electronics

MDPI

Article

Adaptive Whitening and Feature Gradient Smoothing-Based Anti-Sample Attack Method for Modulated Signals in Frequency-Hopping Communication

Yanhan Zhu [1,2,*], Yong Li [2] and Zhu Duan [1]

[1] School of Electronics and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China; duanz@nuist.edu.cn

[2] Sixty-Third Research Institute, National University of Defense Technology, Nanjing 210007, China; liyong17@nudt.edu.cn

* Correspondence: 202212490441@nuist.edu.cn; Tel.: +86-15189561788

Abstract: In modern warfare, frequency-hopping communication serves as the primary method for battlefield information transmission, with its significance continuously growing. Fighting for the control of electromagnetic power on the battlefield has become an important factor affecting the outcome of war. As communication electronic warfare evolves, jammers employing deep neural networks (DNNs) to decode frequency-hopping communication parameters for smart jamming pose a significant threat to communicators. This paper proposes a method to generate adversarial samples of frequency-hopping communication signals using adaptive whitening and feature gradient smoothing. This method targets the DNN cognitive link of the jammer, aiming to reduce modulation recognition accuracy and counteract smart interference. First, the frequency-hopping signal is adaptively whitened. Subsequently, rich spatiotemporal features are extracted from the hidden layer after inputting the signal into the deep neural network model for gradient calculation. The signal's average feature gradient replaces the single-point gradient for iteration, enhancing anti-disturbance capabilities. Simulation results show that, compared with the existing gradient symbol attack algorithm, the attack success rate and migration rate of the adversarial samples generated by this method are greatly improved in both white box and black box scenarios.

Keywords: frequency-hopping communication; modulation recognition; deep neural network; adaptive whitening; feature gradient; adversarial example

check for updates

Citation: Zhu, Y.; Li, Y.; Duan, Z. Adaptive Whitening and Feature Gradient Smoothing-Based Anti-Sample Attack Method for Modulated Signals in Frequency-Hopping Communication. *Electronics* **2024**, *13*, 1784. https://doi.org/10.3390/electronics13091784

Academic Editors: Teodor B Iliev, Lorant Andras Szolga and Gani Balbayev

Received: 2 April 2024
Revised: 29 April 2024
Accepted: 2 May 2024
Published: 5 May 2024

1. Introduction

In recent years, with the rapid development of electronic countermeasure technology, jamming means have become complex and diverse, which puts forward higher requirements for the reliability of communication. Owing to its excellent performance, frequency-hopping communication has become widely utilized and is regarded as a secure method in military applications for hostile environments [1]. However, the emergence of targeted interference has highlighted the limitations of traditional frequency-hopping techniques. To enhance the anti-interference ability of wireless communication systems [2], this paper studies the anti-interference strategy based on Game Theory in frequency-hopping communication to deal with the interference attack in frequency-hopping communication and puts forward new ideas to solve the interference countermeasure problem, which is of great significance to improve the anti-interference ability of frequency-hopping communication systems.

As an important research topic in the field of digital signal processing, modulation recognition of communication signals has shown great potential in military and civil fields. In the military field, modulation recognition provides an important technical means for obtaining enemy intelligence in electromagnetic countermeasures and selecting the best jamming and suppression method. Accurately identifying the modulation mode of frequency-hopping signals can provide strong support for military information warfare by, for example, judging the attributes of enemy and our own targets and jamming enemy signals [3]. Generally, after successfully intercepting enemy communication signals, it is undoubtedly a crucial task in communication countermeasure technology to determine the number levels and extract the feature level of the obtained mixed modulation signals and use the extracted features for further modulation recognition.

Traditional modulation recognition methods usually rely on manually designed features and complex signal processing algorithms, including maximum likelihood estimation based on hypothesis testing [4] and feature extraction based on pattern recognition [5]. These methods tend to perform poorly in the face of complex and variable frequency-hopping signals. In recent years, modulation recognition technology based on deep learning (DL) has attracted the close attention of researchers. Compared with traditional methods, modulation signal recognition based on DL does not need to rely on prior knowledge and can automatically extract features from data and classify them, so it not only has high classification accuracy but also stronger generalization ability in the face of large-scale data training. Mohamed A and others used a convolutional filter to use the basic convolutional neural network Alex Net and a residual neural network for compatibility with a constellation diagram, which significantly improved the accuracy of signal modulation classification [6]. Lihong Guang et al. removed noise from a two-dimensional time–frequency map of a frequency-hopping signal by adaptive Wiener filtering and accurately extracted the time–frequency map of each hop signal by using the algorithm in image processing, which achieved the accurate recognition of the modulation mode of the frequency-hopping signal and achieved good results at -4 db [7]. At present, DNNs are widely used in automatic modulation recognition (AMR) to complete signal detection and demodulation [8], which greatly improves the accuracy of modulation recognition. In communication countermeasures, a jammer can accurately identify the modulation mode being used in a target communication system and decode the frequency-hopping signal by training a DNN to more effectively interfere with and destroy the enemy's communication link.

Although deep learning modulation recognition technology has brought great convenience to people, its anti-interference performance has been questioned since 2013. In 2013, Szegedy et al. [9] found adversarial examples that can attack the neural network model—examples that can make the machine learning model misjudge or misclassify by perturbing the normal examples slightly and imperceptibly. The study indicates that deep neural network (DNN) models are typically characterized by their high complexity and sensitivity, which enable them to detect minute variations within the input space. Exploiting this characteristic, it is possible to enhance resistance to attacks by introducing precisely calibrated minor perturbations to the original samples. This method constructs adversarial examples that can provoke incorrect classifications by the model, thereby demonstrating a critical vulnerability in its predictive accuracy. Goodfellow et al. [10] proposed the fast gradient sign method (FGSM) in 2014. They added adversarial noise to the linear model and observed that when processing high-dimensional data input, the linear model was more vulnerable to the interference of adversarial examples, which overturned the theoretical explanation that the existence of adversarial examples was because the model was highly nonlinear. Kurakin et al. [11] introduced the iterative fast gradient sign method (I-FGSM), building on prior work. This approach incrementally introduces perturbations through multiple iterations and reprojects the currently generated adversarial samples back into a predefined constraint set. Classification outcomes indicate that most of these adversarial examples are misclassified, thereby demonstrating the efficacy of adversarial attacks on neural network classifiers in practical scenarios. Dong et al. [12] proposed a momentum iterative fast gradient sign method (MI-FGSM) to enhance resistance against sample attacks. This method integrates momentum into the gradient and gets rid of the bad local maximum in the iteration process to generate more mobile adversarial examples. Mardy et al. [13] proposed projected gradient descent (PGD), which is different from the clipping operation of I-FGSM. It limits the size of a disturbance by projecting the results of each iteration to the $\in -l_\infty$ field of pure input.

At present, research on adversarial examples is mainly focused on image and audio. In the field of communication signals, communicators can add adversarial examples with specific disturbances to modulated signals. These adversarial examples can attack the modem of a communication system so that the DNN model of the reconnaissance party cannot correctly demodulate the signal or cause wrong decoding results, which significantly improves the ability of the communicators to resist smart interference. Therefore, this paper proposes a frequency-hopping modulation signal adversarial example attack method based on adaptive whitening and feature gradient smoothing to reduce the recognition rate of the modulation signal in the DNN model. The main contributions of this paper can be summarized as follows:

1. Experiments show that the conventional method of generating countermeasure samples has shortcomings when attacking the frequency-hopping modulation recognition model, and, according to the particularity of the frequency-hopping signal and the rich space–time characteristics of the hidden layer of the model, a countermeasure sample generation method AWFGS-MIFGSM suitable for the field of frequency-hopping signal modulation recognition is proposed.
2. The method initially considers that frequency-hopping signals are non-stationary signals whose frequencies change non-linearly over time. This typical time-varying characteristic results in a relatively concentrated energy distribution within a short time frame. To address this, the acquired frequency-hopping signals undergo an adaptive whitening process. This treatment enables a more uniform distribution of energy across frequencies, eliminates correlations between signals, and simplifies the generation of adversarial samples.
3. This method uses the high-dimensional spatial features of the hidden layer of the target model to calculate the gradient to launch the attack, which ensures that the amount of characteristic information of the spectrum signal sample is rich enough. Considering that single-point gradient information might be unreliable due to loss function surface oscillations, the characteristic gradient is smoothed using surrounding sample data to identify the optimal direction for countering disturbances and improving adversarial sample transfer.

Section 2 of this paper introduces the basic principle of adversarial samples and adversarial attack based on DNN modulation recognition. In Section 3, the system model and the generation method of countermeasure samples based on adaptive whitening and feature gradient smoothing are described and analyzed. In Section 4, the experimental setup is explained, and a series of experiments are described from the perspective of white box attack and black box attack, and the experimental results are analyzed. Finally, we discuss and conclude this work in Section 5.

