IEEE Open Access Journal of
Power and Energy

Received 7 January 2025; revised 1 April 2025; accepted 13 June 2025.
Date of publication 16 June 2025; date of current version 2 July 2025.

Digital Object Identifier 10.1109/OAJPE.2025.3580108

Guaranteed False Data Injection Attack Without Physical Model

CHENHAN XIAO[1], (Student Member, IEEE),
NAPOLEON COSTILLA-ENRIQUEZ[2], (Member, IEEE),
AND YANG WENG[1], (Senior Member, IEEE)

[1]School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85281 USA
[2]Electric Power Research Institute, Arizona State University, Tempe, AZ 85281 USA

CORRESPONDING AUTHOR: Y. WENG (yang.weng@asu.edu)

The work was supported in part by the Israel-U.S. Binational Industrial Research and Development (BIRD) Foundation and in part by the U.S. National Science Foundation.

ABSTRACT Smart grids are increasingly vulnerable to False Data Injection Attacks (FDIAs) due to their growing reliance on interconnected digital systems. Many existing FDIA techniques assume access to critical physical model information, such as grid topology, to successfully bypass Bad Data Detection (BDD). However, this assumption is often impractical, as utilities may restrict access to this data, or the evolving nature of distribution grids—particularly with the integration of renewable energy—can render this information unavailable. Current methods that address the absence of physical model lack formal guarantees for BDD evasion. To bridge this gap, we propose a novel physical-model-free FDIA framework that 1) bypasses BDD with formal guarantees and 2) maximizes the attack impact without requiring explicit physical model. Our approach leverages an autoencoder (AE) with a regularized latent space to enforce physical consistency, using historical measurements to replicate the residual error distribution, ensuring BDD evasion. Additionally, we integrate a Generative Adversarial Network (GAN) to explore the measurement manifold and induce the most significant state changes, enhancing the impact of the attack. The key innovation lies in the AE-GAN hybrid model's ability to replicate the residual error distribution while maximizing attack efficacy, offering a performance guarantee that existing methods lack. We validate our method across 11 representative grid systems, using real power profiles simulated in MATPOWER, and demonstrate its consistent ability to bypass BDD by preserving the residual error distribution. The results highlight the robustness and generalizability of the proposed FDIA framework.

INDEX TERMS False data injection attack, state estimation, bad data detector, no physical model, auto-encoder, generative adversarial network.

I. INTRODUCTION

IN MODERN power systems, the integration of digital systems and communication enables real-time monitoring and control. However, such digitization also exposed power systems to vulnerabilities exploitable by malicious attackers [1], as evidenced by incidents such as the 2015 cyber attack on Ukraine's electricity infrastructure [2], the 2018 attack on the U.S. power grid [3], and the recent 2022 cyber-attack on energy entities in U.S. [4]. Nowadays, such attacks are even more prevalent in distribution systems, as data storage and cloud services are increasingly outsourced to third-party companies, and the security and integrity of measurements become more susceptible to breaches [5]. As a consequence, it is critical to study False Data Injection Attacks (FDIAs) so proper defense protocol can be enforced.

In FDIAs, adversaries leverage the leaked system measurements and physical model to compromise state estimation algorithms [6]. Such adversaries lead to severe consequences such as power outages [7], line congestion [8], and economic disruption [9]. Traditionally, many model-based FDIA methods offer theoretical guarantees of stealth by precisely crafting attack vectors that reproduce the residual structure of system measurements [10]. These guarantees rely on complete knowledge of the system model, enabling attackers to manipulate measurements while maintaining the same likelihood of bypassing the Bad Data Detector (BDD) as genuine measurements. However, these model-based attacks necessitate access to fundamental power system details [11], including power system topology, parameters, and a state estimator model. However, this reliance on physical model information

