AIMS Energy

https://www.aimspress.com/journal/energy

AIMS Energy, 12(6): 1271–1293.
DOI: 10.3934/energy.2024058
Received: 17 July 2024
Revised: 20 November 2024
Accepted: 04 December 2024
Published: 17 December 2024

Research article

False data injection attack sample generation using an adversarial attention-diffusion model in smart grids

Kunzhan Li[1], Fengyong Li[1,2], Baonan Wang[1,2] and Meijing Shan[3,*]

[1] College of Computer Science and Technology, Shanghai University of Electric Power, Shanghai 201306, China

[2] Engineering Research Center of Offshore Wind Technology Ministry of Education (Shanghai University of Electric Power), Yangpu District, Shanghai 200090, China

[3] Institute of Information Science and Technology, East China University of Political Science and Law, Shanghai 200042, China

* **Correspondence:** Email: shanmeijing@ecupl.edu.cn; Tel: +86 13148019097.

Abstract: A false data injection attack (FDIA) indicates that attackers mislead system decisions by inputting false or tampered data into the system, which seriously threatens the security of power cyber-physical systems. Considering the scarcity of FDIA attack samples, the traditional FDIA detection models based on neural networks are always limited in their detection capabilities due to imbalanced training samples. To address this problem, this paper proposes an efficient FDIA attack sample generation method by an adversarial attention-diffusion model. The proposed scheme consists of a diffusion model and a GAN model with an attention mechanism (ATTGAN). First, the forward diffusion of the diffusion model was used to add noise to the real data while injecting the attack vector. Then, the ATTGAN model was trained to effectively focus on the information of power grid measurements and topological nodes, while weakening irrelevant information. In the reverse diffusion process, the trained ATTGAN model was combined to predict the noise, and it was further iterated forward step by step and denoised in this process. Finally, a large number of efficient FDIA attack samples can were generated. Extensive experiments have been carried out on IEEE 14, IEEE 39, and IEEE 118 bus systems. The experimental results indicate that the generated attack samples outperform existing state-of-the-art schemes in terms of evasion detection capability, robustness, and attack strength.

Keywords: FDIA; smart grid; diffusion model; attention mechanism; cyber-physical system

1. Introduction

As a key component of modern power systems, smart grids are designed to achieve efficient, flexible, and safe energy distribution. By integrating advanced sensor technology, communication networks, and intelligent control strategies, smart grids have gradually evolved into a highly integrated cyber-physical system (CPS) [1], and can respond to various operating conditions and external events in real time, optimize resource allocation, and improve the system's adaptability and anti-interference capabilities. However, this progress has also brought new security challenges, especially in the face of increasingly complex cyber attacks. A false data injection attack (FDIA) is one of the main security threats facing smart grids. By manipulating the measurement data of the power grid, attackers can mislead the system state estimation and decision-making process, thereby posing a serious threat to the stability and security of the power grid [2]. The stealth and complexity of such attacks make it difficult for traditional security measures to effectively deal with it. It is urgent to develop new defense mechanisms to ensure the data integrity and system security of smart grids.

In recent years, large-scale grid security incidents in which criminals use data tampering mechanisms to attack occur frequently. Take the attack on Ukrenergo as an example. From 2015 to 2016, Ukrenergo suffered two cyber attacks. The attackers injected false data into the monitoring data acquisition system and deleted the original data, causing huge economic losses and social unrest [3]. This has also attracted the attention of many researchers to FDIA attacks. Since false data can be cleverly designed by attackers to evade detection by detection mechanisms, the purpose of intrusion can be achieved, causing irreversible damage to the power system. Therefore, society has carried out a lot of research work to identify and defend against an FDIA. For example, Kumar et al. [4] explored the effectiveness of machine learning algorithms in detecting false data injection attacks (FDIA) in critical infrastructure. Cheng et al. [5] introduced a highly discriminative FDIA detector called k-minimum residual similarity (kSRS) detection. This method achieves a high detection rate and low false alarm rate for an FDIA by characterizing the statistical consistency of measurement residuals in AC state estimation. Qu et al. [6] proposed a novel approach to improve the accuracy of DDIA detection and localization by synthesizing topological correlations in non-Euclidean spatial attributes of grid data. However, the above research is based on traditional machine learning to design FDIA detection solutions, which are limited by the complexity of manual feature extraction, dimensionality disaster, and insufficient generalization ability when detecting FDIA attacks. Therefore, traditional FDIA detection methods have gradually lost their effectiveness in defending against these network attacks.

With the update and iteration of technology, some FDIA detection methods based on machine learning and deep learning have also been developed. In order to correctly detect false data injection attacks in power grids, Fand et al. [7] proposed a one-dimensional convolutional neural network (1DCNN) and long short-term memory (LSTM) network multi-channel fusion network model based on the residual neural network (ResNet) structure, referred to as the channel-fused Res-CNN-LSTM network model. In order to make the detection results more credible, Su et al. [8] proposed an interpretable deep learning FDIA detection method, namely the dual attention multi-head graph attention network (DAMGAT). Takiddin et al. [9] proposed an anomaly detector based on generalized graph neural networks (GNN), which is robust to an FDIA and data poisoning. Li et al. [10] proposed a false data detection method based on graph neural networks, extracting the spatial features of power

grid topology information and operation data through gated graph neural networks (GGNN). Qu et al. [11] proposed an improved adaptive Kalman filter (AKF) and convolutional neural network (CNN) hybrid detection method for power information physical system (CPS) FDIA. Wang et al. [12] proposed a deep learning-based location detection architecture (DLLD) to detect the exact location of an FDIA in real time. Xie et al. [13] proposed a Bayesian deep learning-based method to detect network attacks and maintain the security of smart grids. Although these methods show great advantages in FDIA attack detection, they also have great limitations. Most of the methods rely on large-scale training attack samples, but in the actual scenario, FDIA attack samples are difficult to obtain, and samples are scarce, resulting in extremely unbalanced sample data sets. Although they have strong detection performance, if there are not enough attack samples for model training, the efficiency of the FDIA attack detection model in smart grids will also be reduced. Therefore, the construction of FDIA attack samples and the acquisition of balanced data sets become the first problems to be solved.

Several researchers have conducted extensive research on the construction of FDIA attack vectors. Yan et al. [14] proposed the DoS-WGAN architecture, which uses a Wasserstein generative adversarial network (WGAN) with gradient penalty technology to evade network traffic classifiers. WGAN is used to automatically synthesize attack vectors that can be disguised as normal network traffic in the case of network attacks and bypass the detection of network intrusion detection systems. To address the problem of highly unbalanced data samples, Kumar et al. [15] proposed a Wasserstein conditional generative adversarial network (WCGAN) combined with an XGBoost classifier. The gradient penalty is used together with WCGAN for stable learning of the model. Tian et al. [16] introduced the joint adversarial examples and FDIA (AFDIA) to explore various attack scenarios on power system state estimation. Considering that disturbances directly added to the measurement are likely to be detected by the BDD, a method of adding disturbances to the state variables is proposed to ensure that the attack is invisible to the BDD. Bhattacharjee et al. [17] proposed a scheme for constructing invisible attack vectors based on independent component analysis to identify mixing matrices. Wu et al. [18] proposed a new FDIA sample generation method to construct large-scale attack samples by introducing a mixed Laplace model that can accurately fit the distribution of data changes. Li et al. [19] designed an efficient adversarial model for generating FDIA attack samples at the edge of public and private networks that can effectively bypass detection models and threaten grid systems. Through interactive adversarial learning, efficient FDIA attack samples can be continuously generated. Although the existing sample generation method can appropriately expand the FDIA attack sample library, its generation method is too simple, resulting in low concealment and weak attack capability of the generated samples, which is not conducive to the training of the detection model and the improvement of performance.

With the rise of generative AI technology, a new perspective has been brought to the detection of various network attacks. Among them, the diffusion model has gradually become a new and most advanced model among the deep generation models in recent years, showing superior performance in the research of image generation and multi-modal generation. Compared with traditional deep learning methods, it can generate new attack samples with better concealment and attack capabilities by learning the distribution of data, which provides a new way to understand the strategies that attackers may adopt. Therefore, in response to the above problems, considering the diversity, concealment, and scalability of the data generated by the diffusion model [20] and its powerful generation ability, this paper presents a method to generate FDIA attack samples using an adversarial

attention diffusion model. The proposed AttDiff model consists of a diffusion model and a GAN (ATTGAN) model with an attention mechanism. First, as the time step increases, in the forward diffusion process, the original input data is gradually denoised, the attack vector is injected, and the data state corresponding to each time step and the added noise are recorded. The recorded data is then used to train generative adversarial network (ATTGAN) models with attention mechanisms. The purpose is to allow the ATTGAN model to effectively focus on the information of power grid measurements and topological nodes through adversarial training, while weakening irrelevant information and fully learning the characteristics of real sample data. Afterward, in the reverse diffusion process, the noise-added data processed in the forward diffusion process should be denoised from back to front in combination with the size of the time step, and the trained ATTGAN model should be combined to predict the noise value to be removed. In this process, the purpose is to build the attack sample data closest to the real sample from the noise. The FDIA sample generated by the method in this paper is more diversified, more hidden, and can effectively simulate the behavior of the real attacker, and is less likely to be detected by the detector.