2. Related Literature Review

2.1. Adversarial Example Attack

Adversarial examples refer to the special samples formed by artificially adding subtle disturbances that are difficult to detect by the naked eye or that are visible to the naked eye after processing but that do not affect the overall system in the original data set. These disturbances are not random disturbances in the learning process but artificially constructed disturbances that can deceive the neural network model, as shown in Formula (1):

$$\min_{\delta} ||\delta||_2 \text{ s.t. } C(\mathrm{x} + \delta) = I; x + \delta \in [0,1]^m \tag{1}$$

where δ represents the added disturbance, C represents the neural network classifier, x represents the original image, and I represents the specified class. Since the minimum value of $||\delta||_2$ is not easy to calculate, the loss function is introduced to change Formula (1) to Formula (2):

$$\min_{\delta} C|\delta| + J(X+\delta, I) \text{ s.t. } X+\delta \in [0,1]^m \tag{2}$$

where J is the loss function, which is realized by calculating the cross entropy.

Adversarial samples possess strong camouflage capability, exploiting model vulnerabilities to launch targeted attacks that mislead the model into categorizing these samples into incorrect categories with high confidence. The impact of an adversarial example on modulation recognition is illustrated in Figure 1. By introducing counter disturbance, the signal originally identified as a sine wave with 97.85% confidence is misclassified as a square wave with 99.92% confidence. This demonstrates that despite the incorrect classification results, the waveforms of the two signals are nearly identical.

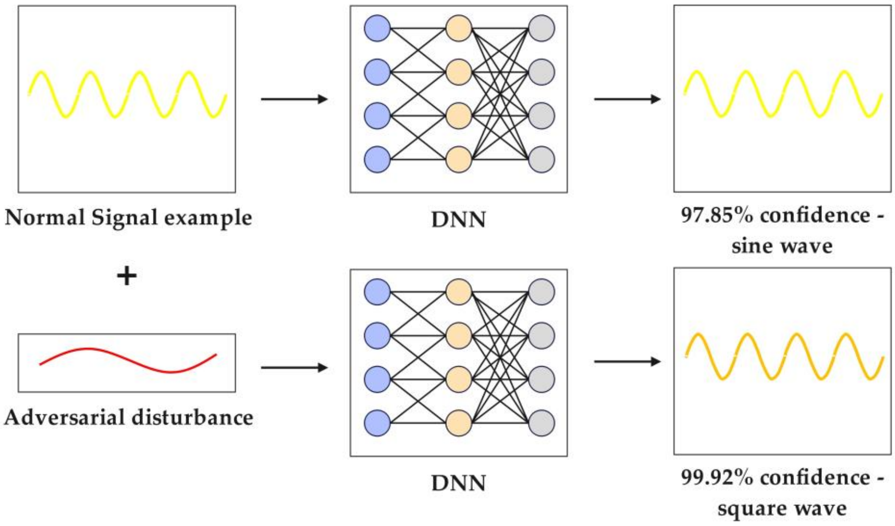

Normal Signal example
DNN
97.85% confidence - sine wave
+
Adversarial disturbance
DNN
99.92% confidence - square wave

Figure 1. Example of modulated signal adversarial example.

Below are descriptions of the four most commonly used methods to generate adversarial examples.

2.1.1. FGSM

FGSM is an efficient and fast adversarial example generation method proposed by Goodfellow [10] that is committed to generating adversarial examples close to original images. The generation formula is shown in Formula (3):

$$x^{adv} = x - \varepsilon \cdot \text{sign}(\nabla_x \mathrm{L}(\theta, x, t)) \tag{3}$$

where ∇_x is the gradient of loss function L to input x, ε is the parameter controlling the size of the disturbance, and t is the target category of the attack, that is, a single gradient iteration is performed in the direction of reducing the loss function corresponding to model category t. When the intention is to launch a no-target attack, the above formula is simply updated as follows:

$$x^{adv} = x + \varepsilon \cdot \text{sign}(\nabla_x \mathrm{L}(\theta, x, y)) \tag{4}$$

where y is the correct category corresponding to the input sample x. The biggest feature of FGSM is its efficient running speed, so it is often widely used in scenarios that need to generate many adversarial examples, such as confrontation training. However, its disadvantage is that the overall performance of the generated adversarial samples is somewhat poor.

2.1.2. I-FGSM

I-FGSM [11] can be regarded as a multiple-iteration version of FGSM. The original FGSM only adds a single-step disturbance along the direction of gradient increase, while I-FGSM makes a multi-step small disturbance along the direction of gradient increase through iteration and cuts the iteration results after each iteration update to ensure that they are kept within the valid interval (for example, it is usually the [0, 1] or [0, 255] interval for image data). Compared with FGSM, I-FGSM can construct more accurate disturbances, but the amount of calculation is increased. This method can be expressed as follows:

$$x_{i+1}{}^{adv} = clip_{x,\alpha}(x_i{}^{adv} - \varepsilon \cdot \mathrm{sign}(\nabla_x L(\theta, x, t))) \tag{5}$$

where the subscript i denotes the number of iteration rounds, $clip_{x,\alpha}(x^{adv}) = \min[1, x + \alpha, \max(0, x - \varepsilon, x^{adv})]$.

2.1.3. MI-FGSM

MI-FGSM [12] attack incorporates momentum into the I-FGSM attack by introducing a small number of gradients generated by the current step while retaining some gradients from the previous step to stabilize the update direction and avoid falling into the local extremum. The improvement of this method is the accumulation of the velocity vector in the gradient direction by using momentum. The formula is as follows:

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J(x_t{}^{adv}, y)}{\left|\left|\nabla_x J(x_t{}^{adv}, y)\right|\right|_1} \tag{6}$$

$$x_{i+1}{}^{adv} = x_i{}^{adv} + \alpha \cdot \mathrm{sign}(g_{t+1}) \tag{7}$$

First, $x_t{}^{adv}$ is input to classifier f to obtain gradient $\nabla_x J(x_t{}^{adv}, y)$; then, the velocity vector is accumulated in the gradient direction through Formula (6) to update g_{t+1}, and $x_{i+1}{}^{adv}$ is updated by applying the symbol gradient in Formula (7), finally generating disturbance F. Compared with FGSM and I-FGSM, MI-FGSM gives higher mobility of adversarial examples.

2.1.4. PGD

Compared with the one-step confrontation of FGSM, PGD [13] adopts the strategy of small-step and multi-step. PGD initializes with uniform random noise to project the gradient and clips the disturbance to a specified range after each iteration. The attack process is shown in Formula (8):

$$x_{t+1}^{adv} = \mathrm{proj}_{x,\varepsilon}(x_t^{adv} + \alpha \cdot sign(\nabla_x J(x_t^{adv}, y, \theta))) \tag{8}$$

where $\mathrm{proj}_{x,\varepsilon}(\cdot)$ is the projection operation.

2.2. Modulation Recognition Adversarial Example Attack Based on a DNN

Modulation recognition can be regarded as a classification problem involving N modulation modes. The signal received by the communication receiver can be expressed as $y = \alpha e^{j(2\pi\omega + \varphi)} x + \sigma$, where x is the signal modulated by the transmitter according to a specific modulation scheme, α transmits the impulse response of the wireless channel, ω is the frequency offset, φ is the phase offset, and σ indicates additive white Gaussian noise (AWGN). The purpose of any modulation classifier is to identify the modulation type $P(x \in N|y)$ of the signal given the received signal y.

Modulation recognition can be categorized into classical and DL-based methods, depending on the use of deep learning algorithms. DL-based modulation recognition automates feature extraction and classification by feeding preprocessed signals directly into the network, significantly reducing the time needed to manually analyze communication signal characteristics. This advantage makes the method better adapted to future situations following the development of wireless communication where the amount of information may increase significantly, and it has higher recognition accuracy. The process is shown in Figure 2.

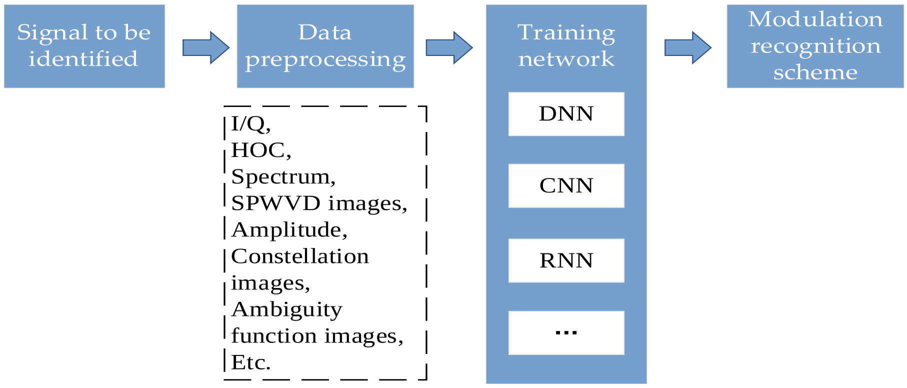


Figure 2. Modulation recognition process based on DL.

