2023

Defending AI-Based Automatic Modulation Recognition Models Against Adversarial Attacks

Haolin Tang

Ferhat Ozgur Catak

Murat Kuzlu
Old Dominion University, mkuzlu@odu.edu

Evren Catak

Yanxiao Zhao

Original Publication Citation

Tang, H., Catak, F. O., Kuzlu, M., Catak, E., & Zhao, Y. (2023). Defending AI-based Automatic Modulation Recognition models against adversarial attacks. *IEEE Access*, *11*, 76629-76637. https://doi.org/10.1109/ACCESS.2023.3296805

IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 27 June 2023, accepted 10 July 2023, date of publication 19 July 2023, date of current version 28 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3296805

RESEARCH ARTICLE

Defending AI-Based Automatic Modulation Recognition Models Against Adversarial Attacks

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

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

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

This work was supported in part by the Commonwealth Cyber Initiative (CCI), an investment in the advancement of Cyber Research and Development, Innovation, and Workforce Development in Virginia.

ABSTRACT Automatic Modulation Recognition (AMR) is one of the critical steps in the signal processing chain of wireless networks, which can significantly improve communication performance. AMR detects the modulation scheme of the received signal without any prior information. Recently, many Artificial Intelligence (AI) based AMR methods have been proposed, inspired by the considerable progress of AI methods in various fields. On the one hand, AI-based AMR methods can outperform traditional methods in terms of accuracy and efficiency. On the other hand, they are susceptible to new types of cyberattacks, such as model poisoning or adversarial attacks. This paper explores the vulnerabilities of an AI-based AMR model to adversarial attacks in both single-input-single-output and multiple-input-multiple-output scenarios. We show that these attacks can significantly reduce the classification performance of the AI-based AMR model, which highlights the security and robustness concerns. Therefore, we propose a widely used mitigation method (i.e., defensive distillation) to reduce the vulnerabilities of the model against adversarial attacks. The simulation results indicate that the AI-based AMR model can be highly vulnerable to adversarial attacks, but their vulnerabilities can be significantly reduced by using mitigation methods.

INDEX TERMS Artificial intelligence, next-generation networks, automatic modulation recognition, adversarial attacks, model poisoning, defensive distillation.

I. INTRODUCTION

In recent years, significant support for next-generation networks has been provided due to the high demand for new-era applications, such as mobile health, self-driving cars, metaverse, digital twins, extended and virtual reality (XR and VR), as well as the requirement for more intelligent networks, ultra-low latency, extremely high data speed, and ability to support a massive number of various mobile and Internet of Things (IoT) devices with extreme density. The high communication performance required by diverse applications can be achieved through advanced communication and networking solutions. Among these technologies, Automatic Modulation Recognition (AMR) plays a critical role in the signal processing chain as it involves identifying the modulation signal efficiently and accurately even when insufficient or lacking prior information exists. AMR can significantly improve overall communication performance. Various AMR solutions using different methods, such as instantaneous features [1], the optimized linear combination of higher-order cumulants [2], wavelet transforms [3], and cyclic spectrum [4], have been proposed in the literature.

Recent studies have demonstrated that Artificial Intelligence (AI) and Machine Learning (ML) based solutions outperform across all aspects of next-generation networks, from the physical layer to the application layer [5], [6], [7].

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

A conceptual model for 6G has been presented in [8], which emphasizes the importance of AI/ML-powered solutions at each layer of the model to meet the requirements of next-generation wireless networks in terms of latency, power allocation, privacy, security, and more. AMR methods can significantly improve next-generation network performance using Deep Learning (DL) methods. AMR is defined as a multi-classes classification problem in ML/DL implementation. The study [9] provides a comprehensive overview of DL-based AMR models for wireless communications and conducts extensive simulations for both Single-Input-Single-Output (SISO) and Multiple-Input-Multiple-Output (MIMO) communication systems to analyze the performance of different DL-based AMR models. Although there has been considerable research on DL-based AMR solutions, little attention has been paid to the security threats these models may encounter, such as adversarial ML attacks or model poisoning. Adversarial attacks are widely used cyberattacks that can severely compromise the accuracy of most ML models. These attacks introduce Adversarial Examples (AEs) or manipulated input with slight differences during training to mislead the model's performance. Although AEs are often imperceptible to humans, they can cause the model to misclassify or be directed erroneously.

In recent research, we have investigated adversarial threats and mitigation methods for various communication systems, including mmWave beamforming [10], channel estimation [11], and Intelligent Reflecting Surfaces (IRS) [12], [13], [14]. This study aims to investigate the susceptibilities of AI-based AMR models to adversarial attacks and propose a mitigation approach that can enhance their resilience for both SISO and MIMO communication systems. We adopt an AI-based AMR model, i.e., Long Short Term Memory (LSTM)-based neural network, from [15] since the main focus of this study is to investigate the vulnerability of AI-based AMR models against adversarial attacks, not to develop new AI-based AMR models. Regarding mitigation methods, this paper employs the defensive distillation method to increase the model's robustness against adversarial attacks. Four different adversarial attack methods, i.e., Fast Gradient Sign Method (FGSM), Momentum Iterative Method (MIM), Basic Iterative Method (BIM), and Projected Gradient Descent (PGD), are applied to models to investigate the attack success ratio. From this, the vulnerability of each model is identified. Simulation results demonstrate that AI-based AMR models are vulnerable to adversarial attacks, but it is shown that the proposed defensive distillation mitigation method can effectively enhance the robustness of LSTM-AMR models against such attacks.

