Adversarial Robustness of Deep-Learning-Based Near-Field Beam Prediction In XL-MIMO Systems

Sneha Singh[1], Harshita Seth[2], Kashish Sharma[3], Prof. Guna Dhondwad[4]
[1,2,3]*MIT ACSC, Pune, India*
[4]*Professor, MIT ACSC, Pune, India*
doi.org/10.64643/ijirt.208499-459

***Abstract*—Deep learning is increasingly used in sixth generation (6G) extremely large-scale MIMO (XL-MIMO) systems to predict beams directly from channel observations to alleviate the overhead of exhaustive near-field beam search. However, these channel-based observations also provide an attack surface to learned beam-selection models. In this paper, we consider the adversarial robustness of a deep learning-based near-field beam predictor for XL-MIMO. We train a fully connected neural network to classify 45 near-field candidate beams generated for a simulated 8x8 planar array and evaluate its vulnerability under the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). The beam classification accuracy drops from 74.07% under clean condition to 28.70% under FGSM and 26.11% under PGD at perturbation budget of ε=0.05. Adversarial training improves the respective accuracies to 36.11% and 33.70% while maintaining comparable accuracy on clean data. Specifically, this classification-level degradation does not yield a proportional calculated communication-level impact: the beamforming gain ratio stays close to unity and the derived effective SNR, theoretical BPSK BER and Shannon-based spectral-efficiency estimate only change slightly, with a maximum reported effective-SNR loss of 0.014 dB. Results show that the near-field XL-MIMO beam predictors are vulnerable at the classification level under the input-space threat model considered, while the impact on communication-level is highly dependent on the spatial structure of the beam codebook.**

***Index Terms*—XL-MIMO, near-field communications, beam prediction, deep learning, adversarial machine learning, FGSM, PGD, adversarial training, beamforming, adversarial robustness.**

I. INTRODUCTION

The increasing use of deep-learning models in XL-MIMO systems for beam prediction raises the security concern of channel-derived inputs.

The evolution toward sixth generation (6G) wireless communications motivates the development of extremely large-scale multiple-input multiple-output (XL-MIMO) systems, in which very large antenna apertures enable high beamforming gain, spatial resolution and communication capacity. However, increasing the physical aperture of an antenna array also affects the propagation characteristics seen by users. In traditional far-field MIMO systems, the electromagnetic waves impinging on the antenna array can in many cases be modeled as planar wavefronts. As the array aperture becomes sufficiently large, users can fall within the electromagnetic near-field region, where spherical-wave propagation becomes important and conventional far-field assumptions are no longer adequate [1]– [3].

The spatial characteristics of the received signal in the near-field region is not only dependent on angular direction of the user but also on the distance from the antenna array. Therefore, near-field beamforming is capable of steering electromagnetic energy to a particular spatial location rather than just an angular direction. This adds a further distance dimension to beam selection and increases the size of the beam search space. Near-field beam training therefore becomes more challenging, as the system must identify suitable beams across both angular and distance dimensions while limiting the training overhead associated with beam searching [2], [3], [6], [7].

To reduce this complexity, deep learning (DL) has been investigated as a data-driven approach to near-field beam training and beam prediction. Instead of exhaustively evaluating every candidate beam, a neural network can learn the relationship between channel-related observations and suitable beam selections from previously generated training data.

Recent studies have demonstrated deep-learning-based approaches for near-field XL-MIMO beam training, including methods that learn near-field beam codewords or infer the beam associated with the user's spatial characteristics [8]– [10]. These studies demonstrate the potential of data-driven methods to reduce beam-training overhead while maintaining effective beamforming performance.

However, replacing an explicitly designed beam-search procedure with a learned model introduces a security dependency: the reliability of the beam decision depends on the integrity of the information provided to the neural network. Deep neural networks are known to be vulnerable to adversarial examples, in which carefully crafted input perturbations can cause a model to produce an incorrect prediction [12]. Adversarial training and robust optimization have subsequently been investigated as approaches for improving the resistance of neural networks to such perturbations [13].

This concern is particularly relevant to wireless systems because machine-learning models may operate on radio-frequency, channel, received-signal-strength, or other signal-derived observations. Adversarial machine-learning research in wireless communications has identified challenges that differ from those in conventional machine-learning applications, including the relationship between the physical communication environment and the inputs observed by a learning model [14]– [16]. Consequently, assessing the security of a wireless learning system requires consideration not only of whether an adversary can alter the model's prediction, but also of how such prediction errors affect the underlying communication process.

Importantly, adversarial attacks against deep-learning-based beam prediction are not an entirely unexplored problem [11],[12].

However, the near-field XL-MIMO setting introduces a different spatial structure from conventional mmWave beam prediction. In near-field systems, beam selection depends on both angular and distance characteristics because the wavefront is spherical and beam focusing is spatially dependent [2], [3]. Therefore, an adversarially induced prediction error may not necessarily produce the same communication-level consequence as an error in a conventional far-field beam-selection system. In particular, a predicted beam that differs from the optimal codebook label may still provide a comparable beamforming gain if its corresponding spatial focusing point is sufficiently close to the optimal one.

This leads to the central research problem addressed in this work: how robust is deep-learning-based near-field beam prediction when its channel-derived input is subjected to adversarial perturbations, and to what extent does prediction degradation translate into communication-level performance degradation? The problem is important because evaluating adversarial robustness solely through classification accuracy may provide an incomplete representation of the practical impact on an XL-MIMO communication system.

Accordingly, this study investigates the vulnerability of a deep-learning-based near-field beam-prediction model to inference-time adversarial perturbations and evaluates adversarial training as a robustness mechanism. The study considers two gradient-based attacks, Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), and compares a conventionally trained model with an adversarially trained model. In addition to beam-prediction accuracy and attack success, the study evaluates beamforming gain, effective signal-to-noise ratio (SNR), theoretical BPSK bit-error rate (BER), and spectral efficiency to examine the relationship between machine-learning-level degradation and communication-level impact.