Comparing with the existing schemes, we make the following novel contributions in FDIA attack sample generation:

- We propose a novel and efficient sample generation method. Our proposed scheme introduces an adversarial attention-diffusion model to generate a large number of attack samples, which can solve the problem of scarce attack samples in FDIA detection. According to the investigation, this is the first time that the diffusion model has been applied to the study of FDIA attack sample generation.
- We use the forward diffusion of the diffusion model to add noise to the real data while injecting the attack vector. Subsequently, the ATTGAN model is trained to effectively focus on the information of power grid measurements and topological nodes, while weakening irrelevant information. In the reverse diffusion process, we combine the trained ATTGAN model to predict the noise, and gradually iterate forward step by step and denoise in this process. A large number of efficient FDIA attack samples can be generated.
- Comprehensive experiments are performed with classical data sets. The experimental results demonstrate that the proposed method can significantly improve the overall performance against FDIA detection, and outperform the existing FDIA attack sample construction methods in terms of attack strength, concealment, and decline of FDIA detection accuracy.

The rest of this paper is organized as follows. In Section 2, related works on FDIAs and diffusion models are introduced. In Section 3, the framework of the proposed attack sample generation method based on the adversarial attention diffusion model and the specific structure of the AttDiff model are introduced in detail. In Section 4, a large number of experiments and corresponding comparative experiments are carried out, and the experimental results are presented. Finally, in Section 5, the work of this paper is summarized and prospected.

2. Related work

2.1. False data injection attack

A fake data injection attack is a kind of network attack against CPS. Its core is to construct and inject fake data to bypass system detection, so as to achieve the purpose of attack. This type of attack is particularly common in critical infrastructure such as smart grids and can have a serious impact on the operational stability, supply reliability, and data accuracy of the power system.

The core principle of constructing an FDIA is to manipulate the key measurement data in the AC power system in order to change the state estimation results of the system. By tampering with the active power on any bus node, an attacker generates a set of attack vectors designed to interfere with the normal operation of the system. These attack vectors are injected into the measured data of the system, which causes the calculated state estimation to deviate from the real operating state of the power grid. In this way, attackers are able to mislead the system's decision-making process for their potential sabotage or manipulation purposes. In DC power flow, the relationship between system state x and measurement data z can be expressed as:

$$z = Hx + e \tag{1}$$

where $z = \{z_1, z_2, \ldots, z_I\}$ represents the measurement data, $x = \{x_1, x_2, \ldots, x_J\}$ represents the state vector, $e = \{e_1, e_2, \ldots, e_I\}$ represents the measurement error vector, I and J represent the number of measurements and the number of state data, respectively, and H represents the measurement Jacobian matrix, which is a matrix representing the grid topology information. The measurement data includes the flow and injection power of different buses.

For DC state estimation, assuming that the per unit voltage of each node in the system is 1, ignoring the influence of line resistance and ground branch, the power between bus i and bus j is:

$$P_{ij} = \frac{(\theta_i - \theta_j)}{x_{ij}} \tag{2}$$

where x_{ij} represents the reactance of the transmission line between bus i and bus j. θ_i and θ_j are the voltage phase angles of bus i and bus j. The minimized objective function is estimated by linear weighted least squares:

$$\min F(x) = (z - Hx)^T R^{-1} (z - Hx) \tag{3}$$

where R is the error covariance matrix of the measurement, and $\hat{x}$ of the minimized objective function is:

$$\hat{x} = (H^T R^{-1} H)^{-1} H^T R^{-1} z \tag{4}$$

The principle of residual detection has been widely used in the traditional bad data detection mechanism (BDD). Residual r is the difference between the measurement value z and the measurement estimate $\hat{z} = H\hat{x}$, i.e.:

$$r = z - \hat{z} \tag{5}$$

BDD can then detect bad data by comparing the Euclidean norm of the measured residuals to a predefined threshold τ, as follows:

$$\|r\| \leq \tau \tag{6}$$

In order to make the injected attack vector invisible, we assume that the measurement error e follows an ideal normal distribution, and use $a = [a_1, a_2, ..., a_m]^T$ to represent the FDIA vector injected by the attacker into the measured value. Then the actual measurement data is $z_a = z + a$, and the error vector of the state variable caused by the FDIA is $c = [c_1, c_2, ..., c_n]^T$. At this time, the estimated state variable $x_a = \hat{x} + c$, and the residual error after attack can be expressed as:

$$\begin{aligned} \|r_a\| &= \|z_a - Hx_a\| \\ &= \|z + a - H(\hat{x} + c)\| \\ &= \|z - H\hat{x} + a - Hc\| \end{aligned} \tag{7}$$

When $a = Hc$, the following formula holds:

$$\begin{aligned} \|r_a\| &= \|z_a - Hx_a\| \\ &= \|z - H\hat{x}\| = \|r\| \ \leq \ \tau \end{aligned} \tag{8}$$

It can be seen that when the above conditions are met, an FDIA can successfully bypass BDD, causing changes and losses to the state estimation of the power system. However, it is worth noting that the basic assumption of an FDIA is that both the attacker and the defender (that is, the power system operator) have complete information about the parameters and topology of the power system, that is, H in Eq (1), and it is also necessary to find highly sparse attack vectors that meet certain conditions to launch attacks. For this, the cost is high.

2.2. Diffusion model

Diffusion models have surpassed the previous leading generative adversarial network (GAN) technology in the field of image creation and have shown great application potential in multiple fields such as computer vision, natural language processing, signal processing, multimodal learning, molecular structure modeling, time series analysis, and adversarial sample purification. In the field of computer vision, diffusion models are used to perform various image restoration tasks, such as image resolution enhancement, image painting, and image translation [21–23]. In natural language processing, diffusion models can generate character level text through a technique called the discrete denoising diffusion probability model (D3PM) [24]. In addition, in terms of time series data generation, diffusion models are also used to fill in the blanks in time series data, as shown in the literature [25, 26]. In the field of data generation, due to the similar structure of time series data and intrusion detection data, both of them are one-dimensional data composed of multiple features, which stimulates the idea of generating FDIA attack sample data in this experiment. According to the survey, there is no literature that uses diffusion models for FDIA attack sample generation to generate a balanced data set.

The existing sample generation methods are too simple, resulting in low concealment and weak attack ability of the generated samples, which is not conducive to the training and performance improvement of the detection model. The use of a diffusion model can construct more robust, more threatening, and more hidden attack samples, so that the training of the detection model can be further optimized and improved, and finally achieve the purpose of enhancing the defense capability of the power system, making its operation safer and more stable.

3. Proposed method

3.1. Proposed FDIA sample generation framework

The proposed framework mainly consists of three parts: the data processing module, AttDiff model, and data generation module. In the data processing module, various interactive data from the virtual power grid and power plant are first collected and processed, and then the flow calculation is performed on these data. Then, the measurement data in the power system is further simulated through simulation. Then, the processed measurement data is input into the AttDiff model and the sample data is normalized to adapt to network training. After initializing the model parameters, the forward diffusion process of the diffusion model is entered. During this process, as the time step increases, noise and attack vectors are gradually added to the original data, and the data in the forward diffusion process is recorded. After this process is completed, the ATTGAN model is trained with the recorded data. In addition, in backpropagation, noise prediction is combined with a trained ATTGAN model, and noise is iteratively removed step by step with time steps. Finally, new attack samples are generated based on the features of the original data to obtain a balanced data set. The complete framework of the proposed attack sample generation method is shown in Figure 1.