The main contributions of this paper are summarized as follows.

- We explore the vulnerabilities of an AI-based AMR model under adversarial attacks in both SISO and MIMO communication systems. An LSTM-based AMR model is adopted for simulations.
- We conduct simulations using four different adversarial attack methods (FGSM, MIM, BIM, and PGD) to analyze the attack performance in terms of attack success ratio and identify the potential weaknesses in the communication system.
- We propose a widely used mitigation method, i.e., defensive distillation, to reduce the model's vulnerabilities against adversarial attacks and compare the robustness of the defended model to the undefended model.

The rest of the paper is organized as follows. Section II introduces the preliminaries in AMR, adversarial attacks, and the mitigation method. Section III introduces the adopted LSTM-AMR model and data preparation. Section IV presents simulation results and observations, and Section V concludes.

II. PRELIMINARIES

A. AUTOMATIC MODULATION RECOGNITION (AMR)

Wireless communication systems typically employ various modulation techniques to modulate signals for efficient data transmission. AMR is an intermediary process between signal detection and signal demodulation that relies solely on received signals to identify the modulation scheme of the transmitted signals without any additional auxiliary information. In the last few decades, the modulation methods have become more complex and diverse to meet the requirements of increasingly complex communication scenarios. Meanwhile, various AMR methods have been developed to achieve effective modulation recognition. AMR methods typically can be classified into likelihood-based AMR and feature-based AMR [9]. Likelihood-based AMR approaches essentially formulate modulation recognition as a multi-hypothesis test, in which the likelihood function of a received signal is compared with a threshold under the assumption of the known probability density function of the signal [16]. However, they suffer from high computational complexity. Feature-based AMR approaches mainly perform two steps: feature extraction and classification. Instantaneous features and/or statistical features are extracted from the received signals, and then decision-making methods are adopted to classify the received signals based on the extracted features. By contrast, feature-based AMR methods provide sub-optimal performance but with low computational complexity.

In the last decade, we have witnessed tremendous achievement by applying DL methods to various challenging applications where traditional methods are not able to provide promising performance, such as computer vision and natural language processing. This also inspires the continuing research of AMR through a learning manner. In one of the early DL based AMR studies [17], the authors proposed a Convolutional Neural Network (CNN) to extract features from a modulated signal for modulation classification. The results showed that the proposed CNN outperforms traditional methods in terms of accuracy. The authors in [18] introduced a DL-based AMR algorithm that utilizes CNN and Recurrent Neural Networks (RNN) to extract representative and effective features automatically. This algorithm was designed to classify various signal modulations, such as BFSK, DQPSK, MSK, GMSK, 4PAM, and 16QAM, under different channel conditions, such as additive white Gaussian noise (AWGN) and Rayleigh fading. In [19], this paper combines the advantages of CNN and the long short-term memory (LSTM) to extract the spatial and temporal features of signals, respectively, to improve the performance of the DL-based AMR method further. In conclusion, DL-based AMR methods provide better performance than traditional algorithms.

B. ADVERSARIAL ATTACKS

Adversarial attacks are a type of cyberattack that aims to reduce the accuracy of a machine-learning model by adding imperceptible perturbations to the input data. Perturbation refers to modifications or alterations made to the input data, which can take different forms. For example, noise can be considered one such form of perturbation, and the magnitude of perturbation, i.e., noise level, can be quantified using different scales, including dB-based scales, depending on the context and domain.

Formally, given an input sample $x \in \mathbb{R}^n$ and a machine learning model $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, an adversarial example x_{adv} can be generated as follows:

1) Choose a perturbation direction $\delta \in \mathbb{R}^n$.
2) Add the perturbation to the input sample: $x_{adv} = x + \delta$.
3) Ensure that the perturbation is small enough to be imperceptible: $|\delta|_p \leq \epsilon$, where ϵ is a small constant and $| \cdot |p$ denotes the p-norm.
4) Ensure that the perturbation causes misclassification: $f(xadv) \neq f(x)$.

There are several different methods for generating adversarial examples, including the following:

1) **Fast Gradient Sign Method (FGSM):** This method generates adversarial examples by adding a small perturbation in the direction of the gradient of the loss function with respect to the input. Formally, the perturbation is given by $\delta = \epsilon$, $\text{sign}(\nabla_x J(x, y))$, where $J(x, y)$ is the loss function, y is the actual label of the input x, and ϵ is a small constant.
2) **Basic Iterative Method (BIM):** BIM is a variant of the FGSM attack where multiple small perturbations are added iteratively to the input. Formally, the perturbation at each iteration i is given by $\delta_i = \epsilon \text{sign}(\nabla_x J(x_{i-1}, y))$, where $J(x_{i-1}, y)$ is the loss function concerning the previous iteration's adversarial example x_{i-1} and y is the true label of the input x. The final adversarial example is obtained by clipping the perturbations to ensure that they remain within the ϵ-ball around the original input: $x_{adv} = \text{clip}\epsilon(x + \sum i = 1^k \delta_i)$, where k is the number of iterations and clip_ϵ is the projection operator onto the ϵ-ball.
3) **Moment Iterative Method (MIM):** MIM is a variant of BIM that adds a momentum term to the iterative updates to smooth out the perturbations and improve the transferability of the adversarial examples. Formally, the update rule at each iteration i is given by $\delta_i = \alpha\delta_{i-1} + \frac{\epsilon}{|\nabla_x J(x_{i-1}, y)|1} \text{sign}(\nabla_x J(xi - 1, y))$, where α is a momentum parameter and $|\cdot|_1$ denotes the L_1 norm. The final adversarial example is obtained in the same way as BIM.
4) **Projected Gradient Descent (PGD):** PGD is a more powerful iterative attack that performs multiple steps of gradient descent with small step sizes, followed by a projection onto the ϵ-ball. Formally, the update rule at each iteration i is given by $x_i = \text{clip}\epsilon(xi - 1 + \alpha \text{sign}(\nabla_x J(x_{i-1}, y)))$, where α is the step size and $\text{clip}\epsilon$ is the projection operator onto the ϵ-ball. The final adversarial example is obtained by taking the output of the last iteration: $xadv = x_k$.

