applied sciences

MDPI

Article

The Detection of False Data Injection Attack for Cyber–Physical Power Systems Considering a Multi-Attack Mode

Buxiang Zhou [1,2], Xuan Li [1,2], Tianlei Zang [1,2,*], Yating Cai [1,2], Jiale Wu [1,2] and Shijun Wang [1,2]

[1] College of Electrical Engineering, Sichuan University, Chengdu 610065, China
[2] Intelligent Electric Power Grid Key Laboratory of Sichuan Province, Sichuan University, Chengdu 610065, China
* Correspondence: zangtianlei@126.com

Abstract: Amidst the evolving communication technology landscape, conventional distribution networks have gradually metamorphosed into cyber–physical power systems (CPPSs). Within this transformative milieu, the cyber infrastructure not only bolsters grid security but also introduces a novel security peril—the false data injection attack (FDIA). Owing to the variable knowledge held by cyber assailants regarding the system's network structure, current achievements exhibit deficiencies in accommodating the detection of FDIA across diverse attacker profiles. To address the historical data imbalances encountered during practical FDIA detection, we propose a dataset balancing model based on generating adversarial network-gated recurrent units (GAN-GRU) in conjunction with an FDIA detection model based on the Transformer neural network. Harnessing the temporal data extraction capabilities of gated recurrent units, we construct a GRU neural network system as the GAN's generator and discriminator, aimed at data balance. After preprocessing, the balanced data are fed into the Transformer neural network for training and output classification to discern distinct FDIA attack types. This model enables precise classification amidst varying FDIA scenarios. Validation involves testing the model on load data from the IEEE 118-bus system and affirming its high accuracy and effectiveness in detecting power systems after multiple attacks.

Keywords: cyber–physical power systems; false data injection attack; attack modeling; generative adversarial network; gated recursive unit; the Transformer neural network detection

Citation: Zhou, B.; Li, X.; Zang, T.; Cai, Y.; Wu, J.; Wang, S. The Detection of False Data Injection Attack for Cyber–Physical Power Systems Considering a Multi-Attack Mode. *Appl. Sci.* **2023**, *13*, 10596. https://doi.org/10.3390/app131910596

Academic Editors: Meng Tian, Ying Zhang and Zhengcheng Dong

Received: 30 August 2023
Revised: 19 September 2023
Accepted: 20 September 2023
Published: 22 September 2023

1. Introduction

The electric power industry is witnessing a significant development trend with the emergence of smart grid cyber–physical power systems (CPPSs). This trend achieves intelligent control and management by integrating traditional power systems with information control equipment, thereby establishing communication and sensing networks to facilitate the seamless interlinking and interoperability of power, information, and physical systems [1,2]. However, in the wake of the burgeoning new energy internet, the realm of smart grids encounters novel security challenges, with cybersecurity threats emerging as a grave concern. A prevailing form of attack in this context is the false data injection attack, which poses a considerable risk to power system stability. By maliciously introducing counterfeit data into the power system, this attack disrupts critical functionalities such as state estimation and economic dispatch. Evidential occurrences such as the Ukraine blackout (June 2015); the compromise of the Utah grid control system (July 2019); the malevolent assault on the Delta Montrose Electric Power Association in Colorado, USA (2021); and the ransomware attack on the Australian Electricity Provider Network Systems Energy (2021) have underscored the tangible threat posed by false data injection attacks on power systems [3,4]. Consequently, the development of robust detection methodologies for identifying such spurious data injection attacks becomes imperative within the purview of smart grid systems. To address these burgeoning security concerns, smart grids are proactively used in advanced technologies, encompassing sensing and measurement techniques,

and incorporated into robust control methodologies. Furthermore, the establishment of bidirectional communication networks is actively pursued to neutralize cybersecurity vulnerabilities. Equally vital is the need to institute preemptive strategies for averting and detecting false data injection attacks and other cyber–physical risks. By judiciously adopting pertinent security mechanisms and deploying discerning attack detection frameworks, the stable operation of the smart grid is assured, thereby mitigating potential disruptions to power systems stemming from cybersecurity threats. We list the abbreviations in this paper in Abbreviations.

The inception of false data injection attacks was initially introduced by [5]. In the work presented, the authors delineate this phenomenon as a prototypical assault on the data integrity of power system state estimation. The realm of CPPS has, as a consequence, engendered substantial scholarly interest in exploring avenues of investigation surrounding concealed and arduously detectable FDIAs. As described in [6–8], extant methodologies for the detection of FDIAs predominantly encompass approaches grounded in model-based prediction and data-driven machine learning paradigms.

1.1. FDIA Detection Based on a Model-Driven Approach

Reference [9] proposes a false data static detection method based on the similar characteristics of buses at a certain moment. This method can detect the false data injection problem that may occur in the power system. Enhanced sensitivity to dynamic changes in the power system is achieved, albeit with the possibility of certain attacks being missed due to the primary focus on static features. Reference [10] proposes an event-triggered fully distributed algorithm, which can effectively reduce communication time. In that study, for FDIAs, the authors constructed a specific model that considers the attacker to maximize the loss and the defender to minimize the loss. The proposed methods require complex communication architectures, and their applicability is constrained by the specific characteristics of the system. Reference [11], on the other hand, proposes an FDIA detection method that combines model predictive control and artificial neural networks. Through the incorporation of neural networks, a better adaptation to nonlinear system characteristics is realized to enhance the detection efficiency of the security control layer. However, a substantial volume of data is needed for neural network training, imposing significant computational resource requirements. To achieve FDIA attack detection, in [12], an interval observer was used to accurately estimate the state values of the internal system; then, an isolation algorithm was constructed, and an interval residual detection criterion was established based on a logical judgment matrix of attack characteristics. Reference [13], on the other hand, proposes a model for the detection and defense of FDIAs in load frequency control systems by combining an evolutionary game model with a Kalman filter algorithm. In practice, the implementation of this approach entails the utilization of more intricate models and algorithms, potentially leading to heightened complexity. Reference [14] proposes an improved detection method based on principal component analysis, which introduced a mathematical transformation principal approach to improve the detection performance, such as the detection rate and false alarm rate. Nonetheless, specific prerequisites regarding data preprocessing and computational resources exist, rendering the method potentially inapplicable to all datasets.

1.2. FDIA Detection Based on a Data-Driven Approach

However, model-based methods for detecting FDIAs have certain limitations, such as the tendency of local convergence during online identification and the difficulty in selecting model thresholds. With the development of CPS, the model-based approach is insufficient to cope with the state estimation problem created by the increasing quantity of data. Therefore, a detection method that integrates data-driven, method-based, and intelligent algorithms has emerged. Reference [15] proposes an online FDIA detection method combining wavelet transform and deep neural networks to detect spatial and temporal data inconsistencies caused when spurious data are injected into state vectors. A superior

ability to handle multi-dimensional data features is demonstrated; however, in certain cases, the fine-tuning of model hyperparameters is imperative to attain optimal performance. Reference [16] proposes an FDIA detection method based on the Kalman filter algorithm and recurrent neural networks (RNNs), which obtains dynamic thresholds based on the fit of observations and predictions to determine whether FDIAs occur. To ensure that FDIAs do not affect the accuracy of state estimation and further improve the accuracy of FDIA detection, effective detection methods need to be developed. The adaptability of detection is bolstered through the utilization of dynamic thresholds that can adjust to changing scenarios. Concurrently, stringent demands are placed on system modeling and parameter adjustment, necessitating real-time threshold updates. A framework for fault identification and diagnosis is introduced in [17]. It harnesses graph-edge conditional convolutional networks. The fundamental objective underlying this framework pertains to the establishment of a mapping relationship, one that aligns measurement estimates with the authentic state of the power system. The consideration of power system topology renders the method suitable for intricate systems. However, the construction of graph convolutional networks requires a substantial volume of training data, resulting in heightened computational complexity. In [18], an online intelligent anomaly and attack detection method is introduced, which makes use of the partially observable Markov decision process. This method is designed to detect cyber-attacks targeting smart grids. The method takes into account the dynamics and incomplete observability inherent to the system, rendering it applicable to complex network environments. Nevertheless, it necessitates high precision in system model accuracy and exhibits sensitivity to model errors. A distributed microgrid control system FDI attack detection structure is presented in [19], which is based on the utilization of a Gaussian process regression and a one-class support vector machine anomaly detection algorithm. Due to its suitability for distributed microgrid control systems characterized by strong distribution and real-time performance, the method may require specialized knowledge and time investment for model parameter selection and adjustment, contributing to heightened implementation complexity. The evaluation of the introduced attack detector encompasses two viewpoints: the assessment of detection loss probability and false alarm probability. The data-driven approach introduced in [20] demonstrates the integration of residual networks and attention-mechanism-based long short-term memory models. This integration is directed toward the achievement of temporal correlation and feature extraction in measurement data. Through the amalgamation of these models, the temporal dependencies inherent in the data are efficiently captured. It should be noted that the practical application of this method might encounter challenges, including the intricacy associated with model fusion and the need for parameter tuning. However, a single deep neural network is susceptible to adversarial attacks, which may lead to poor output stability in the trained network.

Accordingly, rooted in the aforementioned concepts of various FDIA detection methodologies, the focus of this study is directed toward the exploration of the practical implementation of FDIA presence within power system state estimation. In operational scenarios, historical data in terms of practical application may involve instances of sample imbalance due to the scarcity of attack data. This scarcity can subsequently engender noteworthy discrepancies within data-driven detection models, resulting in substantial errors. Owing to the intricate nature of acquiring comprehensive CPPS information, the proposed model in this paper is developed from the perspective of attackers. Consequently, a multi-attack model involving the injection of false data is formulated that is contingent upon the extent of information they possess concerning the network structure. This endeavor aims to establish a closer alignment with real-world application scenarios. Building upon this foundational premise, an FDIA detection method is proposed in this paper, wherein the amalgamation of the GAN-GRU data balanced processing model and the Transformer neural network is employed. This confluence is purposefully designed to effectively address the challenge posed by the diminished attack detection rate, arising from both data imbalance and intricate attack patterns.