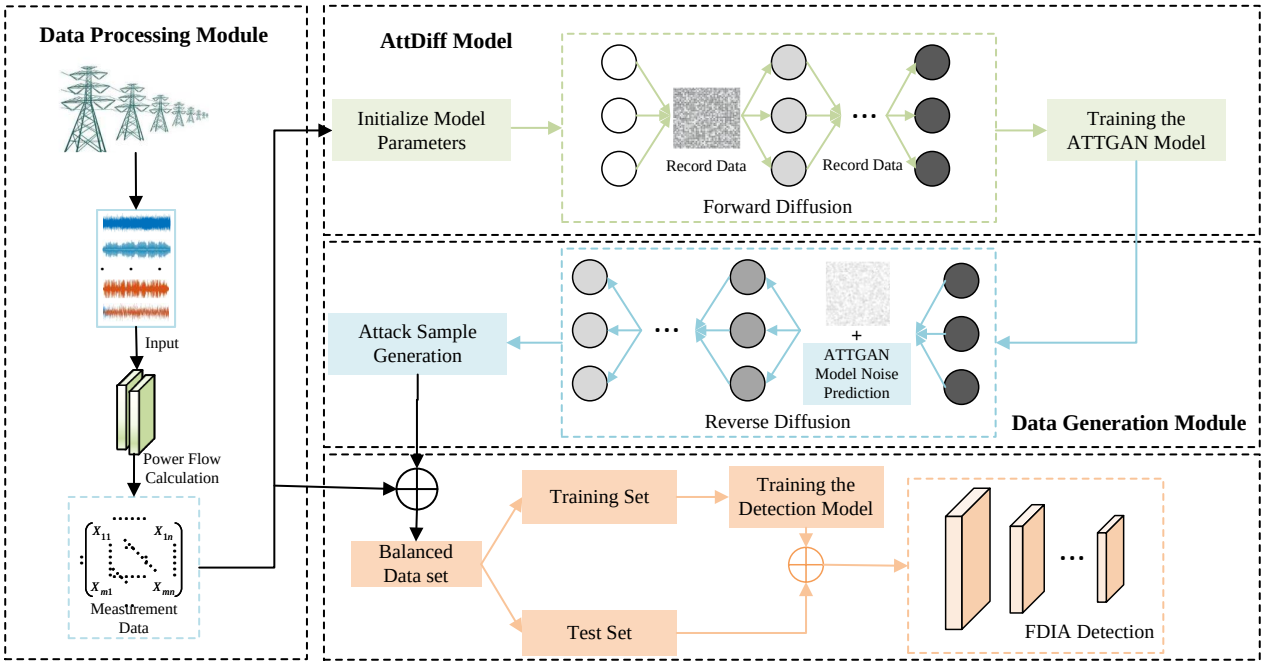

Figure 1. Proposed FDIA attack sample generation framework.

3.2. Data processing module

In a cyber-physical system, data comes from a variety of sources and may include multiple sensors. These data have significant differences in structure, accuracy, transmission delay, and update frequency, so it is necessary to properly preprocess these diverse data. Each column of data in the data set represents a feature of the sensor, and different features correspond to different units of measurement and magnitude, which may lead to the uneven influence of certain attributes in data

analysis and model training. In order to balance this dimensional difference and ensure that each attribute is given equal consideration during analysis and training, the minimum-maximum normalization technique is used to linearly transform the data into a specified interval while preserving the original relative order and distribution characteristics of the data. The specific normalization formula is as follows:

$$y_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{9}$$

where x_i is the eigenvalue of the original data set, y_i is the normalized data value, $\min(x)$ is the minimum eigenvalue, and $\max(x)$ is the maximum eigenvalue. In this way, each feature value x_i is converted into a range of 0 to 1, which helps ensure that different features have the same importance in model training, and reduces the impact of dimension and magnitude.

3.3. Attention-diffusion model

In this section, the diffusion model is combined with the ATTGAN model to complete the forward diffusion and reverse diffusion process of the diffusion model, and finally generate more hidden FDIA attack samples. In the reverse diffusion process, the ATTGAN model is used for noise prediction. The specific structure of the AttDiff model is shown in Figure 2.

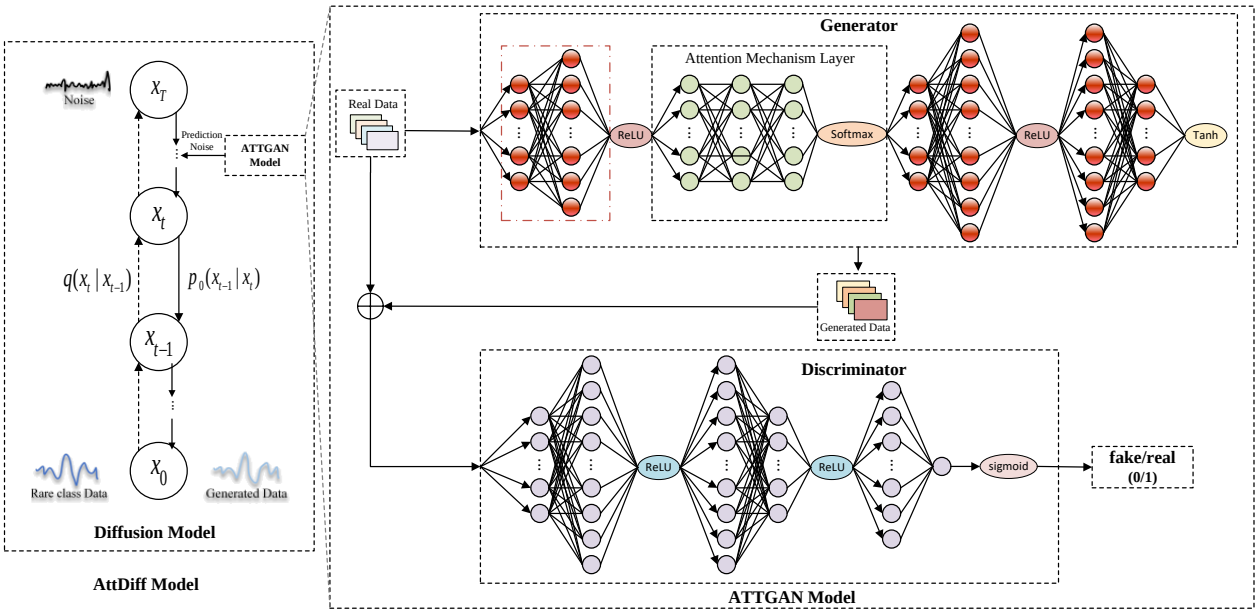

Figure 2. Detailed structure diagram of the proposed attention-diffusion model, where the left indicates the diffusion model, while the right is the attention-based GAN model.

Diffusion models [20] usually contain two key stages: the forward diffusion process and reverse diffusion process. In this model, it is necessary to select a data point from the real data distribution as the starting point $x_0 \sim q(x)$, and gradually add Gaussian noise and attack vectors to x_0 as the time step increases in the forward diffusion process, we finally turn x_0 into standard Gaussian noise $x_T = N(0, 1)$. For each forward step $t \in [1, 2, ..., T]$, the noise disturbance is controlled by the following formula:

$$q(x_t|x_{t-1}) = N(x_t;\ \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{10}$$

where $\{\beta_t \in (0,1)\}_{t=1}^{T}$ is a different variance scale. After the forward diffusion ends, the reverse diffusion process gradually reconstructs the original data x_0 by sampling x_T and learning the diffusion model p_θ.

In addition, the parameter t plays a crucial role in the diffusion model, and can directly affect its performance. Therefore, choosing the right t value in our study is crucial to balance the concealment and quality of the generated attack samples. If t is too large, the model may overfit the noise rather than the true distribution of the data, causing the generated attack sample to lose its correlation to the real data. Conversely, if t is too small, the model may not be trained enough to efficiently generate high-quality attack samples. In the experimental part of this study, we find the optimal t value by adjusting the value of t and observing its effect on the quality of the generated sample to ensure that the generated attack sample is both highly concealed and able to effectively challenge the detection model.

3.3.1. Forward diffusion process

For a specific data point x_0 extracted from the actual data distribution $q(x)$, this experiment constructs a forward diffusion process. The process involves gradually adding a small amount of Gaussian noise to the sample in T steps along with varying degrees of attack vectors to generate a series of noisy data points $x_1, ..., x_T$. The noise variance at each step is controlled by $\{\beta_t \in (0,1)\}_{t=1}^{T}$. As the step size t increases, the original data x_0 will have fewer and fewer recognizable features. When $T \to \infty$, the final data point x_T will tend to be an isotropic Gaussian distribution.

In addition, in the forward diffusion process, this experiment extracted data points from normal sample data, and recorded each time step corresponding to the added random noise and the data value before the current state was noised. The purpose is to train the ATTGAN model with the recorded data after the forward diffusion is over. In the process of inverse diffusion, the added noise corresponding to each time step can be predicted and denoised more accurately, so that the attack sample is closer to the real original normal data, and the invisibility of the attack sample is increased.

One benefit of the above process is that we can use the re-parameterization technique. We can sample x_t in closed form at any time point t. Here we define $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i, i.e.$, which means x_t can be expressed as $x_t = \sqrt{\bar{\alpha}} x_0 + \sqrt{1 - \bar{\alpha}_t}\varepsilon$. In this way, we can directly use the initial data point x_0 to represent x_t.

$$q(x_t|x_0) = N(x_t;\ \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)I) \tag{11}$$

3.3.2. Reverse diffusion process

By reversing the above forward diffusion process and sampling $q(x_{t-1}|x_t)$ at the same time, the real sample data can be recovered from the Gaussian noise input $x_T \sim N(0, I)$. If the variance β_t at each step is small enough, then $q(x_{t-1}|x_t)$ will also take on the characteristics of a Gaussian distribution. However, it is important to note that directly estimating $q(x_{t-1}|x_t)$ is challenging because it often relies on the entire data set. Therefore, the ATTGAN model trained after the forward diffusion process mentioned above is used in this experiment to predict a noise value by inputting the x_t corresponding to the current time step to approximate these conditional probabilities, so as to perform the denoising process of reverse diffusion. According to the nature of the Markov chain, the reverse diffusion process at the current time step t depends only on its previous time step $t - 1$, which means that we have the following formula:

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \sum\nolimits_\theta (x_t, t)) \tag{12}$$

From this, we can know the overall process of generating samples by the diffusion model. The sample generation process of the diffusion model involves gradually adding noise to the data from front to back until the data is completely transformed into standard Gaussian noise. This process is called forward diffusion. Subsequently, in the reverse diffusion process, by learning how to reverse the operation of adding noise, the noise is gradually removed from the back to the front to restore the original sample data. In this way, new data points can be generated from a limited sample data set, and these newly generated sample data are exactly the data needed at the moment.