C. DEFENSIVE DISTILLATION

Defensive distillation is a popular method to defend machine learning models against adversarial attacks [20], [21]. Defensive distillation aims to train a new model less sensitive to small perturbations in the input data. The process involves two main steps: first, a softened probability distribution is used as a label distribution for the original model, and second, a distilled model is trained using the softened labels.

The first step involves using a softened probability distribution p_{soft} as the label distribution for the original model $f(x; \theta)$. The softmax function with temperature τ is used to obtain the softened probabilities:

$$p_{\text{soft}}(y|x; \tau) = \frac{\exp(f_y(x; \theta)/\tau)}{\sum_i \exp(f_i(x; \theta)/\tau)}, \tag{1}$$

where y is the actual label for the input x, the softened probabilities are smoother than the hard probabilities obtained from the regular softmax function. They are effective in defending against certain types of adversarial attacks.

The second step involves training a distilled model $g(x; \phi)$ using the softened labels p_{soft}. The objective function for training the distilled model is given by:

$$\min_{\phi} \mathbb{E}(x, y) \sim D[-\sum i p_{\text{soft}}(i|x; \tau) \log g_i(x; \phi)], \tag{2}$$

where D is the dataset used for training, the objective function encourages the distilled model to predict similar probabilities to the original model while being less sensitive to small perturbations in the input data.

A pseudocode implementation of the defensive distillation is given in Algorithm 1.

In the algorithm, N is the number of training examples, and C is the number of classes in the dataset. The algorithm iteratively updates the parameters of the distilled model by minimizing the loss function, where the loss is defined using the softened probabilities p_{soft} and the predictions of the distilled model $g(x; \phi)$. The number of iterations T is a hyperparameter that can be tuned to achieve the best performance on a validation set.

Algorithm 1 Defensive Distillation

```
Input: Training data $(x_i, y_i)$, original model $f(x; \theta)$, temperature $\tau$, number of iterations $T$
Output: Distilled model $g(x; \phi)$
Initialize distilled model parameters $\phi$
for $t = 1$ to $T$ do
    Compute softened probabilities $p_{\text{soft}}(y|x; \tau)$ for each training example $(x, y)$
    Update distilled model parameters $\phi$ by minimizing the loss:
        $$\mathcal{L}(\phi) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} p_{\text{soft}}(j|x_i; \tau) \log g_j(x_i; \phi)$$
    $iter \leftarrow iter + 1$
end for
return $g(x; \phi)$
```

Defensive distillation is a powerful technique for defending machine learning models against adversarial attacks. However, it is not foolproof and can still be vulnerable to specific attacks. Evaluating the effectiveness of defensive distillation of a particular dataset and attack scenario is crucial.

III. MODEL DESCRIPTION AND DATASET PREPARATION

In this section, we first introduce the dataset preparation for both SISO and MIMO scenarios and then describe the adopted LSTM-based AMR model in the simulations. SISO refers to a communication system with only one antenna at the transmitter and one at the receiver while MIMO represents a communication system with multiple antennas at both the transmitter and the receiver. MIMO systems are used in modern wireless communication standards, such as 4G LTE, 5G, and beyond, to improve the data throughput, increase the range, and enhance the reliability of the communication link. The critical differences between SISO and MIMO systems are antenna configuration, channel capacity, complexity, and performance. MIMO systems typically provide higher data rates, longer ranges, and better reliability than SISO systems, especially in environments with multipath propagation and interference.

A. DATASET PREPARATION FOR SISO SCENARIO

To investigate the performance and vulnerability of the AI-based (i.e., LSTM-based) AMR model in a SISO scenario, the GNU radio ML dataset RML2016.10a [22] is adopted for simulations since this dataset is publicly available and widely used in research as the benchmark. There are 220,000 signal samples in the GNU radio ML dataset RML2016.10a, and each sample is associated with one modulation at a specific Signal-to-Noise Ratio (SNR). Each sample consists of a 256-dimensional vector comprising 128 in-phase and 128 quadrature components. There are 11 different modulations, including BPSK, QPSK, 8PSK, QAM16, QAM64, CPFSK, GFSK, PAM4, WBFM, AM-SSB, and AM-DSB. The data samples are constructed at 20 different SNR levels from -20 dB and 18 dB with an interval of 2 dB.

B. DATASET PREPARATION FOR MIMO SCENARIO

MIMO system with precoding is adopted from [9]. It is a common MIMO system that consists of a transmitter with N_t antennas and a receiver with N_r antennas. The transmitter and receiver are assumed to have full knowledge of the channel, and the transmission is over a flat fading channel. With the MIMO system above, we generate the dataset with three different antenna setting groups: $(N_t = 4, N_r = 2)$, $(N_t = 16, N_r = 4)$ and $(N_t = 64, N_r = 16)$. The signal samples are modulated with six different modulations, i.e., 2PSK, QPSK, 8PSK, 16QAM, 64QAM, and 128QAM, at different SNR levels from -10 dB to 20 dB. 500 samples are prepared per SNR for each modulation and the number of transmitted symbols per signal sample is 128.

