Digital Communications and Networks 10 (2024) 756–764

Contents lists available at ScienceDirect

Digital Communications and Networks

journal homepage: www.keaipublishing.com/dcan

Adversarial attacks and defenses for digital communication signals identification

Qiao Tian [a], Sicheng Zhang [b], Shiwen Mao [c], Yun Lin [b,*]

[a] College of Computer Science and Technology, Harbin Engineering University, Harbin, 150001, China
[b] College of Information and Communication Engineering, Harbin Engineering University, Harbin, 150000, China
[c] Department of Electrical and Computer Engineering, Auburn University, Auburn, AL, 36849, USA

ARTICLE INFO

Keywords:
Digital communication signals identification
AI model
Adversarial attacks
Adversarial defenses
Adversarial indicators

ABSTRACT

As modern communication technology advances apace, the digital communication signals identification plays an important role in cognitive radio networks, the communication monitoring and management systems. AI has become a promising solution to this problem due to its powerful modeling capability, which has become a consensus in academia and industry. However, because of the data-dependence and inexplicability of AI models and the openness of electromagnetic space, the physical layer digital communication signals identification model is threatened by adversarial attacks. Adversarial examples pose a common threat to AI models, where well-designed and slight perturbations added to input data can cause wrong results. Therefore, the security of AI models for the digital communication signals identification is the premise of its efficient and credible applications. In this paper, we first launch adversarial attacks on the end-to-end AI model for automatic modulation classification, and then we explain and present three defense mechanisms based on the adversarial principle. Next we present more detailed adversarial indicators to evaluate attack and defense behavior. Finally, a demonstration verification system is developed to show that the adversarial attack is a real threat to the digital communication signals identification model, which should be paid more attention in future research.

1. Introduction

In pursuit of higher-performance mobile communication, the modern communication system has been rolled out rapidly in recent years. Driven by high-tech technologies such as ultra-dense heterogeneous deployment [1], massive Multi-Input-Multi-Output (MIMO) [2], edge computing [3], and Artificial Intelligence (AI) [4], the fifth generation mobile communication system (5G) has extended traditional wireless systems and made great breakthroughs, and its network connection density and the number of users have also increased explosively. The sixth generation mobile communication system (6G) network has entered the research and development stage and is expected to be deployed in the next decade [5]. The digital communication signal recognition takes an important part in modern communication systems, such as Automatic Modulation Classification (AMC) for spectrum awareness, adaptive transmissions in cognitive radio networks, and user authorization authentication in communication monitoring and management systems [6–8].

The ultra-large-scale connections and ultra-high-density deployment of modern communication system makes the electromagnetic space highly dynamic, difficult to model and generate big data. Traditional technical methods have exhibited many obvious limitations in such a setting [9]. In recent years, AI technology represented by Deep Learning (DL) has become an effective solution to the 5G/6G network due to its automatic feature extraction and modeling capabilities [10–12]. With the support of public collected datasets, AI technology has been widely investigated and applied in the field of AMC [13–15]. Effective methods based on the image domain [16,17], time domain [18], frequency domain [19], and multimodality [20] have been proposed. Furthermore, interesting problems such as few-shot [21], zero-shot [22] and even inter-domain transfer learning [23] have been studied. Lightweight techniques for AMC models have also been studied for industrial applications [24,25].

However, AI technology also faces considerable security risks due to its data-dependence and inexplicability. Therefore, systems based on AI technology mainly face threats to data privacy security and model security [26]. The security threats of AI models are divided into training integrity threats caused by backdoor attacks and data poisoning [27] and test integrity threats caused by adversarial attacks. Adversarial attacks

* Corresponding author.
E-mail address: linyun@hrbeu.edu.cn (Y. Lin).

https://doi.org/10.1016/j.dcan.2022.10.010
Received 31 January 2022; Received in revised form 4 October 2022; Accepted 16 October 2022
Available online 20 October 2022

were first discovered by C. Szegedy et al. [28] in the field of Computer Vision (CV). It has been shown that adding carefully crafted, slight, imperceptible perturbations to the input data can cause an AI model to produce erroneous outputs. The tempered examples are called adversarial examples. Adversarial attacks have become an important security threat to various key technologies and applications based on AI in the modern communication system [29,30], as illustrated in Fig. 1.

Adversarial threats have attracted great attention in the problem domain of AMC. The first study was about the scenario that adversarial attacks were carried out at the receiver after an additive white Gaussian noise channel before applying the AMC model [31]. Various adversarial attack methods were subsequently proposed to challenge AMC models. The authors in Ref. [32] improved the gradient-based attack method for CV applications to attack the AMC model, examining the attack effects of white box and black box attacks, single-step, and iterative attacks on the model. In Ref. [33], the authors investigated optimization-based methods to increase the aggression, while the work in Ref. [34] further studied the influence of the channel effect of white Gaussian noise and Rayleigh channels on adversarial attacks and proposed Channel-Aware Adversarial Attack methods. Under the adversarial security threats, on the one hand, advances have been made in the evaluation system for adversarial examples from the perspective of electromagnetic signal characteristics, such as Perturbation-to-Signal Ratio (PSR) [31], Bit Error Rate (BER) [35], and so on [36]. On the other hand, much effort has been devoted to the investigation of adversarial defense methods. In Ref. [34], the robustness of the model of adversarial examples was effectively improved through adversarial training. In Ref. [37], the authors exploited the traditional features, such as the peak-to-average power ratio of modulated signals, as the verification of DL model results to effectively detect adversarial examples.

To enhance the efficient and credible application of AI-based AMC in modern communication systems, we conduct research on adversarial attacks against end-to-end AMC models in this paper. We also explore adversarial defense mechanisms that are explainable. Moreover, we propose more detailed adversarial evaluation indicators to evaluate the performance of adversarial attacks and defense algorithms. Finally, we develop a demonstration verification system of the adversarial game with electromagnetic signals for a realistic assessment of such threats.

In summary, our contributions are given as follows:

- We launch adversarial attacks against the end-to-end AMC model and explain and present three defense mechanisms based on the adversarial principle.
- We present more detailed adversarial evaluation indicators, including attack success rate, defense success rate, attack mistake rate and defense mistake rate to evaluate the performance of adversarial attack and defense algorithms.

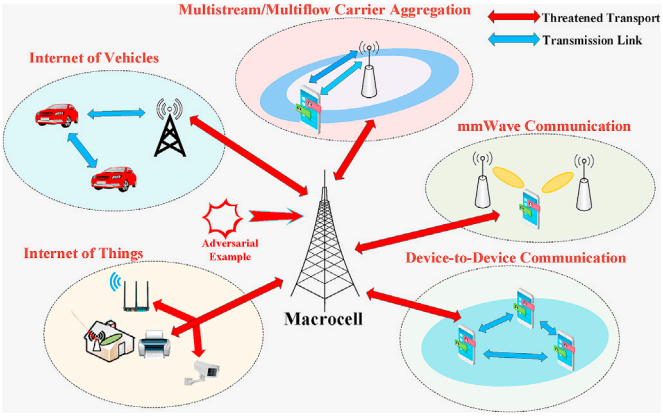


Fig. 1. Scenarios of the threat of adversarial attacks to applications in the AI-based modern communication system.