II. PROBLEM STATEMENT

The combination of these developments creates a specific research problem. Existing research has separately established the importance of near-field modeling and beam training for XL-MIMO systems, the potential of deep learning to reduce near-field beam-training overhead, and the vulnerability of learned wireless systems to adversarial manipulation [2]– [7]. However, the robustness of a deep-learning-based near-field beam-prediction system under adversarial input perturbations, together with the relationship between prediction degradation and communication-level performance, requires further investigation.

A further complication is that an incorrect beam-prediction label does not necessarily correspond to an equally severe communication failure. In a near-field angle-distance codebook, neighboring beams may

correspond to spatially similar focusing locations and can therefore produce comparable beamforming gains. Consequently, evaluating an adversarial attack solely through classification accuracy may provide an incomplete picture of its practical impact. A more informative evaluation should examine both the machine-learning-level effect of the attack and its resulting communication-level consequences.

III. RESEARCH OBJECTIVE

The primary objective of this study is to investigate the vulnerability of deep-learning-based near-field beam prediction to inference-time adversarial perturbations and to evaluate the effectiveness of adversarial training in improving model robustness.
The study considers both machine-learning-level metrics, including beam-prediction accuracy and attack success, and communication-level metrics, including beamforming gain, effective SNR, theoretical BPSK bit-error rate, and spectral efficiency.

Research Questions

The study addresses the following research questions:
RQ1: To what extent do adversarial perturbations degrade the beam-prediction accuracy of a deep-learning-based near-field beam-prediction model in an XL-MIMO setting?

RQ2: How does adversarial degradation in beam prediction translate into changes in beamforming gain, effective SNR, theoretical BER, and spectral efficiency?

RQ3: Can adversarial training improve the robustness of the beam-prediction model against FGSM and PGD attacks while maintaining comparable performance on clean inputs?

Contributions

The main contributions of this study are:
1. Adversarial vulnerability assessment:
An investigation of the vulnerability of a deep-learning-based near-field beam-prediction model to inference-time adversarial perturbations using FGSM and PGD attacks.

2. Robustness evaluation:
A quantitative comparison between conventionally trained and adversarially trained beam-prediction models under clean and adversarial conditions.

3. Communication-level analysis:
An evaluation of whether degradation in beam-prediction accuracy translates into changes in beamforming gain, effective SNR, theoretical BPSK BER, and spectral efficiency.

4. Cross-layer interpretation:
An analysis distinguishing machine-learning-level prediction degradation from its calculated communication-level consequences, thereby examining whether a successful attack against the classifier necessarily results in a comparable loss of communication performance.

IV. RELATED WORK

A. Near-Field XL-MIMO

Near-field propagation has become an important consideration in XL-MIMO systems because the large physical aperture extends the region in which spherical-wave propagation must be considered. Unlike conventional far-field beamforming, near-field beam focusing depends jointly on the angular and distance characteristics of the user. This angle-distance dependence has motivated the development of near-field channel models, beamforming methods, codebook designs, and beam-training strategies specifically for XL-MIMO systems [1]– [5].

B. Near-Field Beam Training

The increased dimensionality of the near-field beam space can lead to substantial training overhead when exhaustive beam search is used. To address this problem, several reduced-search approaches have been proposed. Zhang et al. developed a fast near-field beam-training method that reduces the number of measurements required for beam selection [6], while Wu et al. proposed a two-stage hierarchical beam-training strategy that exploits the structure of the near-field beam space [7]. These studies demonstrate that reducing the search space is an important consideration for practical near-field beam training.

C. Deep-Learning-Based Near-Field Beam Prediction

Deep learning has also been investigated as a means of reducing the overhead associated with near-field beam training. Liu et al. proposed a deep-learning-based beam-training approach for extremely large-scale MIMO systems operating in the near-field domain [8]. Jiang and Qi investigated near-field beam training based on deep learning, using the relationship between user spatial characteristics and the near-field beam space [9]. More recently, Nie et al. investigated deep-learning-based near-field beam training for XL-MIMO systems [10]. These studies demonstrate that learned models can provide an alternative to exhaustive beam search by learning the relationship between channel information and suitable beam selections.

D. Adversarial Machine Learning for Wireless Systems

Deep neural networks are vulnerable to adversarial perturbations, motivating attack methods such as FGSM and PGD and defense mechanisms such as adversarial training [12], [13]. In wireless communications, adversarial machine learning has been investigated in applications including radio-signal classification, power control, and other learning-based wireless functions [14]– [17]. These studies establish the broader security relevance of adversarial learning in wireless systems.

Of particular relevance to beam prediction, Kim *et al.* demonstrated adversarial attacks against a deep-learning-based mmWave beam-prediction system by manipulating received-signal-strength observations [11]. Their results show that a learned beam-selection model can be manipulated at inference time. However, the considered mmWave setting does not explicitly address the angle-distance structure of near-field XL-MIMO beam selection.

E. Research Gap

Existing studies have therefore established three important aspects of the problem: the significance of near-field propagation in XL-MIMO systems, the use of deep learning to reduce near-field beam-training overhead, and the vulnerability of learning-based wireless systems to adversarial manipulation [1]– [17]. However, the adversarial robustness of deep-learning-based near-field beam prediction and the relationship between prediction errors and communication-level performance have received comparatively limited attention.

This study addresses this gap by evaluating FGSM and PGD attacks against a deep-learning-based near-field beam-prediction model and by comparing conventionally trained and adversarially trained models. In addition to classification-level performance, the study evaluates beamforming gain and calculated communication-level metrics to determine whether a successful attack against the beam classifier necessarily produces a comparable degradation in communication performance.