C. MODEL DESCRIPTION

This subsection explains the LSTM-based AMR model adopted from [15]. Recurrent Neural Networks (RNNs) are commonly applied for learning persistent features of sequence data. LSTM is a particular type of RNN that is efficient in learning long-term dependencies and is heavily used for natural language processing and signal processing [23]. The major components in an LSTM cell are three gates, namely the input gate, forget gate, and the output gate, which are used to control how the information propagates in the network. The gating mechanism allows LSTM cells to memorize information for extended periods, thus realizing continuous feature learning. The key equations of an LSTM cell are listed below:

$$\begin{aligned} i_t &= \sigma(x_t U^i + h_{t-1} W^i + b_i) \\ f_t &= \sigma(x_t U^f + h_{t-1} W^f + b_f) \\ o_t &= \sigma(x_t U^o + h_{t-1} W^o + b_o) \\ \hat{C}_t &= tanh(x_t U^g + h_{t-1} W^g + b_c) \\ C_t &= f_t \odot C_{t-1} + i_t \odot \hat{C}_t \\ h_t &= o_t \odot tanh(C_t U^o) \end{aligned} \tag{3}$$

where x_t is input vector, i_t is input gate vector, f_t is forget gate vector, o_t is output gate vector, c_t is cell state vector, h_t is hidden state vector, b_i, b_f, b_o, b_c are bias vectors, U, W is parameter matrices, and σ, $tanh$ are activation functions. $\odot$ denotes the Hadamard product for the element-wise product of matrices.

The adopted LSTM-based AMR model consists of two LSTM layers followed by a fully connected layer and a softmax layer as shown in Figure 1. The in-phase and quadrature components of modulated signals are fed to the model as a two-dimensional vector. The first two LSTM layers have 128 LSTM units each, and the output of the last LSTM layer is a 128-dimensional vector which is passed to the following fully connected linear layer and softmax layer. In the

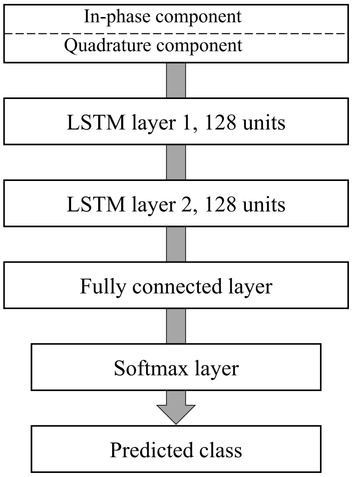


FIGURE 1. The architecture of the LSTM-based AMR model. This model is trained for signal modulation recognition using the amplitude-phase signal.

SISO scenario, the softmax layer maps the features learned from previous layers to one of 11 output classes indicating the 11 modulation schemes. In the MIMO scenario, the softmax layer maps the features learned from previous layers to one of 6 output classes since there are six different modulations in the MIMO dataset. Essentially, the reason to use an LSTM model for signal classification is that signals with different modulation schemes contain different amplitude and phase features, and the LSTM model is capable of learning these temporal features effectively.

To train the LSTM model for modulation recognition, first, the SISO and MIMO datasets are split into training, validation, and test sets at a ratio of 6:2:2 for SISO and MIMO scenarios, respectively. The loss function used is categorical cross-entropy, and the initial learning rate is set to 0.001 with the Adam optimizer. The learning rate will be halved if the validation loss does not decrease within 5 epochs, and the training process will be stopped if the validation loss remains stable for 50 epochs. The batch size is set to 400, and the training process is conducted using an Nvidia GTX 1080Ti GPU and Keras with Tensorflow as the backend.

IV. EXPERIMENTS

This section provides the experimental results for the SISO and MIMO scenarios using LSTM-based AMR undefended and defended models, with the attack success ratio. The attack success ratio refers to the ratio of successfully transmitted malicious data or signals to the total amount of data or signals transmitted. It is also widely used in communication systems to assess their security and provide a measure of how vulnerable the system is to different types of attacks. In this study, the experimental results are obtained by averaging across multiple iterations, i.e., 30 times. The analysis focuses on the attack success ratio of four different adversarial attack methods (BIM, FGSM, MIM, and PGD) with and without the application of a mitigation method (defensive distillation). The objective is to identify potential weaknesses in the communication system.

A grid search approach is employed to determine the optimal parameters for defensive distillation-based adversarial ML attack mitigation. The grid search involved systematically exploring a predefined parameter grid to find the parameter combination that yielded the best performance. The parameters considered in the grid search included the temperature parameter for defensive distillation, the regularization strength, and the learning rate. The combination that resulted in the highest model robustness against adversarial attacks was identified by exhaustively searching the parameter grid. This grid search methodology ensures a comprehensive exploration of parameter space, leading to an informed selection of the optimal parameters for defensive distillation in the context of adversarial ML attack mitigation.

