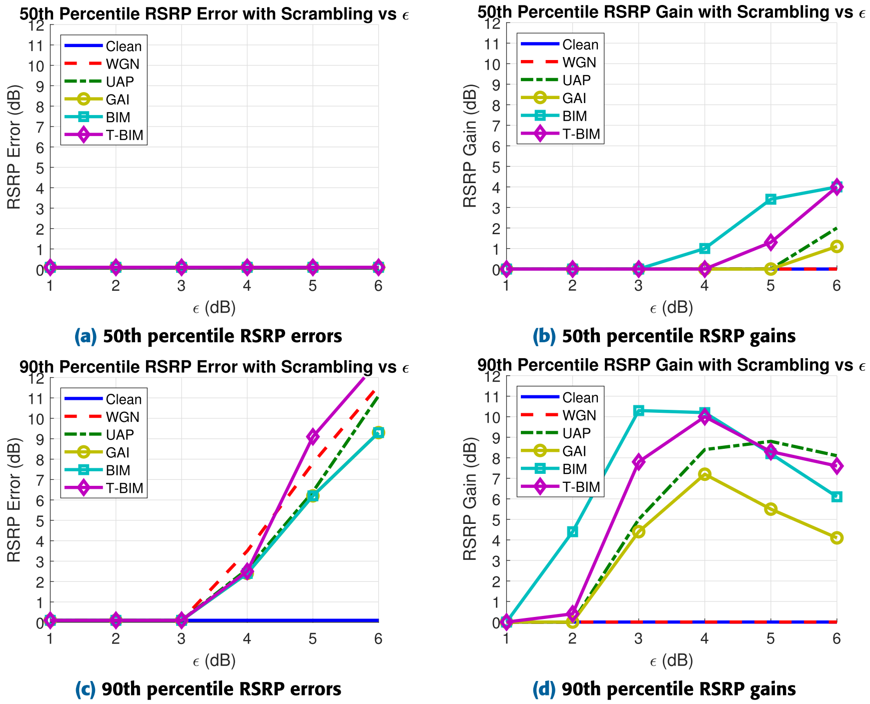
(b) **RSRP gain for $\epsilon = 3$ dB**

(c) **RSRP error for $\epsilon = 6$ dB**

(d) **RSRP gain for $\epsilon = 6$ dB**

FIGURE 13. **RSRP error/gain statistics for $\epsilon = 3, 6$ dB with scrambling.**

attacks become ineffective thanks to the scrambler. As shown in 13b, RSRP errors due to adversarial attacks are decreased by 4 to 11 dB at the 90th percentile value. No RSRP gain is obtained for clean or WGN attacks as there is no input-dependent perturbation in these cases. When we consider $\epsilon = 6$ dB case, as presented by 13b, the median (50th percentile) RSRP errors are again close to zero when we use a scrambler. Even though adversarial attacks can



(a) 50th percentile RSRP errors

(b) 50th percentile RSRP gains

(c) 90th percentile RSRP errors

(d) 90th percentile RSRP gains

FIGURE 14. RSRP errors/gains for various ϵ values with scrambling.

result in some non-negligible errors at higher percentiles, their resulting RSRP errors are close to that of WGN. This result shows that the scrambling operation eliminates the effectiveness of the adversarial attacks over WGN whose perturbations are independent from the input. As indicated by 13d, 4 to 9 dB RSRP enhancements are obtained against adversarial attacks, which can also be validated by comparing Fig. 11 and Fig. 13. As a result, we conclude that by using a scrambler, a substantial enhancement is obtained in the beam selection performance.

In Fig. 14, we observe 50th and 90th percentile RSRP errors and RSRP gains with scrambling as ϵ varies. We conclude that median RSRP errors become negligible for all ϵ values showing the effectiveness of the proposed defense technique. 90th percentile RSRP errors are not negligible for $\epsilon > 3$ dB but as shown in Fig. 14d, at least 4 dB RSRP gain is obtained against any four adversarial attacks. For all methods, the RSRP gains with scrambling have a global maximum for a specific ϵ value. This is because when ϵ is small, the effectiveness of the attacks is limited and hence there is not much space to enhance RSRP errors. When ϵ is sufficiently large, the performance degradation under any attack becomes very high and even the WGN can have substantial effects on the performance. Therefore, the mitigation performance of the scrambler starts degrading. Nevertheless, intermediate values of ϵ are more practical as a small ϵ is ineffective and a large ϵ may cause the attacker to be detected by the network, and hence we conclude that the defense algorithm can successfully mitigate those threats.