3.3.3. ATTGAN model

The ATTGAN model is mainly composed of generators and discriminators. In this experiment, in order to learn different features of the input data more effectively and improve the sensitivity of the model to specific parts of the input data, an attention mechanism layer is added to the generator, thereby improving the performance of the model.

The model's generator consists of six fully connected layers and is equipped with an appropriate activation function at each layer, such as LeakyReLU, to introduce nonlinear properties. In addition, an attention layer is introduced into the structure of the generator, which consists of three fully connected layers and focuses on the salient parts of the input features through a learnable weight vector w. The output of the attention mechanism is combined with the raw output of the generator to enhance the detail and quality of the data. Note that the design of the attention layer is based on the following formula:

$$a = soft\max(W_a \cdot h + b_a) \tag{13}$$

where h is the output of the previous layer, W_a and b_a are the learnable weights and biases, and softmax functions are used to generate a normalized weight vector a.

The discriminator corresponds to the generator, again consisting of six fully connected layers, each of which is followed by an activation function. The goal of a discriminator is to distinguish between real data and data generated by a generator.

The training of the model follows the standard adversarial training process. In the training process, this experiment adopts binary cross entropy loss (BCELoss) to optimize the model. Generators and discriminators are optimized with the following loss functions:

$$\begin{aligned} L_G &= -E_{z\sim p(z)}[\log(D(G(z)))] \\ L_D &= E_{x\sim p_{data}(x)}[\log(D(x))] + E_{z\sim p(z)}[\log(1 - D(G(z)))] \end{aligned} \tag{14}$$

where L_G and L_D are the loss functions of the generator and discriminator, respectively, G and D represent the generator and discriminator, respectively, z is the noise vector sampled from the prior distribution $p(z)$, and x is the data point from the real data distribution $p_{data}(x)$. In this experiment, an Adam optimizer was used to train the model, the learning rate was 0.0001, and the batch size was 64. In the training process, the quality of the generated data and the performance of the discriminator are monitored in real time, and the hyperparameters are carefully adjusted to ensure the effectiveness and stability of the model.

3.4. Data generation module

After the forward diffusion of the diffusion model ends, the reverse diffusion process must be followed. Before the reverse diffusion, the ATTGAN model needs to be trained with data recorded

during the forward diffusion process described above, which is used for noise prediction in the reverse diffusion. After the model is trained, it officially enters the process of reverse diffusion. Combined with the trained ATTGAN model, it iteratively denoises step by step from the back, and the noise value removed is predicted by the ATTGAN model. In the end, large-scale samples close to the real data will be generated, and these samples are the required attack samples, so that the rare sample library can be expanded to form a balanced data set. Finally, the purpose of improving the detection performance of the detector can be achieved, and then the defense capability of the power system can be enhanced. The overall process of FDIA attack sample generation based on the AttDiff model is shown in Algorithm 1.

Algorithm 1: FDIA Sample Generation Process based on the AttDiff Model

Input:	Original measurement data X, number of time steps T;
Output:	Generated FDIA attack samples X_{FDIA};
1	Initialization the number of time steps T, batch size, and other related parameters;
2	Forward Diffusion Process:
3	for each time step t =1 to T do:
4	Generate noise $\varepsilon(t) \sim N(0, \sigma^2)$ and attack vectors of different intensities a;
5	Perform iterative calculation of forward diffusion according to Eq 11;
6	Record noisy data $X_{noise}(t) = X + \varepsilon(t) + a$ and current time step t;
7	Store the current $X_{noise}(t)$ and $\varepsilon(t)$ corresponding to t for subsequent ATTGAN model training;
8	end for
9	Train ATTGAN model using data recorded during forward diffusion process;
10	Reverse Diffusion Process:
11	for each time step $t = T$ to 1 do:
12	Predict noise $\varepsilon_{hat}(t)$ for $X_{noise}(t)$ using the trained ATTGAN model;
13	Denoise the data $X_{denoised}(t) = X_{nosie}(t) - \varepsilon_{hat}(t)$;
14	Perform iterative calculation of reverse diffusion according to formula 12;
15	Update the data state $X_{denoised}(t)$ for the next iteration;
16	end for
17	Output the final denoised data X_{FDIA} after the last iteration of the reverse diffusion process.
18	Return the generated FDIA attack samples X_{FDIA}.

Among them, steps 1–8 are the forward diffusion stage of the diffusion model. According to the size of the time step, noise and attack vectors are added to the original data by iterating from front to back, and the data state and added noise corresponding to each time step are recorded. Step 9 is the training stage of the ATTGAN model. The data recorded in the process of forward diffusion is used for training. The purpose is to enable the ATTGAN model to effectively focus on power grid measurement values and topological node information, while weakening irrelevant information and fully learning

the characteristics of real sample data, so as to play a better role in reverse diffusion. Steps 10–18 are the reverse diffusion stage of the diffusion model. According to the size of the time step, the noisy data is denoised by iterating from back to front. In this process, the diffusion model is combined with the trained ATTGAN model to predict the noise value to be removed, so as to build the attack sample data that is closest to the real sample. Finally, a large number of FDIA attack samples with high quality and high concealment can be obtained.

4. Experimental results and discussions

4.1. Experimental setup

In this study, three well-known power system test platforms, namely the IEEE 14 bus, IEEE 39 bus, and IEEE 118 bus systems [27, 28], were used to evaluate and verify the performance of the proposed method. These bus systems are widely used benchmark models in power system analysis and research. They provide a simplified representation of the power system network, including bus configuration, electrical parameters, generator and load data, etc. Real load data from January 1, 2020, to May 1, 2022, were collected from the New York Independent System Operator (NYISO) as test data. These data cover a time range of more than two years, including seasonal changes throughout the year and load patterns under different weather conditions. By applying these actual load data to the above bus system, we can simulate the measurement data of the power system, and further obtain the load information of each bus state variable and each node, providing real data support for model evaluation.

In this experiment, the above test data are all the data under normal measurement conditions. In order to simulate the real attack behavior, the proposed method is used to attack the test data. A traditional FDIA is a classical method with a certain degree of invisibility and flexibility. The proposed method in this experiment combines the traditional FDIA method to construct an attack sample generation method based on adversarial attention diffusion. By combining the proposed adversarial generation network, the attack sample mixed with normal data features can be generated through reverse diffusion. In order to test the influence of different attack levels, the attack vector is divided into three different attack strengths according to the method in [29]: a) **Weak attack**: the ratio of the mean power injection deviation in c to x is less than 10%; b) **Strong attack**: the ratio of the mean power injection deviation in c to x is greater than 30%; c) **Medium attack**: FDIA samples that do not fall under either of the above attacks.

Since the three bus systems all contain 15,000 pieces of data, the data sets are divided into training sets, verification sets, and test sets according to the ratio of 5:3:2, and experimental comparison tests were carried out under different attack intensity scenarios.

In addition, the experiment uses four indicators, namely precision, recall, accuracy, and F_1 score, as the evaluation criteria for the output results, which are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{17}$$

$$F_1 = \frac{2 \times \mathrm{Pr}\,ecision \times \mathrm{Re}call}{\mathrm{Pr}\,ecision + \mathrm{Re}call} \tag{18}$$

where *true positive* (TP) indicates the amount of incorrect data correctly detected, *true negative* (TN) indicates the amount of normal data detected as normal, *false positive* (FP) indicates the amount of normal data incorrectly detected as incorrect, and *false negative* (FN) indicates the amount of incorrect data incorrectly detected as normal. The higher the accuracy and precision, the worse the concealment of the attack sample and the easier it is to be detected.

4.2. Effectiveness verification for the proposed model

All experimental simulations in this study were performed on devices with an Intel Core i7-11390H CPU, NVIDIA GeForce MX450, and 16 GB RAM. MATPOWER is used to calculate the power flow and estimate the state of the data in MATLAB, and the proposed AttDiff model is used to generate FDIA attack samples in Python. The noise error of power flow calculation simulation measurement data is set to 0.25 and Gaussian noise with mean zero and standard deviation 1 is adopted. The time step of the diffusion model is set to 4.