V. SYSTEM MODEL AND PROBLEM FORMULATION

A. Overall Experimental Framework

The proposed experimental framework investigates the robustness of deep-learning-based near-field beam prediction in an XL-MIMO setting against inference-time adversarial perturbations. The framework consists of five stages: near-field system configuration, near-field channel and beam-codebook generation, deep-learning-based beam prediction, adversarial attack and robust training, and communication-level performance evaluation. A planar 8×8 antenna array is used as a simulation-scale representation of an XL-MIMO near-field environment. Channel realizations are generated for multiple user locations, and an angle-distance beam codebook is constructed. A fully connected neural network is trained to predict the optimal beam from channel-derived input features. The trained models are evaluated under clean conditions and under FGSM and PGD attacks. Finally, the predicted beams are evaluated using beamforming gain ratio, effective SNR, theoretical BPSK BER, and a Shannon-based spectral-efficiency Estimate.

B. XL- MIMO System Configuration

The considered system employs a uniform planar array (UPA) consisting of 8×8 antenna elements, resulting in 64 antenna elements. The operating wavelength is set to $\lambda=0.01$ m, and the inter-element spacing is $d=\lambda/2=0.005$ m. The simulation considers a near-field propagation setting in which the spatial characteristics of the channel depend on both the angular position and the propagation distance of the user. Near-field beam training and beamforming generally require

consideration of these two spatial dimensions [2], [3], [6], [7].

To construct the candidate beam-search space, nine angular positions from −60∘ to 60∘ and five user-distance values from 0.10 m to 0.30 m are considered. The combination of these angular and distance values produces 9×5=45 candidate spatial locations and corresponding candidate beams. Each candidate beam is represented by 64 antenna coefficients corresponding to the elements of the planar array.

Parameter	Configuration
Array Geometry	8×8 planar array
Number of antennas	64
Wavelength	0.01 m
Antenna spacing	0.005 m
Angular positions	9
Angular range	−60° to 60°
Distance positions	5
Distance range	0.10–0.30 m
Candidate beams	45

C. Near-Field Region and Rayleigh Distance

The near-field operating region is verified using the Rayleigh distance, which provides a conventional boundary between the radiative near-field and far-field regions. For the considered 8×8 planar array, the inter-element spacing is d=λ/2=0.005 m and the maximum physical aperture is determined from the diagonal dimension of the array. The aperture is therefore given by

$$D = \sqrt{[(8-1)d]^2 + [(8-1)d]^2}.$$

Substituting d=0.005 m gives

D≈0.0495 m.

The corresponding Rayleigh distance is calculated as

$$R_{\text{Rayleigh}} = \frac{2D^2}{\lambda}.$$

With λ=0.01 m, the resulting Rayleigh distance is approximately RRayleigh≈0.49 m.

The user distances considered in this study range from 0.10 m to 0.30 m. Since these distances are below the calculated Rayleigh distance, the simulated users lie within the radiative near-field region under the adopted criterion. Therefore, the use of a spherical-wave channel model and an angle-distance beam codebook is appropriate for the experimental setting.

D. Near-Field Channel Generation

For each simulated user location, the distance between the user and each antenna element is calculated. A simplified spherical-wave line-of-sight channel model is adopted, in which the channel coefficient associated with the n-th antenna is given by

$$h_n = \frac{1}{r_n} \exp\left(-j\frac{2\pi}{\lambda} r_n\right),$$

where rn denotes the distance between the user and the n-th antenna element and λ denotes the operating wavelength. The resulting channel vector is normalized and represented as a complex-valued vector h∈C64. Since the neural network operates on real-valued features, the real and imaginary components of the 64 complex channel coefficients are separated. Consequently, each channel realization is represented by 128 real-valued input features.

E. Beam Codebook and Ground-Truth Generation

A near-field beam codebook containing 45 candidate beams is used to formulate beam prediction as a supervised classification problem. For each generated channel realization, the response associated with every candidate beam is evaluated. The optimal beam is selected according to

$$b^* = \arg\max_b \left|\mathbf{w}_b^H \mathbf{h}\right|,$$

where wb denotes the beamforming vector corresponding to the b-th candidate beam and h denotes the channel vector. The candidate beam producing the maximum response is assigned as the ground-truth beam label.

The resulting learning problem is therefore formulated as a 45-class classification task in which the neural network receives the 128-dimensional channel representation and predicts the most appropriate beam from the available candidate codebook.

F. Dataset Construction and Preprocessing

The dataset consists of 2,700 channel samples generated across the 45 candidate spatial locations, with 60 samples generated for each location. Each sample contains 128 real-valued channel features and an associated ground-truth beam label. Although user angle and distance are used during channel generation and ground-truth beam determination, only the 128 channel-derived features are provided to the neural network.

For each reference beam location, multiple channel realizations are generated by introducing small variations around the corresponding reference angle and distance. The user angle is randomly perturbed within ±5◦ of the reference angle, while the user distance is randomly perturbed within ±0.015 m of the reference distance. The resulting values are clipped to the considered simulation limits of −60◦ to 60◦ for angle and 0.10 m to 0.30 m for distance. This procedure provides spatial variation around each codebook location while maintaining the predefined simulation region.

The dataset is divided into training and testing subsets using an 80:20 split, resulting in 2,160 training samples and 540 test samples. A fixed random state of 42 is used, and stratification by beam class is applied to preserve the class distribution between the subsets. The input features are standardized using a StandardScaler fitted exclusively on the training data and subsequently applied to both the training and test sets. This prevents information from the test set from influencing the feature-scaling process. Because the split is performed over realizations generated around the same reference beam locations, this protocol primarily evaluates interpolation within the simulated spatial region rather than generalization to completely unseen user locations.