DNNs are central to DL-based modulation recognition technology. They process signal characterization results, analyzing preprocessed and extracted signal data to infer and output the modulation mode. O'Shea et al. [14] achieved the recognition and classification of three analog modulation signals and eight digital modulation signals based on a DNN model for the first time, and the accuracy rate reached 80%, proving the feasibility of applying DNNs to radio data recognition under the condition of a low signal-to-noise ratio. Ali et al. [15] employed IQ samples, constellations, and high-order cumulants to train sparse self-coding for modulation recognition, confirming the DNN's effectiveness in AWGN and flat fading channels via simulations. Xie et al. [16] used high-order cumulants to extract different features of each signal type to train a DNN for modulation recognition. When the signal-to-noise ratio was −5 dB and −2dB, the overall recognition accuracy of the algorithm exceeded 99%. At present, research on the modulation recognition of communication signals mainly focuses on fixed-frequency signals, and there is a big gap in research on the modulation recognition of frequency-hopping signals at home and abroad. For frequency-hopping modulation signal recognition, reference [17] introduced an algorithm that extracted instantaneous features and high-order cumulants from spread spectrum and conventional signals, enhancing recognition accuracy and reliability. Reference [18] developed a method using time–frequency energy spectrum texture features for modulation recognition, employing a support vector machine classifier for training and classification.

Although DNNs have many advantages in the field of signal modulation recognition, there are also some problems and challenges, such as the large amount of data demands and lack of model generalization ability; the deep learning model is also more sensitive to targeted adversary attacks. Small and intentional disturbances may lead to classification errors in the model, which seriously affect the reliability and security of signal recognition.

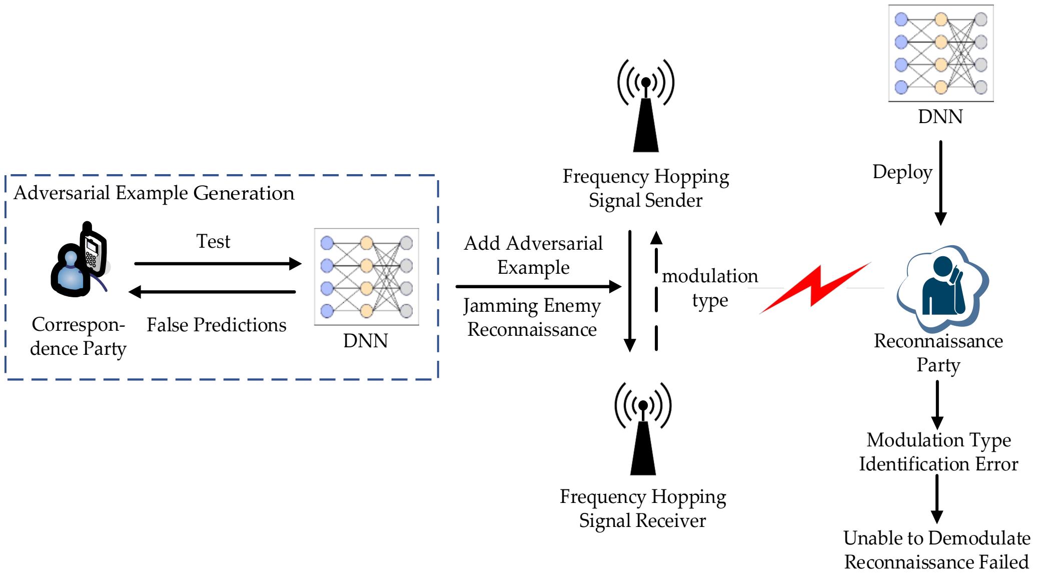
Research on countermeasure samples for modulation recognition started late. In recent years, the academic community has gradually turned its attention to research on countermeasure sample attack methods based on modulation classification. In 2018, Sadeghi [19] and others took the lead in research on countering sample attacks against the modulation recognition model of communication signals based on DL. The research results show that the modulation recognition model based on a DNN automatic encoder is vulnerable to interference. The paper further expounds on how attackers can effectively counterattacks. In 2020, Zhao [20] and others studied and tested counterattack in the process of signal recognition, successfully reduced the recognition accuracy of the model through experiments, and verified the generalization ability of the model. In 2021, Lin et al. [21] analyzed the effects of various gradient-based counterattack methods on modulation recognition; the experimental results showed that when the disturbance intensity was set to 0.001, the prediction accuracy could be reduced by 50%.

At present, the primary goal of counterattacks in modulation recognition is to improve attack performance, but research in the field of communication is still in its infancy, lacking the theoretical interpretation of counter samples. Most of the existing explanations are limited to a hypothetical interpretation and do not fully analyze the characteristics of the communication signal. Furthermore, current methods inadequately address the characteristics and gradient reliability of modulation signals, leading to issues like poor counterattack performance and limited black box adaptability. Improving the processing of the modulation signal's characteristic gradient can significantly enhance both the effectiveness of attacks and the model's security.

3. Anti Attack Method Based on Adaptive Whitening and Feature Gradient Smoothing

3.1. System Model

In the wireless communication environment, both the transmitter and receiver of frequency-hopping signals use the same communication protocol. During the communication process, the sender first modulates the frequency-hopping signal onto a carrier using a particular method to create a frequency-hopping modulation signal, which is then transmitted over the channel. The receiver needs to use the same modulation method as the sender to demodulate and reconstruct the received modulated signal and finally complete the communication process. Considering the existence of the reconnaissance party in the communication process, this party intercepts the communication signal and uses the intelligent DNN model to identify the modulation type of the signal, aiming to capture the content of the frequency-hopping signal. The system model is shown in Figure 3.



Adversarial Example Generation
Test
False Predictions
Correspon-dence Party
DNN
Add Adversarial Example
Jamming Enemy Reconnaissance
Frequency Hopping Signal Sender
modulation type
Frequency Hopping Signal Receiver
DNN
Deploy
Reconnaissance Party
Modulation Type Identification Error
Unable to Demodulate Reconnaissance Failed

Figure 3. System model adversarial example attack in communication system.

To avoid this situation, the communication party needs to add adversarial examples to the communication signal on the premise of ensuring that its own communication is not affected as much as possible. This is done to flexibly attack the reconnaissance party deploying the DNN model and interfere with and mislead the identification results of the DNN model of the reconnaissance party so that the reconnaissance party cannot correctly identify the modulation type or demodulate and recover the intercepted signal, achieving the purpose of anti-reconnaissance. In this paper, an anti-attack method based on adaptive whitening and feature gradient smoothing (AWFGS) is proposed. Initially, the obtained frequency-hopping signal is adaptively whitened to enhance the useful features of the signal and facilitate subsequent feature extraction. Subsequently, the hidden layer feature extracted by the DNN model is used as the attack object, which significantly improves the attack accuracy and produces more refined adversarial examples, and the generated countermeasure samples have higher mobility.

3.2. Adaptive Whitening

Blind source separation refers to the process of recovering the source signal only by using the observed signal according to the statistical characteristics of the signal without any prior knowledge of the source signal and transmission channel. It has important applications in wireless communication and voice signal and digital image processing [22]. As a necessary preprocessing step of blind source separation, whitening can identify the mixing matrix and directly realize the blind separation of non-stationary signals.

Currently, the whitening algorithm can be divided into a batch algorithm and an adaptive algorithm. The batch processing algorithm has good robustness, but it cannot meet the requirements of the system for real-time signal processing. The adaptive whitening algorithm, which is less complex, supports the online processing of mixed signals with effective real-time performance and has therefore been widely adopted and researched [23]. Therefore, when whitening the original signal, the whitening algorithm with an adaptive form [22] is often used. Since signal processing often involves processing signals with different characteristics and statistical properties, and these signals may have different distributions in time and frequency domains, adaptive whitening can better process different types and properties of signal data by adjusting the characteristics and statistical properties of the signal to preprocess the data, thus enhancing the overall effectiveness and quality of signal processing. Moreover, in feature extraction and pattern recognition, adaptive whitening can enhance the useful features in the signal, which is helpful for subsequent pattern recognition, classification, or prediction. Its structure is shown in Figure 4 [24].

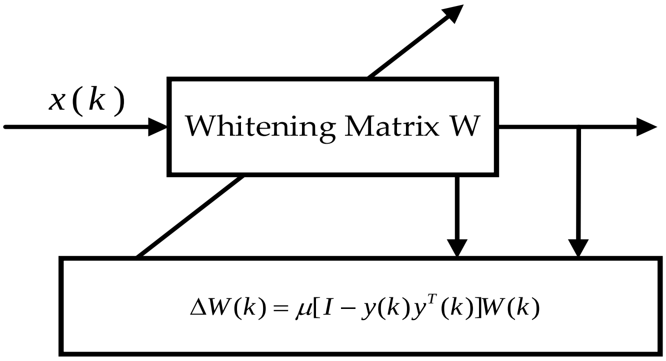


Figure 4. Structure of adaptive whitening algorithm.