The contribution of this paper is as follows:

1. An attacker-centric perspective is embraced in this study to comprehensively investigate the system's security, with an intention to think and analyze from the viewpoint of potential attackers. Thus, within this investigation, diverse tiers of information are taken into account, influencing the creation of a spectrum of attack patterns. These patterns encompass distinct quantities of attack buses with varying numbers and intensities. Through the manipulation of both the count and strength of attack buses, an array of plausible attack scenarios can be simulated, facilitating a holistic comprehension of the ramifications these attacks might impose upon the day-to-day operation of the system.
2. To address the challenge posed by the imbalance in historical measurement data in practical scenarios, a data-balancing processing model based on GAN-GRU is introduced in this study. Within this model, a GRU neural network is incorporated into the GAN framework, serving as integral components of the generator and discriminator, respectively. Through jointly training both the GAN network and the GRU network, a recurrent network is constructed with the purpose of generating a limited segment of attack vector data within an imbalanced sample set. This process facilitates the establishment of balance among the quantities of diverse sample classes in the training dataset, ultimately leading to enhanced accuracy and reliability of the models when subjected to FDIA assessment.
3. In response to the challenge posed by false data injection across various attack scenarios, a false data detection model is constructed utilizing the Transformer neural network. The equilibrium dataset is divided into distinct training and test subsets, enabling the application of the Transformer to the detection and classification of false data injection attacks. Following this, a confusion matrix is generated, contrasting projected and actual values, and comparison experiments are performed against the enhanced convolutional neural network-gated recurrent unit (CNN-GRU), long short-term memory network (LSTM), and support vector machine (SVM). The simulation findings underscore the commendable scalability of the proposed data-balancing processing model based on GAN-GRU and the detection algorithm based on the Transformer neural network when scrutinizing the FDIA within the IEEE 118-bus system. Furthermore, the outcomes demonstrate a substantial enhancement over conventional algorithms in terms of precision in detecting attacks.

This paper is divided into five sections, the rest of which are organized as follows: In Section 2, the fundamental aspects of false data injection attacks are explained, encompassing system state estimation and the modeling of attacks in consideration of partial network information. The data-balancing processing model based on GAN-GRU and the FDIA detection model based on the Transformer neural network presented in this paper are described in Section 3. Then, Section 4 presents illustrative examples and the experimental outcomes pertaining to the models introduced in this study. This encompasses the enhancements observed in the detection mechanism following the implementation of data balancing, as well as the accuracy and performance assessment of the FDIA detection mechanism under standard operating conditions. Additionally, it encompasses comparative experiments involving multiple detection techniques subject to multi-attack mode. Finally, Section 5 serves as the concluding segment of this paper, summarizing the contributions and findings presented within.

2. Problem Description

2.1. Theoretical Foundations of CPPS State Estimation

State estimation entails the examination of instrument measurement data within a SCADA system [21,22], enabling the deduction of the operational state of the power system. This paper addresses the issue of state estimation in CPPS, encompassing a model characterized by n-dimensional measurement vectors and m-dimensional system state vectors. The procedure involves the analysis, processing, and amalgamation of

measurement data to derive a precise estimate of the system's state. State estimation plays a pivotal role in power system operation and monitoring, facilitating the timely identification of anomalies by operators, aiding in decision making and fault diagnosis, and affording a comprehensive comprehension of the system's operational condition [21]. By consistently enhancing and optimizing the algorithms and methodologies used for state estimation, the accuracy and robustness of the power system's state can be augmented, subsequently elevating the security and reliability of the system. The state estimation problem can be formulated as follows:

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

where $z = (z_1, z_2, \ldots, z_n)^T$ represents the n-dimensional measurement vector of the system, H represents the Jacobi matrix, $x = (x_1, x_2, \ldots, x_m)^T$ represents the system state vector, and $e = (e_1, e_2, \ldots, e_n)^T$ represents the measurement error.

In the detection process, the traditional residual comparison method based on BDD plays a significant role. This method discerns the presence of subpar measurements—those that could be either faulty or compromised by malicious attacks—by contrasting the 2-norm of measurement residuals against a predetermined threshold. When the residual value surpasses this established threshold, the detector signals the occurrence of an attack. This technique offers an effective approach to identifying attacks and uncovering anomalies that might potentially jeopardize the system. Consequently, the BDD method relying on residual comparison has emerged as a pivotal element in fortifying the system's security. Through the continuous enhancement and optimization of this detection methodology, the system's attack detection capability can be further elevated, consequently reinforcing the assurance of system security and reliability. The detection approach through residual comparison can be formulated as follows:

$$R = \|z - Hx\|_2^2 \geq \tau, \tag{2}$$

where R represents the 2-norm of residuals, and τ represents the residual test threshold set by the CPPS.

2.2. A Multi-Attack Mode

In this study, a multi-attack mode is formulated based on the theoretical foundations of CPPS state estimation. This model is composed of various configurations, characterized by distinct numbers of attacking buses and varying levels of attack intensities. The former configuration is referred to as FDIA construction based on the partial grid information known, while the latter is quantified through the utilization of 2-norm metrics.

2.2.1. FDIA Construction Based on the Partial Grid Information Known

This paper takes the perspective of the attacker into account during the formulation of attack vectors. Given the attacker's limited comprehension of the electric power CPS network structure, we postulate that only partial network information is accessible to the attacker. Based on the varying degrees of information available, diverse attack vectors are formulated [23]. To enhance the comprehension of the security issues inherent to the target power grid, the grid is divided into two critical regions: the attack region α and the unattacked region β, as depicted in Figure 1. In Figure 1, the loads and transformers in the two regions are represented by L and T, respectively, and a partial depiction of the network structure is illustrated in Figure 1.

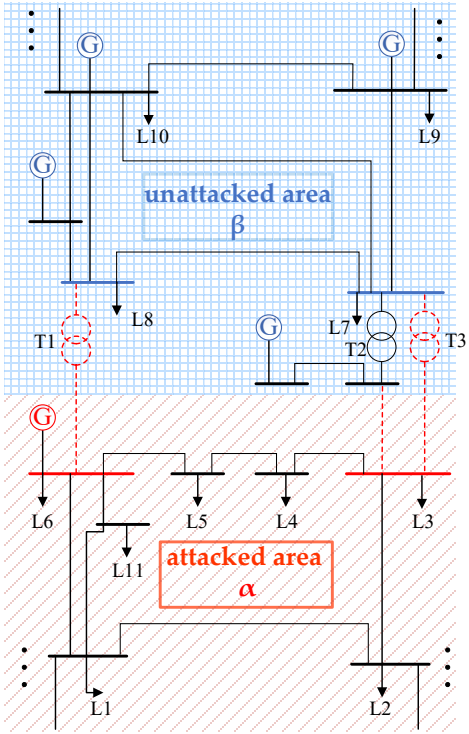


Figure 1. Diagram of the division of the attacked and unattacked areas.

The attackers possess a restricted understanding of the network information affiliated with the attack region, and their aim is to perpetrate the attack by manipulation of measurement data within that area. To effectively evade detection by the BDD detection module, the attackers must guarantee that their manipulative actions do not result in an elevation of the measurement residuals, thereby evading the system's suspicion. The relationship between the target grid's state variables and the measurement data can be formulated as follows:

$$z = \begin{bmatrix} z_\alpha \\ z_\beta \end{bmatrix} = \begin{bmatrix} \boldsymbol{H}_{\alpha\alpha} & \boldsymbol{H}_{\alpha\beta} \\ 0 & \boldsymbol{H}_{\beta\beta} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{x}}_\alpha \\ \hat{\boldsymbol{x}}_\beta \end{bmatrix} + \begin{bmatrix} \boldsymbol{e}_\alpha \\ \boldsymbol{e}_\beta \end{bmatrix}, \tag{3}$$

where z, z_α, and z_β represent the number of system measurements for the total system, the attacked area, and the unattacked area; $\hat{\boldsymbol{x}}_\alpha$ and $\hat{\boldsymbol{x}}_\beta$ are the estimated values of the state variables for the corresponding areas, respectively; $\boldsymbol{e}_\alpha$ and $\boldsymbol{e}_\beta$ are the corresponding measurement errors; and $\boldsymbol{H}_{\alpha\alpha}$, $\boldsymbol{H}_{\alpha\beta}$, and $\boldsymbol{H}_{\beta\beta}$ are the Jacobi matrices associated with the state variables. Expanding the aforementioned equation in accordance with the provided expression results in:

$$z_\alpha = \boldsymbol{H}_{\alpha\alpha}\hat{\mathrm{x}}_\alpha + \boldsymbol{H}_{\alpha\beta}\hat{\boldsymbol{x}}_\beta + \boldsymbol{e}_\alpha, \tag{4}$$

$$z_\beta = \boldsymbol{H}_\beta\hat{\boldsymbol{x}}_\beta + \boldsymbol{e}_\beta, \tag{5}$$

The attacker constructs the following false data based on the knowledge of the network in area α:

$$z'_\alpha = \boldsymbol{H}_{\alpha\alpha}\boldsymbol{x}_\alpha + \boldsymbol{H}_{\alpha\beta}\hat{\boldsymbol{x}}_\beta, \tag{6}$$

At this point, the measurement vector of the system becomes $z' = [z'_\alpha z_\beta]^T$, and the state vector becomes $\hat{\boldsymbol{x}}' = [\boldsymbol{x}_\alpha \hat{\boldsymbol{x}}_\beta]^T$. The 2-norm residual r' test of the measurement data satisfies the following derived relationship:

$$r' = \left\| z' - \boldsymbol{H}(\hat{\boldsymbol{x}}')\hat{\boldsymbol{x}}' \right\|_2 = \left\| \begin{matrix} z'_\alpha - (\boldsymbol{H}_{\alpha\alpha}\boldsymbol{x}_\alpha + \boldsymbol{H}_{\alpha\beta}\hat{\boldsymbol{x}}_\beta) \\ z_\beta - \boldsymbol{H}_{\beta\beta}\hat{\boldsymbol{x}}_\beta \end{matrix} \right\|_2 = \left\| \begin{matrix} 0 \\ \boldsymbol{e}_\beta \end{matrix} \right\|_2 = \| \boldsymbol{e}_\beta \|_2 < r = \left\| \begin{matrix} \boldsymbol{e}_\alpha \\ \boldsymbol{e}_\beta \end{matrix} \right\|_2, \tag{7}$$

Based on the information available to the attackers concerning the network within the attack area, the formulation of false data becomes a straightforward process, as depicted in

Equation (6). Once these false data points infiltrate the system, as described by Equation (7), the resultant measurement residuals of the system exhibit reduced magnitudes compared to the residuals corresponding to scenarios without the presence of false data. Consequently, the construction of a false data attack based on partial network information is not only viable but also tactically discrete. This form of attack introduces a significant latent peril to the security and dependable operation of power systems.