A. SIMULATION RESULTS IN SISO SCENARIO

In a SISO scenario, the transmitter sends a single signal, which is received by the receiver over a single channel. This type of system is commonly used in simple point-to-point communication links, such as those between a mobile phone and a base station. Fig. 2 illustrates the attack success ratio of the undefended SISO model for each adversarial attack, i.e., BIM, FGSM, MIM, and PGD. According to the figure, the developed model is not robust under BIM, MIM, and PGD attacks, i.e., the attack success ratio can go up to 1.0 even under attack powers $\epsilon < 0.06$. However, the FGSM attack has a low success ratio compared to other attack methods, i.e., the maximum attack success ratio is 0.6 under a heavy attack power $\epsilon = 1.0$. It means the developed AMR model is robust against FGSM attacks. In some cases, FGSM attacks may be less effective than other more sophisticated attacks, such as BIM, MIM, and PGD attacks. Therefore, it is important to carefully consider the threat model and evaluate the effectiveness of different attack methods under different scenarios.

Table 1 shows the attack success ratio of different types of attacks along with different levels of attack strength for the undefended SISO model in detail. The first row shows the attack strength, ranging from 0.01 to 1.0, and the first column shows the names of the attack types, i.e., BIM, FGSM, MIM, and PGD. According to the table, BIM, MIM, and PGD have a high success ratio for most attack powers (ϵ), while FGSM has a lower success ratio. This indicates that the other three attacks may be more effective than FGSM in generating adversarial attacks. For example, the BIM attack had a success ratio of 0.38, the FGSM attack had a success ratio of 0.06, the MIM attack had a success ratio of 0.75, and the PGD attack had a success ratio of 0.50 at an attack strength of 0.01. On the other hand, the values of the success ratio go up to 1.00, 0.58, 1.00, and 1.00 for BIM, FGSM, MIM, and PGD attacks at the highest attack power ($\epsilon = 1.0$), respectively.

TABLE 1. Attack success ratio of the undefended SISO model.

Attack	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
BIM	0.38	0.92	1.00	1.0	1.00	1.0	1.00	1.00	1.00	1.00	1.0	1.00	1.00	1.00	1.0	1.0	1.00	1.0	1.00
FGSM	0.06	0.00	0.08	0.0	0.07	0.0	0.17	0.23	0.11	0.11	0.1	0.35	0.41	0.47	0.4	0.5	0.53	0.6	0.58
MIM	0.75	1.00	1.00	1.0	1.00	1.0	1.00	1.00	1.00	1.00	1.0	1.00	1.00	1.00	1.0	1.0	1.00	1.0	1.00
PGD	0.50	0.87	1.00	1.0	0.94	1.0	1.00	1.00	1.00	1.00	1.0	1.00	1.00	1.00	1.0	1.0	1.00	1.0	1.00

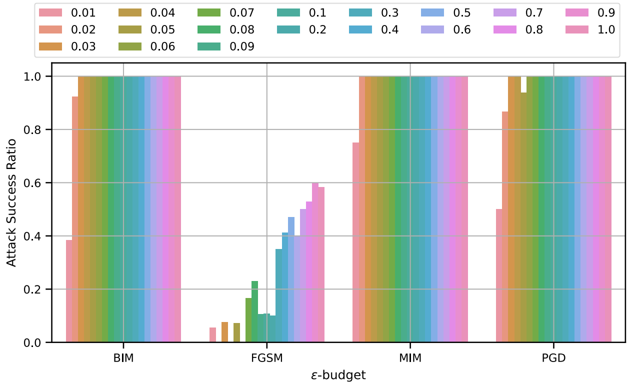


FIGURE 2. Attack success ratio of the undefended SISO model.

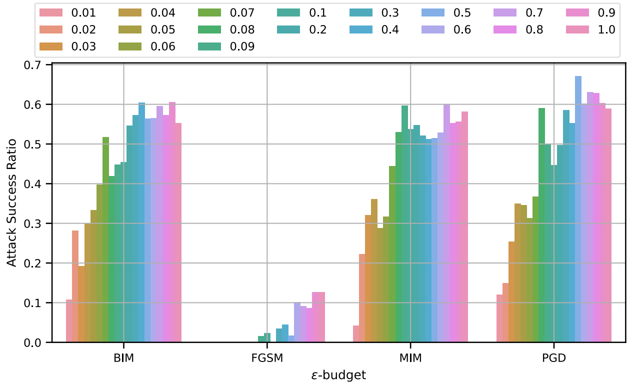


FIGURE 3. Attack success ratio of the defended SISO model.

Fig. 3 shows the attack success ratio of the defended SISO model under the selected attacks, i.e., BIM, FGSM, MIM, and PGD. According to the figure, all attack success ratio values decrease under all attack types compared to the undefended model. BIM, MIM, and PGD show similar trends. However, the attack success ratio values vary between around 0.1 to 0.6 under light ($\epsilon = 0.01$) and heavy attack powers ($\epsilon = 1.0$). As expected, the FGSM attack has a low success ratio compared to other attack methods; i.e., the maximum ratio is around 0.1 under all attack powers ϵ, and the developed model is more robust against FGSM attacks.

Table 2 provides detailed information about the performance of different attack methods on a machine learning model in terms of attack success ratio at different levels of attack power. The table is organized in a grid format, with the rows indicating the attack methods (BIM, FGSM, MIM, and PGD) and the columns indicating the strength of the attack (ranging from 0.01 to 1.0). Each cell in the table represents the success ratio of the corresponding attack method at the corresponding level of attack power. For instance, the success ratio of the BIM at 0.01 attack power is 0.11, while the attack success ratio at 1.0 attack power is 0.55. Similarly, the attack success ratio of MIM attack at 0.1 power is 0.04, while its attack success ratio at 1.0 attack power is 0.58. Note that the FGSM attack method has the least impact on the machine learning model, as its success ratio is consistently low at all levels of attack power. The minimum success ratio for FGSM is 0.00, while the maximum success ratio is 0.13 (at attack power of 0.9 and 1.0). On the other hand, the BIM, MIM, and PGD attack methods are more effective at compromising the model's performance. For BIM/MIM/PGD, the minimum success ratios are 0.11, 0.04, and 0.12 (at an attack power of 0.01), respectively. The maximum success ratios for these methods are 0.61, 0.60, and 0.67 at a high attack power, respectively.