- We develop a demonstration verification system of the adversarial game of AMC, which demonstrates and evaluates the game process from both the communication domain and the adversarial domain.

The remainder of the paper is organized as follows. The classification of adversarial attacks on electromagnetic signals from different aspects is introduced in Section 2. Adversarial attacks and defense algorithms, as well as evaluation indicator are described in Section 3. The experimental results, discussions, and demonstrations of the verification system are presented in Section 4. A summary of this work and future prospects are given in Section 5.

2. Preliminaries

In the adversarial attacks to AMC models, the adversary introduces carefully designed and slight perturbations to the input data, which can make the tempered data sample easily cross the decision boundary of the AMC model to a wrong decision area without being noticed by the operator or the system. The modified data is called an adversarial example for AMC models, while the algorithms that can detect adversarial examples or make AMC models more robust to such attacks are called adversarial defenses. Compared to traditional interference methods, the adversarial attack is oriented to the AI model instead of the communication system, and the design process is more sophisticated and highly covert with a higher interference efficiency. The following aspects are introduced for adversarial attacks in electromagnetic space.

2.1. Position of adversarial attack

The openness of the electromagnetic space gives wireless devices the right to share the electromagnetic spectrum. However, it also provides more opportunities for adversaries in the electromagnetic space. The adversary can launch adversarial attacks in three locations: the receiving-side, the transmitting-side, and the channel-side, corresponding to the direct, indirect, and superimposing modes, respectively, as illustrated in Fig. 2.

2.1.1. Receiving-side

On the receiving-side, the adversary directly designs and calculates perturbations based on the input data, and superimposes the perturbation to the data to obtain adversarial examples. The original perturbation can be injected into the AMC model without being influenced by the channel effect. Adversarial attacks at this position have higher certainty, but the precondition is very difficult to achieve, which requires the authority to access and modify the input data of the target model.

2.1.2. Transmitting-side

On the transmitting-side, the adversary introduces perturbations into the signal to be transmitted. The perturbation will pass through the channel with the signal and then be injected into the target model. The channel effect between the transmitter and receiver is an important factor to consider in the design of adversarial perturbations.

2.1.3. Channel-side

On the channel-side, the adversary, independent to the transmitter and receiver, radiates adversarial disturbances into the electromagnetic space where they are located. The channel effect in the current electromagnetic space should be comprehensively considered to generate effective adversarial perturbations. Adversarial attacks at this position are more flexible and can perform multiple adversarial tasks, such as one-to-many or many-to-many ones.

2.2. The diversity of adversarial attacks

The principles of attack methods can be divided into two categories: gradient-based and optimization-based methods [38]. On the one hand,

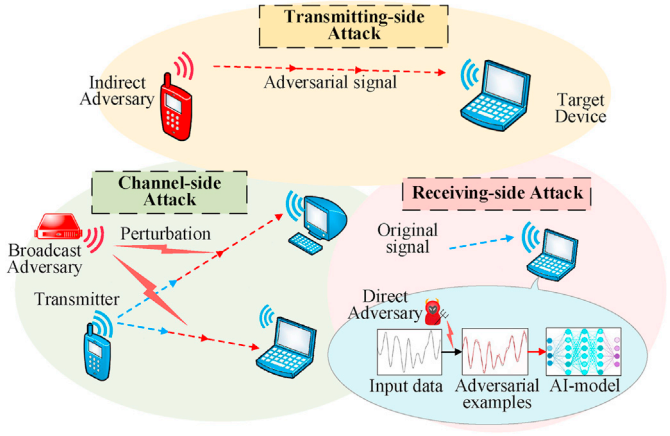


Fig. 2. Schematic diagram of the three attack positions.

there are many optimization algorithms used in the training process of the AI model, but the key is to use the gradient of the loss function to reduce the loss value. Therefore, the opposite process of the training is considered an adversarial attack method, which means, the smallest numerical change can be used in exchange for the largest loss improvement in the gradient ascent direction. On the other hand, adversarial attacks can be viewed as an optimization problem with two optimization objectives. One is that the difference between the adversarial example and the original example should be as small as possible, and the other is that the loss value caused by the adversarial example should be improved as much as possible. Under the same constraints, optimization-based methods can find more satisfactory solutions than gradient-based attack methods.

2.3. Prior information of adversarial attack

Adversarial attacks can be classified into white-box attacks, black-box attacks and grey-box attacks, according to the amount of information the adversary has acquired concerning the target model [39]. The adversarial attack under full information of the target model is called white-box attack, while the attack under no information of the target model is called black-box attack. A black-box attack usually requires the cooperation of surrogate models with similar input-output relations. Attacks between white box and black box attacks are called grey-box attacks.

2.4. Inducibility of adversarial attack

The inducibility of adversarial attacks can be divided into non-targeted and targeted attacks. In non-targeted attacks, the adversary only needs to add perturbations to make the target model produce wrong results, and there is no other requirements. However, in targeted attacks, the adversary not only needs to make the AMC model produce wrong results but also to bias the wrong results toward an intended target. The data from classes with higher classification confidence are generally more difficult to attack [40]. The overall difficulty of targeted attacks is greater than that of non-targeted attacks. In the process of adversarial attacks, the adversary may not launch attacks against all categories. The category that the attacker attacks against is called the Source Category (SC). The expected category that the adversarial attack classifies the model is called the Target Category (TC).

3. Methodology

3.1. Adversarial attack algorithms

The author in Ref. [28] first found that adding a small, carefully constructed perturbation, which cannot be recognized by human eyes, can cause the AI model to output wrong results. Such input data with an added perturbation are called adversarial examples. By adding a perturbation to the signal waveform, a signal adversarial example can be generated. The following methods will be used in this paper to generate adversarial examples of modulated signals.

3.1.1. The Fast Gradient Sign Method

The Fast Gradient Sign Method (FGSM) is an efficiently adversarial attack method, which generates adversarial examples within the l_∞-norm distance of the original sample [41]. The FGSM simultaneously updates all sampling points in a single step along the direction of the gradient sign against the model loss. The update process is given by

$$x' = x + \delta \cdot \mathrm{sign}(\nabla_x L_\theta(x, y)) \tag{1}$$

where x is the original sample of the modulated signal, y is the true label of x, x' is the adversarial example, θ is the parameter of the model, and δ is the perturbation intensity, which limits the l_∞-norm of the perturbation. The gradient of the adversarial loss is calculated by $\nabla_x L_\theta(x, y)$ and sign(·) represents the sign function. It is worth noting that the FGSM is mainly used for the fast generation of adversarial examples, but not for the minimum perturbation. The FGSM will be used to implement an attack and the adversarial robustness of the model will be evaluated in this paper. The process of generating adversarial examples by the FGSM is presented in Algorithm 1.