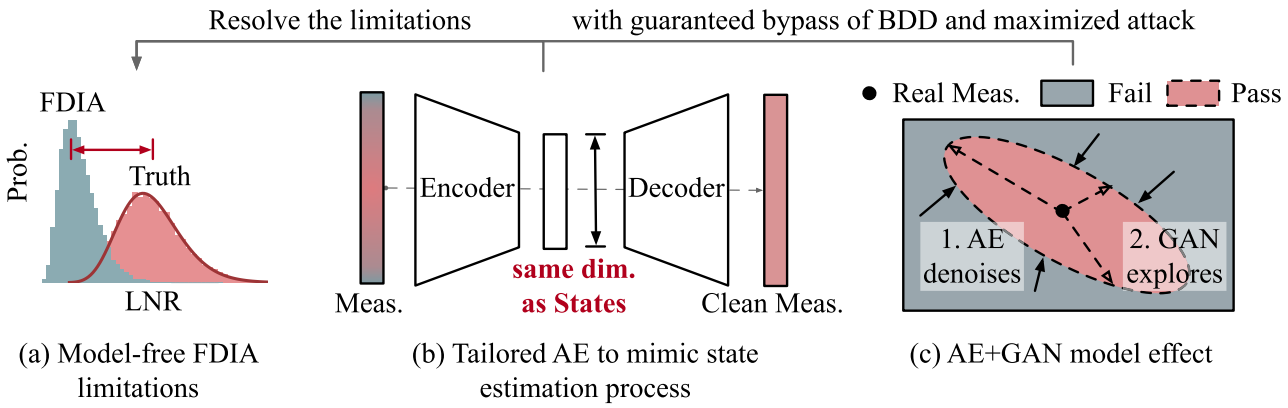


FIGURE 1. Overview of the proposed FDIA model: (a) Traditional physical-model-free FDIA methods face challenges in preserving the Chi-squared distribution of LNR values, thus lack a guaranteed bypass of BDD. (b) The tailored AE structure whose latent layer possesses the same dimensionality as the system states to mimic the state estimation process in terms of denoising and projecting noisy measurements onto a physically meaningful manifold. (c) The principle how the hybrid model of AE and GAN can generate fake measurements with guaranteed bypass of BDD and maximized attack impact.

is increasingly impractical in modern power grids, where system operators diligently safeguard such information to prevent leakage [12]. Additionally, in some distribution grids, even the operators may lack complete knowledge of physical model due to outdated information, evolving infrastructure, and infrequent or missing updates [13]. This challenge is further exacerbated by the rapid integration of renewable energy sources [14]. This is because some of them belong to third party, which does not synchronize information with the utility [5].

In contrast to traditional model-based attacks, physical-model-free attacks, also called data-driven attacks, are crafted without prior system knowledge of the power grid [15], [16]. In this field, the key is to estimate the system topology inexplicitly. For example, there is work that estimates the system Jacobian matrix through linear independent component analysis [17], PCA approximation [18], [19], and low-rank matrix approximation [20]. Recently, matrix reconstruction using eigenvalue decomposition is also utilized [21], [22] to generate attack measurements. Furthermore, machine learning approaches are also utilized to generate false measurements, such as employing auto-encoders [23], [24] and generative adversarial networks [25], [26], [27].

While these physical-model-free attacks have shown advancements in reducing the largest normalized residual (LNR) value in attack measurements to bypass the BDD, there were no guarantees of their effectiveness. Specifically, they typically involve a learning or estimation process of the unknown physical model, which usually suffers from error propagation issues. Eventually, such error propagation can disrupt the inherent Chi-squared distribution that LNR values typically follow, leading to an unguaranteed bypass of the BDD. This issue is demonstrated in Fig. 1(a) and discussed in [28], compromising the effectiveness of physical-model-free FDIA stealthiness [29]. For instance, [20] presented an attack using reduced measurement information based on a low-rank matrix approximation. However, this attack can only achieve a success rate of bypassing the BDD lower than the original data.