In this scenario, the attacker leverages their knowledge of the network within the attack area to specifically manipulate the system's measurement data. By meticulously devising deceptive data and inserting them into the system, the attacker effectively evades conventional methods of detecting erroneous data. Since the introduction of false data leads to a reduction in the measurement residuals of the system, the nature of this attack is notably surreptitious and poses challenges in timely detection.

2.2.2. Attack Intensities

Hence, the construction of FDIAs predicated upon partial network information represents a formidable assault, imperiling the security and dependable operation of power systems. To counter this jeopardy, the imperative lies in the exploration and formulation of advanced false data detection algorithms and methodologies, geared toward unearthing and preempting these surreptitious attacks. This research centers on a 118-bus system comprising 180 measurement instruments [24]. The perspective of the attacker drives the generation of vector configurations for random attacks targeting 30, 65, and 100 measurement gauges. Attack vectors are primarily employed to circumvent the system's residual test by manipulating the 2-norm of the measurement vectors. Thus, in this study, various attack intensities are depicted by configuring different 2-norms of the attack vectors. Standard experimentation introduces the 2-norm of 1, while comparative investigations span 2-norm values ranging from 0.5 to 2.5. A visual representation juxtaposes the attack vectors with the unaltered normal data, the outcome of which stems from the random generation of 30-bus attack vectors, as shown in Figure 2.

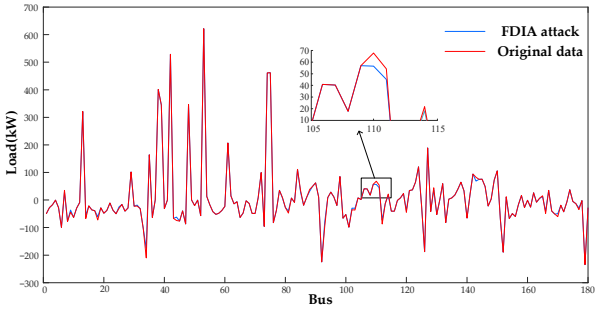


Figure 2. Comparison of attacked data and original load data.

Meticulous scrutiny exposes that the devised 30-bus attack vectors evince insubstantial disparities in relation to the original normal data vectors. This observation underscores the ease with which attack vectors can elude the conventional BDD false detection method, contingent upon residual comparison.

This configuration engenders a novel set of challenges to power system security. Attackers stand poised to exploit these inconspicuous deviations to construct pernicious attack vectors, subsequently infusing them into the system. The intelligent construction of these attack vectors facilitates their effective evasion of traditional residual detection methods, thus circumventing the system's ability to accurately discern the presence of an attack.

3. FDIA Detection Model Based on GAN-GRU and Transformer

To address the challenge posed by the scarcity of attack data in historical records of real power grids, a data-balancing processing model based on GAN-GRU is proposed in this study. This model aims to rectify the imbalance by aiding the classifier in reducing

the false alarm rate associated with the identification of FDIAs. Furthermore, it enhances the stability of the FDIA detection model based on the Transformer neural network in the context of unbalanced data. The approach involves partitioning the balanced dataset into training and detection sets, maintaining a ratio of 3:1. Each of the four distinct data categories—an unattacked category and three types of attacked data—is assigned unique labels. These labeled datasets are subsequently utilized in training the Transformer network model. Subsequently, the models are assessed using the test set. This section proceeds by providing an initial exposition of the data-balancing processing model based on GAN-GRU and the classification detection model based on the Transformer neural network. It is followed by a presentation of the structured framework for conducting FDIA detection as outlined in this research.

3.1. GAN-GRU-Based Data-Balancing Process

3.1.1. Network Structure of GAN-GRU

Our data-balancing processing model utilizes generators and discriminators of generative adversarial network discriminators to balance the entire dataset by generating synthetic samples with the class balance to increase the number of samples in a few classes. At the same time, we employ generators and discriminators with gated loop units to capture temporal dependencies in the sequential data to ensure that the generated samples are temporally consistent. In this way, we are able to effectively deal with the imbalance problem in the historical measurement data, thus improving the performance of the fault-checking model. In order to deal with a small number of attack vectors in the training dataset, they are preprocessed and used as inputs to the GAN, which are trained through continuous iterations of the generator and discriminator, with the ultimate goal of generating a balanced dataset.

The gated recurrent unit initially introduced in [25] and referenced as a structure of a gated recurrent unit represents an evolved iteration of the conventional recurrent neural network. Its effectiveness lies in efficiently capturing semantic correlations within extended sequences, thus addressing the challenge of gradient vanishing or explosion. In parallel, similar to the long short-term memory unit, GRU also incorporates gating units to regulate information flow. Nevertheless, GRU differs from LSTM in that it lacks an independent storage unit, exhibiting a more streamlined structure and computational process. Its operation involves a reset gate and an update gate, forming a core structure that can be expressed as follows:

$$z_t = \sigma\left(W^{(z)}x_t + U^{(z)}h_{t-1}\right), \tag{8}$$

$$r_t = \sigma\left(W^{(r)}x_t + U^{(r)}h_{t-1}\right), \tag{9}$$

$$\widetilde{h}_t = \tanh(Ur_t \circ h_{t-1} + Wh_{t-1}), \tag{10}$$

$$h_t = (1 - z_t) \circ h_{t-1} + z_t \circ \widetilde{h}_t, \tag{11}$$

where x_t is the input information at the current moment; the variants denoted as update gate z_t and reset gate r_t function as intermediary entities within this framework; σ is the sigmoid activation function; the summation of both the input and the past hidden layer state is denoted by $\widetilde{h}_t$ with h_{t-1} signifying the latter; $W^{(z)}$, $U^{(z)}$, $W^{(r)}$, $U^{(r)}$, U, and W are the trainable parameter matrices; and $\circ$ is the Hadamard product and tanh is the tanh activation function.

The utilization of GRU as both the generator and discriminator within the context of a GAN for balanced datasets provides a multitude of benefits and advantages. Firstly, in its role as a generator, the inherent ability of GRU for recursive modeling facilitates the generation of high-quality and diverse synthetic samples, thereby effectively augmenting the number of samples within specific categories. Acting as a discriminator, GRU's proficiency in classification and discrimination lends accuracy to the assessment of generated sample quality, thus enhancing the effectiveness of balanced datasets. This process is further opti-

mized through appropriate training and fine-tuning, enabling GRU to strike a harmonious equilibrium between its roles as generator and discriminator. This equilibrium significantly contributes to further improving the dataset balance and ultimately enhances the performance of the fault detection model. The strategic integration of GRUs as generators and discriminators within GANs presents a robust solution for data-balancing processing, thereby offering substantial improvements in data processing efficacy and quality within practical applications. The data-balancing processing model based on GAN-GRU is shown in Figure 3.

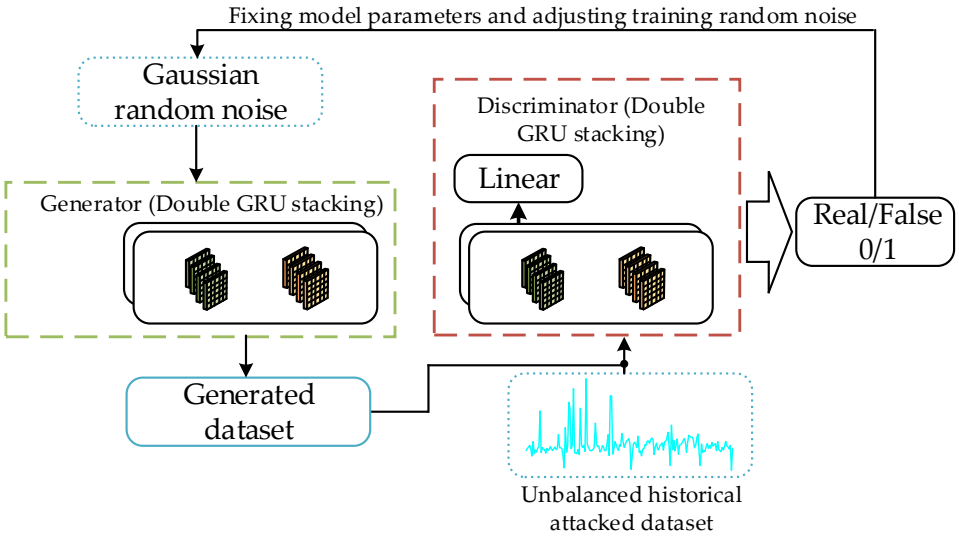


Figure 3. Sample balanced structure of false data based on GAN-GRU.

Both the generator module and the discriminator constitute the structural underpinning of a GRU. To facilitate the mapping of GRU hidden states to the dimensions of the output sequence, a linear layer is introduced. This pivotal linear layer plays a key role in generating specific data through the generator and distinguishing between real and generated samples via the discriminator. During the forward propagation process, the input sequence is initially processed by the GRU layer, generating both an output and a final hidden state. Subsequently, the last time step within the output sequence is selected and directed to the linear layer for transformation, yielding the generated result. Finally, this generated result is conveyed as output. In the pursuit of enhancing the generator's performance, an approach involves the introduction of random Gaussian noise as input to the generated data. Throughout the GAN-GRU training process, the noise value size is systematically trained, with adjustments being made based on the monitored fluctuations in the loss values of both the generator and discriminator. These adjustments are instrumental in aligning the generated data distribution more closely with that of authentic data, thereby amplifying the generative ability of the generator. Consequently, this approach effectively elevates the output quality of the generator, ensuring that the generated data exhibit statistical properties and distribution characteristics that closely mirror those of real data.