$W \in R^{n\times m}$ is the whitening matrix with full rank, and the output whitening vector $y(k)$ meets the following characteristics:

$$E\left\{y(k)y^T(k)\right\} = WE\left\{x(t)x(t)^T\right\}W^T = I \quad (9)$$

$$R_{xx} = E\left\{x(k)x(k)^T\right\} = V_x H_x V_x^T \quad (10)$$

where $x(k)$ is the observation signal, I is the identity matrix, R_{xx} is the autocorrelation matrix of signal $x(k)$, and V_x and H_x are the eigenvector matrix and eigenvalue matrix of R_{xx}, respectively.

The adaptive whitening algorithm has excellent tracking performance and conditions for real-time signal processing. Its estimation of the whitening matrix $W(k)$ can be obtained by minimizing the cost function of Equation (11):

$$J(k) = -\frac{1}{2}\{\log[\det(W^T(k)W(k))] - \sum_{i=1}^{n} \mathrm{E}(y_i^2(k))\} \tag{11}$$

where $\det(Z)$ represents the determinant operation on matrix Z. On the derivation of the instantaneous estimation of $W(k)$ over $J(k)$, there are the following:

$$\frac{\partial J(k)}{\partial W(k)} = -\left[I - y(k)y^T(k)\right]W(k) \tag{12}$$

Based on Equation (12), the updated formula of whitening matrix $W(k)$ in the adaptive algorithm can be obtained as follows:

$$W(k+1) = W(k) - \mu\frac{\partial J(k)}{\partial W(k)} = W(k) + \mu[I - y(k)y^T(k)]W(k) \tag{13}$$

where $y(k) = Wx(k)$ is the whitening signal and μ is the step size parameter. In order to ensure convergence, its value should meet $0 < \mu < \frac{2}{\sqrt{\lambda_{x-\max}}(1+\sqrt{\lambda_{y-\max}})}$, where $\lambda_{x-\max}$ and $\lambda_{y-\max}$ respectively represent the maximum eigenvalues of matrices R_{xx} and $R_{yy} = E\{y(k)y^T(k)\}$.

Different from the waveform of constant-frequency continuous signals, the waveform of frequency-hopping signals shows significant discontinuity, which leads to the inaccurate extraction of frequency-hopping signal features directly using the original signal and then affects the subsequent signal processing. However, gradient features are generally represented by high-dimensional data with high correlations and much redundant information, which not only increases the difficulty of data processing and model training but also reduces the amount of information on features, resulting in some gradient features being affected by abrupt points in the signal when representing modulated signals, making gradient calculation unstable. To solve the above problems, an adaptive whitening algorithm is introduced to minimize the interference between frequencies, effectively remove the correlation between data, improve the independence of sample features, and facilitate the accurate feature extraction of subsequent models. Additionally, the reduction in correlation reduces the dependence of the model on specific features, so the adversarial examples remain effective between different models, that is, there is a higher attack success rate between different models.

3.3. Feature Gradient Smoothing

Reference [25] pointed out that the local non-smoothness of the loss surface impairs the transferability of generated adversary samples. To solve this problem, this study used the local average gradient instead of the original gradient to generate countermeasure samples, as shown in Figure 5.

Source model a was used to generate countermeasure samples to attack target model B. g_A and G_A respectively represent the gradient of a corresponding point on the loss function surface of the two models. It can be seen that the loss function curve of model a showed an obvious oscillation phenomenon, which made the direction difference between g_A and g_B larger, which meant that the countermeasure samples generated on g_A could not effectively attack model B, and the migration of countermeasure samples was low. If the gradient smoothing process was applied to model a, the local average gradient G_A was obtained to replace the original g_A-generated countermeasure samples to attack model B. Since the directions of G_A and g_B were closer, the migration of countermeasure samples could be higher, and the attack performance for model B was stronger, that is, $\left\langle \hat{G}_A, \hat{g}_B \right\rangle > \left\langle \hat{g}_A, \hat{g}_B \right\rangle$.

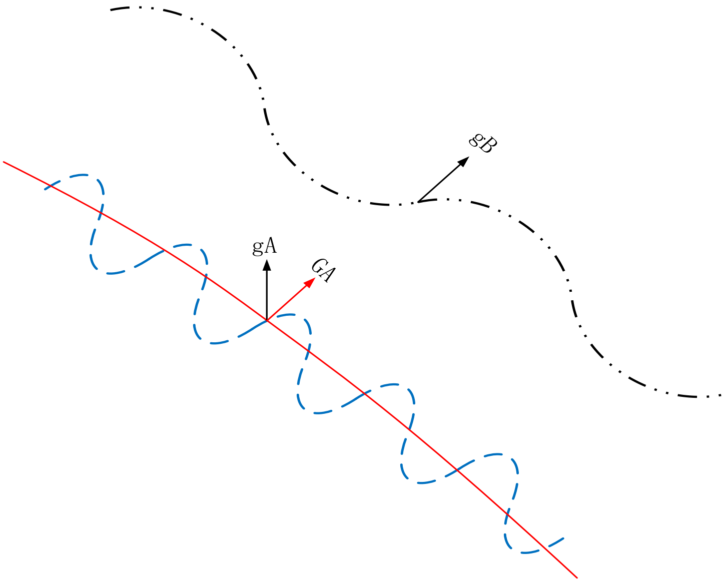


Figure 5. Example of non-smoothed lossy surface.

In this study, we approximated the mathematical expectation of the gradient in the neighborhood by sampling n times in the neighborhood of the sample x_s:

$$E_{\varepsilon'}(x_s) = \frac{1}{N}\sum_{i=1}^{n} f_L(x_s^i;\theta) \tag{14}$$

where $E_{\varepsilon'}(x_s)$ is the average value of characteristics in the x_s neighborhood and ε' is the upper boundary of the x_s neighborhood, set as $\varepsilon' = \beta\varepsilon$, where β is the super parameter.

At present, although modulated signals based on gradient have destructiveness against attacks, they also have a series of limitations and challenges. Compared with high-dimensional data such as pictures, the amount of information in the spectrum signal sample is smaller, and the high-dimensional vector of the middle layer of the deep learning model can magnify the key features of the input sample. If the middle layer features extracted by the DNN model are used as the attack object, and the average gradient of its neighborhood is used to replace its single-point gradient, the surface oscillation of the loss function can be effectively smoothed, the accuracy of the attack can be improved, and a more refined modulated signal can be generated against the sample. In addition, for the same type of modulated signal samples, after different DNN models are trained, the output characteristics of the intermediate layer usually show some similarity, and the characteristics of the samples are transferable. Therefore, the disturbance generated by the counterattacks based on the characteristics of the middle layer should have better mobility.

3.4. Description of Attack Methods

Algorithm 1 introduces the process of generating countermeasure samples based on adaptive whitening and feature gradient smoothing. Firstly, the signal samples are adaptively whitened before the original signal input model, so that the sample features extracted after the input model are more effective, and the gradient can be calculated by using the rich space–time features in the hidden layer of the DNN model. Then, n samples are taken within a certain domain of the current data point x_n^{adv}, n x_n^{adv} samples are input into the intercepted hidden layer model f_L, $E_{\varepsilon'}(x_n^{adv})$ is calculated according to Formula (14), and then the mathematical expectation $E_{\varepsilon'}(x)$ of the gradient in the neighborhood of the data point is used to replace the gradient value of the point for subsequent iterations to reduce unstable factors, avoiding the algorithm falling into local extreme points and effectively smoothing the oscillation of the loss function surface. Then, a new loss function J_L is constructed by Formula (15), and the characteristic gradient is calculated and the attenuation factor g_n is updated. Finally, x_{n+1}^{adv} is continuously updated to obtain the required countermeasure sample x^{adv}. The complete block diagram of the algorithm is shown in Figure 6.