Dataset Parameter	Value
Total samples	2700
Input features/sample	128
Classes	45
Training samples	2160
Testing samples	540
Train/test ratio	80/20
Random state	42
Class stratification	Yes
Feature scaling	StandardScaler

G. Deep-Learning Beam Prediction Model

A fully connected neural network is employed for near-field beam prediction. The model receives the 128-dimensional real-valued channel representation as input and consists of a fully connected layer with 128 neurons, a ReLU activation function, a second fully connected layer with 64 neurons, a second ReLU activation function, and a final output layer containing 45 neurons. Each output corresponds to one candidate beam in the near-field codebook.

The predicted beam is obtained by selecting the output class with the highest model response. This can be expressed as

$$\mathbf{x} \rightarrow f_\theta(\mathbf{x}) \rightarrow \hat{b},$$

where x denotes the 128-dimensional channel input, fθ denotes the trained neural network, and b^ denotes the predicted beam index. Two models are evaluated: a conventionally trained baseline model and an adversarially trained model intended to improve robustness against adversarial perturbations.

Layer	Output
Input	128
Fully connected	128
ReLU	—
Fully connected	64
ReLU	—
Output	45

Training Configuration

The beam-prediction model was trained as a supervised multi-class classification model, where the input feature vector contains 128 features and the output layer corresponds to 45 candidate near-field beams. The model was trained using the cross-entropy loss function and the Adam optimization algorithm. The learning rate was set to 0.001, with a batch size of 32 and a maximum of 50 training epochs. The trained baseline model was evaluated on 540 test samples.

For robustness evaluation, adversarial examples were generated using the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). The perturbation strengths considered in the adversarial evaluation were $\epsilon \in \{0.001, 0.005, 0.010, 0.020, 0.050\}$. Adversarial training was subsequently applied to obtain the adversarially trained model, and its performance was compared with that of the baseline model under clean, FGSM, and PGD conditions.

The considered attacks are untargeted adversarial attacks. Their objective is to cause the beam-prediction model to produce an incorrect beam class rather than forcing the model toward one particular target beam. The adversarial perturbations are applied to the input feature representation at inference time, while the model parameters and training dataset remain unchanged during attack generation.

H. Adversarial Threat Model

The security evaluation considers an inference-time adversary capable of perturbing the channel-derived input features presented to the trained beam-prediction model. The attacker is assumed not to modify the neural-network parameters or the training dataset. Instead, the attacker introduces a constrained perturbation to the model input with the objective of inducing an incorrect beam prediction.

The adversarial input is represented as

$$\mathbf{x}_{adv} = \mathbf{x} + \boldsymbol{\delta},$$

where x denotes the original model input and δ denotes the adversarial perturbation. The perturbation is constrained by a maximum magnitude ϵ, with ϵ=0.05 used for the principal attack evaluation.

Thus, the evaluated threat model targets the inference-time model input rather than the model parameters or training dataset. Accordingly, the present study evaluates model-level robustness to constrained perturbations in standardized input-feature space; it does not establish end-to-end physical-layer attack realizability or an over-the-air attack power constraint.

I. Adversarial Attack Methods

Attack success rate (ASR) is defined as the fraction of test samples that are correctly classified under clean conditions but become incorrectly classified after the corresponding adversarial perturbation is applied. Two gradient-based attacks are considered: the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD).

FGSM generates an adversarial perturbation using the gradient of the classification loss with respect to the model input:

$$\mathbf{x}_{adv} = \mathbf{x} + \epsilon \ \mathrm{sign}\left(\nabla_{\mathbf{x}} L(\theta, \mathbf{x}, y)\right),$$

where L denotes the classification loss, θ represents the model parameters, y denotes the ground-truth beam label, and ϵ controls the perturbation magnitude. The principal FGSM evaluation uses ϵ=0.05.

PGD extends the gradient-based attack through multiple iterative updates. The implemented PGD configuration uses a perturbation budget of ϵ=0.05, 10 iterations, and a step size of α=0.005. After each update, the perturbation is projected back into the permitted perturbation region.

In this study, the perturbation budget is defined in standardized input-feature space because the channel features are normalized using statistics estimated from the training data. Consequently, the selected ε values quantify perturbations relative to the standardized feature representation rather than directly representing a physical change in the wireless channel or a realizable over-the-air perturbation power.

J. Adversarially Robust Training

To examine whether exposure to adversarial examples improves beam-prediction robustness, a second model is trained using adversarial examples in addition to the conventionally trained baseline model. The adversarially trained model is evaluated using the same test set and attack configurations applied to the baseline model. Clean accuracy is first evaluated to determine whether adversarial training affects normal prediction performance, followed by FGSM and PGD evaluations to measure the resulting robustness under adversarial inputs. This provides a direct comparison between conventional training and adversarial training under identical evaluation conditions. During adversarial training, adversarial examples are generated from the training batches using FGSM with a perturbation budget of ϵ=0.010. The original clean samples and their corresponding adversarial samples are then combined and used jointly for model optimization. Both versions retain the same ground-truth beam label. The model is trained using cross-entropy loss, allowing the network to learn from both clean and perturbed channel representations.

K. Robustness Evaluation

The robustness of the beam-prediction models is evaluated by varying the adversarial perturbation strength. The perturbation parameter ϵ is evaluated at 0.001, 0.005, 0.010, 0.020, and 0.050. Beam-prediction accuracy and attack success rate are recorded for both FGSM and PGD at each perturbation level. This evaluation characterizes the sensitivity of the prediction models to increasing adversarial perturbation strength.