In Table 3, we present the 90% RSRP errors under the blackbox setting for WGN and the best attack, and the corresponding RSRP gains when the defense method is applied. (RSRP errors/gains are rounded to the nearest integer for a better illustration.) We only consider $\epsilon = 3, 4, 5, 6$ dB values as attacks are not effective at smaller ϵ values. We observe that depending on the scenario (1 or 2) and the amount of perturbation (ϵ), the most disruptive attack method and the related performance gain when the scrambling is applied changes. In all cases, it is clearly seen that the proposed defense technique relying on scrambling prior to AI operation significantly enhances the performance. After the RSRP gain by defense, we obtain nearly the same RSRP errors as in the WGN attack case, showing that the defense technique can successfully eliminate the disruptive effects of the smart adversarial attacks.

TABLE 3. Summary of numerical results.

Scenario	ϵ (dB)	RSRP err. (WGN) (dB)	RSRP err. (Best) (dB)	RSRP gain by defense (dB)
1	3	0	5 (UAP)	5
1	4	4	11 (UAP)	8
1	5	9	15 (UAP)	9
1	6	13	19 (UAP)	8
2	3	0	8 (BIM)	10
2	4	3	12 (BIM)	10
2	5	9	18 (T-BIM)	8
2	6	13	20 (T-BIM)	7

By applying the scrambling operation, our method can obfuscate how the RSRP values corresponding to different RUs are input to the AI model running at the CP. The permutation of RUs establishes a secret, which makes it more challenging for an attacker to predict the relationship between the input and output of the AI system. In particular, the permutation makes it almost impossible for an attacker to identify the boundaries in the space of measurement values where the determination changes, and thus makes it very challenging to design an effective perturbation to attack the AI model, i.e., makes it very challenging to modify the measurement values to mislead the AI model as an operational controller of the wireless access network.

IV. CONCLUSION

In this work, we investigate potential vulnerabilities of AI-driven beam selection functionality in a D-MIMO network. We demonstrate a practical scenario where a potential malware infects UE and performs smart adversarial attacks, thereby lowering the network performance. We proposed four different adversarial attack methods, two of which make use of only partial knowledge about the RSRP values of forwarded beams. We experimentally showed that adversarial attacks can lead to a considerable degree of RSRP error for beam selection tasks. And the amount of error introduced by these adversarial attacks is much larger than conventional attacks. These results show that effective defensive strategies should not be ignored when using AI for D-MIMO tasks. To that aim, we present a simple but effective mitigation solution against adversarial attack threats providing up to 10 dB better signal strengths by selecting more accurate RU/beam pairs, support our proposal with detailed simulation results, and finally provide a potential deployment option for our proposed solution. As a future work, one can investigate the effectiveness of a similar defense mechanism against adversarial threats on other AI-driven tasks including power control and RU serving subset selection in D-MIMO.

REFERENCES

[1] Ö. T. Demir, E. Björnson, and L. Sanguinetti, "Foundations of user-centric cell-free massive MIMO," *Found. Trends Signal Process.*, vol. 14, nos. 3–4, pp. 162–472, 2021.

[2] J. Shikida, K. Muraoka, T. Takeuchi, and N. Ishii, "Inter-access point coordinated user and beam selection for mmWave distributed MIMO systems," in *Proc. IEEE 96th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2022, pp. 1–5.

[3] Z. Lin, M. Lin, T. de Cola, J.-B. Wang, W.-P. Zhu, and J. Cheng, "Supporting IoT with rate-splitting multiple access in satellite and aerial-integrated networks," *IEEE Internet Things J.*, vol. 8, no. 14, pp. 11123–11134, Jul. 2021.