The previous discussion illustrates that, in the GAN model employing GRU as both the generator and discriminator, the processing of input sequences by the generator module is accomplished through the utilization of the GRU layer. The conversion of the resultant hidden states into the dimensions of the generated outcomes is carried out by means of a linear layer, leading to the production of specific data. Conversely, the discriminator module exhibits a similar architecture whereby authentic and synthetic samples are distinguished. This distinction is achieved by subjecting the input sequences to the GRU layer and subsequently mapping the resultant hidden states to the dimensions of the evaluation outcomes using a linear layer. This design facilitates enhanced generation of high-quality

samples and precise differentiation between samples, thus contributing to the overall performance and effectiveness of the GAN model.

3.1.2. Loss Function

For the generator, we define the mean square error (MSE) as the loss function L_G:

$$L_{\mathrm{G}} = \frac{1}{n}\sum_{i=1}^{n}[y_i - p(y_i)]^2, \tag{12}$$

where n represents the number of data samples, y_i represents the actual value, and $p(y_i)$ indicates the predicted value.

The MSE function serves as a metric for quantifying the extent of separation between two vectors. This function enables a more efficient assessment of the dissimilarity between the generated data and the authentic data, thereby enhancing the proximity of the attack vectors produced by the generator to the original attack vectors.

Shifting the focus to the discriminator, the initial step encompasses the introduction of the authentic signal into the discriminator, enabling forward propagation. Subsequently, the computation of the true loss is computed, which is subsequently juxtaposed with the reference value of 1. Following this, the introduction of random noise culminates in the generation of a spurious signal through the generator. This synthetic signal is then presented to the discriminator, subjected to forward propagation, and culminates in the computation of the false loss. The latter is subsequently juxtaposed with the reference value of 0. In this context, the true loss delineates the discriminator's quantification of loss between the actual signal output and the target label. Conversely, the false loss encapsulates the discriminator's quantification of loss between the synthetic signal output generated by the generator and the target label. To accomplish this objective, the loss function employed by the discriminator utilizes a binary cross-entropy loss function L_D:

$$L_D = -\frac{1}{n}\sum_{i=1}^{n}\{y_i \times log[1 - p(y_i)] + (1 - y_i) \times log[1 - p(y_i)]\}, \tag{13}$$

where y_i is the binary label 0 or 1, and $p(y_i)$ is the sigmoid function representing the probability that the output belongs to the label y_i:

$$p(y_i) = \frac{1}{1 + exp(-y_i)}, \tag{14}$$

The binary cross-entropy loss function is aptly employed to normalize the discriminator's output, treating it as a probability distribution. This encourages the discriminator to acquire the skill of allocating real data to the high-probability realm (proximate to 1) and attributing generated data to the low-probability realm (proximate to 0). Consequently, enhanced discrimination between these data types is facilitated by the discriminator. The primary objective of categorizing the input data into two classes, namely, real data and generated data, is fulfilled using the discriminator. The utilization of a binary cross-entropy loss function is instrumental in securing the stability of gradients during the training process. This stability assumes paramount importance in the context of GAN training, where the adversarial interplay between the discriminator and the generator can lead to training instability. Notably, binary cross-entropy typically yields gradients that exhibit a relatively smooth profile, thereby streamlining the training process. Hence, the adoption of binary cross-entropy emerges as a natural choice, given its widespread application as a loss function in binary classification scenarios.

The loss function assumes a pivotal role within the training procedure of both the generator and the discriminator. The magnitude of the loss value is computed, thereby facilitating the adoption of corresponding optimization strategies to modify the gradient of the discriminator and generator, subsequently updating the network parameters. This

configuration and training regimen within GAN-GRU engenders the optimization of the generator and discriminator performance, culminating in the ultimate achievement of generating balanced datasets.

Through the amalgamation of the mechanisms outlined above, the enactment of a competitive interplay between the generator and the discriminator is facilitated. This dynamic empowers the generator to yield false signals that closely approximate real data, whereas the discriminator acquires enhanced accuracy in discerning authentic from fabricated signals. This training strategy and judicious selection of the loss function for GAN-GRU furnish an effective approach, affording the generation of balanced datasets and robust support for ensuing investigations within the domain of the classification task based on the Transformer. A comprehensive comparison between the accuracy yielded by balanced and unbalanced datasets when inputted into the Transformer detector will be addressed in Chapter 4.

3.2. FDIA Detection Model Based on the Transformer Neural Network

The neural network model known as the Transformer is rooted in an attention mechanism. Its conceptualization originated from Google and was first elucidated in 2017 by Vaswani et al., as documented in [26]. This model was conceived to address intricate natural language processing undertakings encompassing machine translation, text summarization, and speech recognition. In contrast to conventional recurrent neural network models, such as LSTM and GRU, the Transformer model boasts enhanced parallel computation capabilities and abbreviated training durations, rendering it adept at managing protracted sequence tasks. This attribute has endowed the Transformer model with widespread utility within the realm of natural language processing.

However, different from its traditional applications in natural language processing and speech recognition, our study applies the Transformer model to the realm of false data classification in CPPS loads. Figure 4 illustrates the architecture of the Transformer network architecture as delineated within this paper.

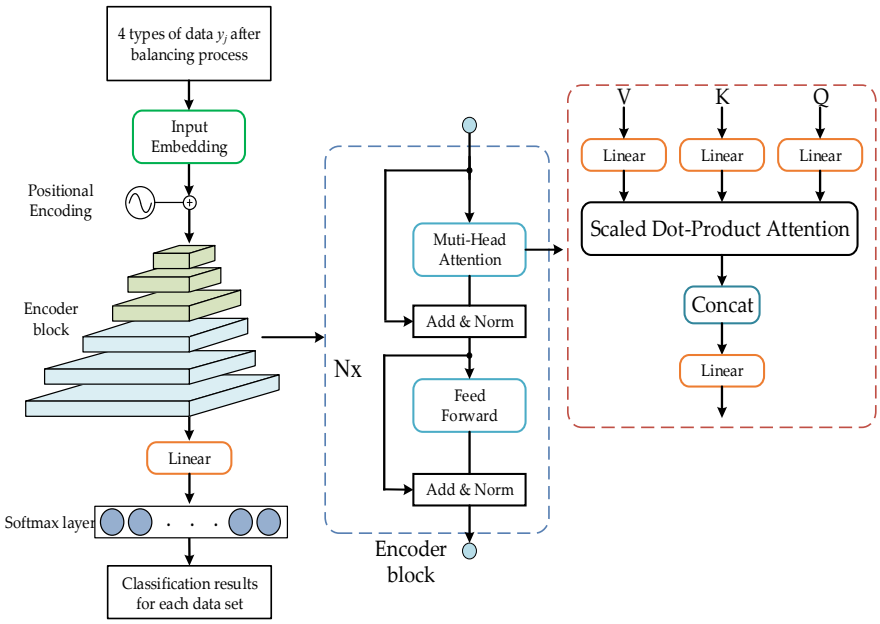


Figure 4. FDIA detection process based on the Transformer neural network.

The innovative application of the Transformer model in the context of CPPS applications carries considerable significance. This expansion of the Transformer model into the realm of load false data detection presents the opportunity to leverage its intrinsic self-attention mechanism and robust parallel computing prowess to discern and analyze the distinctive attributes of counterfeit data. As a result, a novel and effective avenue emerges

to address the escalating predicament of false data within real-world scenarios. In our methodology, the initial step entails the utilization of balanced normal data in conjunction with three distinct attack vector datasets as inputs. Consequently, categorical labels are assigned to both the normative dataset and the three distinct attack datasets as follows:

$$y_j = \begin{cases} 0 \text{ normal operation} \\ 1\ 30 - \text{node attack vectors} \\ 2\ 65 - \text{node attack vectors} \\ 3\ 100 - \text{node attack vectors} \end{cases}, \tag{15}$$

Considering the Transformer's ability to process sequence data, we consider each set of data as a sequence and classify it using the Transformer network. Specifically, we input the sequence data into the encoder and decoder for processing.

3.2.1. Encoder Layer

In the encoder, we first implement the feature transformation by mapping the discrete input features into a continuous vector space through the embedding layer. Next, these vectors are fed into a multi-layered Transformer encoder layer for processing. Each encoder layer consists of multiple Transformer encoder units, and in this paper's experiments, six encoder layers are stacked according to the setup in reference [26]. In each Transformer encoder unit, the data are processed using a multi-head self-attention mechanism and a forward propagation network to extract the semantic information of the input sequence. The multi-head self-attention mechanism interacts with the input symbols with other symbols to obtain a set of attention weights that are used to indicate the importance of the symbols to other symbols.

The forward propagation network is processed through multiple fully connected layers to extract higher-level semantic information. By stacking multiple Transformer encoder layers, the global semantic information is integrated and a rich set of feature vectors is obtained. In classification tasks, the use of a decoder is usually not required, so in this experiment, we only need to focus on the encoder part and do not need to use a decoder.

3.2.2. Multi-Head Attention

The scaled dot-product attention within multi-head attention represents a distinctive form of attention mechanism. In practical applications, the attention function for a set of queries is computed concurrently, and these computations are aggregated into a matrix $\boldsymbol{Q}$. The encoding of keys and values into matrices $\boldsymbol{K}$ and $\boldsymbol{V}$ follows a similar approach. The computation of the output matrix is then formulated as follows:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V}, \tag{16}$$

where d_k represents queries and keys of dimension.

In contrast to the isolated application of dimensional keys, values, and queries to a single attention function, we determined that projecting queries, keys, and values linearly to the $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ dimensions, respectively, yields considerable benefits. Consequently, for each projected version of the query, key, and value, we concurrently execute the attention function, generating results encompassing $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ dimensional output values. These output values are subsequently concatenated and subjected to further projection to generate the ultimate output values, as elucidated in Figure 4.

The incorporation of a multi-head attention mechanism empowers the model with the capacity to collectively attend to distinct subspaces associated with diverse positions. However, for an individual attention head, this effect is mitigated through averaging, as computed below:

$$\text{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_\text{h})\boldsymbol{W}^{\boldsymbol{O}}, \tag{17}$$

$$\text{head}_i = \text{Attention}\left(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}}, \boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}}, \boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{V}}\right), \tag{18}$$

where MultiHead$(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$ represents the result of the calculation of the multi-head attention, and h = 8 represents the number of parallel attention layers in the model. $\boldsymbol{W}_i^{\boldsymbol{Q}}$, $\boldsymbol{W}_i^{\boldsymbol{K}}$, $\boldsymbol{W}_i^{\boldsymbol{V}}$, and $\boldsymbol{W}^{\boldsymbol{O}}$ represent the projection parameter matrices corresponding to $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$, along with the computed results.