L. Communication-Level Evaluation

The communication-level evaluation considers a single-user transmission scenario. Accordingly, inter-user interference is neglected and the interference power is set to zero. The resulting SINR therefore depends on the effective beamformed signal power and the modeled noise power. The communication

metrics should consequently be interpreted as a single-link theoretical evaluation rather than a multi-user system-level assessment. In particular, the BER and spectral-efficiency quantities are analytical estimates derived from the effective SNR, not Monte Carlo measurements from a transmitted bitstream.

The beamforming stage uses the beam selected by the neural network to combine the received signal along the predicted spatial direction. For each channel realization, the selected beam is applied to the corresponding channel to obtain the beamformed signal gain. This gain is then compared with the gain obtained using the optimal beam from the predefined codebook, which serves as the reference for evaluating the beam-selection performance.

To determine whether adversarial degradation in beam prediction results in a measurable communication-level effect, the beam selected by the neural network is evaluated against the corresponding channel. For each test sample, the beamforming response obtained using the predicted beam is compared with the response obtained using the optimal codebook beam. The normalized response ratio is defined as

$$G_{\text{ratio}} = \frac{G_{\text{predicted}}}{G_{\text{optimal}}},$$

where Gpredicted denotes the beamforming response obtained using the beam selected by the neural network, and Goptimal denotes the response obtained using the optimal codebook beam. Thus, Gratio represents the relative beamforming response retained by the model-selected beam compared with the optimal beam; for consistency with the figures, this quantity is referred to as the beamforming gain ratio.

The response ratio is then used as a multiplicative factor in the simplified effective-SNR model:

$$\text{SNR}_{\text{eff}} = \text{SNR}_{\text{input}} \times G_{\text{ratio}},$$

where SNRinput denotes the input SNR and Gratio denotes the normalized beamforming response ratio. The calculation is performed in the linear domain before conversion to decibels.

This is a simplified analytical mapping used to quantify the communication-level consequence of beam selection; it is not intended to replace a full physical-layer link-budget or waveform simulation. The resulting SNReff is subsequently used to calculate the theoretical BPSK BER and the Shannon-based spectral-efficiency estimate. The evaluation considers input SNR values of 5, 10, 15, 20, 25, and 30 dB.

M. BER and Shannon-Based Spectral-Efficiency Estimate

The communication-level evaluation further considers the theoretical BPSK bit-error rate (BER) and a Shannon-based spectral-efficiency estimate. For the theoretical coherent-BPSK calculation, the effective SNR is used in the standard expression

$$BER = \frac{1}{2}\,\text{erfc}\left(\sqrt{\text{SNR}_{\text{eff}}}\right),$$

where SNReff denotes the effective signal-to-noise ratio corresponding to the selected beam, and erfc(·) denotes the complementary error function. The Shannon-based theoretical spectral-efficiency estimate is calculated as

$$SE = \log_2(1 + \text{SNR}_{\text{eff}}),$$

where SE denotes the Shannon-based theoretical spectral-efficiency estimate in bits/s/Hz.

These metrics provide calculated estimates of the communication-level consequences associated with beam selection under clean and adversarial conditions. A higher effective SNR corresponds to a lower theoretical BER and a higher Shannon-based spectral-efficiency estimate, providing an additional perspective on beam-selection performance.

N. Evaluation Procedure

The evaluation is performed in two stages. First, the vulnerability of the beam-prediction models is assessed at the machine-learning level using clean accuracy, adversarial accuracy, perturbation-strength analysis, and attack success rate. The baseline and adversarially trained models are evaluated under both FGSM and PGD attacks.

Second, the predicted beams are evaluated at the communication level using beamforming gain ratio, effective SNR, theoretical BPSK BER, and theoretical spectral efficiency across the considered SNR range. This two-stage evaluation distinguishes degradation in beam classification from its corresponding calculated impact on communication performance.

VI. RESULTS AND DISCUSSION

A. Adversarial Beam-Prediction Performance

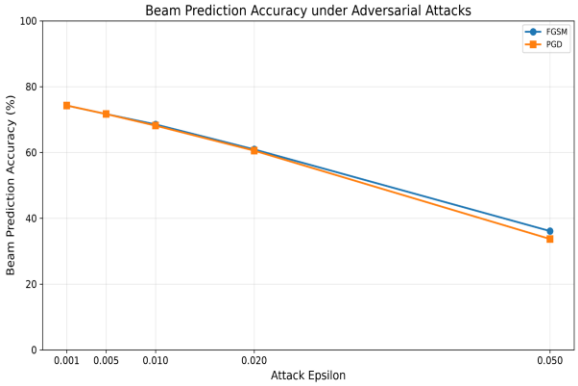

Fig. 1. Beam prediction accuracy of the adversarially trained model under FGSM and PGD attacks for different perturbation strengths.

Figure 1 shows a consistent decrease in beam-prediction accuracy as the adversarial perturbation strength increases. At the highest evaluated perturbation strength of ε=0.05, accuracy decreases to 36.11% under FGSM and 33.70% under PGD. The lower accuracy under PGD indicates that the iterative attack is more effective than FGSM under the evaluated configuration.

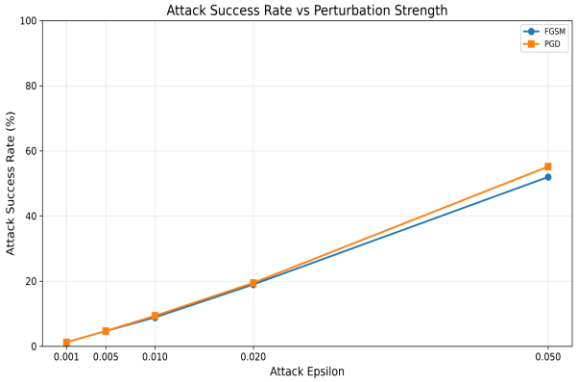

Fig. 2. Attack success rate of the adversarially trained beam-prediction model under FGSM and PGD attacks for different perturbation strengths.