B. SIMULATION RESULTS IN MIMO SCENARIO

In a MIMO scenario, multiple signals are transmitted simultaneously over multiple channels, and the receiver uses advanced signal processing techniques to separate and decode the signals. Fig. 4 shows the attack success ratio of the defended MIMO model under the selected adversarial attacks. According to the figure, BIM/MIM/PFG attacks are very effective, and the attack success ratio values can achieve 1.0 (i.e., 100%) even at mid-level attack powers, $\epsilon >= 0.5$. As in the previous scenario, the FGSM attack has a low attack success ratio compared to other attack methods, i.e., the maximum attack success ratio is around 0.4 under heavy attack powers $\epsilon = 1.0$. It is obvious that the attack success ratio increases with the attack power in parallel. The details will be investigated in the following table.

Table 3 presents the attack success ratio for the selected four adversarial attacks (FGSM, BIM, MIM, and PGD) on the developed undefended MIMO model at different levels of attack powers (from 0.01 to 1.0). According to the table, all attack types except FGSM seem very effective, as they achieve 100% attack success ratio on several high attack powers. Among them, BIM and MIM are the most effective attack methods against the model, as they achieve a high success ratio across a wide range of strength levels. On the other hand, FGSM is not very effective at lower strength levels (0.01 and 0.1), but becomes more effective as the strength level increases.

TABLE 2. Attack success ratio of the defended SISO model.

Attack	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
BIM	0.11	0.28	0.19	0.30	0.33	0.40	0.52	0.42	0.45	0.45	0.55	0.57	0.60	0.56	0.57	0.60	0.57	0.61	0.55
FGSM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.04	0.04	0.02	0.10	0.09	0.09	0.13	0.13
MIM	0.04	0.22	0.32	0.36	0.29	0.32	0.44	0.53	0.60	0.54	0.55	0.52	0.51	0.52	0.53	0.60	0.55	0.56	0.58
PGD	0.12	0.15	0.25	0.35	0.35	0.31	0.37	0.59	0.50	0.45	0.50	0.59	0.55	0.67	0.60	0.63	0.63	0.60	0.59

TABLE 3. Attack success ratio of the undefended MIMO model.

Attack	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
BIM	0.48	0.67	0.79	0.87	1.00	1.00	1.00	0.94	1.00	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FGSM	0.00	0.02	0.00	0.02	0.01	0.04	0.01	0.07	0.06	0.05	0.09	0.15	0.20	0.25	0.32	0.32	0.36	0.37	0.41
MIM	0.53	0.73	0.76	0.67	0.82	0.92	0.94	1.00	1.00	1.00	1.00	0.96	0.90	1.00	0.85	0.97	1.00	1.00	1.00
PGD	0.35	0.71	0.88	0.89	0.90	0.87	0.86	0.95	1.00	1.00	1.00	1.00	0.96	1.00	0.96	1.00	1.00	0.97	0.96

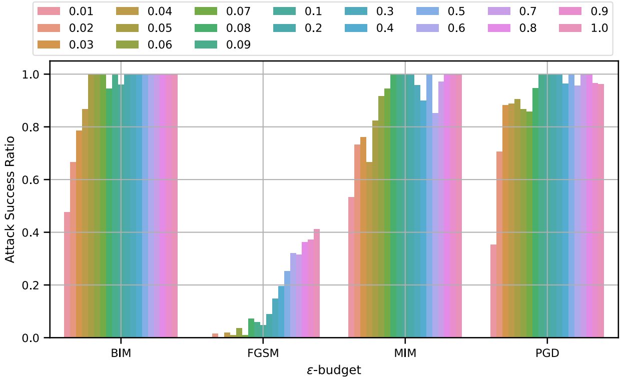


FIGURE 4. Attack success ratio of the undefended MIMO model.

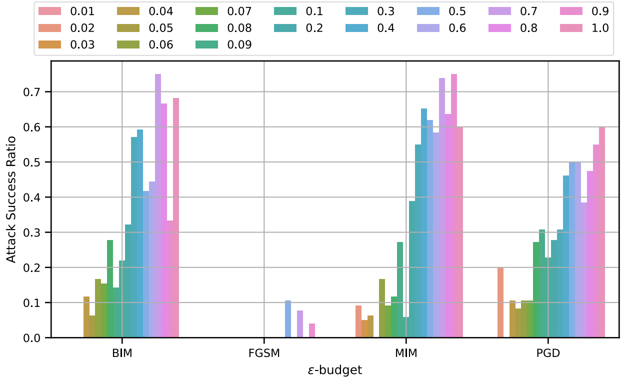


FIGURE 5. Attack success ratio of the defended MIMO model.