In this study, large-scale experiments are carried out on IEEE 14, IEEE 39, and IEEE 118 bus systems to verify the performance of the attack samples generated by the proposed method. Two advanced deep learning detection models are used to detect the generated samples, namely the CNN-based detection model [12] and the LSTM-based detection model [30]. The three types of high, medium, and low attack intensity are subdivided into 9 attack intensities with low intensity is subdivided into 2%, 5%, and 10%, medium intensity is subdivided into 15%, 20%, and 25%, and high intensity is subdivided into 30%, 40%, and 50% for comparison. The corresponding experimental results are shown in Tables 1–6, in which Tables 1 and 2 are the test results of the IEEE 14 bus system, Tables 3 and 4 are the test results of the IEEE 39 bus system, and Tables 5 and 6 are the test results of the IEEE 118 bus system.

Table 1. Performance comparison of three attack sample generation schemes (traditional FDIA sample generation [8], mixed Laplacian model-based FDIA sample generation (LMM-FDIA) [18], and adversarial attention diffusion model-based FDIA sample generation (AttDiff)). This test uses a CNN-based FDIA attack detection model [12] on the IEEE 14 bus system.

Attack Level	Attack Strength	Traditional FDIA [8]			LMM-FDIA [18]			AttDiff-FDIA		
		Precision	Recall	F_1-Score	Precision	Recall	F_1-Score	Precision	Recall	F_1-Score
Weak	2%	0.6883	0.6992	0.6819	0.6740	0.6958	0.6730	0.6469	0.7548	0.6892
	5%	0.7289	0.7233	0.7130	0.6831	0.7482	0.7045	0.6664	0.7173	0.6755
	10%	0.7601	0.7829	0.7602	0.7079	0.6778	0.6768	0.6937	0.7017	0.6860
Medium	15%	0.8205	0.8101	0.8062	0.7351	0.6716	0.6867	0.7081	0.6757	0.6777
	20%	0.8469	0.8647	0.8483	0.7514	0.7095	0.7146	0.7129	0.6529	0.6661
	25%	0.8772	0.8817	0.8731	0.7656	0.6992	0.7159	0.7209	0.6941	0.6958
Strong	30%	0.9092	0.9111	0.9050	0.7857	0.7062	0.7314	0.7324	0.7014	0.7037
	40%	0.9223	0.9423	0.9289	0.7995	0.7286	0.7503	0.7561	0.7280	0.7297
	50%	0.9313	0.9490	0.9371	0.8117	0.7547	0.7702	0.7789	0.7156	0.7339

Table 2. Performance comparison of three attack sample generation schemes (traditional FDIA sample generation [8], mixed Laplacian model-based FDIA sample generation (LMM-FDIA) [18], and adversarial attention diffusion model-based FDIA sample generation (AttDiff)). This test uses an LSTM-based FDIA attack detection model [30] on the IEEE 14 bus system.

Attack Level	Attack Strength	Traditional FDIA [8]			LMM-FDIA [18]			AttDiff-FDIA		
		Precision	Recall	F_1-Score	Precision	Recall	F_1-Score	Precision	Recall	F_1-Score
Weak	2%	0.7038	0.7208	0.7117	0.6859	0.7404	0.7119	0.6669	0.7540	0.7064
	5%	0.7167	0.7437	0.7298	0.6927	0.7707	0.7293	0.6828	0.7456	0.7127
	10%	0.7671	0.7915	0.7602	0.7235	0.7583	0.7404	0.6972	0.7551	0.7248
Medium	15%	0.7958	0.8235	0.8062	0.7431	0.7351	0.7383	0.7203	0.6961	0.7060
	20%	0.8277	0.8976	0.8610	0.7547	0.6959	0.7240	0.7220	0.6695	0.6935
	25%	0.8679	0.9059	0.8864	0.7565	0.7253	0.7402	0.7334	0.6325	0.6787
Strong	30%	0.9075	0.9317	0.9194	0.7715	0.7086	0.7379	0.743	0.6395	0.6862
	40%	0.9484	0.9361	0.9421	0.8039	0.7286	0.7641	0.7651	0.6792	0.7190
	50%	0.9601	0.8864	0.9213	0.8248	0.6981	0.7558	0.7957	0.7144	0.7527

As can be seen from the table above, the attack sample generation method proposed in this experiment can obtain the lowest accuracy and lower F_1 score under different attack intensifications. When the CNN-based FDIA detection on the IEEE 14 bus system is used for low-intensity attacks of 5%, the proposed method has a precision reduction of 6.256% and 1.676% compared with the traditional FDIA [8] method and the hybrid Laplace method [18], and a score reduction of 3.744% and 2.896% in F_1, respectively. For medium-intensity attacks of 20%, the proposed method has a precision reduction of 13.4% and 3.854% compared with the traditional FDIA [8] method and the mixed Laplacian method [18], and a score reduction of 18.22% and 4.852% in F_1, respectively. For high-intensity attacks of 40%, the proposed method has a precision reduction of 16.63% and 4.344% compared with the traditional FDIA [8] method and hybrid Laplacian method [18], and a score reduction of 19.91% and 2.06% in F_1, respectively.

Table 3. Performance comparison of three attack sample generation schemes (traditional FDIA sample generation [8], mixed Laplacian model-based FDIA sample generation (LMM-FDIA) [18], and adversarial attention diffusion model-based FDIA sample generation (AttDiff)). This test uses a CNN-based FDIA attack detection model [12] on the IEEE 39 bus system.

Attack Level	Attack Strength	Traditional FDIA [8]			LMM-FDIA [18]			AttDiff-FDIA		
		Precision	Recall	F_1-Score	Precision	Recall	F_1-Score	Precision	Recall	F_1-Score
Weak	2%	0.7187	0.7181	0.7078	0.7199	0.7223	0.7074	0.6934	0.6977	0.6815
	5%	0.7664	0.7799	0.7650	0.7200	0.7521	0.7242	0.7042	0.7200	0.6990
	10%	0.8594	0.8605	0.8549	0.7404	0.7263	0.7215	0.7208	0.7177	0.7074
Medium	15%	0.9094	0.9250	0.9141	0.7549	0.7109	0.7159	0.7119	0.7214	0.7034
	20%	0.9507	0.9437	0.9452	0.7672	0.7272	0.7342	0.7205	0.7132	0.7052
	25%	0.9737	0.9609	0.9660	0.7743	0.7122	0.7285	0.7309	0.7130	0.7082
Strong	30%	0.9801	0.9789	0.9787	0.7850	0.7346	0.7485	0.7439	0.7227	0.7211
	40%	0.9922	0.9817	0.9864	0.7836	0.7219	0.7385	0.7629	0.7115	0.7247
	50%	0.9919	0.9900	0.9906	0.8051	0.7451	0.7639	0.7943	0.7208	0.7456

Table 4. Performance comparison of three attack sample generation schemes (traditional FDIA sample generation [8], mixed Laplacian model-based FDIA sample generation (LMM-FDIA) [18], and adversarial attention diffusion model-based FDIA sample generation (AttDiff)). This test uses an LSTM-based FDIA attack detection model [30] on the IEEE 39 bus system.

Attack Level	Attack Strength	Traditional FDIA [8]			LMM-FDIA [18]			AttDiff-FDIA		
		Precision	Recall	F_1-Score	Precision	Recall	F_1-Score	Precision	Recall	F_1-Score
Weak	2%	0.7414	0.7723	0.7555	0.7737	0.7234	0.7472	0.7180	0.7508	0.7328
	5%	0.8171	0.9069	0.8593	0.7976	0.8309	0.8138	0.7277	0.6972	0.7112
	10%	0.8814	0.9246	0.9011	0.8270	0.7857	0.8054	0.7461	0.7098	0.7273
Medium	15%	0.9326	0.9334	0.9325	0.8317	0.8007	0.8157	0.7526	0.7548	0.7533
	20%	0.9562	0.9651	0.9605	0.8549	0.7868	0.8194	0.7721	0.7269	0.7485
	25%	0.9835	0.9819	0.9826	0.8705	0.8009	0.8340	0.7866	0.7460	0.7656
Strong	30%	0.9867	0.9881	0.9874	0.8828	0.8098	0.8442	0.8135	0.7936	0.8024
	40%	0.9993	0.9954	0.9973	0.9039	0.7797	0.8371	0.8417	0.7673	0.8024
	50%	1.0000	0.9987	0.9993	0.9234	0.7419	0.8220	0.8860	0.7651	0.8195

In addition, on the IEEE 39 and IEEE 118 bus systems, both the accuracy and F_1 score achieved the lowest values under the three attack intensities. This is enough to show that the samples generated by the method proposed in this experiment have a stronger ability to evade detection, and have better concealment and attack capabilities. In the comparison of the three methods, the reason why the method proposed in this experiment can achieve better performance results mainly depends on the following two reasons: First, this experiment introduces an attention mechanism in the application of the diffusion model, which makes the model itself pay more attention to the node features that have a greater impact on the results, thereby promoting the model's efficient learning of node data features; and second, by combining adversarial training in the entire diffusion process, it helps the model learn more robust feature representations, improves the generalization ability of the model, and then generates high-quality, indistinguishable attack sample data.