To address these limitations, we propose an alternative autoencoder (AE) design that eliminates the need for physical model while providing a theoretical guarantee of BDD evasion in distribution, by replicating the residual behavior of genuine measurements. The core idea involves designing a special AE that mimics the power system state estimation process in terms of denoising and projecting noisy measurements onto a physically meaningful manifold. This mimicking allows the model to produce denoised measurements with small LNR values, a key metric for bypassing the BDD. Specifically, we tailor the latent layer of AE model to possess the same dimensionality as the system state. This process is demonstrated in Fig. 1(b). In doing so, the AE model can capture the essential low-dimensional features to reconstruct the system measurements, while effectively denoises the measurements. This is because the state of the power system is defined as the minimum number of variables that can recover the measurements. Overall, the denoising function of the AE model will remove measurement noises similar to the state estimation process, eventually producing similar LNR values for attack measurements.

Then, we engineer the attack model to maximize attack impact by utilizing a Generative adversarial network (GAN) module. Specifically, we train the GAN model on historical measurement data. This enables the GAN to generate diverse, yet realistic, attack measurements that adhere to the inherent distribution of legitimate system measurements. Additionally, a penalty term is incorporated to discourage deviations between the original and attack measurements, aiming to maximize the impact of the attack. Fig. 1(c) illustrates the operation of our proposed hybrid AE-GAN model.

When presented with real system measurements, the AE confines the generated fake measurements to a region that effectively bypasses the BDD by leveraging its denoising capabilities. Subsequently, the GAN module, incorporating a penalty term, explores this permissible region and identifies the position that leads to the most significant attack impact. In mathematical terms, by carefully integrating these components, the hybrid model generates attack measurements whose residuals replicate the original Chi-squared distribution of genuine measurements. This forms the basis of our statistical guarantee of BDD evasion, ensuring that, the attacks are as likely to bypass the BDD as legitimate measurements.

To assess the efficacy of our proposed FDIA approach across diverse system configurations, we conducted simulations using MATPOWER [30] on 11 testbeds encompassing both transmission and distribution grids. Our evaluation quantitatively validates the advantage of our method with respect to other physical-model-free FDIA baselines, based on metrics of the BDD passing rates. The rest of the paper is organized as follows: Section II introduces the preliminaries of the FDIA problem, Section III presents our proposed physical-model-free FDIA model, Section IV shows numerical experiments and Section V concludes the paper.

II. PRELIMINARIES

Before detailing our FDIA strategy, this section provides an overview of state estimation, bad data detection, and the limitations of traditional FDIA. We highlight that, while our FDIA approach is applicable to both AC and DC systems, for this paper, we focus on demonstrating its application to DC systems, i.e., linearized power flow systems, to facilitate theoretical derivations.

A. DC POWER FLOW STATE ESTIMATION

In FDIAs, the attacker aims to inject malicious data into the grid measurements $\mathbf{z} = (z_1, \cdots, z_m) \in \mathbb{R}^m$ to compromise the accuracy of the state estimation process [31]. For DC state estimation, measurements $\mathbf{z}$ are determined as $\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{e}$, where $\mathbf{x} = (x_1, \cdots, x_n) \in \mathbb{R}^n$ are system states such as voltage angles and magnitudes, and $\mathbf{H} \in \mathbb{R}^{m\times n}$ is the Jacobian matrix defined by the grid topology. Measurement noise $\mathbf{e} \in \mathbb{R}^m$ captures sensor-related disturbances (e.g., SCADA or PMU noise) during the measurement collection process. The noise is typically assumed to be Gaussian distributed as $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$, where $\mathbf{R} = \text{diag}(\sigma_1^2, \cdots, \sigma_m^2)$ is a diagonal covariance matrix and σ_i^2 is the variance of i-th noise [32]. When system operators collect measurements $\mathbf{z}$, they recover the states $\mathbf{x}$ by solving the state estimator [32] as:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \sum_{i=1}^{m} \frac{(z_i - \mathbf{H}_i\mathbf{x})^2}{\sigma_i^2}, \tag{1}$$

where $\mathbf{H}_i$ is the i-th row of matrix $\mathbf{H}$. Furthermore, the solution to Eq. (1) can be explicitly written as [32]

$$\hat{\mathbf{x}} = (\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{z}. \tag{2}$$