Fig. 5 illustrates the attack success ratio of the defended MIMO model for the same attacks and attack powers as in the previous scenario. The figure shows that the attack success ratio values significantly decrease for the defended MIMO model, especially for mid-level attack power. BIM/MIM/PGD attacks exhibit similar trends, i.e., having a low attack success ratio at low attack power and a high attack success ratio at high attack powers. As expected, the FGSM attack method has the least impact, i.e., almost none, as its success ratio is consistently low at all levels of attack power, i.e., around 0.1. Some results show a zero (0) attack success ratio, meaning the attack success ratio is very low or almost 0.

Table 4 provides more detailed information regarding the attack success ratio of different adversarial attack methods on the defended MIMO model at different attack powers. According to the table, the FGSM attack has almost no impact on the defended MIMO model at all attack powers, i.e., the maximum attack success ratio is 0.11. Other attack types (BIM/MIM/PGD) still impact the defended model. For example, looking at the BIM attack, at 0.01 attack power, the success ratio is 0.0, meaning the attack was not successful. However, at 0.7 attack power, the success ratio jumps to 0.75. For the MIM attack, the success ratio for 0.01 attack power is 0.0, but it increases to 0.09 for 0.02 attack power. The success ratio remains low for the following attack powers but increases substantially for higher attack powers, reaching a maximum success ratio of 0.75 for 0.9 attack power. For the PGD attack, the success ratio remains 0.0 for 0.01 attack power, but it increases to 0.20 for 0.02 attack power. The success ratio then varies between 0.0 and 0.5 for different attack powers and reaches a maximum value of 0.60 for 1.0 attack power.

C. OBSERVATIONS

This study aims to investigate the performance and vulnerabilities of AI-based AMR models under popular adversarial attacks, such as FGSM, BIM, MIM, and PGD, as well as the impact of the selected mitigation method (defensive distillation) on performance improvement. The simulation results indicate that AI-based AMR models are vulnerable to model poisoning attacks, but the impact can be reduced or eliminated with mitigation methods. Based on the findings, the following observations can be made:

Observation 1: Adversarial attacks are effective in compromising the accuracy of deep learning models, with attack success ratios ranging from 0% to over 100% depending on the attack method and power.

Observation 2: The attack success ratio of adversarial attacks tends to increase with the attack power. In most cases, attack

TABLE 4. Attack success ratio of the defended SISO model.

Attack	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
BIM	0.0	0.00	0.00	0.12	0.06	0.17	0.15	0.28	0.14	0.22	0.32	0.57	0.59	0.42	0.44	0.75	0.67	0.33	0.68
FGSM	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.08	0.00	0.04	0.00
MIM	0.0	0.09	0.05	0.06	0.00	0.17	0.09	0.12	0.27	0.06	0.39	0.55	0.65	0.62	0.58	0.74	0.64	0.75	0.60
PGD	0.0	0.20	0.00	0.11	0.08	0.11	0.11	0.27	0.31	0.23	0.28	0.31	0.46	0.50	0.50	0.38	0.47	0.55	0.60

success ratios increase rapidly as the attack power goes from 0.01 to 0.1, but then plateau or increase more slowly for larger attack powers.

Observation 3: Mitigation methods can reduce the attack success ratios of adversarial attacks, but their effectiveness varies depending on the attack method and power.

Observation 4: MIMO models can provide better defense against adversarial attacks compared to SISO models.

Observation 5: Adversarial attacks significantly impact both undefended and defended SISO/MIMO models in terms of attack success ratio, particularly for BIM/MIM/PGD attacks.

Observation 6: FGSM attack method has the least impact on models, as its success ratio is consistently low at all levels of attack power.

Observation 7: PGD is the most effective attack against the defended SISO model, with an attack success ratio of 0.67.

Observation 8: MIM is the most effective attack against the defended MIMO model, with an attack success ratio of 0.75.

V. CONCLUSION

As we continually integrate AI/DL technologies into AMR to improve communication performance, it has also aroused security issues that do not receive sufficient attention in the literature. The main objective of this study is to evaluate the performance of AI-based AMR models and their robustness against various adversarial attacks (i.e., FGSM, BIM, PGD, and MIM) with and without the selected mitigation method (defensive distillation). The experimental results demonstrate that both undefended and defended SISO/MIMO models are vulnerable to adversarial attacks, with attack success ratio values significantly increasing at high attack power. In the defended SISO model, the PGD attack has the highest success ratio, followed by BIM and MIM attacks. In the defended MIMO model, the MIM attack has the highest success ratio, followed by BIM and PGD attacks. The FGSM attack had minimal impact on the attack success ratio for both undefended and defended SISO/MIMO models compared to other adversarial attack types due to its simplicity and limitations, i.e., linear approximation, limited perturbation strength, and knowledge of the model. The experimental results also reveal that mitigating methods significantly impact model robustness, reducing the attack success ratio of all attacks. These findings highlight the need to develop more secure and robust AI-based models for next-generation communication technologies to protect against adversarial attacks.

In future work, we will focus on adversarial attack detection in AI-based models in communications, which is the necessary step before attack mitigation. Furthermore, we will attempt to develop better defense mechanisms against adversarial attacks for the AI-based AMR model, improving the security of machine learning systems. While the current study provides valuable insights into the effectiveness of defensive distillation for defending AI-based AMR models against adversarial attacks, it is acknowledged that further comparisons and sensitivity/stability analyses are warranted. These additional analyses, planned as part of future work, will enable a more comprehensive evaluation of the proposed approach, including comparisons with alternative mitigation methods and assessment of the model's sensitivity to different attack scenarios and stability over varying conditions. This will provide a more robust and convincing evaluation of the AMR model defense's proposed defensive distillation methodology.