[4] Z. Lin, H. Niu, K. An, Y. Wang, G. Zheng, S. Chatzinotas, and Y. Hu, "Refracting RIS-aided hybrid satellite-terrestrial relay networks: Joint beamforming design and optimization," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 58, no. 4, pp. 3717–3724, Aug. 2022.

[5] Z. Lin, M. Lin, B. Champagne, W.-P. Zhu, and N. Al-Dhahir, "Secrecy-energy efficient hybrid beamforming for satellite-terrestrial integrated networks," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 6345–6360, Sep. 2021.

[6] Z. Lin, H. Niu, K. An, Y. Hu, D. Li, J. Wang, and N. Al-Dhahir, "Pain without gain: Destructive beamforming from a malicious RIS perspective in IoT networks," *IEEE Internet Things J.*, vol. 11, no. 5, pp. 7619–7629, Mar. 2024.

[7] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, Aug. 2019.

[8] Y. Zhao, I. G. Niemegeers, and S. H. De Groot, "Power allocation in cell-free massive MIMO: A deep learning method," *IEEE Access*, vol. 8, pp. 87185–87200, 2020.

[9] N. Rajapaksha, K. B. Shashika Manosha, N. Rajatheva, and M. Latva-Aho, "Deep learning-based power control for cell-free massive MIMO networks," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2021, pp. 1–7.

[10] F. E. Kadan and Ö. Haliloglu, "A performance bound for maximal ratio transmission in distributed MIMO," *IEEE Wireless Commun. Lett.*, vol. 12, no. 4, pp. 585–589, Apr. 2023.

[11] C. M. Yetis, E. Björnson, and P. Giselsson, "Joint analog beam selection and digital beamforming in millimeter wave cell-free massive MIMO systems," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 1647–1662, 2021.

[12] V. Ranasinghe, N. Rajatheva, and M. Latva-aho, "Graph neural network based access point selection for cell-free massive MIMO systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2021, pp. 01–06.

[13] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, 2014.

[14] Ö. F. Tuna, F. E. Kadan, and L. Karaçay, "Practical adversarial attacks against AI-driven power allocation in a distributed MIMO network," in *Proc. IEEE Int. Conf. Commun.*, May 2023, pp. 759–764.

[15] Y. E. Sagduyu, Y. Shi, and T. Erpek, "Adversarial deep learning for over-the-air spectrum poisoning attacks," *IEEE Trans. Mobile Comput.*, vol. 20, no. 2, pp. 306–319, Feb. 2021.

[16] B. Kim, Y. Sagduyu, T. Erpek, and S. Ulukus, "Adversarial attacks on deep learning based mmWave beam prediction in 5G and beyond," in *Proc. IEEE Stat. Signal Process. Workshop (SSP)*, Jul. 2021, pp. 590–594.

[17] Z. Zhang, M. Tian, C. Li, Y. Huang, and L. Yang, "Poison neural network-based mmWave beam selection and detoxification with machine unlearning," *IEEE Trans. Commun.*, vol. 71, no. 2, pp. 877–892, Feb. 2023.

[18] M. Zolotukhin, P. Miraghaie, D. Zhang, T. Hämäläinen, W. Ke, and M. Dunderfelt, "Black-box adversarial examples against intelligent beamforming in 5G networks," in *Proc. IEEE Conf. Standards for Commun. Netw. (CSCN)*, Nov. 2022, pp. 64–70.

[19] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015.

[20] M. Kuzlu, F. O. Catak, S. Sarp, U. Cali, and O. Gueler, "A streamlit-based artificial intelligence trust platform for next-generation wireless networks," in *Proc. IEEE Future Netw. World Forum (FNWF)*, Jul. 2022, pp. 94–97.

[21] J. Zhang, Y. Dong, M. Kuang, B. Liu, B. Ouyang, J. Zhu, H. Wang, and Y. Meng, "The art of defense: Letting networks fool the attacker," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 3267–3276, 2023.