B. BAD DATA DETECTOR (BDD)

In FDIAs, attackers introduce falsified data into the measurements $\mathbf{z}$, thereby disrupting the accurate estimation of states $\mathbf{x}$ as outlined in Eq. (1). Given that numerous power system operations (e.g. economic dispatch and contingency analysis) depend on accurate state estimation results [33], compromised estimations can result in erroneous system control decisions. In practice, in order to assess if $\mathbf{z}$ contains bad or wrong data due to telecommunication failures, meter errors, or even FDIAs [32], [34], the system operators often calculate the squared measurement residual error

$$\|\mathbf{z} - \hat{\mathbf{z}}\|_2^2 = \|\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}\|_2^2 = \|\mathbf{S}\mathbf{z}\|_2^2, \tag{3}$$

where $\mathbf{S} = \mathbf{I} - \mathbf{H}(\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{R}^{-1}$ is the residual sensitivity matrix [32] and has the property $\mathbf{S}\mathbf{H} = \mathbf{0}$.

If the measurement $\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{e}$ does not contain bad data, the *largest normalized residual* (LNR) approximately follows a Chi-squared distribution of $m - n$ degrees of freedom [32]:

$$\text{LNR}(\mathbf{z}) := \sum_{i=1}^{m} \frac{(\mathbf{S}_i\mathbf{z})^2}{\sigma_i^2} = \sum_{i=1}^{m} \frac{(\mathbf{S}_i\mathbf{e})^2}{\sigma_i^2} \sim \chi_{m-n}^2, \tag{4}$$

where the Chi-squared distribution arises from the Gaussianity assumption of the noise $\mathbf{e}$. The degrees of freedom are attributed to the fact that, given the necessity for at least n measurements to satisfy power balance equations, a maximum of $m - n$ measurement noises can be linearly independent [32]. As Chi-squared test is formed by Gaussian noises, we make the following Assumption 1 to facilitate the derivation of our theoretical results which does not impact the fundamental design principles of our proposed FDIA. Similar assumptions have been adopted in various studies [35], [36].

Assumption 1: In a power grid, the power measurements $\mathbf{z}$ follows a Gaussian distribution as $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma} + \boldsymbol{R})$, where $\mathbf{\Sigma} = \text{diag}(\delta^2, \cdots, \delta^2)$ and $\mathbf{R} = \text{diag}(\sigma^2, \cdots, \sigma^2)$ are diagonal covariance matrices.

System operators implement a bad data detector (BDD) utilize the Chi-squared distribution test as follows. (1) Choose a significance level, denoted as α (e.g., 0.05). (2) Evaluate the normalized residual error LNR($\mathbf{z}$) and compare it to the critical value $\tau = \chi^2_{(m-n),1-\alpha}$ obtained from the Chi-squared distribution table. If LNR($\mathbf{z}$) $\geq \tau$, this raises suspicion of bad data; otherwise, the measurements are considered free from the influence of bad data.

C. CHALLENGES AND PROBLEM DEFINITION

In model-based FDIAs [10], [11] where attackers have access to the Jacobian matrix $\mathbf{H}$ (also referred to as the physical model), they can modify the measurements $\mathbf{z}$ without affecting the LNR value [10]. Specifically, they can inject attack data as $\mathbf{z}_a = \mathbf{z} + \mathbf{H}\mathbf{c}$, where $\mathbf{c}$ is an arbitrary vector [11], [37]. From Eq. (3), the residual error of $\mathbf{z}_a$ remains identical to that of the original measurement, i.e., $\|\mathbf{S}(\mathbf{z} + \mathbf{H}\mathbf{c})\|_2^2 = \|\mathbf{S}\mathbf{z} + \mathbf{0}\|_2^2 = \|\mathbf{S}\mathbf{z}\|_2^2$. Consequently, the manipulated measurements $\mathbf{z}_a$ will bypass the BDD if the original measurements $\mathbf{z}$ bypass the BDD.