Lastly, the feature vectors are aligned to the same dimensions for classification tasks by means of a linear layer. Each dataset, equipped with labels 0, 1, 2, and 3, undergoes processing to yield a conclusive output probability distribution. This distribution serves as the basis for predicting the classification outcomes concerning manipulated load data. To enhance interpretability, a normalized confusion matrix representing the distribution outcomes can be generated. Through the self-attention mechanism and parallel computing capabilities inherent to the Transformer, our methodology effectively harnesses sequence data insights and extracts elevated-level feature representations. This proficiency enables the precise classification of altered load data within power systems.

3.3. FDIA Detection Based on GAN-GRU Data Balancing and the Transformer Classifier

In this paper, a false data detection method is proposed, as depicted in Figure 5. To address the challenge of a low detection rate resulting from imbalanced data, GAN-GRU and the Transformer techniques are employed. The approach functions as a generator and discriminator by integrating a GRU neural network within the GAN framework. Throughout the training process, both the GAN network and the GRU network construct recurrent neural networks to generate a limited number of data samples, thereby achieving a balanced representation of diverse sample types in the training dataset. The balanced dataset is divided into training and test subsets, and the Transformer technique is utilized to detect and classify anomalous data. Ultimately, a normalized confusion matrix is obtained through the comparison of predicted and actual values.

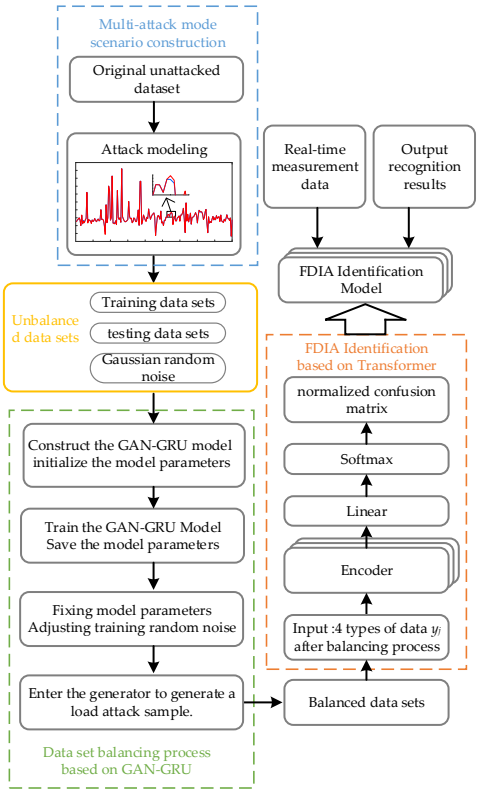


Figure 5. Flowchart of FDIA detection based on GAN-GRU data balancing and the Transformer classifier.

By synergistically employing data balancing and classification models, we attain an enhanced capacity to discern and differentiate various forms of FDIAs. The data-balancing processing model effectively equalizes the dataset through the generation of synthetic attack data, thus providing ample training samples. Leveraging the Transformer classification-based detection model allows us to gain insights into the interrelations among distinct attack types and normal data, enabling precise classification determinations. This elevates the efficacy and resilience of the FDIA identification model, contributing to the amelioration of security challenges in power systems. This approach not only curtails the rate of false alarms but also enhances the system's adeptness at handling disparate data distributions, thus furnishing a robust underpinning for the secure and dependable operation of power systems.

4. Simulation Test and Result Analysis

4.1. Experimental Platform and Data

In this section, the performance of the proposed FDIA detection mechanism is evaluated in the context of IEEE 118-bus power systems. The grid topologies were acquired from MATPOWER [27]. All simulations were executed on a computer system featuring an AMD Ryzen 5 5600H with Radeon Graphics CPU, NVIDIA GeForce RTX 3050 Laptop GPU, AMD Radeon (TM) Graphics, and 64 GB of RAM. The GAN-GRU and Transformer neural network were constructed utilizing the Torch to enhance computational efficiency.

In this experiment, the original load dataset was derived from [24]. To broaden the scope of authentic data and generate unattacked datasets, a manual approach was employed to simulate the load distribution across individual buses. This approach serves a dual purpose, not only augmenting the diversity and volume of datasets but also furnishing a more comprehensive array of training samples to underpin our classification endeavor. When generating these load data points, a normal distribution was assumed, whereby the mean load data are aligned with the baseline load value, and the standard deviation equates to one-sixth of the baseline load value. This meticulous manual simulation process allowed us to generate unattacked datasets that faithfully represent the load distribution across individual buses. By assuming a normal distribution, where the mean load data correspond to the baseline load value, and the standard deviation is set as one-sixth of the baseline load value, we ensured that our generated load data points adhered to realistic distribution patterns. This rigorous approach significantly enriched the diversity and volume of our datasets, laying a solid foundation for our classification efforts and enhancing the authenticity of the data we employed for this experiment.

Consequently, a dataset encompassing 9098 sets of unattacked normal original data samples was effectively assembled. These samples were drawn from 180 m within the IEEE 118-bus network, representing load data reflecting typical operational conditions. The meticulous procurement of these data samples assumes a paramount role in ensuring experimental precision and dependability. Through the execution of the aforementioned experimental design and subsequent data processing, a dependable and representative dataset was meticulously formulated, thereby constituting a robust cornerstone for our research pursuits. Meanwhile, in order to simulate the historical data's inherent imbalance as encountered in real-world applications, an imbalanced sample dataset was formulated from the existing 9098 arrays, tailored to generate vectors of false data injections, spanning a spectrum of distinct attack modes.

Furthermore, adhering to the guidelines presented in Section 3, explicit labels were assigned to each data category to uphold the precision and feasibility of the experiments. These labels, serving as inputs to the Transformer neural network, concurrently form an integral segment of the network's output classification. A detailed account of the quantity of datasets employed is provided in Table 1.

Table 1. Classification and grouping of original datasets.

Data Category	Label	Sample Size
normal operation	0	3698
30 attacked buses	1	2400
65 attacked buses	2	1800
100 attacked buses	3	1200

4.2. Test Content and Assessment Metrics

The experimental facet of this paper encompasses three pivotal sections devised for the purpose of exploring and appraising the effectiveness and performance of distinct methodologies:

(1) Data-balancing experiments, predicated on the application of GAN-GRU, constitute the first section. This innovative methodology was employed to address dataset imbalances, and its effectiveness in performing classification tasks was explored.
(2) Throughout the process of Transformer training, the 2-norm was set to 1.0, and the consequent variations in accuracy and loss values were meticulously observed across the training epochs. An in-depth analysis was undertaken to comprehensively investigate the impact of this 2-norm on the training process of the Transformer model, thereby assessing its resilience and stability.
(3) Comparative experiments were conducted, involving 2-norm values ranging from 0.5 to 2.5. By systematically manipulating the 2-norm level, the performance of the models was meticulously compared across varying degrees of attack intensity. Extensive analysis was performed to identify the implications and distinctive features of these differing 2-norm levels.

For the purpose of conducting a comprehensive comparison, we performed a comparative experiment in which the Transformer classifier was juxtaposed against three alternative classifiers: the enhanced CNN-GRU [28], the conventional LSTM deep learning approach [29,30], and the SVM machine learning method [31–33]. This approach facilitates a holistic assessment of the Transformer classifier's performance and effectiveness.

To expedite network convergence and mitigate the risk of overfitting, the mini-batch gradient descent technique was employed. In our model, each mini-batch encompassed 128 data samples. During each iteration, a specified number of samples were randomly selected from the training set to construct a mini-batch, which was then utilized for gradient computation and parameter updates. Following established practices in machine learning, the batch was divided into training and test sets, with three-fourths of the data allocated for training and the remaining one-fourth for testing. For model optimization, we employed the Adam optimizer with an initial learning rate of 0.001. Additionally, the patience parameter was set to five to monitor performance and enable the early termination of training when deemed appropriate.

The algorithm performance was evaluated by calculating metrics using confusion matrices, including accuracy $\lambda_{\text{Accuracy}}$, detection rate $\lambda_{\text{Precision}}$, recall λ_{Recall}, and the harmonized mean value of detection rate and completeness λ_{F1}.

$$\lambda_{\text{Accuracy}} = \frac{T_{\text{P}} + T_{\text{N}}}{T_{\text{P}} + T_{\text{N}} + F_{\text{P}} + F_{\text{N}}}, \tag{19}$$

$$\lambda_{\text{Precision}} = \frac{T_{\text{P}}}{T_{\text{P}} + F_{\text{P}}}, \tag{20}$$

$$\lambda_{\text{Recall}} = \frac{T_{\text{P}}}{T_{\text{P}} + F_{\text{N}}}, \tag{21}$$

$$\lambda_{\text{F1}} = \frac{2 \times \lambda_{\text{Precision}} \times \lambda_{\text{Recall}}}{\lambda_{\text{Precision}} + \lambda_{\text{Recall}}}, \tag{22}$$

where T_{P} (true positive) indicates the number of samples with a true value of FDIA and a predicted value of FDIA; F_{N} (false negative) indicates the number of samples with a true

value of FDIA but a predicted value of normal run; F_P (false positive) indicates the number of samples with a true value of normal run but a predicted value of FDIA; and T_N (true negative) indicates the number of samples where both the true value and the predicted value are normal run.

4.3. Data-Balancing Tests

To enhance the training of the FDIA detection model, our attention was drawn to the relatively limited presence of false data samples within the foundational dataset, which could potentially impact the model's efficacy. In response to this data imbalance, an evaluation of data-balancing processing was undertaken using the GAN-GRU framework. This evaluation was conducted under the premise of a standard working condition involving an attack 2-norm of 1. Within this evaluation, synthetic false data samples were generated via the GAN-GRU framework to augment the scarcity of false data within the original dataset. The generator and discriminator of the GAN-GRU model were formulated as neural network architectures endowed with a certain level of complexity and expressive capacity. The generator's objective is to produce synthetic data that closely resemble authentic false data samples, whereas the discriminator's task is to ascertain whether the input data are genuine or artificially generated.