Algorithm 1 The FGSM

Input: Input example x; true label y; recognition model parameters θ; loss function L of the recognition model; perturbation intensity δ;

Output: Adversarial example x';

1: Get the loss $L(x, y)$ after forward propagation;
2: Get the gradient $\nabla_x L_\theta(x, y)$ of the input example;
3: Take the sign function for the obtained gradient and get sign $(\nabla_x L_\theta(x, y))$;
4: Calculate the adversarial perturbation:

$$\eta = \delta \cdot \mathrm{sign}\,(\nabla_x L_\theta(x, y))\,; \tag{2}$$

5: Add the adversarial perturbation to the input example and obtain the adversarial example as in (1);
6: **return** x';

3.1.2. The Projected Gradient Descent

The Projected Gradient Descent (PGD) attack can be regarded as a BIM but without the $\alpha T = \delta$ constraint, and the PGD can be initialized randomly using any point within the distance of l_∞-norm of the original sample [42]. Unlike the one-step attack FGSM, the PGD launches multiple iterations. Each time a small step is taken, and each iteration will project the perturbation into the specified range as

$$x_{t+1} = \prod_{x+S}(x_t + \alpha \cdot \mathrm{sign}(\nabla_x L(x_t, y, \theta)))\tag{3}$$

where α is the length of each step and $S = r \in R^d$ is the perturbation set. Moreover, perturbation r satisfies $\|r\|_\infty < \delta$, while $\prod_{x+S}$ represents the projection on the ϵ-neighbor range sphere. That is to say, if the perturbation amplitude is too large, the excess part will be pulled back to the

boundary surface. The PGD will be used to generate a modulation adversarial example and the process is given in Algorithm 2.

Algorithm 2 The PGD

Input: Input example x; true label y; recognition model parameters θ; loss function L of the recognition model; perturbation intensity δ; the total number of iterations T;

Output: Adversarial example x';

Get the perturbation level $\alpha = \delta/T$ of each iteration;
Initialize $x'_0 = x$;
for epoch $t = 0$ to T - 1 **do**
  Get the loss $L(x, y)$ after forward propagation;
  Get the gradient $\nabla_x L_\theta(x, y)$ of the input example;
  Take the sign function for the obtained gradient and get sign $(\nabla_x L_\theta(x, y))$;
  Calculate the adversarial perturbation of each iteration and get $\alpha \cdot \text{sign}(\nabla_x L_\theta(x, y))$;
  Use Proj$\{\cdot\}$ to project the adversarial example in the $\alpha - l_\infty$ neighbor of the original example after each iteration as:

$$x'_{t+1} = \text{Proj}_{t,\alpha}\{x'_t + \alpha \cdot \text{sign}(\nabla_x L(x'_t, y))\}; \quad (4)$$

end for
return $x' = x'_{T-1}$;

Compared with the single-step attack algorithm FGSM, the iterative attack algorithm PGD has more flexibility, so it also has a greater adversarial attack effect. However, it is obvious that the time complexity of the PGD algorithm is higher, and it requires several iteration steps while the FGSM only needs one.

3.2. Adversarial defense mechanisms

Next we introduce several typical adversarial defense mechanisms. Identifying the causes of adversarial examples will render the design of defense mechanisms rule-based and rational. There are different opinions on the causes of adversarial examples, which leading to different defense mechanisms.

3.2.1. Adversarial training

The author in Ref. [43] put forward an explanation from the perspective of model generalization through a large number of experimental analyses, indicating that the example complexity of the standard dataset is far from enough to represent the natural data model, and the lack of feature richness of the training set is the reason for the existence of adversarial examples. The transferability of the adversarial example is also in line with such an interpretation to a certain extent [28]. The DL is a fitting problem, and in theory, a sufficiently large network can fit any complex models. Therefore, the most obvious defense mechanism is to train the DL model with adversarial examples to make it robust to adversarial examples.

The general principle of adversarial training can be summarized as a Max-Min problem as

$$\min_\theta \mathbb{E}_{(\mathbf{Z},y)\sim\mathcal{D}}\left[\max_{\|\boldsymbol{\eta}\|\le\delta} L(f_\theta(\mathbf{X}+\boldsymbol{\eta}), y)\right] \quad (5)$$

where $\mathbf{X}$ represents the input example, $f_\theta(\cdot)$ is a model function whose parameter is θ, y is the right label of $\mathbf{X}$, $L(\cdot, \cdot)$ is the loss function, $\mathcal{D}$ represents the distribution of the dataset, and $\mathbf{Z}$ represents the adversarial examples. The Max-Min problem consists of two optimization goals. The first goal is to find the adversarial example that maximizes the loss between the model output and the label within a certain perturbation range. Based on that, the second goal is, to reduce the average loss value between the output result of the effective adversarial example and the label by adjusting the network parameters.

In summary, the adversarial training is a training process that makes the model more robust to effective adversarial attack examples. The flow chart of the adversarial training algorithm is shown in Fig. 3. In order to ensure that the adversarial training model maintains the recognition performance of the original data, 60% of each batch of data is generated as an adversarial example during training.

3.2.2. Noise training

Obviously, the adversarial training requires a pre-trained model. This model can be the model to be optimized or an alternative model trained on the same dataset as the target model. Unlike the adversarial training, the idea of the noisy training is to improve the completeness of the dataset features without using a pre-trained model and to obtain a more robust model. The implementation method is to add slight random noise with a certain probability distribution to the original dataset. The data with random noise is still in the neighborhood of the original data. The expanded dataset possesses more abundant features than the original data set, and the model obtained this way will also be more robust. However, this kind of unguided expanded dataset will inevitably add a large number of features that are not beneficial for improving the robustness of the model, and these data examples will increase the training burden.

3.2.3. Network binarization

The authors in Ref. [28] provided an explanation with experiments that the high degree of nonlinearity of the DL model leads to a large number of local over-fitting phenomena, such that the decision boundary of the model produces a multitude of low-probability pockets in the manifold space. The data in the pocket obviously belongs to the type, but it is decided by the DL model to be another type. The authors in Ref. [41] argue against this view and explain experimentally that the cause of adversarial examples is the network is extremely linear. The slight perturbation applied to the input data is amplified by the high-dimensional linear network layer by layer, and the level of this perturbation will not change the original attributes of the data, and produce a large deviation in the output result. Analyzing the above two points of view, we can see that the over-fitting problem and extreme linearization of deep learning networks are the biggest possibilities against examples. The binary network can avoid these two problems. In the deterministic binary network in the naive binary network category, the weight and input of the binary layer network are binarized as [44].

$$w^b = \begin{cases} +1 & w \ge 0 \\ -1 & \text{otherwise} \end{cases} \quad (6)$$

where w represents the original parameter and w^b represents the binarized parameter. The sign$(\cdot)$ function can be used for the quantization operation.

This binarization rule results in the obvious compression of the degree of freedom space of the network parameters, which effectively reduces the fitting ability of the network and avoids the problem of over-fitting. In addition, the binarization rule is a severe non-linear function, which prevents small perturbations at the input from being linearly amplified layer by layer.

The difference between adversarial training and noisy training

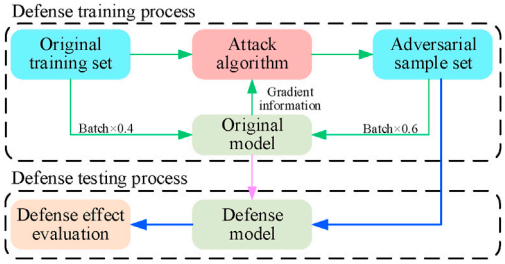


Fig. 3. Algorithm flow chart of the adversarial training method.