However, a more realistic scenario in modern power grids is that attackers lack access to physical model information (e.g., line parameters, grid topology), as many utilities diligently safeguard such information [12]. Motivated by this challenge, we formally define this paper's study as follows.

- **Given**: Historical time-series system measurements $\mathbf{z}$ without physical model $\mathbf{H}$.
- **Generate**: Attack measurements $\mathbf{z}_a$ whose LNR value still follows the Chi-squared distribution χ^2_{m-n}.

By generating LNR values adhering to the same Chi-squared distribution as the real data, we aim to provide a performance guarantee of bypassing the BDD. This is because past FDIA methods without physical model lack such a guarantee [17], [18], [19], [20], [24], [25], [26], [27] and we want to bridge this gap.

III. PERFORMANCE GUARANTEED ATTACK WITHOUT PHYSICAL MODEL

To guarantee a successful FDIA, the tampered measurements need to bypass the BDD, as discussed in Section II-B. The key is to ensure that the tampered data possesses approximately the same LNR value as real measurements. While model-based approaches (see Section II-C) can rely on physical model information (i.e., the Jacobian matrix $\mathbf{H}$) to achieve so, we lack such information in model-free scenarios.

A. STATE-PRESERVED RESIDUAL ERROR MINIMIZATION

We note that the residual error in Eq. (3) is equivalent to the **minimal distance of z to the manifold** $\mathcal{H}$ where $\mathcal{H} : \{\mathbf{Hx}|\mathbf{x} \in \mathbb{R}^n\}$ is defined by the Jacobian matrix $\mathbf{H}$. Model-based FDIAs leverage the matrix $\mathbf{H}$ to construct an attack vector $\mathbf{a} = \mathbf{Hc} \in \mathcal{H}$ ensuring that the minimal distance (i.e., residual error) of $\mathbf{z}$ and $\mathbf{z} + \mathbf{a}$, to the manifold $\mathcal{H}$, remains unchanged. While we don't have the knowledge of matrix $\mathbf{H}$, we recognize there could be alternative mappings $\mathbf{H}'$ that defines the same manifold $\mathcal{H}$ as long as the span of the columns of $\mathbf{H}'$ remains the same as that of $\mathbf{H}$.

This recognition leads us to utilize the auto-encoder (AE), which represents a machine-learning based method to learn the manifold $\mathcal{H}$ from historical measurements $\mathbf{z}$. In AE, an encoder network maps the input measurements to the latent "states" space, and a decoder network tries to reconstruct the input measurements from the latent "states" space [38]. Denoting the overall AE model by a function $\mathrm{AE}(\cdot)$, it is trained with the loss function:

$$\min_{\theta_{\mathrm{AE}}} \mathbb{E}_{\mathbf{z}} \|\mathbf{z} - \mathrm{AE}(\mathbf{z}; \theta_{\mathrm{AE}})\|^2 \tag{5}$$

with network parameters θ_{AE}. Upon convergence of such training, the decoder mapping in AE model is expected to span the manifold $\mathcal{H}$, thus producing small residual errors for attack measurements to bypass the BDD.

To achieve so, we tailor the latent space in the AE model to possess the same dimensionality as the real system state. This information, unlike the exact Jacobian matrix $\mathbf{H}$, if often available to attackers in many scenarios. By doing so, the latent "states" will form a manifold and align with the definition of states in power systems, e.g., the minimum set of variables that can uniquely define all the measurements in the systems. This design is shown in Fig.1 (b). Mathematically, we use the linear AE model as an example to provide a proof of the reduced residual error of attack measurements.