In the course of our experimentation, the parameters for the GAN's generator and discriminator were established as follows: An input size of 1, 128 hidden neurons, and the incorporation of two GRU layers were initialized in a batch-first dimensionality manner. Specifically, the generator undertakes the role of generating synthetic data to emulate authentic data samples. By utilizing a noise vector with an input size of 1 as the generator's input, a sequence of neural network layers and activation functions were employed to yield a final output resembling synthetic data akin to genuine data samples. The inclusion of 128 hidden neurons within the generator implies that 128 hidden units, instrumental in capturing potential features of the input data, were incorporated. Meanwhile, the discriminator, responsible for discerning between genuine and synthetically generated data, similarly adopted a network architecture featuring a GRU layer with 128 hidden neurons. Through training, the discriminator is endowed with the capacity to accurately classify both genuine and synthetic data.

The loss function for the generator was defined as the mean-squared error, serving as a measure of the discrepancy between the synthetic data generated by the generator and the authentic data. In essence, it quantifies the mean-squared error between the output data and the target data. Conversely, the discriminator's loss function was stipulated as the binary cross-entropy loss. This loss function gauges the discriminator's precision in categorizing the input data by contrasting the predictions made for genuine and synthetic data with their corresponding labels.

With these parameter configurations and loss function formulations, the GAN models were trained, facilitating the generation of high-quality synthetic data by the generator while enabling the discriminator to aptly discern disparities between genuine and synthetic data. During testing, the generator and discriminator's loss functions were continuously recorded, and their fluctuations were plotted against the number of training iterations. These graphical representations furnish insights into the loss function's behavior throughout model training, consequently aiding the comprehension of the models' learning progression and performance, as shown in Figure 6.

By plotting the loss functions of the generator and discriminator against epochs, the training progression and performance trajectory of the model can be effectively observed. The oscillatory behavior exhibited by these two loss functions illustrates the dynamic interplay inherent in the adversarial training process. As the training unfolds, both the generator's and discriminator's loss functions gradually stabilize at elevated levels. This convergence signifies the successful optimization of both the generator and the discriminator over the course of the training process.

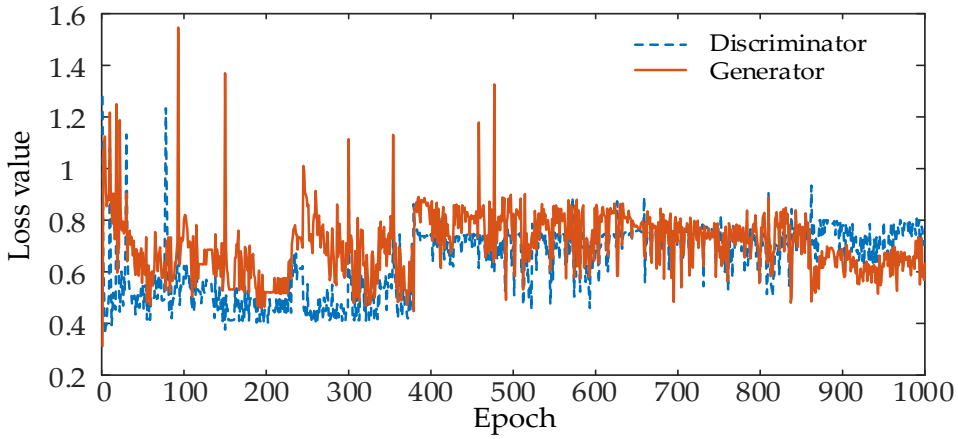


Figure 6. Loss value variation in data during the GRU-GAN training process.

In order to gauge the effectiveness of data-balancing experiments in augmenting the accuracy of FDIA detection, a sequence of comparative experiments was conducted, encompassing both the dataset subjected to the balancing procedure and the unbalanced dataset for FDIA detection. The accuracy curve depicting training iterations is illustrated in Figure 7, serving as an intuitive portrayal of the accuracy trend across epochs. Furthermore, Table 2 presents a comprehensive array of metric results, facilitating an exhaustive evaluation of model performance. Through the orchestration of these comparative experiments, the impact of data-balancing interventions on FDIA detection accuracy was thoroughly explored.

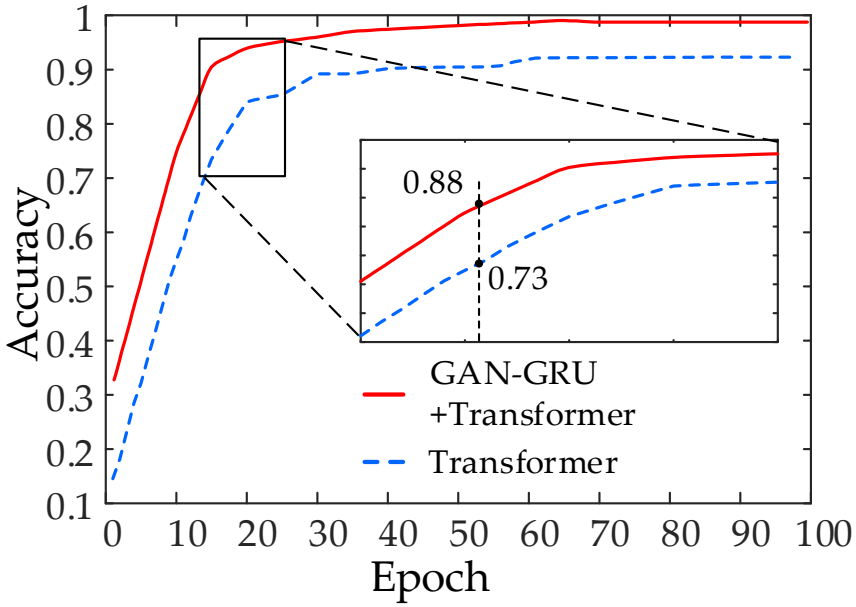


Figure 7. Comparison of FDIA detection experiment accuracy indices before and after GAN-GRU data-balancing processing.

Table 2. Comparison of FDIA detection experiment results before and after GAN-GRU data-balancing processing.

Arithmetic	$\lambda_{\text{Accuracy}}$	$\lambda_{\text{Precision}}$	λ_{Recall}	λ_{F1}
GAN-GRU +Transformer	0.9832	0.9837	0.9815	0.9824
Transformer	0.8436	0.8527	0.8395	0.8412

Upon a meticulous examination of the experimental outcomes, a conspicuous enhancement in FDIA detection accuracy is discerned subsequent to the implementation of data-balancing experiments. This discernible enhancement underscores the pivotal role played by data-balancing experiments in ameliorating the predicament of data imbalance intrinsic to FDIA detection. By augmenting the representation of false data injection attack samples, the model's ability to discern and assimilate the distinguishing attributes of genuine and spurious data is significantly enhanced, thereby culminating in an elevation of the accuracy in FDIA detection. This empirical endeavor attests to the effectiveness of the data-balancing approach based on GAN-GRU within the realm of FDIA detection.

4.4. Experiments on FDIA Detection of Grid Loads

4.4.1. FDIA Testing Experiments under Standard Operating Conditions

To assess the FDIA detection model based on the Transformer, the balanced dataset was divided into training and testing sets at a ratio of 3:1. As described in Section 3, distinct labels were allocated to each dataset. The segmentation outcomes are presented in Table 3, wherein the training set encompasses 2773 instances of normal data along with each of the three attack categories. With respect to the Transformer test set, comparative experiments were conducted. One set of experiments involved the utilization of the test dataset following GAN-GRU balancing, while the other set involved the unbalanced test dataset, which corresponds to real-world scenarios. These experiments were undertaken to assess the performance of the model proposed in this paper under varying conditions. This meticulous dataset segmentation strategy optimally harnesses the samples within the dataset. Such a stratagem not only aims to facilitate robust model evaluation and performance validation but also ensures a precise estimation of the model's capacity to extrapolate its efficacy across diverse data subsets. The original dataset utilized in this study comprised 9098 sets of unattacked load datasets recorded during normal operation. This dataset was divided into three distinct categories of attack vectors, each contingent on the number of targeted attack nodes. Consequently, an imbalanced dataset was created to simulate real-world scenarios where historical data often contain a scarcity of attack samples.

Table 3. Statistics of sample quantity in each stage.

Sample Type	Before Data Balancing	After Data Balancing	Training Set	Balanced Data Test Set	Unbalanced Data Test Set
0	3698	3698	2773	925	925
1	2400	3698	2773	925	300
2	1800	3698	2773	925	200
3	1200	3698	2773	925	100

In subsequent experiments, we rectified this imbalance by employing a data-balancing processing model rooted in GAN-GRU. This strategic adjustment ensured that each category of data achieved a state of complete balance, thus enhancing the efficacy of model training in the subsequent FDIA detection phase. A comprehensive overview of the dataset changes and divisions throughout this entire process is presented in Table 3.

Following the achievement of sample balancing, the overarching dataset imbalance was effectively alleviated to a level of 0. This intervention markedly enhanced the precision and dependability of the FDIA detection model. In the course of conducting experiments on this dataset, the attack 2-norm was established under the standard operational parameters of 1. Four distinct detection methods were employed. Table 4 delineates the performance metrics of the diverse methods in relation to the accuracy $\lambda_{\text{Accuracy}}$, detection rate $\lambda_{\text{Precision}}$, recall λ_{Recall}, and F1 value λ_{F1}:

Table 4. Algorithm performance comparison with other deep learning models.