ACKNOWLEDGMENT

For more information about CCI, visit www.cyberinitiative.org.

REFERENCES

[1] A. K. Nandi and E. E. Azzouz, "Algorithms for automatic modulation recognition of communication signals," *IEEE Trans. Commun.*, vol. 46, no. 4, pp. 431–436, Apr. 1998.

[2] M. Wei, Z. Wei, J. Yang, and L. Sang, "Automatic modulation recognition of digital signal based on auto-encoding network in MIMO system," in *Proc. IEEE 18th Int. Conf. Commun. Technol. (ICCT)*, Oct. 2018, pp. 1017–1021.

[3] K. Hassan, I. Dayoub, W. Hamouda, and M. Berbineau, "Automatic modulation recognition using wavelet transform and neural networks in wireless systems," *EURASIP J. Adv. Signal Process.*, vol. 2010, no. 1, pp. 1–13, Dec. 2010.

[4] H. Zhao, Y. Zhou, B. Sun, L. Tian, and J. Shi, "Cyclic spectrum based intelligent modulation recognition with machine learning," in *Proc. 10th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2018, pp. 1–6.

[5] M. H. Alsharif, A. H. Kelechi, M. A. Albreem, S. A. Chaudhry, M. S. Zia, and S. Kim, "Sixth generation (6G) wireless networks: Vision, research activities, challenges and potential solutions," *Symmetry*, vol. 12, no. 4, p. 676, Apr. 2020.

[6] Y. Wu, H.-N. Dai, H. Wang, Z. Xiong, and S. Guo, "A survey of intelligent network slicing management for industrial IoT: Integrated approaches for smart transportation, smart energy, and smart factory," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 2, pp. 1175–1211, 2nd Quart., 2022.

[7] V. P. Rekkas, S. Sotiroudis, P. Sarigiannidis, S. Wan, G. K. Karagiannidis, and S. K. Goudos, "Machine learning in beyond 5G/6G networks—State-of-the-art and future trends," *Electronics*, vol. 10, no. 22, p. 2786, Nov. 2021.

[8] J. Kaur, M. A. Khan, M. Iftikhar, M. Imran, and Q. E. U. Haq, "Machine learning techniques for 5G and beyond," *IEEE Access*, vol. 9, pp. 23472–23488, 2021.

[9] F. Zhang, C. Luo, J. Xu, Y. Luo, and F.-C. Zheng, "Deep learning based automatic modulation recognition: Models, datasets, and challenges," *Digit. Signal Process.*, vol. 129, Sep. 2022, Art. no. 103650.

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

[11] F. O. Catak, M. Kuzlu, E. Catak, U. Cali, and O. Guler, "Defensive distillation-based adversarial attack mitigation method for channel estimation using deep learning models in next-generation wireless networks," *IEEE Access*, vol. 10, pp. 98191–98203, 2022.

[12] F. O. Catak, M. Kuzlu, H. Tang, E. Catak, and Y. Zhao, "Security hardening of intelligent reflecting surfaces against adversarial machine learning attacks," *IEEE Access*, vol. 10, pp. 100267–100275, 2022.

[13] H. Tang, Y. Zhao, and W. Wang, "Angle of arrival based signal classification in intelligent reflecting surface-aided wireless communications," in *Proc. Int. Conf. Mobile Comput., Appl., Services*. Cham, Switzerland: Springer, Jul. 2023, pp. 57–66.

[14] H. Tang, S. Sarp, Y. Zhao, W. Wang, and C. Xin, "Security and threats of intelligent reflecting surface assisted wireless communications," in *Proc. Int. Conf. Comput. Commun. Netw. (ICCCN)*, Jul. 2022, pp. 1–9.

[15] S. Rajendran, W. Meert, D. Giustiniano, V. Lenders, and S. Pollin, "Deep learning models for wireless signal classification with distributed low-cost spectrum sensors," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 3, pp. 433–445, Sep. 2018.

[16] G. B. Tunze, T. Huynh-The, J.-M. Lee, and D.-S. Kim, "Sparsely connected CNN for efficient automatic modulation recognition," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 15557–15568, Dec. 2020.

[17] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Trans. Cogn. Commun. Netw.*, vol. 3, no. 4, pp. 563–575, Dec. 2017.

[18] C. Yang, Z. He, Y. Peng, Y. Wang, and J. Yang, "Deep learning aided method for automatic modulation recognition," *IEEE Access*, vol. 7, pp. 109063–109068, 2019.

[19] Z. Zhang, H. Luo, C. Wang, C. Gan, and Y. Xiang, "Automatic modulation classification using CNN-LSTM based dual-stream structure," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 13521–13531, Nov. 2020.

[20] 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.

[21] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

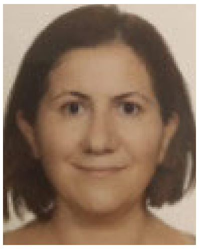
[22] T. J. O'Shea and N. West, "Radio machine learning dataset generation with gnu radio," in *Proc. GNU Radio Conf.*, 2016, vol. 1, no. 1, pp. 1–6.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

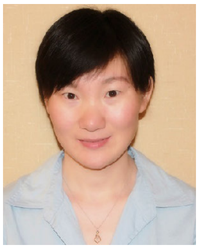
HAOLIN TANG received the B.S. degree in computer science and technology from Yunnan Normal University, China, in 2018. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Virginia Commonwealth University. His research interests include next-generation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

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

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



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



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