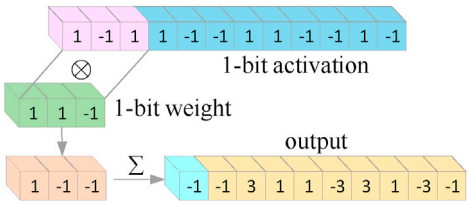


Fig. 4. Schematic diagram of the convolutional layer operation rules in the binary network.

algorithms is only the training data used. Compared to the model training, the difference in time between the two algorithms for generating training data is negligible. Therefore, both algorithms have the same time and space complexity. For the binarized network, the 32-bit floating point numbers of the model parameters are replaced by 1-bit binary numbers, so the network achieves a 32-fold reduction in space. In addition, another advantage of the binary network is the fast computation. In the binary convolutional layer, the input and weight can only have two values: $+1$ and -1. Thus the multiplication in the convolutional layer can be replaced by the *XNOR* operation, and the accumulation can be replaced by the *Bitcount* operation. The forward calculation rules are illustrated in Fig. 4.

3.3. Evaluation indicators

3.3.1. Attack success rate

The adversary uses the adversarial attack algorithm to add perturbations to the original dataset to obtain the adversarial attack dataset. In a non-targeted attack, the TC corresponding to each SC is all other categories except itself. In the process of adversarial attacks, the ratio of the number of adversarial examples that are originally correctly classified into the SC, but are classified into the TC after adding the adversarial perturbation to the total number of SC is called the Attack Success Rate (ASR). The ASR of a certain category is $ASR_{o-\mathbf{T}}$ and the average ASR is *ASR*, given by

$$ASR_{o-\mathbf{T}} = \frac{1}{N_o}\sum_{x\in o}\text{bool}(f(x^{'}) = t, t \in \mathbf{T} | f(x) = o) \tag{7}$$

$$ASR = \frac{\sum_o^{\mathbf{O}} ASR_{o-\mathbf{T}} \cdot N_o}{\sum_o^{\mathbf{O}} N_o} \tag{8}$$

where bool$(\cdot)$ is bool function, $f(\cdot)$ is the DL model, x is the original example, $o \in \mathbf{O}$ represents the SC, and $t \in \mathbf{T}$ represents the TC corresponding to category o, and N_o is the number of category o in the dataset.

3.3.2. Defense success rate

The purpose of the defense is to correct the incorrectly classified examples caused by the attack algorithm on the correct label. The Defense Success Rate (DSR) of the defense mechanism for a certain category DSR_o is the ratio of the number of examples, which are adversarial examples previously misclassified by the original network and then correctly classified by the defense mechanism to the total number of examples in that category. The average DSR is denoted by *DSR*. DSR_o and *DSR* are defined as

$$DSR_o = \frac{1}{N_o}\sum_{x\in o}\text{bool}(f^{'}(x^{'}) = o | f(x^{'}) \neq o) \tag{9}$$

$$DSR = \frac{\sum_o^{\mathbf{O}} DSR_o \cdot N_o}{\sum_o^{\mathbf{O}} N_o} \tag{10}$$

where $f^{'}(\cdot)$ represents the defense model.

3.3.3. Attack mistake rate

During the attack process, there may be some examples that were originally misclassified, but then are correctly classified after the perturbation is introduced. This phenomenon is a mistake caused by the adversarial attack algorithm, termed Attack Mistake Rate (AMR). The attack mistake rate for a certain category is $AMR_{o-\mathrm{T}}$ and the average attack mistake rate is *AMR*, which is given by

$$AMR_{o-\mathrm{T}} = \frac{1}{N_o}\sum_{x\in o}\text{bool}(f(x^{'}) = o | f(x) = t, t \in \mathbf{T}) \tag{11}$$

$$AMR = \frac{\sum_o^{\mathbf{O}} AMR_{o-\mathbf{T}} \cdot N_o}{\sum_o^{\mathbf{O}} N_o} \tag{12}$$

3.3.4. Defense mistake rate

The defense process could also make a mistake. That is, the attack examples that are originally correctly classified may be misclassified by the defense mechanism. Such mistakes need not be mistaken for the adversary's TC. The proportion of such examples is called the Defense Mistake Rate (DMR). The defense mistake rate for a certain category is DMR_o and the average defensive mistake rate is *DMR*, which is given by

$$DMR_o = \frac{1}{N_o}\sum_{x\in o}\text{bool}(f^{'}(x^{'}) \neq o | f(x^{'}) = o) \tag{13}$$

$$DMR = \frac{\sum_o^{\mathbf{O}} DMR_o \cdot N_o}{\sum_o^{\mathbf{O}} N_o} \tag{14}$$

4. Experimental study and discussions

In order to investigate the security issues of AMC models, this paper conducts extensive adversarial attack and defense experiments with the end-to-end AMC model. Although channel-side and transmitter-side attacks are more in line with actual attack methods in the adversarial attack and defense process, the research on receiver-side attacks attracts more researchers to analyze the adversarial safety issues of electromagnetic signal recognition models without other interference factors. Such a study will provide the useful knowledge for further studies on channel-side and transmitter-side attacks. Therefore, this paper studies, evaluates and analyzes the above adversarial attack methods and defense mechanisms on the receiver side.

4.1. Experiment methodology

The data used in the experiment is the public dataset RML2016.10A [13]. The dataset contains 11 modulation types, including 8 digital modulation methods and 3 analog modulations. Channel effects such as center frequency offset, phase difference, and sampling rate offset are considered. The Signal-to-Noise Ratio (SNR) ranges from -20dB to 18 dB, with an interval of 2 dB. The total number of samples is 220,000, and the number of samples for each type of modulation and each SNR level is the same. Each data sample contains In-phase and Quadrature (I/Q) data in the form of 2×128. The dataset is randomly divided into the training set, the validation set, and the test set with the ratio of 8:1:1.

This experiment uses the Pytorch DL framework to build and train AI models. And the test library for the adversarial example attack and defense selected in the experiment is Cleverhans, which supports Tensorflow and Pytorch. Three models are created. The first is the baseline model, also known as the original model, which is used as a modulation signal recognition model. The second is a surrogate model with a different structure from the baseline network, which is used to carry out black-box attacks. The last one is a binary model with the same structure as the baseline model, which is used to analyze the defense effect compared to the original model, instead of itself [45,46]. The network structure refers to the basic Convolutional Neural Network (CNN), which