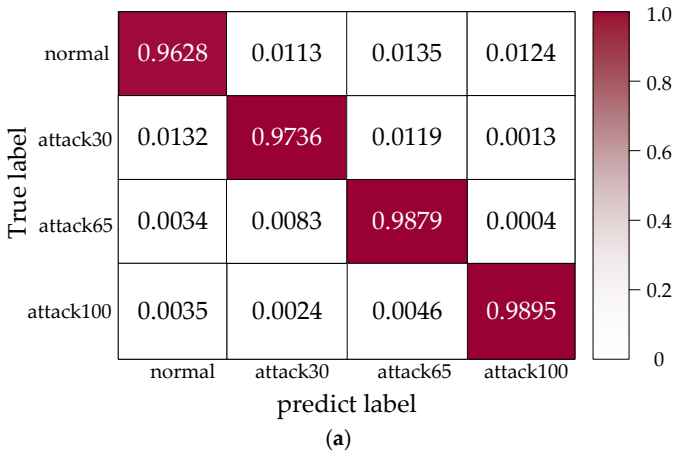
Test Set Type	Detection Methods	$\lambda_{Accuracy}$	$\lambda_{Precision}$	λ_{Recall}	λ_{F1}
Balanced data test set	Transformer	0.9832	0.9837	0.9815	0.9824
	CNN-GRU	0.9324	0.9436	0.9213	0.9284
	LSTM	0.8532	0.8412	0.8564	0.8523
	SVM	0.7935	0.8016	0.7932	0.7993
Unbalanced data test set	Transformer	0.9691	0.9581	0.9426	0.9502
	CNN-GRU	0.8351	0.8396	0.8475	0.8435
	LSTM	0.7924	0.8016	0.7869	0.7942
	SVM	0.7496	0.7534	0.7519	0.7526

Based on our analysis of the experimental findings, it can be deduced that, within the scope of this study, the Transformer classifier emerges as the most adept performer, achieving the highest values in terms of accuracy, overall detection rate, recall, and F1 score. The improved convolutional neural network CNN-GRU classifier is the second-best model, followed by the LSTM classifier, while the SVM classifier achieves the least favorable performance.

Upon evaluating the performance of the trained Transformer with an unbalanced dataset, it is observed that there are decreases in accuracy, detection rate, recall, and F1 value by 1.41%, 2.56%, 3.89%, and 3.22%, respectively. A similar trend is noted across all metrics for other deep learning methods. Nevertheless, the Transformer continues to exhibit high-performance metrics overall. To enhance the comprehensibility of the prediction results generated by the Transformer model, normalized confusion matrices for both the balanced dataset and the imbalanced dataset, reflecting real-world conditions, are presented in Figure 8.

The Transformer model harnesses a self-attention mechanism, facilitating the modeling of diverse positions within the input sequence and the capture of long-range dependencies. This attribute furnishes the Transformer with remarkable parallel computational proficiency. When juxtaposed against enhanced convolutional neural networks like CNN-GRU and conventional models like LSTM and SVM, the Transformer is superior in its heightened efficiency in handling voluminous datasets, thus accelerating both training and inference. By further enhancing its capabilities, the multi-layer stacking architecture of the Transformer facilitates more in-depth feature extraction and abstraction, thereby augmenting the model's ability to define variables. This capability positions the Transformer at an advantage when grappling with intricate multi-label classification tasks, thereby enhancing its capacity to discern varying data types.

Additionally, the variation in the loss values of the Transformer neural network and the method used in the reference [28,29,32] during the training process is depicted in Figure 9.



(**a**)

Figure 8. *Cont.*

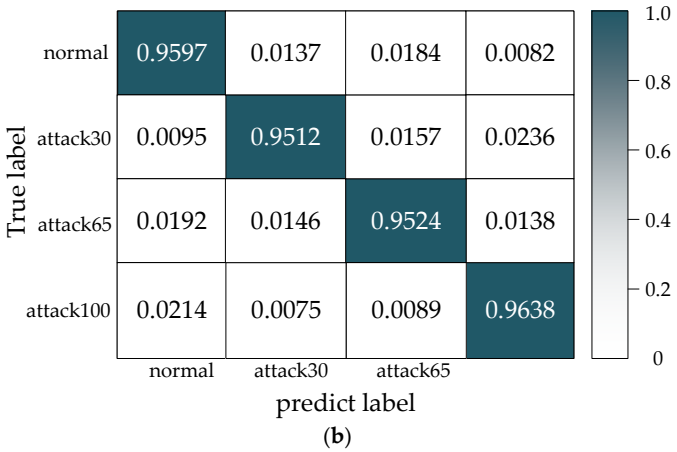


Figure 8. Confusion matrix of normalized Transformer detection results: (**a**) balanced test dataset; (**b**) unbalanced test dataset.

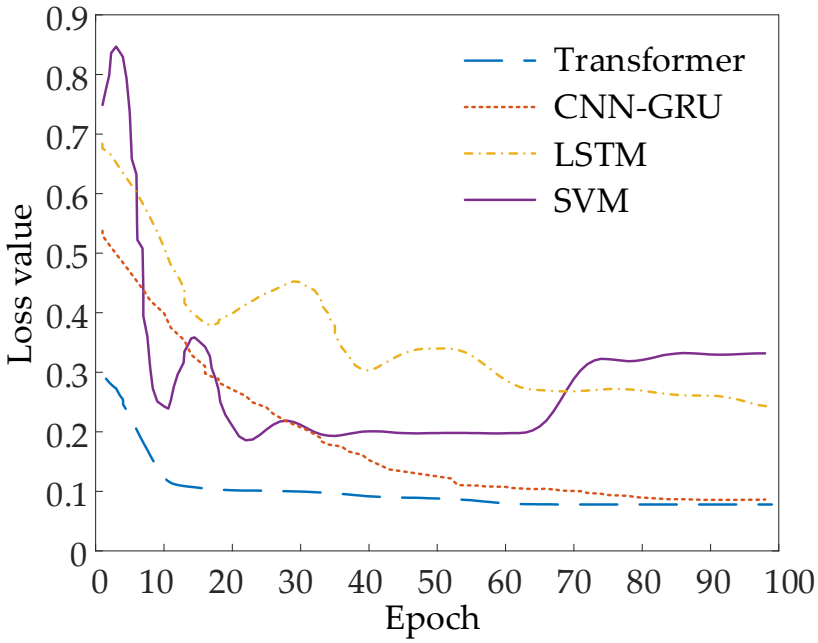


Figure 9. Changes in loss values during the Transformer training process.

The confusion matrix functions as a visual instrument for portraying the classifications generated using a given model across distinct categories. It juxtaposes the model's predictions against the true labels on a categorical basis, thereby providing a matrix that explains the classification accuracy.

The metrics encompassed in Table 4, alongside the depictions in Figure 8, indicate that the Transformer model consistently demonstrates heightened accuracy in discerning false data injection attacks across a spectrum of attacked bus counts. In Figure 9, it is evident that, within the scope of this study, the introduced Transformer algorithm has superior performance with regard to the loss function. Our model consistently achieves lower loss values when addressing the given task in comparison to the other three algorithms. This outcome serves as an indicative measure of our algorithm's exceptional performance in the context of FDIA prediction, thereby offering a more effective solution to the problem at hand. Such findings further substantiate the superiority and potential of our algorithm. This model excels in the precise identification and accurate classification of samples tainted with false data. This underscores the Transformer model's capability to maintain an elevated level of accuracy, notwithstanding the varying degrees of malicious actor behaviors.

4.4.2. Different Attack Intensity

In this experiment, a baseline 2-norm value of 1 was employed, and attack vectors characterized by five distinct 2-norm intensities were developed, ranging from 0.5 to 2.5. The primary objective was the exploration of how these varied 2-norm magnitudes impact the accuracy of the four distinct detection methods. Figure 10 illustrates the fluctuation in accuracy across the four detection methods, as the 2-norm values undergo alteration.

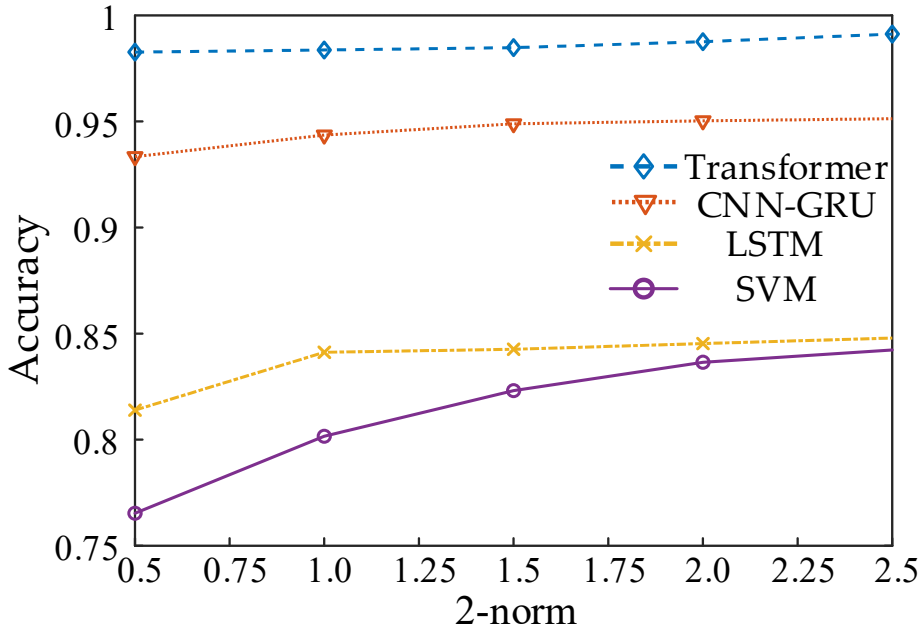


Figure 10. Accuracy of four deep learning methods under different attack intensities.

The experimental results reveal a clear trend: As the 2-norm value of the attack vectors increases, the accuracy of the detection method gradually increases. This can be explained by the fact that the increase in attack intensity leads to an increase in the differentiation between normal and false data. When the 2-norm intensity is low, the two types of data may be more similar, thus increasing the likelihood of misclassification. However, as the attack intensity increases, the distortion characteristics of the attack data are more pronounced, making it easier for the detection method to distinguish them from normal data, and the accuracy of the detection increases. Therefore, this method overcomes the shortcomings of traditional state estimation methods and is able to detect FDIAs efficiently, and the experiments verify the advantages of the method in terms of detection efficiency and classification accuracy.

5. Conclusions

In this study, a data-balancing model based on GAN-GRU and an FDIA detection model with the Transformer architecture was introduced. Specifically, false data injection attacks were formulated with multiple attack modes contingent upon the quantum of network information at the disposal of potential attackers. To address the prevalent issue of data imbalance in practical scenarios, a data-balancing model rooted in GAN-GRU was employed. Through this approach, the generation of requisite small-scale data samples is facilitated, culminating in a balanced dataset. The outcomes of this treatment led to a pronounced improvement in the accuracy and dependability of the FDIA detection model. In the subsequent phase of experimentation, the FDIA detection model based on the Transformer was established, utilizing diverse attack vectors as inputs. A multi-label classification strategy was employed by the model to assign suitable labels to each dataset. Subsequently, the FDIA detection and classification performance of the Transformer was assessed independently using both balanced and unbalanced test datasets. The empirical

findings substantiated the superior performance and detection rate of the Transformer-based detection model when contrasted against traditional counterparts.