Table 5. Performance comparison of three attack sample generation schemes (traditional FDIA sample generation [8], mixed Laplacian model-based FDIA sample generation (LMM-FDIA) [18], and adversarial attention diffusion model-based FDIA sample generation (AttDiff)). This test uses a CNN-based FDIA attack detection model [12] on the IEEE 118 bus system.

Attack Level	Attack Strength	Traditional FDIA [8]			LMM-FDIA [18]			AttDiff-FDIA		
		Precision	Recall	F_1-Score	Precision	Recall	F_1-Score	Precision	Recall	F_1-Score
Weak	2%	0.7134	0.7293	0.710	0.6996	0.7330	0.7039	0.6880	0.7415	0.7043
	5%	0.7462	0.7644	0.7455	0.7174	0.7304	0.7107	0.7035	0.7280	0.7039
	10%	0.8235	0.8068	0.8072	0.7407	0.7309	0.7208	0.7106	0.7240	0.7049
Medium	15%	0.8652	0.8707	0.8628	0.7644	0.7457	0.7439	0.7106	0.7357	0.7118
	20%	0.9282	0.9162	0.9190	0.7856	0.7318	0.7461	0.7329	0.7134	0.7118
	25%	0.9493	0.9322	0.9382	0.7990	0.7221	0.7467	0.7424	0.7187	0.7194
Strong	30%	0.9571	0.9603	0.9571	0.8110	0.7554	0.7733	0.7513	0.7469	0.7381
	40%	0.9746	0.9741	0.9734	0.8227	0.7357	0.7691	0.7991	0.7441	0.7601
	50%	0.9836	0.9773	0.9797	0.8422	0.7668	0.7925	0.8364	0.7577	0.7861

Table 6. Performance comparison of three attack sample generation schemes (traditional FDIA sample generation [8], mixed Laplacian model-based FDIA sample generation (LMM-FDIA) [18], and adversarial attention diffusion model-based FDIA sample generation (AttDiff)). This test uses an LSTM-based FDIA attack detection model [30] on the IEEE 118 bus system.

Attack Level	Attack Strength	Traditional FDIA [8]			LMM-FDIA [18]			AttDiff-FDIA		
		Precision	Recall	F_1-Score	Precision	Recall	F_1-Score	Precision	Recall	F_1-Score
Weak	2%	0.7263	0.6764	0.6960	0.7159	0.8198	0.7640	0.6963	0.7457	0.7192
	5%	0.7541	0.7241	0.7382	0.7353	0.8149	0.7730	0.7179	0.6841	0.6994
	10%	0.8359	0.8269	0.8305	0.7637	0.8015	0.7819	0.7204	0.6691	0.6928
Medium	15%	0.8815	0.8458	0.8630	0.7829	0.7701	0.7761	0.7219	0.6527	0.6853
	20%	0.9153	0.9364	0.9255	0.7971	0.7372	0.7657	0.7323	0.6584	0.6920
	25%	0.9341	0.9515	0.9427	0.8090	0.7355	0.7704	0.7456	0.7242	0.7327
Strong	30%	0.9538	0.9624	0.9580	0.8270	0.7113	0.7648	0.7608	0.7020	0.7272
	40%	0.9776	0.9731	0.9753	0.8409	0.7082	0.7688	0.8002	0.7195	0.7574
	50%	0.9873	0.9737	0.9804	0.8630	0.7151	0.7812	0.8269	0.7501	0.7861

4.3. Performance comparison of different parameters

The main framework adopted in this experiment is the diffusion model. As we all know, the diffusion model is closely related to the time step, and the performance of the samples generated by the diffusion model is also closely related to the time step. In order to verify that the time step used is the most cost-effective value, a series of experiments were conducted to observe the effect of different time steps on the detection results of the final generated attack samples. Comparative experiments were carried out on IEEE 14, IEEE 39, and IEEE 118 bus systems using detection models based on CNN and LSTM. The result is shown in Figure 3.

This comparative experiment discusses the cases where the time step of the diffusion model is set to 2, 4, 6, 8, and 10, respectively. From the above experimental results, it can be seen that the detection accuracy on the three bus systems will decrease slightly with the increase of the time step. On the IEEE 14 bus test system, the detection accuracy of the CNN detector at different time steps is 0.6953, 0.6715, 0.6751, 0.673, and 0.6679, respectively. The detection accuracy of the LSTM detector at different time steps is 0.7051, 0.6788, 0.6744, 0.6727, and 0.6681, respectively. When the time step is 2, the detection accuracy based on the two detectors is the highest, and when the time step is 10, the detection accuracy is the lowest. However, when the time steps are 4, 6, 8, and 10, the accuracy only decreases in a small range, and there is not much difference.

In addition, for the comparison of the IEEE 39 and IEEE 118 bus systems, the detection accuracy based on the two detectors eventually showed a downward trend. For the diffusion model, the time step has a direct impact on the final detection result. Within a reasonable range, the larger the time step, the more detailed control the model will have, and the more opportunities to learn and approximate the true distribution of the data, and better capture and recover the complex patterns and structures in the data, so as to improve the diversity and authenticity of the generated samples. However, a larger the time step is not the better, because the increase of the time step will lead to a large increase in the amount of computation and training time, and too high of a time step may also cause the model to overfit the training data, and can not be well generalized to new data. In order to guarantee the quality of attack

samples, the calculation cost is reduced as much as possible by analyzing the comparative experiments on three bus systems with different time steps. Therefore, the most cost-effective time step setting was chosen for this experiment, that is, the time step was set to 4. This setting can not only effectively control the time overhead, but also ensure the high quality of the generated samples.

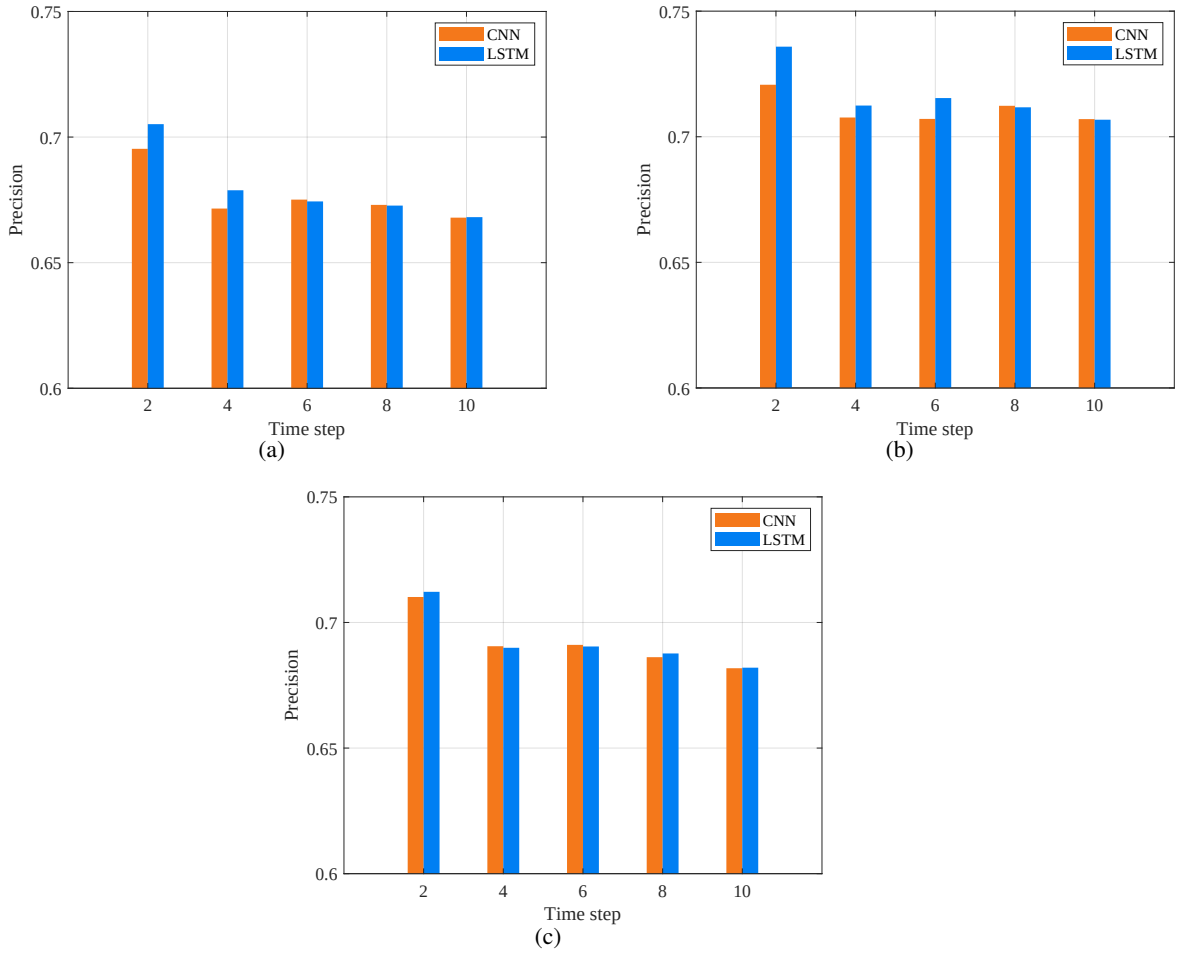

Figure 3. Comparison of the performance of the proposed sample generation scheme under different time steps: (a) comparison based on the IEEE 14 bus system; (b) comparison based on the IEEE 39 bus system; (c) comparison based on the IEEE 118 bus system.