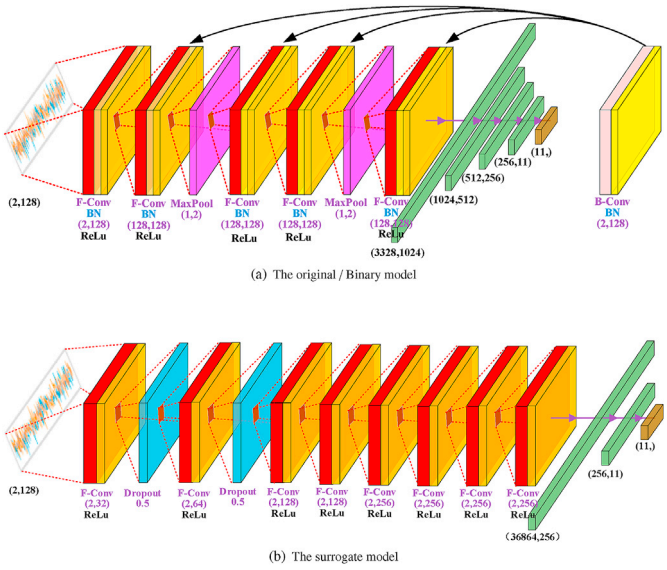


Fig. 5. The network structure of the original model, the surrogate model, and the binary model.

means, the convolutional layers are in the front and the fully connected layers are in the back. The network structure of the three models are depicted in Fig. 5. Fig. 5(a) is for the original model, including five convolutional layers and four fully connected layers. The structure of the binary model is the same as the original model, but binary blocks replace all convolutional layers except the first layer. Fig. 5(b) is for the surrogate model, where eight convolutional layers are used and two dropout layers are passed in between. The number of fully connected layers is reduced to two.

4.2. Evaluation of network performance

Before launching the evaluation of adversarial attacks and defenses, we first evaluate the recognition performance of the three models on the original signal dataset. The loss function used in this experiment is the cross-entropy loss function, and the training optimizer is Root Mean Square prop (RMSprop). The batch size is 256.

The test accuracy of the three models under different SNR data is presented in Fig. 6. It can be seen that the recognition accuracy of the three models all increase as the SNR increases. The recognition performance of the original model and the surrogate model is almost consistent. Therefore, the surrogate model is qualified for black-box attacks. Since the degree of freedom of the parameters of the binarized model is

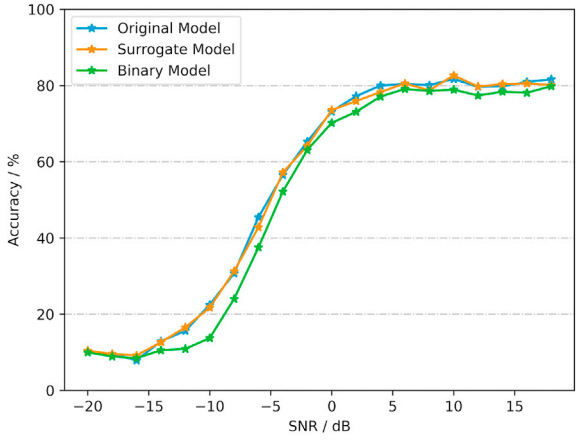


Fig. 6. Recognition performance of the three network models.

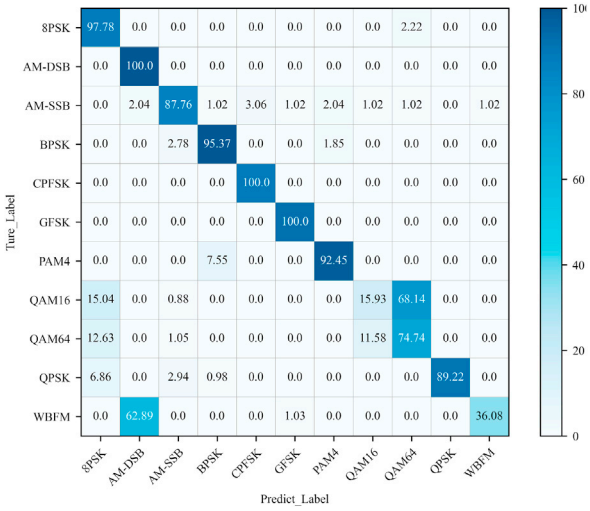


Fig. 7. Confusion matrix of the original model under 12 dB SNR.

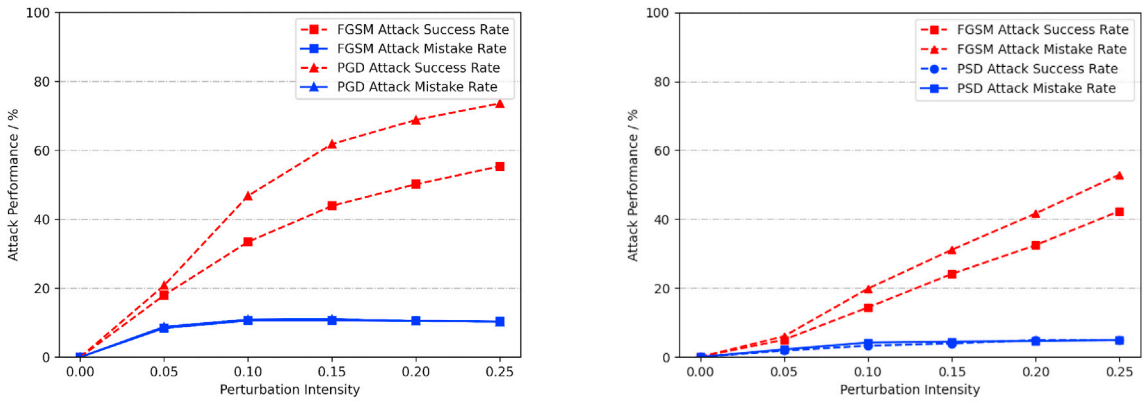
severely reduced, it is expected that its recognition performance will be poorer than the original model. Under an extremely low SNR, all three models do not work well due to the severe damage of the separability of data. Under a higher SNR, the separability of the data becomes high, resulting in little difference between the performance of the binary model and the original model. Under intermediate SNR levels, the data is more difficult to classify and the fitting ability of the binary model becomes weaker, so the accuracy is also lower.

Therefore, the dataset at a higher SNR is the best choice for further analysis of adversarial attacks and defense because the three networks are on the same starting line. The confusion matrix of the original model on the 12 dB test dataset is provided in Fig. 7 with a lot of information. First, the data between major categories is not easy to be confusing. Second, within the same major category, the data between low order types is often not easy to be confusing. Finally, within the same major category, the data between high order types is easy to be confusing.

4.3. Evaluation of attack algorithms

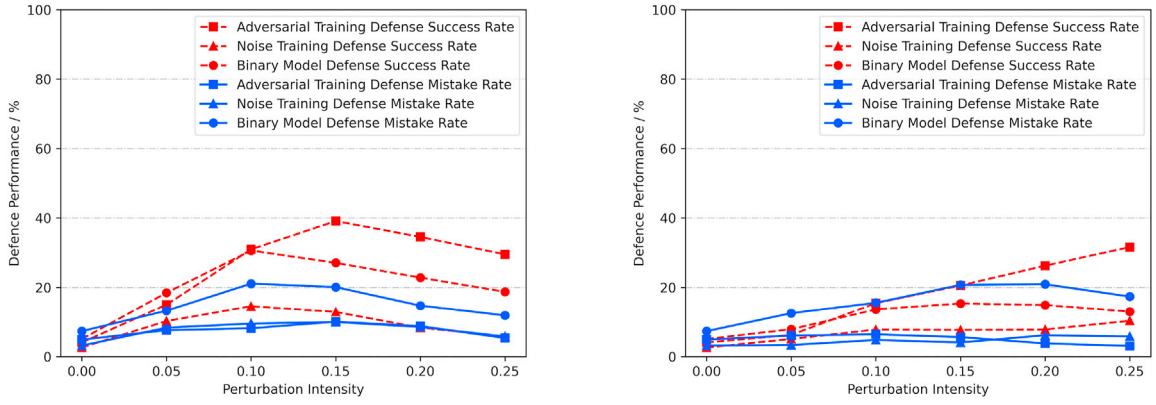
For the evaluation of attack algorithms, the FGSM and the PGD attack algorithms are used as the representatives of single-step attacks and iterative attacks, respectively. Therefore, the two attack algorithms and their corresponding white-box and black-box attacks will be used to attack the original model. The perturbation intensity of the attacking algorithm will be set from 0 to 0.25 at the step size of 0.05. In order to enable the PGD algorithm to fully search for adversarial perturbations within a given perturbation intensity, we set the number of iteration steps to 15. The dataset under 12 dB SNR is still selected as the target data.