In summary, these experimental investigations affirm the effectiveness of the data-balancing model based on GAN-GRU and the FDIA detection model based on the Transformer by providing evidence of its increased accuracy and dependability. This substantiates their potential for pertinent applications and paves the way for future research in the realm of network security. Nonetheless, it is important to note that while this study focuses on the performance of the FDIA detection model, practical applications necessitate a consideration of factors such as real-time responsiveness and scalability of the model.

Abbreviations

Abbreviation	Full Name
CPPS	Cyber–physical power systems
FDIA	False data injection attack
GAN-GRU	Generating adversarial network-gated recurrent units
RNN	Recurrent neural networks
CNN-GRU	Convolutional neural network-gated recurrent unit
LSTM	Long short-term memory network
SVM	Support vector machine
MSE	Mean-squared error

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

Funding: This research was supported in part by the National Science Foundation of China (No. 51907097) and the National Key R&D Program of China (No. 2021YFB4000500).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Hasan, M.K.; Habib, A.A.; Shukur, Z.; Ibrahim, F.; Islam, S.; Razzaque, M.A. Review on Cyber-Physical and Cyber-Security System in Smart Grid: Standards, Protocols, Constraints, and Recommendations. *J. Netw. Comput. Appl.* **2023**, *209*, 103540. [CrossRef]
2. Dibaji, S.M.; Pirani, M.; Flamholz, D.B.; Annaswamy, A.M.; Johansson, K.H.; Chakrabortty, A. A Systems and Control Perspective of CPS Security. *Annu. Rev. Control.* **2019**, *47*, 394–411. [CrossRef]
3. Tian, J.; Wang, B.; Li, T.; Shang, F.; Cao, K.; Guo, R. TOTAL: Optimal Protection Strategy Against Perfect and Imperfect False Data Injection Attacks on Power Grid Cyber–Physical Systems. *IEEE Internet Things J.* **2021**, *8*, 1001–1015. [CrossRef]
4. Liang, G.; Weller, S.R.; Zhao, J.; Luo, F.; Dong, Z.Y. The 2015 Ukraine Blackout: Implications for False Data Injection Attacks. *IEEE Trans. Power Syst.* **2017**, *32*, 3317–3318. [CrossRef]
5. Liu, Y.; Ning, P.; Reiter, M.K. False Data Injection Attacks against State Estimation in Electric Power Grids. *ACM Trans. Inf. Syst. Secur.* **2011**, *14*, 1–33. [CrossRef]
6. Ahmed, M.; Pathan, A.-S.K. False Data Injection Attack (FDIA): An Overview and New Metrics for Fair Evaluation of Its Countermeasure. *Complex Adapt. Syst. Model.* **2020**, *8*, 4. [CrossRef]
7. Liang, G.; Zhao, J.; Luo, F.; Weller, S.R.; Dong, Z.Y. A Review of False Data Injection Attacks Against Modern Power Systems. *IEEE Trans. Smart Grid* **2017**, *8*, 1630–1638. [CrossRef]
8. Musleh, A.S.; Chen, G.; Dong, Z.Y. A Survey on the Detection Algorithms for False Data Injection Attacks in Smart Grids. *IEEE Trans. Smart Grid* **2020**, *11*, 2218–2234. [CrossRef]
9. Li, X.; Li, X.; Lu, Z. Static Detection of False Data in the Power Grid by Fusing Structure and Attributes of Node. *J. Electr. Eng. Technol.* **2023**, 1–12. [CrossRef]

10. Zhen, R.; Lin, L.; Youbo, L.; Jie, L.; Diangang, W.; Huang, L. Coordinated attack model of cyber-physical power system consisering false load data injection. *Electr. Power Autom. Equip.* **2019**, *39*, 181–187.
11. Habibi, M.R.; Baghaee, H.R.; Blaabjerg, F.; Dragicevic, T. Secure MPC/ANN-Based False Data Injection Cyber-Attack Detection and Mitigation in DC Microgrids. *IEEE Syst. J.* **2022**, *16*, 1487–1498. [CrossRef]
12. Wang, X.; Luo, X.; Zhang, Y.; Guan, X. Detection and Isolation of False Data Injection Attacks in Smart Grids via Nonlinear Interval Observer. *IEEE Internet Things J.* **2019**, *6*, 6498–6512. [CrossRef]
13. Zhang, Z.; Hu, J.; Lu, J.; Cao, J.; Alsaadi, F.E. Preventing False Data Injection Attacks in LFC System via the Attack-Detection Evolutionary Game Model and KF Algorithm. *IEEE Trans. Netw. Sci. Eng.* **2022**, *9*, 4349–4362. [CrossRef]
14. Cui, J.; Gao, B.; Guo, B. A Novel Detection and Defense Mechanism against False Data Injection Attack in Smart Grids. *IET Gener. Transm. Distrib.* 2023, *Early View*. [CrossRef]
15. Yu, J.J.Q.; Hou, Y.; Li, V.O.K. Online False Data Injection Attack Detection With Wavelet Transform and Deep Neural Networks. *IEEE Trans. Ind. Inf.* **2018**, *14*, 3271–3280. [CrossRef]
16. Wang, Y.; Zhang, Z.; Ma, J.; Jin, Q. KFRNN: An Effective False Data Injection Attack Detection in Smart Grid Based on Kalman Filter and Recurrent Neural Network. *IEEE Internet Things J.* **2022**, *9*, 6893–6904. [CrossRef]
17. Chen, B.; Wu, Q.H.; Li, M.; Xiahou, K. Detection of False Data Injection Attacks on Power Systems Using Graph Edge-Conditioned Convolutional Networks. *Prot. Control. Mod. Power Syst.* **2023**, *8*, 16. [CrossRef]
18. Wang, L.; Zhu, Y.; Du, W.; Fu, B.; Wang, C.; Wang, X. A Novel Model-Based Reinforcement Learning for Online Anomaly Detection in Smart Power Grid. *Int. Trans. Electr. Energy Syst.* **2023**, *2023*, 6166738. [CrossRef]
19. Choi, J.; Roshanzadeh, B.; Martínez-Ramón, M.; Bidram, A. An Unsupervised Cyberattack Detection Scheme for AC Microgrids Using Gaussian Process Regression and One-class Support Vector Machine Anomaly Detection. *IET Renew. Power Gener.* **2023**, *17*, 2113–2123. [CrossRef]
20. Gao, X.; Yang, X.; Meng, L.; Wang, S. Fast Economic Dispatch with False Data Injection Attack in Electricity-Gas Cyber–Physical System: A Data-Driven Approach. *ISA Trans.* **2023**, *137*, 13–22. [CrossRef]
21. Hu, P.; Gao, W.; Li, Y.; Hua, F.; Qiao, L.; Zhang, G. Detection of False Data Injection Attacks in Smart Grid Based on Joint Dynamic and Static State Estimation. *IEEE Access* **2023**, *11*, 45028–45038. [CrossRef]
22. Zhao, J.; Qi, J.; Huang, Z.; Meliopoulos, A.P.S.; Gomez-Exposito, A.; Netto, M.; Mili, L.; Abur, A.; Terzija, V.; Kamwa, I.; et al. Power System Dynamic State Estimation: Motivations, Definitions, Methodologies, and Future Work. *IEEE Trans. Power Syst.* **2019**, *34*, 3188–3198. [CrossRef]
23. Mode, G.R.; Calyam, P.; Hoque, K.A. False Data Injection Attacks in Internet of Things and Deep Learning Enabled Predictive Analytics. *arXiv* **2019**, arXiv:1910.01716.
24. Wang, S.; Bi, S.; Zhang, Y.-J.A. Locational Detection of the False Data Injection Attack in a Smart Grid: A Multilabel Classification Approach. *IEEE Internet Things J.* **2020**, *7*, 8218–8227. [CrossRef]
25. Cho, K.; van Merrienboer, B.; Bahdanau, D.; Bengio, Y. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *arXiv* **2014**, arXiv:1409.1259.
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2023**, 30. [CrossRef]
27. Zimmerman, R.D.; Murillo-Sanchez, C.E.; Thomas, R.J. MATPOWER: Steady-State Operations, Planning, and Analysis Tools for Power Systems Research and Education. *IEEE Trans. Power Syst.* **2011**, *26*, 12–19. [CrossRef]
28. He, Y.; Li, L.; Qian, H.; Yao, S. CNN-GRU Based Fake Data Injection Attack Detection Method for Power Grid. In Proceedings of the 2022 2nd International Conference on Electrical Engineering and Control Science (IC2ECS), Nanjing, China, 16 December 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 408–411.
29. Zhao, Y.; Jia, X.; An, D.; Yang, Q. LSTM-Based False Data Injection Attack Detection in Smart Grids. In Proceedings of the 2020 35th Youth Academic Annual Conference of Chinese Association of Automation (YAC), Zhanjiang, China, 16 October 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 638–644.
30. Yang, L.; Zhai, Y.; Li, Z. Deep Learning for Online AC False Data Injection Attack Detection in Smart Grids: An Approach Using LSTM-Autoencoder. *J. Netw. Comput. Appl.* **2021**, *193*, 103178. [CrossRef]
31. Chen, Z.; Zhu, J.; Li, S.; Luo, T. Detection of False Data Injection Attack in Automatic Generation Control System with Wind Energy Based on Fuzzy Support Vector Machine. In Proceedings of the IECON 2020 The 46th Annual Conference of the IEEE Industrial Electronics Society, Singapore, Singapore, 18 October 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 3523–3528.
32. Cabelin, J.D.; Alpano, P.V.; Pedrasa, J.R. SVM-Based Detection of False Data Injection in Intelligent Transportation System. In Proceedings of the 2021 International Conference on Information Networking (ICOIN), Jeju Island, Republic of Korea, 13 January 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 279–284.
33. Xiong, X.; Hu, S.; Sun, D.; Hao, S.; Li, H.; Lin, G. Detection of False Data Injection Attack in Power Information Physical System Based on SVM–GAB Algorithm. *Energy Rep.* **2022**, *8*, 1156–1164. [CrossRef]