Algorithm 1 AWFGS-MIFGSM adversarial example attacks

```
Input: Raw modulated signal sample $x_S$, Truncate hidden layer model $f_L$, New loss function $J_L$, Norm constraint $p$, Momentum decay factor $\mu$, Disturbance size $\varepsilon$, Sampling times $N$, Iterations $T$, Attenuation factor $g_n$, Neighborhood range size $\beta$.
Output: Optimize adversarial example $x^{adv}$.
Iteration step $\alpha = \varepsilon / T$, neighborhood boundary $\varepsilon' = \beta \cdot \varepsilon$
$g_0 = 0$, $x_0^{adv} = x_S$
For t = 0 to T − 1 do
    $(x_n^{adv})_{whitened}$ is obtained by adaptive whitening of $x_n^{adv}$
    Take N samples randomly for $\varepsilon'$ neighborhood of $(x_n^{adv})_{whitened}$
    Input N samples into the hidden layer model $f_L$ and obtain $E_{\varepsilon'}(x_n^{adv})_{whitened}$ according to Formula (13)
    Calculate new loss function $J_L(x_n^{adv};\theta) = \left|\left|E_{\varepsilon'}(x_n^{adv})_{whitened}\right|\right|_p$
    Calculate characteristic gradient, update $g_{n+1}$
        $g_{n+1} = \mu \cdot g_n + \frac{\nabla_{x_n^{adv}} J_L(x_m^{adv},\theta)}{||\nabla x_n^{adv} J_L(x_m^{adv},\theta)||_1}$
    Update $x_{n+1}^{adv} = \mathrm{Clip}_{\varepsilon}\left\{x_n^{adv} + \alpha \cdot \mathrm{sign}(\mathrm{g}_{n+1})\right\}$
End for
Obtain optimized adversarial example $x^{adv}$
```

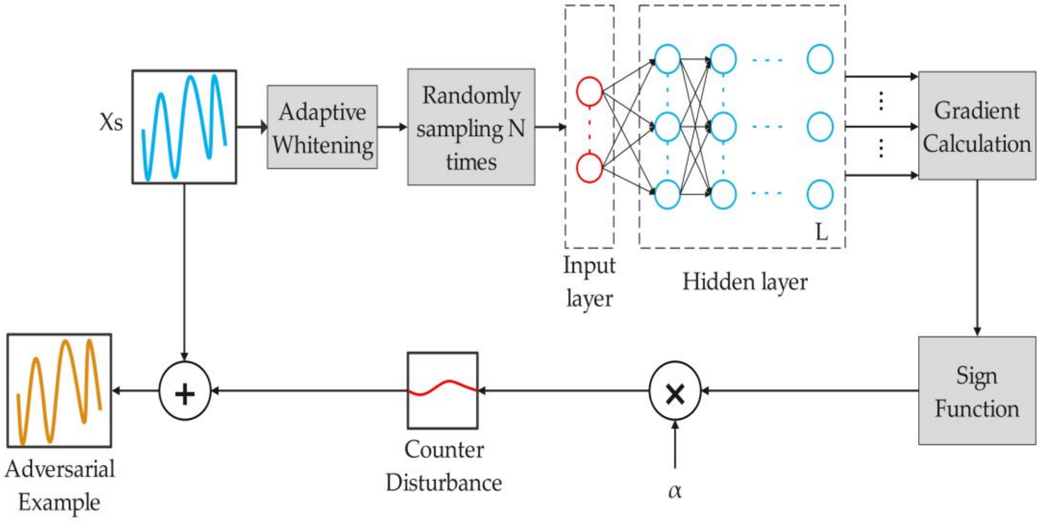


Figure 6. Anti sample attack based on adaptive whitening and feature gradient smoothing.

After obtaining the original signal sample feature $(x_S)_{whitened}$ after whitening, the average feature information is obtained. Different loss functions can be designed by using different p ($p = 0, 1, \ldots, \infty$) norms to constrain the features, as shown in Formula (15).

$$J_L(x_S,\theta) = ||E_{\varepsilon'}(x_s)||_p \tag{15}$$

To verify the effectiveness of the experimental method, AWFGS is introduced into MI-FGSM to obtain the momentum iteration fast gradient sign method AWFGS-MIFGSM, which is based on adaptive whitening and feature gradient smoothing. The pseudo-code of the algorithm is shown in Algorithm 1.

3.5. Analysis of Attack Methods

1. In machine learning, input data typically consist of various measurements, and there is a significant correlation between adjacent sampling points. If unprocessed data are fed into the network, this creates excessive redundancy and lowers the network's training efficiency. A whitening operation before feature extraction can decrease data correlation and streamline the feature extraction process. Subsequently, the gradient calculated from these processed features is used to attack the DNN model. This approach enhances the mobility of the generated adversarial examples, increasing their attack success rate across different models. The reduction in correlation diminishes the model's reliance on specific features, thereby increasing the likelihood that an adversarial example will be effective across various models.
2. Most of the attacks based on label gradients are methods that attackers try to maximize the gradient of the loss function with respect to the input data so that the model can produce a false classification of the adversarial example. In this process, the optimization goal is to maximize the classification loss. Adjusting the input data thus generates classification errors in the adversarial examples. The proposed algorithm does not use the classification loss as the optimization goal but uses extensive high-dimensional feature data in the DNN hidden layer to design adversarial examples, which not only makes the obtained sample signal features richer but also produces finer disturbances.
3. At present, most of the methods that have been used to combat sample attacks use the single-point data gradient value on the optimized path. Because the surface oscillation of the loss function leads to the unreliability of the single-point gradient information, the method proposed in this paper helps the model make full use of the data point neighborhood gradient information by whitening and neighborhood sampling, making the gradient direction on the loss function of the source model and the attack model closer so that the disturbance generated by this has better mobility and the success rate of black box attacks is higher.

4. Experimental Results and Analysis

4.1. Experimental Setup

All experiments were calculated on NVIDIA GeForce GTX 1650 GPU and implemented by Tensorflow2.8 and cuda12.1.

4.1.1. Data Set

In this study, the frequency-hopping modulation signal was generated by MATLAB R2024a software simulation as the experimental data set. The data set covered four common modulation methods of frequency-hopping signals and simulated the Gaussian white noise in the real channel environment, which better restored the signals collected by the real communication. Using this data set, the recognition of the model for the basic modulation type signals and noise interference environment could be compared. The four modulation modes of the data set were divided into two digital modulation modes (QPSK and MFSK) and two analog modulation modes (AM and SSB). The frequency-hopping signal sampling rate was 40 KHz, the hopping speed was set to 500 hop/s, and eight frequency-hopping points were set, 250 points for each hopping. With the background of Gaussian white noise, the signal-to-noise ratio ranged from −20 dB to 18 dB, with an interval of 2 dB. The frequency-hopping signal of each modulation type generated 300 samples under each signal-to-noise ratio, including 24,000 signal samples in total. The data set was divided into a 70% training set, 20% verification set, and 10% test set.

4.1.2. DNN Model

Considering the characteristics of signal samples, model parameters, recognition effects under normal conditions, and other factors, ResNet, CLDNN, and LSTM were selected as modulation recognition models in this study. Each model was trained 500 times, and the learning rate was set to 0.001. If the training loss of the test set did not decrease for five consecutive times, the learning rate was halved.

4.1.3. Hyperparameter Settings

During the model training phase, to ensure the training's efficiency and consistency, the number of iterations, learning rate, and other hyperparameters were kept consistent. To set the maximum disturbance reference [19], PNR (perturbation-to-noise ratio) controlled ε under different signal-to-noise ratios, iteration times $M = 10$, the momentum attenuation factor $\mu = 0.7$, the step size of adaptive whitening was set to 0.001, sampling times N in the neighborhood was 30, and the neighborhood range size β was 11.

4.1.4. Evaluation Index

To effectively evaluate the AWFGS-MIFGSM method, three evaluation metrics—total attack time, attack success rate, and black box migration rate—were defined for the generated countermeasure samples.

The total attack time was represented by the time required by different algorithms in the white box scenario to complete the white box targetless attack on all samples on the model.

The attack success rate was expressed by the model recognition accuracy (MRA). The lower the recognition accuracy rate, the higher the attack success rate against the sample. If the total number of test samples was m and the number of samples successfully identified by the model was n, then the MRA was as follows:

$$MRA = \frac{N}{M} \times 100\% \tag{16}$$

Black box mobility (BBM) refers to the ratio between the number of samples that can deceive both the white box model and the black box model in the countermeasure samples and the number of successful deceptions of the white box model. Let the number of samples that successfully deceive the white box model be D_w, and the number of samples that can also deceive the black box model among the samples that deceive the white box model be D_b. Then, BBM is as follows:

$$BBM = \frac{D_b}{D_w} \tag{17}$$

4.2. Analysis of Results in Different Experimental Environments

In this study, the data set was tested with three models, and the modulation recognition accuracy of the three models was obtained, as shown in Figure 7. With the increase in the signal-to-noise ratio, the accuracy of the three models showed an upward trend and then tended to be stable. When the signal-to-noise ratio was negative because the noise power exceeded the signal power, the characteristics of the signal waveform itself were distorted, so the recognition rates of the three models were generally lower than those under the positive signal-to-noise ratio. When the signal-to-noise ratio was greater than 2 dB, the recognition effect was the best, and the curve tended to be stable. The recognition accuracies of CLDNN and ResNet were close at 2–18 dB.