For demonstrating that the linear AE model can reduce the residual error, we start from exploring its connection to Principal Component Analysis (PCA). A linear AE model contains a linear network $\boldsymbol{A} \in \mathbb{R}^{n\times m}$ as encoder and another linear network $\boldsymbol{B} \in \mathbb{R}^{m\times n}$ as decoder. Notice that the dimension of the latent space is set to n, i.e., the number of real states. Suppose the historical measurements matrix $\boldsymbol{Z} = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_T] \in \mathbb{R}^{m\times T}$ contains $T > m$ data points and is properly zero-centered and scaled, the linear AE model wishes to reconstruct the original data as $\boldsymbol{Z} \approx \boldsymbol{BAZ}$. This AE model is trained as $\min_{\boldsymbol{A},\boldsymbol{B}} \|\boldsymbol{Z} - \boldsymbol{BAZ}\|_2^2$. The work in [39] identified the connection between AE model and PCA. When PCA uses the top eigenspace of $\boldsymbol{XX}^\top$ to approximate the dataset, any $\boldsymbol{B}$ at a local minimizer recovers the top rank-n eigenspace of $\boldsymbol{ZZ}^\top$ under mild nondegeneracy conditions. This is presented in Lemma 1.

Lemma 1: [Equivelancy of autoencoder and PCA]. Suppose that $\boldsymbol{Z} \in \mathbb{R}^{m\times T}$ (with $T > m$) satisfies that $\boldsymbol{ZZ}^\top$ has distinct eigenvalues. Then, at any local minimizer of the optimization

$$\min_{\boldsymbol{A}\in\mathbb{R}^{n\times m},\boldsymbol{B}\in\mathbb{R}^{m\times n}} \|\boldsymbol{Z} - \boldsymbol{BAZ}\|_2^2, \tag{6}$$

$\boldsymbol{B}$ spans the top rank-n eigenspace of $\boldsymbol{ZZ}^\top$.

To explicitly calculate the residual error of attack measurements crafted from the decoder network, we utilize the singular value decomposition (SVD) of $\boldsymbol{Z}$. Here, we assume that the data has been properly centered and scaled for this analysis. Suppose the SVD of $\boldsymbol{Z}$ is expressed as $\boldsymbol{Z} = \boldsymbol{U\Sigma V}^\top = \sum_{i=1}^{m} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, where $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{V} \in \mathbb{R}^{T\times m}$ are unitary matrices, and $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times m}$ is a diagonal matrix with non-negative singular entries. Due to Lemma 2, we note that the decoder network essentially reconstruct a rank-n approximation by keeping the leading n singular values and vectors and discarding the rest: $\mathrm{AE}(\boldsymbol{Z}) \approx \boldsymbol{Z}_{(n)} = \boldsymbol{U}_{(n)}\boldsymbol{\Sigma}_{(n)}\boldsymbol{V}_{(n)}^\top = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$. Here, $\boldsymbol{U}_{(n)} \in \mathbb{R}^{m\times n}$ is the truncated $\boldsymbol{U}$ matrix, $\boldsymbol{V}_{(n)} \in \mathbb{R}^{T\times n}$ is the truncated $\boldsymbol{V}$ matrix, and $\boldsymbol{\Sigma}_{(n)} \in \mathbb{R}^{n\times n}$ is the truncated $\boldsymbol{\Sigma}$ with the leading n singular values. In Lemma 2, we explicitly analyze the distribution of the resulted residual error of the proposed AE model.