Figure 2 shows that the attack success rate increases with perturbation strength for both attack methods. At ε=0.05, the reported success rates reach 51.97% for FGSM and 55.17% for PGD.

The trend is consistent with the reduction in beam-prediction accuracy as the perturbation budget increases.

B. Comparison of Baseline and Robust Models

Model	Clean	FGSM	PGD
Baseline	74.07%	28.70%	26.11%
Adversarially trained	75.19%	36.11%	33.70%

The baseline model achieves 74.07% accuracy under clean conditions, while the adversarially trained model achieves 75.19%. Under FGSM, adversarial training improves accuracy from 28.70% to 36.11%, an increase of 7.41 percentage points. Under PGD, accuracy improves from 26.11% to 33.70%, an increase of 7.59 percentage points. Although adversarial training provides a measurable robustness improvement, the substantial reduction from clean to adversarial accuracy demonstrates that the evaluated attacks remain effective.

C. Communication-Level Impact

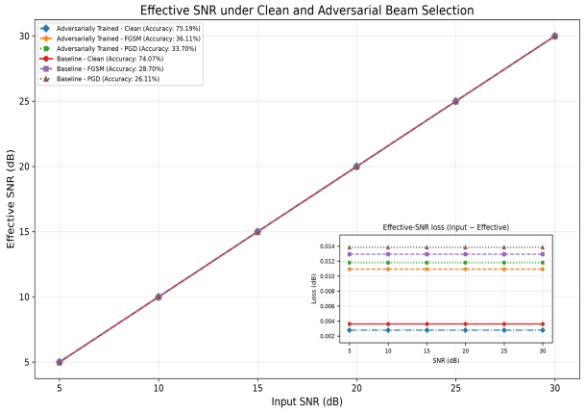

Fig. 3. Effective SNR under clean and adversarial beam-selection conditions.

Figure 3 compares the calculated effective SNR for clean and adversarial beam-selection conditions. The curves remain closely grouped across the investigated SNR range, indicating that the calculated effective-SNR degradation is small despite the substantial degradation observed in beam-prediction accuracy. The largest reported effective-SNR loss is approximately 0.014 dB for the baseline PGD condition.

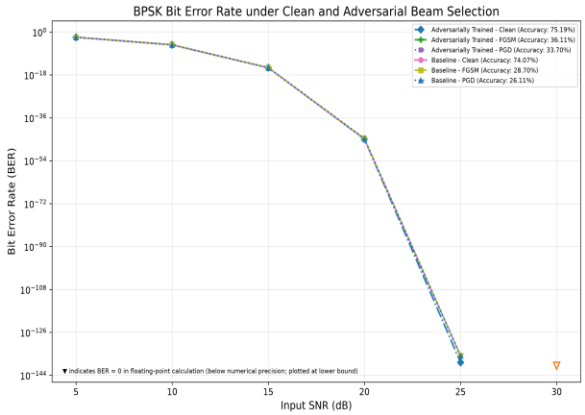

Fig. 4. Theoretical BPSK bit error rate under clean and adversarial beam-selection conditions.

Figure 4 presents the theoretical BPSK BER as a function of input SNR. The BER decreases rapidly as SNR increases, with clean and adversarial conditions exhibiting closely similar trends. At high SNR values, extremely small theoretical BER values approach the numerical precision limit of floating-point computation and may therefore appear as zero in the plotted results.

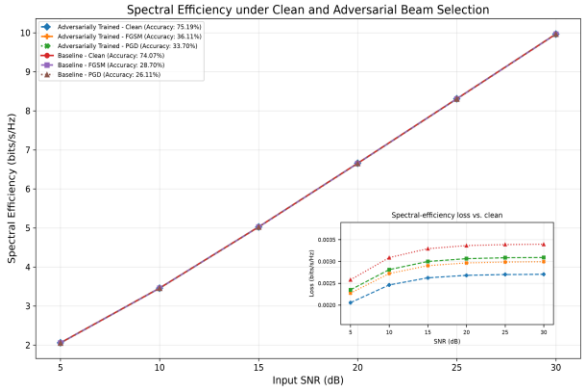

Fig. 5. Shannon-based spectral-efficiency estimate under clean and adversarial beam-selection conditions.

Figure 5 shows the Shannon-based spectral-efficiency estimate as a function of input SNR. The estimate increases monotonically with SNR for all evaluated conditions. The curves remain closely grouped, while the magnified comparison indicates a limited reduction under adversarial beam-selection conditions. Because this quantity is obtained from log2(1+SNR), it should be interpreted as a Shannon-based theoretical estimate rather than the actual spectral efficiency of the BPSK waveform.

D. Beamforming Performance

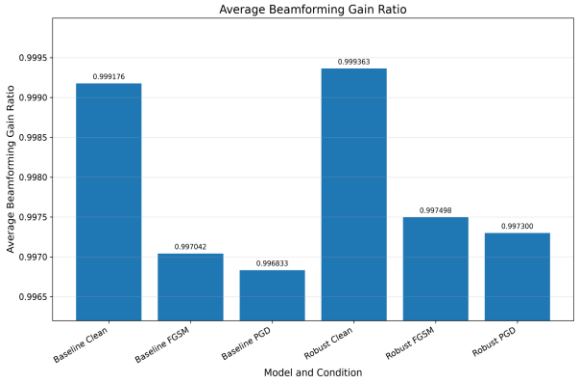

Fig. 6. Average beamforming gain ratio under clean, FGSM, and PGD conditions.