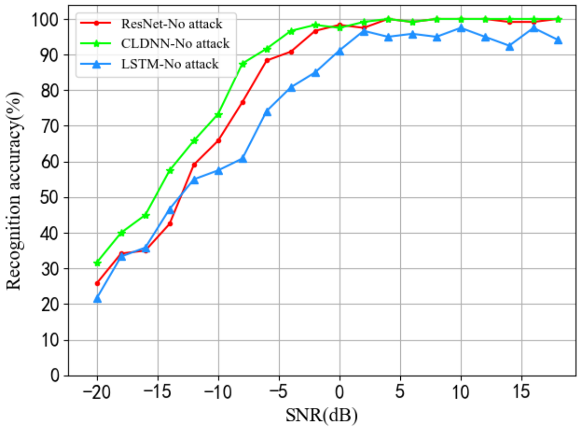


Figure 7. Modulation recognition accuracy of three models without attack.

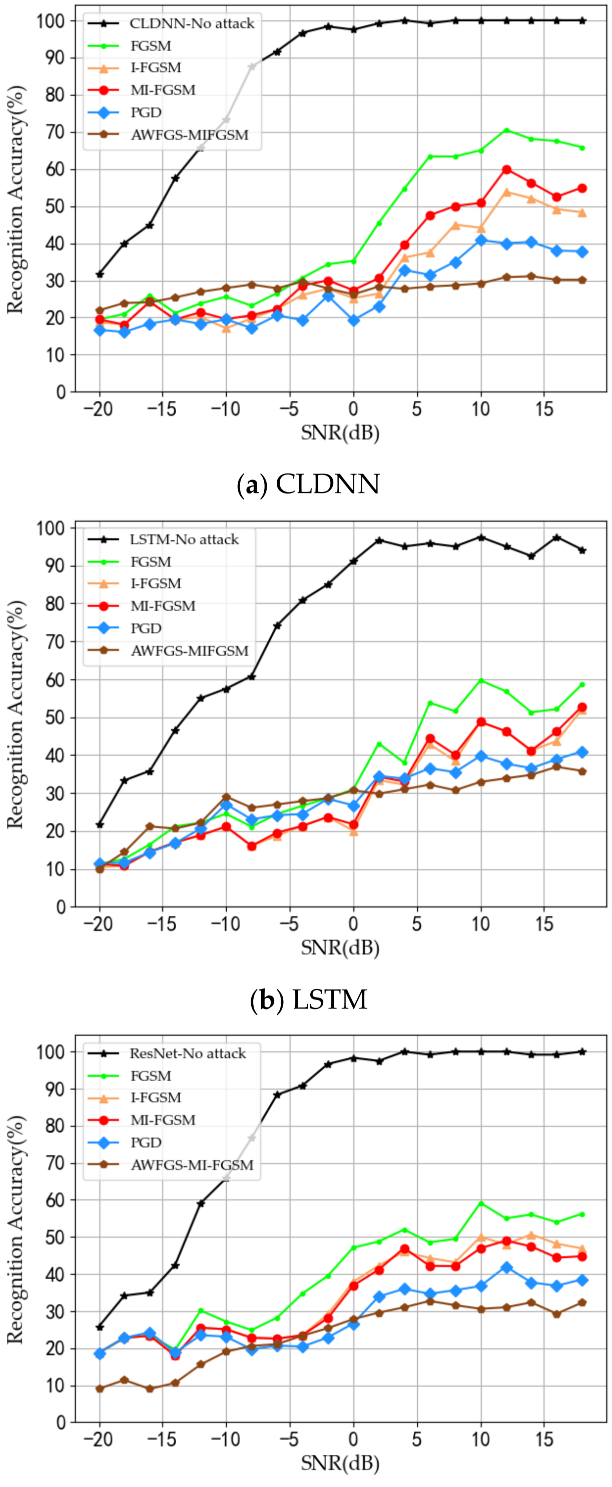
4.2.1. Analysis of White Box Environment Experiment

The core purpose of generating adversarial examples was to cause the recognition model to misclassify the original samples. Based on this feature, this method only attacked the samples correctly classified in the original samples when carrying out white box attacks. In addition, considering the characteristics of modulated signals with different signal-to-noise ratios, it was also necessary to attack one by one for different signal-to-noise ratios when performing a counter sample attack.

Figure 8 shows the modulation recognition accuracy of the three models after generating adversarial examples from FGSM, I-FGSM, MI-FGSM and the algorithm AWFGS-MIFGSM proposed in this paper under the condition of no target in the white box. A comparison showed the following:

1. The white box no-target attack was relatively simple. Under low SNR, each attack method could make the recognition rate of the model reach less than 25%, but, under high SNR, the FGSM attack effect was the worst, and the accuracy rate of the model only decreased by about 35% at 10 dB, which was significantly weaker than other attack methods. The reason may be that FGSM covers single-step attacks, and the gradient direction of the generated disturbance was inaccurate for the nonlinear model.
2. MI-FGSM introduces momentum into I-FGSM to correct the gradient. Theoretically, the attack effect should be better than that of I-FGSM. However, in the LSTM and ResNet models, the attack effect of the two models was almost the same under low SNR. In the CLDNN model, the attack effect was not as good as that of I-FGSM when it was more than −2 dB. On the one hand, the difference in the model structure had an impact on the output. On the other hand, the amount of information in a single signal sample may have been too small, so there was no qualitative change to the gradient correction, or even the opposite effect.
3. PGD is recognized as the most effective first-order attack in the industry. It can be seen from the algorithm that PGD randomly added some noise to the attack destination samples and projected the gradient obtained in each iteration, which could retain more useful disturbance information. As a result, the attack effect was significantly better than that of FGSM, I-FGSM, and MI-FGSM, and the model recognition rate could be reduced to about 40%.
4. At 10 dB, the recognition rate of the CLDNN model, LSTM model, and ResNet model decreased by 71%, 66%, and 69%, respectively. However, at low SNR, the recognition rate of the CLDNN and LSTM models was slightly higher than that of the other attack methods. On the one hand, whitening and gradient smoothing may cause the method to generate more diversified countermeasure samples. At low SNR, this diversity may make it easier for countermeasure samples to escape from the detection or recognition system, thereby improving the recognition rate. On the other hand, the CLDNN and LSTM models may have certain robustness when processing sequence data, and

this method "corrects" some noise in the original signal to some extent at low SNR, making it easier to identify at low SNR.



(**a**) CLDNN

(**b**) LSTM

(**c**) ResNet

Figure 8. Recognition accuracy of three models under white box no-target attack.

From the previous analysis, the attack effect of FGSM was the worst, but, due to its one-step calculation characteristics, it took the least time to generate adversarial examples. I-FGSM, MI-FGSM, and PGD were all iterative attacks. With the same number of iterations, it took a relatively long time to generate adversarial examples, but the difference was not significant. Because the proposed algorithm AWFGS-MIFGSM needed to sample samples and calculate the feature gradient smoothing, it took the longest time to generate adversblearial examples, but the attack success rate was the highest (see Table 1).

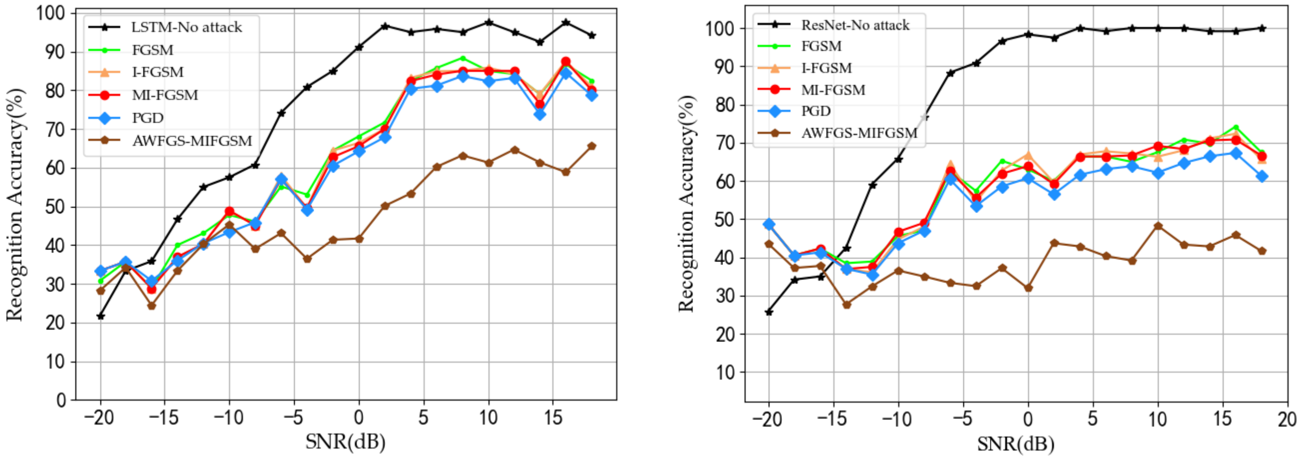
Table 1. The time consumption of the three models to generate countermeasure samples under the white box no-target attack.