The attack performance of the FGSM and the PGD for white-box and black-box attacks under 12 dB SNR is presented in Fig. 8. Some interesting phenomena can be observed. For white-box and black-box attacks, the ASRs of the two attacking algorithms both increase as the perturbation intensity increases. The AMR shows the same trend, eventually stabilizing within a small range. In white-box or black-box scenarios, the AMR caused by the two algorithms is almost the same. However, the ASR of the PGD is significantly higher than that of the FGSM. The ASR of a single-step attack is significantly weaker than that of an iterative attack with the same perturbation intensity. This is because iterative attacks can find more refined and efficient perturbations through multiple iterations and obtain a higher attack performance. Moreover, the performance of black-box attacks is obviously weaker than that of white-box attacks. Although the white-box attack has a relatively higher attack mistake rate, the overall attack effect of the white-box attack is significantly stronger



(a) White-box attack (b) Black-box attack

Fig. 8. Attack performance of FGSM and PGD for white-box and black-box attacks under 12 dB SNR.



(a) White-box attack (b) Black-box attack

Fig. 9. Defense performance of the three methods against white-box and black-box attacks using PGD under 12 dB SNR.

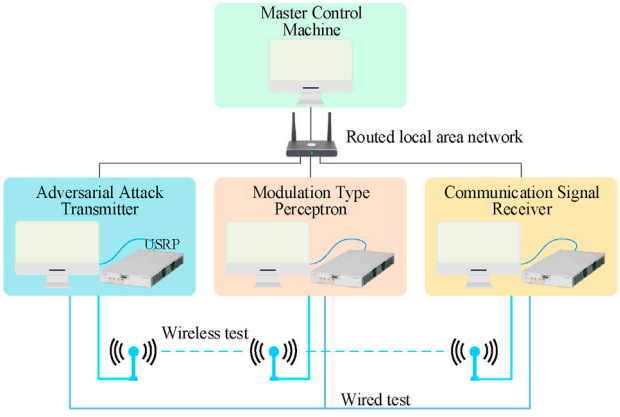


Fig. 10. Structure diagram of the demonstration verification system.

than that of the black-box attack. It can be seen that the transferability of adversarial examples between models cannot completely maintain their robustness. Therefore, black-box attacks can only test the general trend of the robustness of the model.

4.4. Evaluation of defense mechanisms

Attack algorithms with stronger attack performances are more likely to reveal more problems of the defense mechanism. Therefore, we choose the PGD attack algorithm in this experiment to evaluate the defensive performance. In adversarial training defense mechanisms, the adversarial example set of the PGD attack with a disturbance intensity of 0.2, an iteration step size of 0.05 and 15 iteration steps is used to train the original model to obtain a defense model, which is robust to adversarial attacks. In the noise training defense mechanism, the Gaussian white noise is selected; the mean value is set to 0 and the variance is set to 0.1. These will ensure that the generated example is within the neighborhood of 0.2 with a probability of 95.45%, and the probability of occurrence of perturbations of different sizes is close. The binary model will be directly used as a defense model without adversarial training and noise training.

The defense performance of the three defense mechanisms against the white-box and black-box attacks using the PGD under 12 dB SNR is shown in Fig. 9. As can be seen from Fig. 9(a), in terms of the DSR, the trends of the three defense model are almost identical. The DSRs of the three methods all reach the highest when the perturbation intensity is around 0.2. For adversary training and noise training, this is because the selected training perturbation intensity is 0.2, and the model is more robust to adversarial examples with a perturbation intensity of 0.2. While for the binary model, when the attack intensity exceeds a certain level, the defense performance will decline rapidly. In terms of the DMR, the binary network has not been specially trained, and thus makes more visible mistakes. Overall, the defense performance of adversarial training is the best, followed by the binary model and noise training. It can be concluded that targeted defense mechanisms are more effective.

It can be seen from Fig. 9(b) that black-box attacks can only test the general trend of the robustness of the model. In terms of the DSR, there is a similar trend for all the three models. The defense performance of the

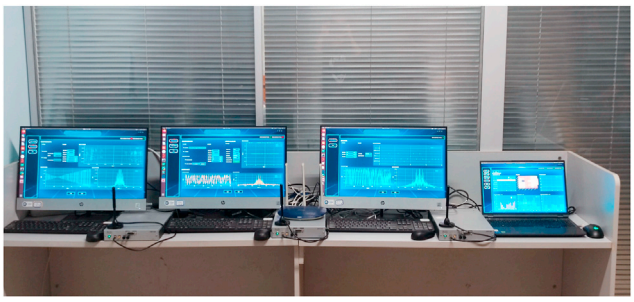


Fig. 11. A picture of demonstration verification system in operation.

Fig. 12. Display of real-time results of the demonstration verification system.

three models achieves a higher DSR under a higher disturbance intensity. While in terms of the DMR, the binary model will make more mistakes. Therefore, the binary model, as a kind of active defense, is less effective than the passive defense with targeted special training. The resistance of the binary model to white-box attacks is better than that to black-box attacks.

4.5. Demonstration verification system

In order to demonstrates and evaluate adversarial attacks and the defense game process, we develop a demonstration verification system of the adversarial game for the AMC. As shown in Fig. 10, the system includes an Adversarial Attack Transmitter (AAT), a Modulation Type Perceptron (MTP), a Communication Signal Receiver (CSR), and a Master Control Machine (MCM). The AAT is responsible for transmitting communication signals with adversarial perturbations. The MTP is used as the target of the AAT and is responsible for recognizing the type of communication signals the AAT transmits. It is undesirable that the superimposed perturbation affects the communication performance of the original signal. Therefore, a CSR is leveraged to evaluate whether the characteristics of the adversarial signal are damaged. Each of the three sub-devices is composed of a computer and a Universal Software Radio Peripheral (USRP). The computer executes the algorithm and baseband signal processing, and the USRP implements the conversion between the baseband signal and the radio frequency signal. The system supports both offline and online operating modes. In the offline mode, the three sub-devices are equipped with a human-computer interaction interface for display and control. In the online mode, the three sub-devices exchange data and control messages with the MCM through a local area network, while the MCM is responsible for the overall control and status display of the system.