Figure 6 compares the average beamforming gain ratio of the baseline and adversarially trained models under clean and adversarial conditions. The gain ratios remain close to unity across all evaluated cases. The baseline model achieves gain ratios of 0.999176, 0.997042, and 0.996833 under clean, FGSM, and PGD conditions, respectively, while the adversarially trained model achieves 0.999363, 0.997498, and 0.997300. These results indicate that the calculated average beamforming gain changes only marginally under the evaluated attacks.

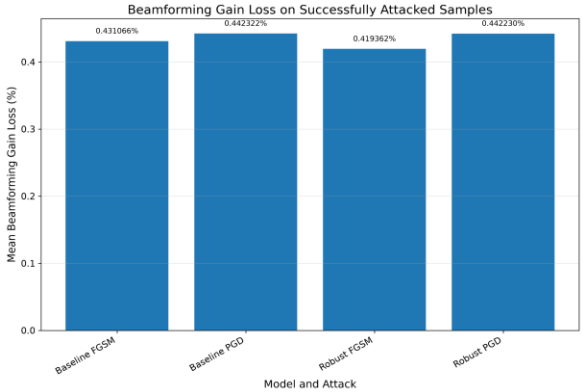

Fig. 7. Mean beamforming gains loss on successfully attacked samples.

Figure 7 presents the mean beamforming gain loss calculated over successfully attacked samples. The calculated loss ranges from approximately 0.419% to 0.442% across the evaluated model and attack combinations. The adversarially trained model exhibits a slightly lower loss under FGSM, while its PGD loss remains comparable to that of the baseline model. Overall, the evaluated attacks produce a

relatively small calculated reduction in average beamforming gain despite their substantial effect on beam-classification accuracy.

Overall, the experimental results show a clear distinction between machine-learning-level vulnerability and calculated communication-level impact. The baseline model accuracy decreases from 74.07% under clean conditions to 28.70% under FGSM and 26.11% under PGD, while adversarial training improves the corresponding adversarial accuracies to 36.11% and 33.70%. Thus, adversarial training provides a measurable but incomplete robustness improvement. In contrast, the average beamforming gain ratio remains close to unity, and the calculated effective SNR, theoretical BPSK BER, and Shannon-based spectral-efficiency estimate show comparatively small changes under the evaluated conditions. This mismatch indicates that classification accuracy alone can overstate the communication impact of beam-selection errors when neighboring codebook beams provide similar spatial focusing and beamforming gains.

VII. LIMITATIONS AND FUTURE WORK

The present study demonstrates the vulnerability of a deep-learning-based near-field beam-prediction model to adversarial input perturbations and shows that adversarial training can provide a partial improvement in robustness. However, several limitations should be considered when interpreting the results. First, the evaluation is based on a simulated 8×8 planar array with a fixed set of angular and distance locations. The results may therefore not directly generalize to larger arrays, different carrier frequencies, propagation environments, or user distributions. Second, the random train-test split is performed over realizations generated around the same reference beam locations; consequently, the evaluation primarily measures interpolation within the simulated spatial region rather than generalization to completely unseen user locations. Third, the adversarial evaluation is limited to input-space gradient-based attacks, specifically FGSM and PGD, and does not represent the full range of possible adversarial strategies or physically constrained wireless attacks. Fourth, although adversarial training improves robustness against the evaluated attacks, the resulting model remains vulnerable, indicating that adversarial training alone is insufficient to eliminate the security risk. Finally, the communication-level analysis is based on the simulated channel and beamforming model adopted in this study, and the reported results use a single fixed train-test split rather than repeated independent experimental runs.

Future work should investigate larger and more diverse XL-MIMO configurations, more realistic propagation environments, and adversarial attacks constrained by the physical wireless channel. Additional defense mechanisms, including alternative robust-training methods and attack-detection approaches, should also be considered. Robustness evaluation across different perturbation budgets, attack types, channel conditions, and unseen user locations would provide a stronger assessment of model generalization. Repeated experiments with independent random seeds would also strengthen statistical confidence in the reported results. Experimental validation using measured or real-world channel data would further improve the practical relevance of the security analysis. In addition, the present channel model assumes a single line-of-sight path and does not capture multipath propagation. Incorporating multipath and other realistic channel effects would allow future studies to examine whether the relationship between beam-classification errors and communication-level degradation remains similar under more realistic propagation conditions.

In addition, the adversarial perturbation budgets are defined in standardized feature space and therefore do not directly correspond to a physically realizable perturbation power or over-the-air attack constraint.

VIII. CONCLUSION

This study investigated the adversarial robustness of deep-learning-based near-field beam prediction in XL-MIMO systems under an inference-time input-space threat model. A simulated 8×8 planar array with 45 angle-distance candidate beams was used to formulate beam prediction as a supervised classification problem. A conventionally trained model and an adversarially trained model were evaluated under clean conditions and against FGSM and PGD attacks. The results show that adversarial perturbations can substantially degrade beam-prediction accuracy. The baseline model achieves 74.07% accuracy under clean

conditions, decreasing to 28.70% under FGSM and 26.11% under PGD at ε=0.05. Adversarial training improves the corresponding accuracies to 36.11% and 33.70% while maintaining comparable clean-data performance. These results demonstrate a measurable but incomplete robustness improvement.

More importantly, the large classification-level degradation does not produce a proportional calculated communication-level degradation in the considered codebook. The average beamforming gain ratio remains close to unity, while the calculated effective SNR, theoretical BPSK BER, and Shannon-based spectral-efficiency estimate show only small changes. This result highlights the importance of evaluating learned beam-selection systems using both machine-learning and communication-level metrics rather than classification accuracy alone.