DNN Model	**ATS (min)**				
	FGSM	**I-FGSM**	**MI-FGSM**	**PGD**	**AWFGS-MIFGSM**
CLDNN	1.05	7.13	7.35	8.08	9.23
LSTM	0.91	7.31	7.47	7.44	48.09
ResNet	1.21	2.79	2.85	3.19	15.26

4.2.2. Experimental Analysis of Black Box Environment

In the real electronic warfare environment, the information related to the target model is often unknown to the communicators, that is, it is usually the case of a black box attack. At this time, the adversarial examples generated by the communicators must have good mobility.

Different from the traditional attack method of using an alternative model to replace the target black box model, to better verify the transferability of countermeasure samples, this study directly migrated the countermeasure samples generated in the CLDNN white box attack to the LSTM and ResNet models to execute the black box attack. The experimental results are shown in Figure 9.



(**a**) LSTM

(**b**) ResNet

Figure 9. Recognition accuracy of two models under black box non-target attack.

As can be seen from Figure 9, due to the unknown black box information, the effect of all the anti-attack methods was reduced to varying degrees, but the attack performance of the method proposed in this paper remained optimal whether in low SNR or high SNR. It can be seen from Figure 9a that for the black box model of LSTM, the attack methods that could significantly reduce the recognition rate of the CLDNN model migrated to the LSTM model, and the attack effect was significantly worse. At 10 dB, FGSM, I-FGSM, and MI-FGSM only reduced the recognition rate of the LSTM model by about 13%. Although PGD achieved good results in the white box attack, its attack effect in the black box model was also unsatisfactory. In contrast, the attack method proposed in this paper still had a good effect. At 10dB, the recognition rate of the LSTM model was reduced by 37%.

It can also be seen from Figure 9b that the adversarial examples generated by the attack method based on adaptive whitening and feature gradient smoothing still had a strong attack effect when migrated to the ResNet black box model. At 10 dB, FGSM, I-FGSM, and MI-FGSM only reduced the recognition rate of the ResNet model by about 31%. The PGD method only reduced the recognition rate of the ResNet model by 36%. The recognition rate of the RESNET model was reduced by 51% by the proposed method.

Figure 10 shows the proportion of the counter samples that successfully attacked the white box model but also successfully attacked the black box model, that is, the comparison of black box mobility. It is evident that the black box mobility of the adversarial examples generated by the AWFGS-MIFGSM method was higher than that generated by the traditional method, whether using the LSTM model or the ResNet model, which shows that the adversarial example attack method proposed in this paper has superior attack migration performance, significantly improving the robustness of adversarial examples.

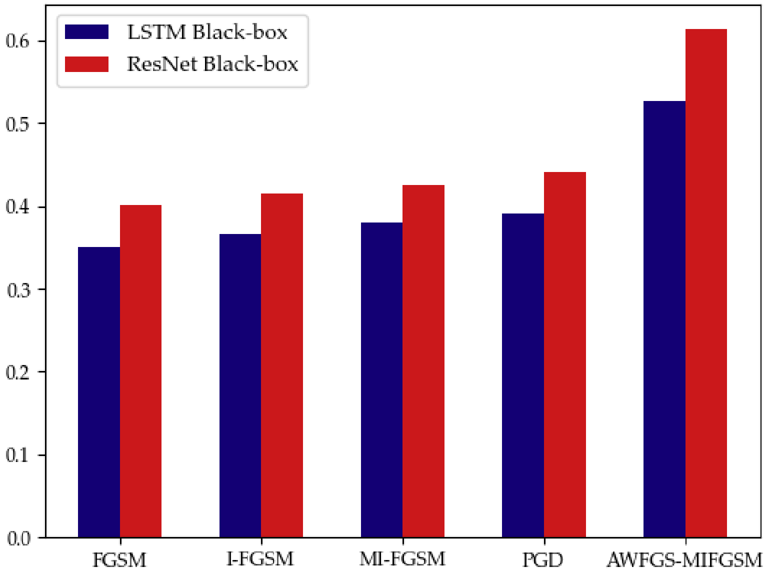


Figure 10. Black box mobility.

4.2.3. Experimental Performance Analysis under Mixed Signal-to-Noise Ratio

In the actual battlefield environment, the signal-to-noise ratio of the received frequency-hopping signal is not fixed. In order to better verify the attack performance of the method proposed in this paper under a mixed signal-to-noise ratio, the data set was processed as follows: 3000 samples were randomly extracted from each type of signal, and the training set and verification set were formed according to 4:1. The above three models of ResNet, CLDNN, and LSTM were trained, and the test set was composed of 10% of all test samples divided by 4.1. A comparison of the recognition results of the three models in the white box environment is shown in Table 2.

Table 2. Comparison of average recognition accuracy of three models under mixed SNR.

DNN Model	Recognition Accuracy under Different Attack Modes (%)					
	No Attack	FGSM	I-FGSM	MI-FGSM	PGD	AWFGS-MIFGSM
ResNet	82.08	39.77	34.43	33.61	28.67	23.66
CLDNN	84.17	42.67	31.57	34.66	26.50	27.75
LSTM	77.24	35.25	28.58	29.17	28.15	28.26

It can be seen from the results in Table 2 that for the mixed signal-to-noise ratio data, the three models had good recognition performance for frequency-hopping modulated signals under conventional conditions, and the average recognition accuracy was more than 77%. The FGSM, I-FGSM, and MI-FGSM attack methods significantly reduced the recognition rate of the model, and the PGD method had a better attack effect than the first three methods in the three models, but the attack effect of the AWFGS-MIFGSM method proposed in this paper was stronger than that of PGD method in the ResNet model, and was almost the same as that of the PGD method in the CLDNN and ResNet models, indicating that this method still had strong attack performance for mixed signal-to-noise ratio data.

4.2.4. Hyperparameters Analysis

The control variable method was used to study the hyperparameters. Therefore, except for the number of samples in neighborhood N and the size of neighborhood β, other parameters remained unchanged. At the same time, all signal samples under 10 dB were extracted from the original data set, and 600 signal samples with a signal-to-noise ratio of 10 dB were randomly selected from the training set to form a new test set, which was subjected to white box non-target attack.

First, the neighborhood size β was analyzed and the sampling times N were set in the neighborhood to 30. The experimental results are shown in Figure 11. It can be seen from the figure that when $\beta = 0$, the recognition accuracy was the highest and the attack effect was the worst. When increasing the value of β, the recognition rate of the model continued to decline until $\beta = 11$, when the curve reached the inflection point, but the recognition rate increased. According to this, when the neighborhood size β was set to 11 in the experiment, the attack effect was the best.

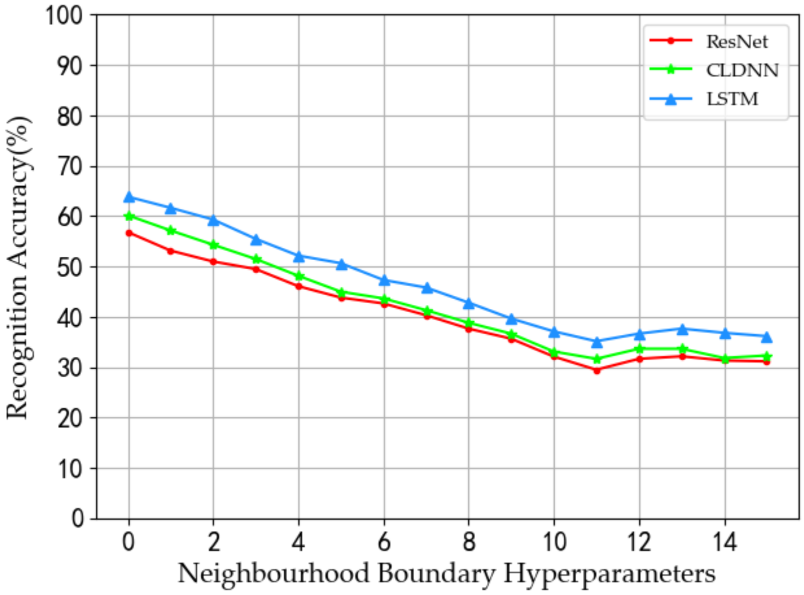


Figure 11. Relationship between neighborhood boundary and recognition accuracy at 10 dB.

We then analyzed the sampling times N within a specified neighborhood, setting the neighborhood range size to $\beta = 11$. The experimental results are shown in Figure 12. It can be seen from the figure that when $N = 0$, the recognition accuracy was the highest and the attack effect was the worst. When increasing the value of N, the recognition rate of the model continued to decline until $N = 30$, when the curve reached the inflection point, and the recognition rate tended to be stable. Considering that the larger N, the greater the computational overhead and time cost of the experiment, N was set to 30 in the experiment.