Lemma 2: Suppose the collected power measurements $\mathbf{z}$ satisfy Assumption 1 in a power grid with m measurements and n system states. Let a linear autoencoder, trained via Eq. (5), have a hidden layer of dimension n. Then, the autoencoder behaves equivalently to principal component analysis (PCA) to recover the top rank-n ($n < m$) eigenspace of the measurement data, thus producing the LNR value (Eq. (4)) as

$$\mathrm{LNR}_{\mathrm{AE}} = \frac{1}{\sigma^2}\|\mathbf{Sz}_n\|_2^2 \sim \mathrm{Gamma}(\frac{m-n}{2}, 2\frac{n}{m}). \tag{7}$$

Proof: The LNR of the n-rank measurement is given as

$$\mathrm{LNR}_{\mathrm{AE}} = \frac{1}{\sigma^2}\|\mathbf{S}\mathbf{z}_n\|_2^2 = \frac{1}{\sigma^2}\|\mathbf{S}\mathbf{U}_{(n)}\mathbf{R}_{(n)}\mathbf{V}_{(n)}^{\top}\|_2^2 \tag{8}$$

$$= \frac{1}{\sigma^2}\|\mathbf{S}\mathbf{U}_{(n)}\mathbf{R}_{(n)}\|_2^2 = \frac{n}{m}\frac{1}{\sigma^2}\|\mathbf{S}\mathbf{U}_{(m)}\mathbf{R}_{(m)}\|_2^2 \tag{9}$$

$$= \varepsilon \sim \frac{n}{m}\chi^2(m-n) := \mathrm{Gamma}(\frac{m-n}{2}, 2\frac{n}{m}). \tag{10}$$

■

Lemma 2 demonstrates that the residual error of measurements generated by the AE is significantly reduced compared to the original measurements. This process essentially transforms the measurements from the gray area (likely failing the BDD) to the pink area (likely bypassing the BDD) as illustrated in Fig. 1(c). However, this approach also has a problem: the AE-generated attack measurements might become too similar to the original measurements, potentially rendering the attack undetectable but also ineffective. As discussed previously, we choose to utilize the generative adversarial network to explore the "bypass" region to identify its boundary and thereby maximize the attack impact.

B. ATTACK IMPACT MAXIMIZATION

To generate new data conform to specific underlying distributions, [40] introduced the idea of the generative adversarial network (GAN). GAN is a framework to implicitly learn the training data distribution so that one can sample from it and generate new data from that same distribution, in our case, the power system measurement distribution.

In this paper, to address well-documented challenges with GANs, such as vanishing gradients and the lack of convergence guarantees, we leverage the Wasserstein GAN proposed by [41]. Specifically, GAN conducts adversarial training between a generator G and a discriminator D using

$$\min_{G}\max_{D\in\mathcal{D}} \quad \mathbb{E}_{\mathbf{z}\sim\mathbb{P}_z}\mathbb{E}_{\boldsymbol{c}\sim\mathbb{P}_c}\left[D\left(\mathbf{z}\right) - D\left(G(\boldsymbol{c})\right)\right], \tag{11}$$

where $\mathcal{D}$ is the set of 1-Lipschitz functions [41]; $\mathbb{P}_z$ is the real measurement distribution; $\boldsymbol{c}$ is the noise sampled from standard Gaussian distribution $\mathbb{P}_c$. Eventually, the generator G has the ability to converge [41] to learn the real measurement distribution $\mathbb{P}_z$ from the set of historically observed measurements $\mathcal{Z} = \{\mathbf{z}_i = \mathbf{H}\mathbf{x}_i + \mathbf{e}_i \in \mathbb{R}^m\}_{i=1}^{L}$, where L is the size of the dataset. We note that, despite GAN's ability to recover the distribution of measurements $\mathbf{z}$, it does not inherently ensure the recovery of the associated residual error distribution. This phenomenon is illustrated in Lemma 3.