The key implication is that an adversarially induced beam-classification error is not necessarily equivalent to a severe communication failure when neighboring near-field codebook beams correspond to spatially similar focusing locations. Thus, the practical impact of an adversarial attack depends on the structure of the beam codebook and the relationship between classification labels and physical beamforming performance.

Overall, the study establishes that deep-learning-based near-field beam prediction is vulnerable to adversarial input perturbations under the evaluated threat model, while also showing that the corresponding communication-level consequences can remain limited. Future work should test physically realizable attacks, larger arrays, multipath and measured channels, unseen user locations, repeated experimental runs, and stronger defense mechanisms to determine whether this classification-to-communication gap persists in realistic XL-MIMO deployments.

REFERENCES

[1] Z. Wang, J. Zhang, H. Du, W. E. I. Sha, B. Ai, D. Niyato, and M. Debbah, "Extremely Large-Scale MIMO: Fundamentals, Challenges, Solutions, and Future Directions," *IEEE Wireless Communications*, vol. 31, no. 3, pp. 117–124, Jun. 2024, doi: 10.1109/MWC.132.2200443.

[2] M. Cui, Z. Wu, Y. Lu, X. Wei, and L. Dai, "Near-Field MIMO Communications for 6G: Fundamentals, Challenges, Potentials, and Future Directions," IEEE Communications Magazine, vol. 61, no. 1, pp. 40–46, Jan. 2023, doi: 10.1109/MCOM.004.2200136.

[3] Y. Liu, Z. Wang, J. Xu, C. Ouyang, X. Mu, and R. Schober, "Near-Field Communications: A Tutorial Review," IEEE Open Journal of the Communications Society, vol. 4, pp. 1999–2049, 2023, doi: 10.1109/OJCOMS.2023.3305583.

[4] H. Lu and Y. Zeng, "Near-Field Modeling and Performance Analysis for Multi-User Extremely Large-Scale MIMO Communication," IEEE Communications Letters, vol. 26, no. 2, pp. 277–281, Feb. 2022, doi: 10.1109/LCOMM.2021.3129317.

[5] X. Wei and L. Dai, "Channel Estimation for Extremely Large-Scale Massive MIMO: Far-Field, Near-Field, or Hybrid-Field?" IEEE Communications Letters, vol. 26, no. 1, pp. 177–181, Jan. 2022, doi: 10.1109/LCOMM.2021.3124927.

[6] Y. Zhang, X. Wu, and C. You, "Fast Near-Field Beam Training for Extremely Large-Scale Array," IEEE Wireless Communications Letters, vol. 11, no. 12, pp. 2625–2629, Dec. 2022, doi: 10.1109/LWC.2022.3212344.

[7] C. Wu, C. You, Y. Liu, L. Chen, and S. Shi, "Two-Stage Hierarchical Beam Training for Near-Field Communications," IEEE Transactions on Vehicular Technology, vol. 73, no. 2, pp. 2032–2044, Feb. 2024, doi: 10.1109/TVT.2023.3311868.

[8] W. Liu, H. Ren, C. Pan, and J. Wang, "Deep Learning Based Beam Training for Extremely Large-Scale Massive MIMO in Near-Field Domain," IEEE Communications Letters, vol. 27, no. 1, pp. 170–174, Jan. 2023, doi: 10.1109/LCOMM.2022.3210042.

[9] G. Jiang and C. Qi, "Near-Field Beam Training Based on Deep Learning for Extremely Large-Scale MIMO," IEEE Communications Letters, vol. 27, no. 8, pp. 2063–2067, Aug. 2023, doi: 10.1109/LCOMM.2023.3289513.

[10] J. Nie, Y. Cui, Z.-H. Yang, W. Yuan, and X. Jing, "Near-Field Beam Training for Extremely Large-Scale MIMO Based on Deep Learning," IEEE Transactions on Mobile Computing, vol. 24, no. 1, pp. 352–362, Jan. 2025, doi: 10.1109/TMC.2024.3462960.

[11] B. Kim, Y. E. Sagduyu, T. Erpek, and S. Ulukus, "Adversarial Attacks on Deep Learning Based mmWave Beam Prediction in 5G and Beyond," in Proc. IEEE Statistical Signal Processing Workshop (SSP), 2021, pp. 590–594, doi: 10.1109/SSP49050.2021.9513738.

[12] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in Proc. International Conference on Learning Representations (ICLR), 2015.

[13] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and I. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," in Proc. International Conference on Learning Representations (ICLR), 2018.

[14] D. Adesina, C.-C. Hsieh, Y. E. Sagduyu, and L. Qian, "Adversarial Machine Learning in Wireless Communications Using RF Data: A Review," IEEE Communications Surveys & Tutorials, vol. 25, no. 1, pp. 77–100, 2023, doi: 10.1109/COMST.2022.3205184.

[15] J. Liu, M. Nogueira, J. Fernandes, and B. Kantarci, "Adversarial Machine Learning: A Multilayer Review of the State-of-the-Art and Challenges for Wireless and Mobile Systems," IEEE Communications Surveys & Tutorials, vol. 24, no. 1, pp. 123–159, 2022, doi: 10.1109/COMST.2021.3136132.

[16] M. Sadeghi and E. G. Larsson, "Adversarial Attacks on Deep-Learning Based Radio Signal Classification," IEEE Wireless Communications Letters, vol. 8, no. 1, pp. 213–216, Feb. 2019, doi: 10.1109/LWC.2018.2867459.

[17] B. Kim, Y. Shi, Y. E. Sagduyu, T. Erpek, and S. Ulukus, "Adversarial Attacks against Deep Learning Based Power Control in Wireless Communications," in Proc. IEEE GLOBECOM Workshops, 2021, doi: 10.1109/GCWkshps52748.2021.9682097.