[22] Y. Wang, X. Ma, J. Bailey, J. Yi, B. Zhou, and Q. Gu, "On the convergence and robustness of adversarial training," in *Proc. 36th Int. Conf. Mach. Learn.*, vol. 97, 2019, pp. 6586–6595.

[23] A. Raghunathan, S. M. Xie, F. Yang, J. Duchi, and P. Liang, "Understanding and mitigating the tradeoff between robustness and accuracy," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 7909–7919. [Online]. Available: https://proceedings.mlr.press/v119/raghunathan20a.html

[24] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, 2018.

[25] 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)*, May 2016, pp. 582–597.

[26] M. Kuzlu, F. O. Catak, Y. Zhao, S. Sarp, and E. Catak, "Security and privacy concerns in next-generation networks using artificial intelligence-based solutions: A potential use case," in *Wireless Networks*. Cham, Switzerland: Springer, 2023, pp. 205–226.

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

[28] H. Tang, F. O. Catak, M. Kuzlu, E. Catak, and Y. Zhao, "Defending AI-based automatic modulation recognition models against adversarial attacks," *IEEE Access*, vol. 11, pp. 76629–76637, 2023.

[29] O. F. Tuna, F. O. Catak, and M. T. Eskil, "Unreasonable effectiveness of last hidden layer activations for adversarial robustness," in *Proc. IEEE 46th Annu. Comput., Softw., Appl. Conf. (COMPSAC)*, Jun. 2022, pp. 1098–1103.

[30] D. Meng and H. Chen, "MagNet: A two-pronged defense against adversarial examples," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2017, pp. 135–147.

[31] D. Adesina, C.-C. Hsieh, Y. E. Sagduyu, and L. Qian, "Adversarial machine learning in wireless communications using RF data: A review," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 77–100, 1st Quart., 2023.

[32] K. Ren, T. Zheng, Z. Qin, and X. Liu, "Adversarial attacks and defenses in deep learning," *Engineering*, vol. 6, no. 3, pp. 346–360, Mar. 2020.

[33] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, 2017.

[34] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," 2016, *arXiv:1610.08401*.

[35] S. Gatlan. (2021). *Qualcomm Vulnerability Impacts Nearly 40% of All Mobile Phones*. [Online]. Available: https://www.bleepingcomputer.com/news/security/qualcomm-vulnerability-impacts-nearly-40-percent-of-all-mobile-phones/

[36] 3GPP. (Mar. 2022). *Study on Channel Model for Frequencies From 0.5 to 100 GHz*. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/38_series/38.901/38901-h00.zip

ÖMER FARUK TUNA received the B.S. and M.S. degrees in electrical and electronics engineering from Boğaziçi University, Istanbul, Turkey, in 2004 and 2007, respectively, and the Ph.D. degree in computer engineering from Işık University, Istanbul, in 2023. Previously, he was with various multinational technology companies as an engineer and the technical manager. He is currently with Ericsson Research, Turkey, as a Senior Researcher, focusing on trustworthy AI. He has authored numerous research papers that have been published in peer-reviewed international conferences and journals. He has been involved in several EU-funded research projects. His research interest includes the intersection of security and privacy of AI-driven systems for 6G networks. He was a recipient of the IEEE International Black Sea Conference on Communications and Networking Best Paper Award, in 2023.

FEHMI EMRE KADAN (Member, IEEE) received the B.S. degree in mathematics and in electrical and electronics engineering and the M.S. and Ph.D. degrees in electrical and electronics engineering from Middle East Technical University, Ankara, Turkey, in 2013, 2015, and 2021, respectively. From 2013 to 2021, he was an algorithm designer with various defense industry companies. Since 2022, he has been a Senior Researcher with Ericsson Research, Turkey. His research interests include wireless communications, signal processing, optimization, and beamforming applications. He was a recipient of the Silver Medals in the International Mathematical Olympiad, in 2008 and 2009; the Bronze and Gold Medals in the Balkan Mathematical Olympiad, in 2008 and 2009; and the IEEE International Black Sea Conference on Communications and Networking Best Paper Award, in 2023.