The prototype system implements the above two attack methods, including FGSM and PGD, their corresponding targeted attacks and black-box and white-box attacks. It supports the evaluation of adversarial example generation time, real-time batch accuracy, difference in waveform fit of adversarial examples, perturbation intensity, model defense performance, and communication error rate. The communication signal types include 2DPSK, 4DPSK, and 8DPSK, and the coding types include Reed Solomon coding and convolutional coding. Communication parameters such as signal bandwidth, carrier frequency, and transmission gain are adjustable. It can display time domain waveforms, bandwidth, constellation diagrams, and so on. In summary, the system can demonstrate the attack and defense game process in both wireless and wired environments in the lab.

Fig. 11 shows a picture of the demonstration verification system in the operation, while Fig. 12 shows the displayed results of the demonstration verification system. The upper left corner of the figure shows the real-time recognition results of the MTP. For the convenience of comparison, the modulation types of AAT are also plotted in the figure. The depth of the display window is 20, that is, the latest 19 historical data are displayed in the window at the same time. The confusion matrix is shown in the upper middle part of the figure, which includes all the data after this startup. The upper right corner of the figure shows the difference in fit for single data, while the lower middle part of the figure shows the batch perturbation intensity. The lower left corner of the figure shows the real-time batch accuracy comparison of the original model and the defense model. The lower right corner of the figure shows the improved recognition accuracy brought by the defense model under the adversarial attack signal, compared to the original model. Each interval batch is a variety of signals with and without adversarial attacks. When there is no adversarial attack, the recognition accuracy of the original model and the defense model are almost the same. However, when there is an adversarial attack, the recognition accuracy of the original model is greatly reduced, while that of the defense model is significantly higher. These results prove that adversarial attacks are a real threat to AMC models.

5. Conclusions and future work

The main goal of this work is to analyze the adversarial threats to AMC models to ensure the efficient and credible application in modern communication systems. Various adversarial attacks to the end-to-end AMC model were investigated. Furthermore, based on the explanation principle of adversarial attacks, adversarial training, noise training, and binary model defense mechanisms were designed, and their defense effects on typical attack algorithms were evaluated in this paper, which is very helpful to study active defense mechanisms. The results of this investigation showed that the adversarial training achieves the highest defensive performance improvement. The noise training introduces a lot of training burden, and its effectiveness in improving the robustness of the network is extremely poor. The binary model as a kind of active defense is less effective than the passive defense with targeted special training. The resistance of the binary model to white-box attacks is better than that of black-box attacks. This study has provided a deeper insight into the adversarial threat of AMC models. In addition, this work contributed to the existing knowledge of adversarial evaluation by defining useful performance metrics. The demonstration and verification system proposed in this paper proved that the adversarial attack is a real threat to AMC models and helped to focus more on the integration of the adversarial threat and electromagnetic field characteristics.

However, there is still much room for improvement in this paper. For example, the effects of a black-box attack should be extended to multiple models rather than just one model. The effects of more defensive algorithms should be displayed for comparison. Furthermore, based on this work, the following research directions can be tackled in future work: (i) Channel effect: in the actual adversarial attack and defense process of communications field, the channel effect is an important factor that cannot be ignored. Although the channel effect has attracted great attention, more theoretical analyses, experiments, and prototype systems are needed. (ii) Specific attacks: investigations on issues such as attention mechanisms have shown that AI models are more sensitive to locally important features in the data. Therefore, adding adversarial perturbations to local features in the signal rather than the full signal segment is a

direction for future research, which will improve the stealth and energy efficiency of adversarial attacks. (iii) Active defense: most of the current defense methods are heuristic approaches. The defense models tailored for specific attack methods should be more effective. Designing active and provable defense methods and establishing a unified defense mechanism should be investigated in future work.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (61771154) and the Fundamental Research Funds for the Central Universities (3072022CF0601). This work is also supported by Key Laboratory of Advanced Marine Communication and Information Technology, Ministry of Industry and Information Technology, Harbin Engineering University, Harbin, China.