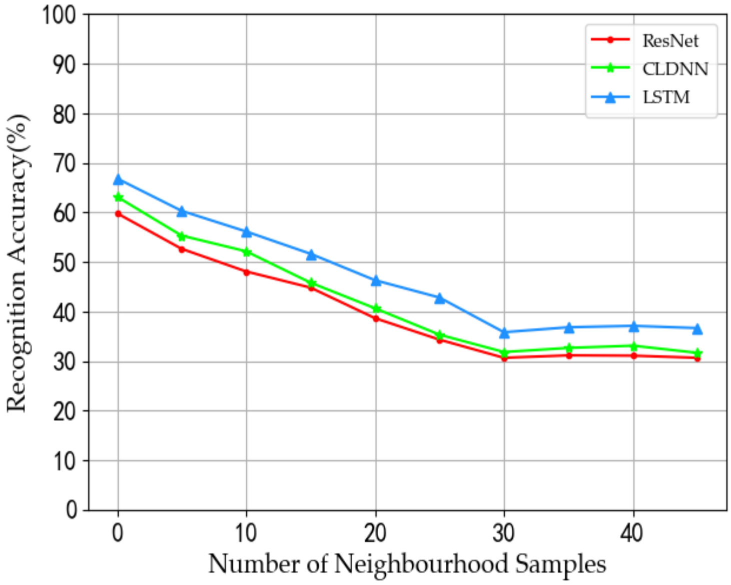


Figure 12. Relationship between sampling times and recognition accuracy in neighborhood under 10 dB.

5. Conclusions

Addressing the big gap in research on frequency-hopping modulation recognition in the field of anti-attack, this paper proposed a method of frequency-hopping modulation signal adversarial example attack based on adaptive whitening and characteristic gradients. Different from the conventional gradient attack, this study did not use classification confidence as the backpropagation data; instead, it employed the high-dimensional characteristics of the middle layer of the model to design the corresponding countermeasures. The simulation results show that the algorithm proposed in this paper achieves excellent results in both white box attack and black box attack. Although the algorithm has some advantages, there are still some further improvements: (1) like most adaptive algorithms, the adaptive whitening algorithm also has a contradiction between convergence speed and steady-state performance, which often cannot meet the requirements of the system for time-varying environment tracking ability, algorithm convergence speed, and steady-state error; (2) compared with the method of directly using tags to attack, the method proposed in this paper needs more time. How to reduce the complexity of the algorithm and improve efficiency is the main problem to be considered in the future.

Author Contributions: Conceptualization, Y.Z. and Y.L.; methodology, Y.Z., Y.L. and Z.D.; software, Y.Z. and Y.L.; validation, Y.Z., Y.L. and Z.D.; formal analysis, Y.Z., Y.L. and Z.D.; investigation, Y.Z. and Y.L.; resources, Y.Z.; data curation, Y.Z.; writing—original draft preparation, Y.Z.; writing—review and editing, Y.Z. and Y.L.; visualization, Y.Z.; supervision, Y.Z.; project administration, Y.Z.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 2022-SKJJ-B-112.

Data Availability Statement: The data presented in this study are not available due to privacy.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gummadi, R.; Wetherall, D.; Greenstein, B.; Seshan, S. Understanding and mitigating the impact of RF interference on 802.11 networks. In Proceedings of the ACM SIGCOMM Computer Communication Review, Kyoto, Japan, 27–31 August 2007; pp. 385–396.
2. Gao, Y.; Xiao, Y.; Wu, M.; Xiao, M.; Shao, J. Game theory-based anti-jamming strategies for frequency hopping wireless communications. *IEEE Trans. Wirel. Commun.* **2018**, *17*, 5314–5326. [CrossRef]

3. Zhang, J.; Yu, L. Frequency hopping signal modulation identification based on time-frequency characteristics. *J. Terahertz Sci. Electron. Inf.* **2022**, *20*, 40–46.
4. Panagiotou, P.; Anastasopoulos, A.; Polydoros, A. Likelihood ratio tests for modulation classification. In Proceedings of the MILCOM 2000 Proceedings. 21st Century Military Communications. Architectures and Technologies for Information Superiority (Cat. No. 00CH37155), Los Angeles, CA, USA, 22–25 October 2000; pp. 670–674.
5. Zhao, Y.; Jiang, H.; Qin, Y.; Xie, H.; Wu, Y.; Liu, S.; Zhou, Z.; Xia, J.; Zhou, F. Preserving minority structures in graph sampling. *IEEE Trans. Vis. Comput. Graph.* **2020**, *27*, 1698–1708. [CrossRef] [PubMed]
6. Abdel-Moneim, M.A.; Al-Makhlasawy, R.M.; Abdel-Salam Bauomy, N.; El-Rabaie, E.S.M.; El-Shafai, W.; Farghal, A.E.; Abd El-Samie, F.E. An efficient modulation classification method using signal constellation diagrams with convolutional neural networks, Gabor filtering, and thresholding. *Trans. Emerg. Telecommun. Technol.* **2022**, *33*, e4459. [CrossRef]
7. Li, H.G.; Guo, Y.; Sui, P. Convolutional neural network frequency hopping modulation identification based on time-frequency features. *J. Zhejiang Univ. (Eng. Ed.)* **1945**.
8. Liang, Y.X.; Tan, J.J.; Dusit, N. Research overview of smart wireless communication technology. *J. Commun.* **2020**, *41*, 1–17.
9. Szegedy, C.; Zaremba, W.; Sutskever, I. Intriguing properties of neural networks. In Proceedings of the 2nd International Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2013; pp. 1–10.
10. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
11. Kurakin, A.; Goodfellow, I.J.; Bengio, S. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*; Chapman and Hall/CRC: London, England, 2018; pp. 99–112.
12. Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; Li, J. Boosting adversarial attacks with momentum. In Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9185–9193.
13. Madry, A.; Makelov, A.; Schmidt, L. Towards Deep Learning Models Resistant to Adversarial Attacks. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018. Available online: https://openreview.net/forum?id=rJzIBfZAb (accessed on 7 December 2023).
14. O'Shea, T.J.; Corgan, J.; Clancy, T.C. Convolutional radio modulation recognition networks. In Proceedings of the Engineering Applications of Neural Networks: 17th International Conference, EANN 2016, Aberdeen, UK, 2–5 September 2016; pp. 213–226.
15. Ali, A.; Yang, Y.F.; Liu, S. Automatic modulation classification of digital modulation signals with stacked autoencoders. *Digit. Signal Process.* **2017**, *71*, 108–116. [CrossRef]
16. Xie, W.; Hu, S.; Yu, C.; Zhu, P.; Peng, X.; Ouyang, J. Deep learning in digital modulation recognition using high order cumulants. *IEEE Access* **2019**, *7*, 63760–63766. [CrossRef]
17. Zhan, J.M.; Zhao, Z.J. Modulation recognition algorithm for conventional modulation signals and spread spectrum signals. *Signal Process.* **2020**, *36*, 511–519.
18. Li, H.G.; Guo, Y.; Sui, P. Frequency hopping modulation mode identification based on time-frequency energy spectrum texture characteristics. *J. Commun.* **2019**, *40*, 20–29.
19. Sadeghi, M.; Larsson, E.G. Adversarial attacks on deep-learning based radio signal classification. *IEEE Wirel. Commun. Lett.* **2018**, *8*, 213–216. [CrossRef]
20. Zhao, H.; Lin, Y.; Gao, S.; Yu, S. Evaluating and improving adversarial attacks on DNN-based modulation recognition. In Proceedings of the GLOBECOM 2020-2020 IEEE Global Communications Conference, Taipei, Taiwan, 7–11 December 2020; pp. 1–5.
21. Lin, Y.; Zhao, H.; Ma, X.; Tu, Y.; Wang, M. Adversarial attacks in modulation recognition with convolutional neural networks. *IEEE Trans. Reliab.* **2020**, *70*, 389–401. [CrossRef]
22. Cichocki, A.; Amari, S. *Adaptive blind signal and image processing: Learning algorithms and applications*; John Wiley & Sons: Hoboken, NJ, USA, 2002.
23. Coviello, C.M.; Yoon, P.A.; Sibul, L.H. Source separation and tracking for time-varying systems. *IEEE Trans. Aerosp. Electron. Syst.* **2008**, *44*, 1198–1214. [CrossRef]
24. Ou, S.F.; Gao, Y.; Zhao, X.H. Adaptive whitening algorithm for variable factors based on random gradients. *J. Autom.* **2012**, *38*, 1370–1374.
25. Wu, L.; Zhu, Z.; Tai, C. Understanding and enhancing the transferability of adversarial examples. *arXiv* **2018**, arXiv:1802.09707.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.