4.4. Influence of different noise environments

In order to verify the reliability of the proposed sample generation method, a series of experiments were carried out to study the influence of different noise disturbances on the measured data of the power system. Because the actual power system measurement data is often disturbed by noise, it is crucial to understand the impact of these disturbances on the proposed approach. This experiment simulates this kind of interference by adding Gaussian noise with different variance to the data of each node in the power system. In addition, on the IEEE 14, IEEE 39, and IEEE 118 bus systems, the detection models based on CNN and LSTM were used to conduct comparative tests on the attack samples generated by the traditional FDIA method, the mixed Laplacian method, and the method

proposed in this experiment. The test results are shown in the figure below. Figures 4, 5, and 6 are the tests based on the IEEE 14 bus system, the IEEE 39 bus system, and the IEEE 118 bus system, respectively.

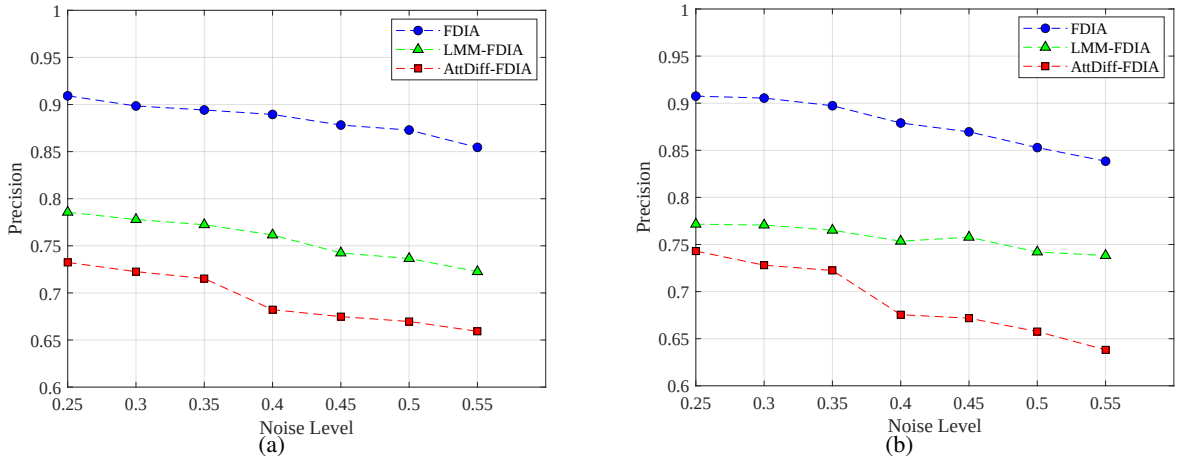

Figure 4. Comparison of three attack sample generation schemes under different noise levels based on the IEEE 14 bus system. (a) Detection model based on CNN. (b) Detection model based on LSTM.

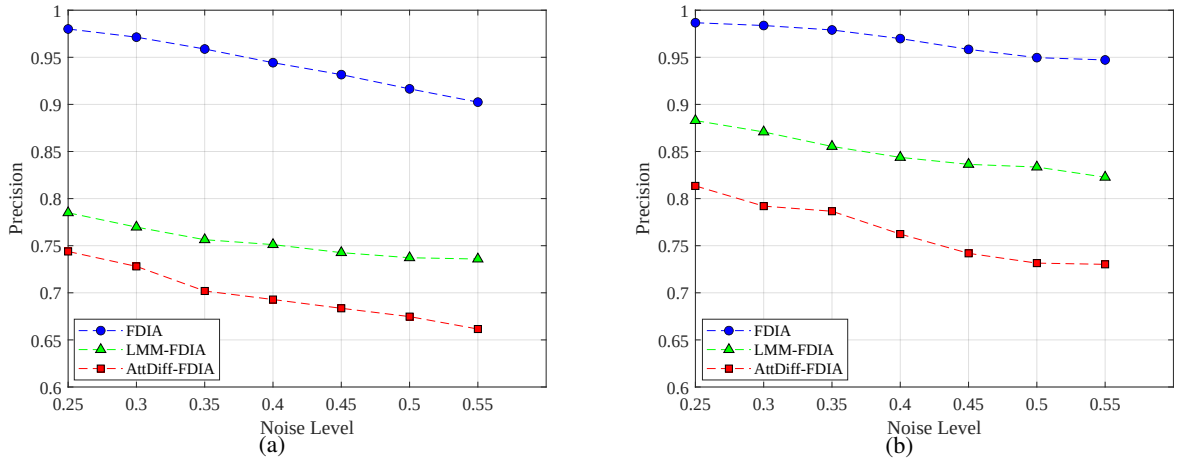

Figure 5. Comparison of three attack sample generation schemes under different noise levels based on the IEEE 39 bus system. (a) Detection model based on CNN. (b) Detection model based on LSTM.

In this comparative experiment, seven different noise variance levels were set for the measurement data: 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, and 0.55. By observing the above experimental results, it can be found that the precision of all models shows a decreasing trend with the increase of the noise variance level. Therefore, the noise variance level in the simulated power grid environment in this study is set at 0.25. This decrease in detection accuracy is mainly due to the fact that the increase in noise variance

makes it more difficult to distinguish between normal and attack samples. Noise interference makes it easier for pure data to be masked by it. In addition, the experimental data also show that under the two detection models, the attack sample generation method proposed in this experiment performs better than the traditional FDIA sample generation method and the mixed Laplacian sample generation method under various noise levels. This proves that the sample generation method proposed in this study has more obvious advantages over the existing technology in terms of robustness and adaptability. In general, the method proposed in this paper can generate more stable and anti-jamming FDIA attack samples, and has stronger ability to evade detection. This provides a large number of efficient attack sample resources for constructing and training FDIA detection models based on deep learning and AI.

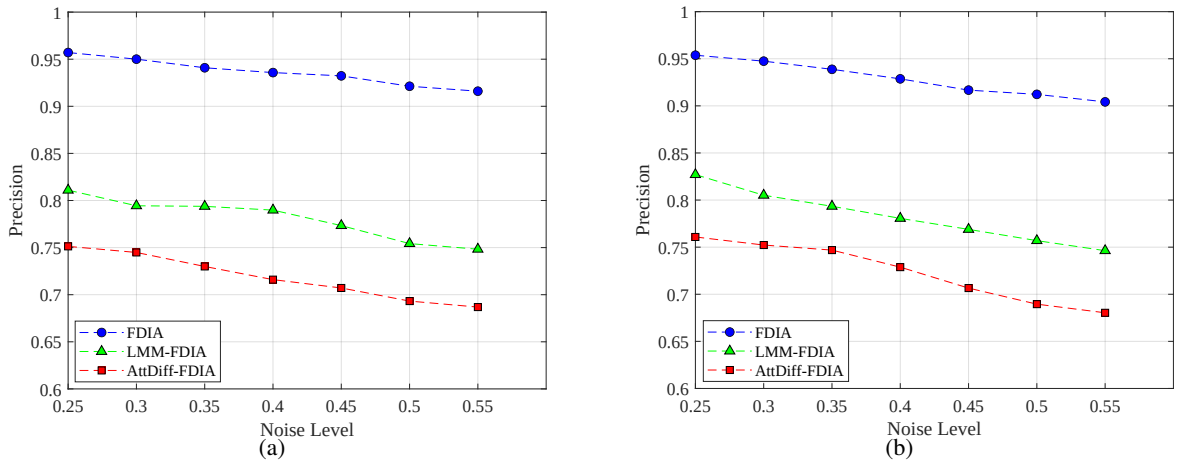

Figure 6. Comparison of three attack sample generation schemes under different noise levels based on the IEEE 118 bus system. (a) Detection model based on CNN. (b) Detection model based on LSTM.