References

[1] A. Zhu, M. Ma, S. Guo, S. Yu, L. Yi, Adaptive multi-access algorithm for multi-service edge users in 5g ultra-dense heterogeneous networks, IEEE Trans. Veh. Technol. 70 (3) (2021) 2807–2821.
[2] Y. I. A. Al-Yasir, A. M. Abdulkhaleq, N. O. Parchin, I. T. Elfergani, J. Rodriguez, J. M. Noras, R. A. Abd-Alhameed, A. Rayit, R. Qahwaji, GReen and highly efficient MIMO transceiver system for 5g heterogenous networks, IEEE Trans. Green Commun. Netw. To appear. 6(1)(2022)500-511.
[3] Y. Siriwardhana, P. Porambage, M. Liyanage, M. Ylianttila, A survey on mobile augmented reality with 5G mobile edge computing: architectures, applications, and technical aspects, IEEE Commun. Surv. Tutor. 23 (2) (2021) 1160–1192.
[4] C. Zhang, Y.-L. Ueng, C. Studer, A. Burg, Artificial intelligence for 5g and beyond 5G: implementations, algorithms, and optimizations, IEEE J. Emerg. Select. Top. Circ. Syst. 10 (2) (2020) 149–163.
[5] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, M. Zorzi, Toward 6g networks: use cases and technologies, IEEE Commun. Mag. 58 (3) (2020) 55–61.
[6] N.A. Mousa, S.B. Sadkhan, Identif. digit. modul. signal cogn. radio netw. a surv. (2021) 311–314, https://doi.org/10.1109/BICITS51482.2021.9509927.
[7] S. Peng, S. Sun, Y.-D. Yao, A survey of modulation classification using deep learning: signal representation and data preprocessing, IEEE Transact. Neural Networks Learn. Syst. 33 (12) (2022) 7020–7038.
[8] Z. Lv, A.K. Singh, J. Li, Deep learning for security problems in 5G heterogeneous networks, IEEE Netw. 35 (2) (2021) 67–73.
[9] V. Monga, Y. Li, Y.C. Eldar, Algorithm unrolling: interpretable, efficient deep learning for signal and image processing, IEEE Signal Process. Mag. 38 (2) (2021) 18–44.
[10] C. Luo, J. Ji, Q. Wang, X. Chen, P. Li, Channel state information prediction for 5G wireless communications: a deep learning approach, IEEE Trans. Netw. Sci. Eng. 7 (1) (2020) 227–236.
[11] M.-Y. Chen, M.-H. Fan, L.-X. Huang, Ai-based vehicular network toward 6g and iot: deep learning approaches, ACM Trans. Manage. Inf. Syst. 13 (1)(2022)1-12.
[12] M.E. Morocho Cayamcela, W. Lim, Artificial intelligence in 5G technology: a survey, in: Proc. 2018 International Conference on Information and Communication Technology Convergence, Jeju Island, Korea, 2018, pp. 860–865.
[13] T.J. O'Shea, J. Corgan, T.C. Clancy, Convolutional radio modulation recognition networks, in: C. Jayne, L. Iliadis (Eds.), Engineering Applications of Neural Networks, Cham, 2016, pp. 213–226.
[14] T.J. O'Shea, T. Roy, T.C. Clancy, Over-the-air deep learning based radio signal classification, IEEE Journal of Selected Topics in Signal Processing 12 (1) (2018) 168–179.
[15] Y. Tu, Y. Lin, H. Zha, J. Zhang, Y. Wang, G. Gui, S. Mao, Large-scale real-world radio signal recognition with deep learning, Chin. J. Aeronaut. doi:https://doi.org/10.1016/j.cja.2021.08.016. URL https://www.sciencedirect.com/science/article/pii/S1000936121002934.
[16] S. Peng, H. Jiang, H. Wang, H. Alwageed, Y. Zhou, M.M. Sebdani, Y.-D. Yao, Modulation classification based on signal constellation diagrams and deep learning, IEEE Transact. Neural Networks Learn. Syst. 30 (3) (2019) 718–727.
[17] Y. Lin, Y. Tu, Z. Dou, L. Chen, S. Mao, Contour stella image and deep learning for signal recognition in the physical layer, IEEE Trans. Cogn. Commun. Netw. 7 (1) (2021) 34–46.
[18] Z. Ke, H. Vikalo, Real-time radio technology and modulation classification via an LSTM auto-encoder, IEEE Trans. Wireless Commun. 21 (1) (2022) 370–382.
[19] M.M.T. Abdelreheem, M.O. Helmi, Digitalmmodulation classification through time and frequency domain features using neural networks, in: Proc. 2012 IX International Symposium on Telecommunications, No. Oct., Bosnia and Herzegovina, Sarajevo, 2012, pp. 1–5.
[20] P. Qi, X. Zhou, S. Zheng, Z. Li, Automatic modulation classification based on deep residual networks with multimodal information, IEEE Trans. Cogn. Commun. Netw. 7 (1) (2021) 21–33.
[21] Q. Zhou, R. Zhang, J. Mu, H. Zhang, F. Zhang, X. Jing, AMCRN: few-shot learning for automatic modulation classification, IEEE Commun. Lett. 26 (3) (2022) 542–546.
[22] Y. Dong, X. Jiang, H. Zhou, Y. Lin, Q. Shi, SR2CNN: zero-shot learning for signal recognition, IEEE Trans. Signal Process. 69 (2021) 2316–2329.
[23] M. Wang, Y. Lin, Q. Tian, G. Si, Transfer learning promotes 6G wireless communications: recent advances and future challenges, IEEE Trans. Reliab. 70 (2) (2021) 790–807.
[24] Y. Lin, Y. Tu, Z. Dou, An improved neural network pruning technology for automatic modulation classification in edge devices, IEEE Trans. Veh. Technol. 69 (5) (2020) 5703–5706.
[25] S. Zhang, Y. Lin, Y. Tu, S. Mao, Electromagnetic signal modulation recognition technology based on lightweight deep neural network, J. Commun. 41 (11) (2020) 12.
[26] Y. Hu, W. Kuang, Z. Qin, K. Li, J. Zhang, Y. Gao, W. Li, K. Li, ARtificial intelligence security: threats and countermeasures, ACM Comput. Surv. 55 (1).
[27] Y. Yao, H. Li, H. Zheng, B.Y. Zhao, Latent backdoor attacks on deep neural networks, in: Proc. 2019 ACM SIGSAC Conference on Computer and Communications Security, London, UK, 2019, pp. 2041–2055.
[28] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing Properties of Neural Networks, arXiv E-Prints, 2013 arXiv:1312.6199arXiv:1312.6199.
[29] Y.E. Sagduyu, T. Erpek, Y. Shi, Adversarial Machine Learning for 5G Communications Security, arXiv E-Prints, 2021 arXiv:2101.02656arXiv:2101.02656.
[30] Z. Bao, Y. Lin, S. Zhang, Z. Li, S. Mao, Threat of adversarial attacks on dl-based iot device identification, IEEE Internet Things J. 9 (11) (2022) 9012–9024.
[31] M. Sadeghi, E.G. Larsson, Adversarial attacks on deep-learning based radio signal classification, IEEE Wireless Commun. Lett. 8 (1) (2019) 213–216.
[32] Y. Lin, H. Zhao, X. Ma, Y. Tu, M. Wang, Adversarial attacks in modulation recognition with convolutional neural networks, IEEE Trans. Reliab. 70 (1) (2021) 389–401.
[33] M. Usama, M. Asim, J. Qadir, A. Al-Fuqaha, M.A. Imran, Adversarial machine learning attack on modulation classification, in: Proc. 2019 UK, China Emerging Technologies, Glasgow, UK, 2019, pp. 1–4.
[34] B. Kim, Y.E. Sagduyu, K. Davaslioglu, T. Erpek, S. Ulukus, Channel-aware adversarial attacks against deep learning-based wireless signal classifiers, IEEE Trans. Wireless Commun. 21 (6) (2022) 3868–3880.
[35] B. Kim, Y.E. Sagduyu, K. Davaslioglu, T. Erpek, S. Ulukus, How to make 5G communications "invisible": adversarial machine learning for wireless privacy, in: Proc. 2020 54th Asilomar Conference on Signals, Systems, and Computers, Virtual Conference, 2020, pp. 763–767.
[36] D. Adesina, C.-C. Hsieh, Y.E. Sagduyu, L. Qian, Adversarial Machine Learning in Wireless Communications Using RF Data: A Review, arXiv E-Prints, 2020 arXiv:2012.14392.
[37] S. Kokalj-Filipovic, R. Miller, G. Vanhoy, Adversarial examples in rf deep learning: detection and physical robustness, in: PRoc. IEEE GlobalSIP'19, Canada, Ottawa, 2019, pp. 1–5.
[38] Z. Yuan, J. Zhang, Y. Jia, C. Tan, T. Xue, S. Shan, Meta Gradient Adversarial Attack, 2021, pp. 7748–7757.
[39] R. Sahay, D.J. Love, C.G. Brinton, Robust Automatic Modulation Classification in the Presence of Adversarial Attacks, 2021, pp. 22955–22967.
[40] Y. Lin, H. Zhao, Y. Tu, S. Mao, Z. Dou, Threats of adversarial attacks in DNN-based modulation recognition, in: Proc. IEEE INFOCOM, vol. 2020, Virtual Conference, 2020, pp. 2469–2478.
[41] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and Harnessing Adversarial Examples, arXiv E-Prints, 2014 arXiv:1412.6572.
[42] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards Deep Learning Models Resistant to Adversarial Attacks, arXiv E-Prints, 2017 arXiv:1706.06083.
[43] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, A. Madry, Adversarially Robust Generalization Requires More Data, arXiv E-Prints, 2018 arXiv:1804.11285.
[44] H. Qin, R. Gong, X. Liu, X. Bai, J. Song, N. Sebe, Binary neural networks: a survey, Pattern Recogn. 105 (2020) 107281.
[45] S. Zhang, Y. Lin, Z. Bao, J. Fu, A Lightweight Modulation Classification Network Resisting White Box Gradient Attacks, Hindawi Security and Communication Networks, 2021.
[46] A. Galloway, G.W. Taylor, M. Moussa, Attacking Binarized Neural Networks, arXiv E-Prints, 2017 arXiv:1711.00449.