5. Conclusions

The existing FDIA detection methods have been limited by the problem of sample scarcity and data set imbalance. This paper presents a method to generate FDIA attack samples using an adversarial attention diffusion model. The overall model of the method consists of two parts: the diffusion model and GAN model with an attention mechanism (ATTGAN). First, the original power flow data is noised by the forward diffusion of the diffusion model combined with the time step. In this process, the original measurement data and added noise corresponding to each time step are recorded for the subsequent model training. After the forward diffusion is completed, the data recorded during the forward diffusion process is used to train the proposed ATTGAN model, which is used for noise prediction during the reverse diffusion process. Finally, in the reverse diffusion process, the trained ATTGAN model is combined with the diffusion model, and according to the size of the time step, the iterative denoising is done step by step from back to front to generate FDIA attack samples with stronger concealment and attack capabilities. This experiment was conducted on a large scale on the IEEE 14, IEEE 39, and IEEE 118 bus systems. The experimental results show that the attack samples generated by the proposed sample generation method have better ability to evade

detection after detection by two kinds of detectors, thus effectively reducing the detection precision and accuracy.

Although our proposed attention-diffusion method can generate more covert and disguised attack samples, thereby expanding the attack sample base, it also has some limitations. The main limitation is the relatively large time overhead of the model, which may affect its efficiency in real-time applications. In addition, model performance is sensitive to parameter selection and requires careful tuning to achieve optimal performance, potentially adding complexity to model deployment. To overcome these limitations, future work could explore algorithmic optimization and hardware acceleration to improve the computational efficiency of the model, as well as the application of hyperparameter optimization techniques to automate the parameter selection process to reduce the effort of manual adjustment and potentially discover a better combination of parameters. Future work could also further explore the explainability of the model, improve user trust in the model output, and extend the model's application to support broader smart grid security research.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research is supported by grants from the Engineering Research Center of Offshore Wind Technology Ministry of Education (Shanghai University of Electric Power).

Conflict of interest

The authors declare no conflicts of interest.

Author contributions

Conceptualization, Kunzhan Li; methodology, Kunzhan Li and Fengyong Li; validation, Kunzhan Li; formal analysis, Fengyong Li and Baonan Wang; writing—original draft preparation, Kunzhan Li; writing—review and editing, Fengyong Li and Baonan Wang; supervision, Meijing Shan. All authors have read and agreed to the published version of the manuscript.

References

1. Baheti R, Gill H (2011) Cyber-physical systems. *The Tmpact of Control Technology* 12: 161–166. Available from: www.ieeecss.org.
2. Aoufi S, Derhab A, Guerroumi M (2020) Survey of false data injection in smart power grid: Attacks, countermeasures and challenges. *J Inf Secur Appl* 54: 102518. https://doi.org/10.1016/j.jisa.2020.102518

3. Li Z, Tong W, Jin X (2016) Construction of cyber security defense hierarchy and cyber security testing system of smart grid: Thinking and enlightenment for network attack events to national power grid of Ukraine and Israel. *Autom Electr Power Syst* 40: 147–151. https://doi.org/10.7500/AEPS20160313005

4. Kumar A, Saxena N, Jung S, et al. (2021) Improving detection of false data injection attacks using machine learning with feature selection and oversampling. *Energies* 15: 212. https://doi.org/10.3390/en15010212

5. Cheng G, Lin Y, Zhao J, et al. (2022) A highly discriminative detector against false data injection attacks in AC state estimation. *IEEE Trans Smart Grid* 13: 2318–2330. https://doi.org/10.1109/TSG.2022.3141803

6. Qu Z, Dong Y, Li Y, et al. (2024) Localization of dummy data injection attacks in power systems considering incomplete topological information: A spatio-temporal graph wavelet convolutional neural network approach. *Appl Energy* 360: 122736. https://doi.org/10.1016/j.apenergy.2024.122736

7. Fang Z (2024) Detection of false data injection attacks in power grid based on Res-CNN-LSTM with channel fusion. *Electr Eng* 25: 11–17. Available from: www.cesmedia.cn.

8. Su X, Deng C, Yang J, et al. (2024) Damgat based interpretable detection of false data injection attacks in smart grids. *IEEE Trans Smart Grid* 15: 4182–4195. https://doi.org/10.1109/TSG.2024.3364665

9. Takiddin A, Ismail M, Atat R, et al. (2023) Robust graph autoencoder-based detection of false data injection attacks against data poisoning in smart grids. *IEEE Trans Artif Intell* 5: 1287–1301. https://doi.org/10.1109/TAI.2023.3286831

10. Li X, Wang Y, Lu Z (2023) Graph-based detection for false data injection attacks in power grid. *Energy* 263: 125865. https://doi.org/10.1016/j.energy.2022.125865

11. Qu Z, Bo X, Yu T, et al. (2022) Active and passive hybrid detection method for power CPS false data injection attacks with improved AKF and GRU-CNN. *IET Renewable Power Gener* 16: 1490–1508. https://doi.org/10.1049/rpg2.12432

12. Wang S, Bi S, Zhang YJA (2020) Locational detection of the false data injection attack in a smart grid: A multilabel classification approach. *IEEE Int Things J* 7: 8218–8227. https://doi.org/10.1109/JIOT.2020.2983911

13. Xie J, Rahman A, Sun W (2024) Bayesian gan-based false data injection attack detection in active distribution grids with DERs. *IEEE Trans Smart Grid* 15: 3223–3234. https://doi.org/10.1109/TSG.2023.3337340

14. Yan Q, Wang M, Huang W, et al. (2019) Automatically synthesizing DoS attack traces using generative adversarial networks. *Int J Mach Learn Cybern* 10: 3387–3396. https://doi.org/10.1007/s13042-019-00925-6

15. Kumar V, Sinha D (2023) Synthetic attack data generation model applying generative adversarial network for intrusion detection. *Comput Secur* 125: 103054. https://doi.org/10.1016/j.cose.2022.103054

16. Tian J, Wang B, Wang Z, et al. (2021) Joint adversarial example and false data injection attacks for state estimation in power systems. *IEEE Trans Cybern* 52: 13699–13713. https://doi.org/10.1109/TCYB.2021.3125345

17. Bhattacharjee A, Mondal AK, Verma A, et al. (2022) Deep latent space clustering for detection of stealthy false data injection attacks against AC state estimation in power systems. *IEEE Trans Smart Grid* 14: 2338–2351. https://doi.org/10.1109/TSG.2022.3216625

18. Wu Y, Zu T, Guo N, et al. (2023) Laplace-domain hybrid distribution model based FDIA attack sample generation in smart grids. *Symmetry* 15: 1669. https://doi.org/10.3390/sym15091669

19. Li F, Shen W, Bi Z, et al. (2024) Sparse adversarial learning for FDIA attack sample generation in distributed smart grids. *CMES-Comput Model Eng Sci* 139. https://doi.org/10.32604/cmes.2023.044431

20. Tang B, Lu Y, Li Q, et al. (2023) A diffusion model based on network intrusion detection method for industrial cyber-physical systems. *Sensors* 23: 1141. https://doi.org/10.3390/s23031141

21. Batzolis G, Stanczuk J, Schönlieb CB, et al. (2021) Conditional image generation with score-based diffusion models. *arXiv preprint arXiv*:2111, 13606. https://doi.org/10.48550/arXiv.2111.13606

22. Deng J, Dong W, Socher R, et al. (2009) Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition,* Miami, FL, USA, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

23. Esser P, Rombach R, Ommer B (2021) Taming transformers for high-resolution image synthesis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12873–12883. https://doi.org/10.1109/CVPR46437.2021.01268

24. Austin J, Johnson DD, Ho J, et al. (2021) Structured denoising diffusion models in discrete state-spaces. *Adv Neural Inf Process Syst* 34: 17981–17993. Available from: www.proceedings.neurips.cc.

25. Park SW, Lee K, Kwon J (2021) Neural markov controlled SDE: Stochastic optimization for continuous-time data. *International Conference on Learning Representations*. Available from: https://openreview.net/pdf?id=7DI6op61AY.

26. Tashiro Y, Song J, Song Y, et al. (2021) Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Adv Neural Inf Process Syst* 34: 24804–24816. Available from: www.proceedings.neurips.cc.

27. Wu Y, Wang Q, Guo N, et al. (2023) Efficient multi-source self-attention data fusion for FDIA detection in smart grid. *Symmetry* 15: 1019. https://doi.org/10.3390/sym15051019

28. Wu Y, Sheng Y, Guo N, et al. (2022) Hybrid deep network based multi-source sensing data fusion for fdia detection in smart grid. *2022 Asia Power and Electrical Technology Conference (APET)*, 310–315. https://doi.org/10.1109/APET56294.2022.10072807

29. Li Y, Wei X, Li Y, et al. (2022) Detection of false data injection attacks in smart grid: A secure federated deep learning approach. *IEEE Trans Smart Grid* 13: 4862–4872. https://doi.org/10.1109/TSG.2022.3204796

30. Musleh A, Chen G, Dong Z, et al. (2023) Attack detection in automatic generation control systems using LSTM-based stacked autoencoders. *IEEE Trans Ind Inf* 19: 153–165. https://doi.org/10.1109/TII.2022.3178418

AIMS Press