Lemma 3: Suppose the collected power measurements $\mathbf{z}$ satisfy Assumption 1 and are used to train the Generative Adversarial Network (GAN) defined in Eq. (11). Given sufficient training, the generator in the GAN can produce false measurements that follow the distribution $\tilde{\mathbf{z}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma} + \boldsymbol{R})$. Thus, the LNR value of false measurements satisfy

$$\mathrm{LNR}_{\mathrm{GAN}} = \frac{1}{\sigma^2}\|\mathbf{S}\tilde{\mathbf{z}}\|_2^2 \sim \mathrm{Gamma}(\frac{m-n}{2}, 2\frac{\delta^2+\sigma^2}{\sigma^2}). \tag{12}$$

Proof: The LNR of the original measurement is $\frac{1}{\sigma^2}\|\mathbf{S}\mathbf{z}\|_2^2 = \frac{1}{\sigma^2}\|\mathbf{S}(\mathbf{H}\mathbf{x}+\mathbf{e})\|_2^2 = \frac{1}{\sigma^2}\|\mathbf{S}\mathbf{e}\|_2^2 \sim \chi^2(m-n)$, where noise $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{R})$. Since $\tilde{\mathbf{z}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma} + \boldsymbol{R})$, the new LNR is

$$\frac{1}{\sigma^2}\|\mathbf{S}\tilde{\mathbf{z}}\|_2^2 \sim \frac{\delta^2+\sigma^2}{\sigma^2}\chi^2(m-n) \tag{13}$$

$$:= \mathrm{Gamma}(\frac{m-n}{2}, 2\frac{\delta^2+\sigma^2}{\sigma^2}). \tag{14}$$

■

The rationale underlying Lemma 3 is that while GAN can generate fake measurements $\tilde{\mathbf{z}}$ that match the exact original measurement distribution, the resulting data $\tilde{\mathbf{z}}$ lacks genuine power physical model, resembling pure noise. Consequently, the new residual error for $\tilde{\mathbf{z}}$ is notably larger compared to the original residual error, which depends only on the noise $\mathbf{e}$—a small component of the original data $\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{e}$. This increase is proportional to $\frac{\delta^2+\sigma^2}{\sigma^2}$, determined by the signal-to-noise ratio in the measurements $\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{e}$.

From Lemma 2 and Lemma 3, it's important to note that the application of AE yields a reduced LNR value compared to the original, while GAN tends to produce a higher LNR. With this insight and explicit derivation of LNR distributions when utilizing AE and GAN individually, an opportunity emerges: to effectively bypass the BDD with guarantees, it becomes evident that a hybrid model integrating the strengths of both AE and GAN is a natural progression. To connect the AE and GAN model, we design a hybrid loss

$$\min_{G}\max_{D\in\mathcal{D}} \quad \mathbb{E}_{\mathbf{z}\sim\mathbb{P}_z}\mathbb{E}_{\boldsymbol{c}\sim\mathbb{P}_c}\Big[D\left(\mathbf{z}\right) - D\left(G(\boldsymbol{c})\right) + \lambda_{\mathrm{AE}}\cdot\|G(\boldsymbol{c}) - \mathrm{AE}^*\left(G(\boldsymbol{c})\right)\|^2\Big], \tag{15}$$

where AE^* denotes the well-trained AE model and λ_{AE} serves as the hyperparameter balancing the contributions of AE and GAN. In Eq. (15), we configure the generator $G(\cdot)$ to use a real measurement as input and generate a tampered version as the output by setting the noise space to be the measurement space, i.e., $\mathbb{P}_c = \mathbb{P}_z$. That is, $\tilde{\mathbf{z}} = G(z)$ are the attack measurements modified from real measurements $\mathbf{z}$. This approach of feeding the generator real data as input has been explored in previous research [42] and has shown to be just as effective as using random noise as input. In this hybrid model, the collaboration between AE and GAN is two-fold: the AE works to reduce noise in the measurements and diminish the residual error, while the GAN emulates the distributional characteristics of the measurements. The equilibrium achieved in this hybrid model is elucidated in Theorem 1.

Theorem 1: Suppose the collected power measurements $\mathbf{z}$ satisfy Assumption 1, and let the hybrid model consist of an autoencoder (AE) and a generative adversarial network (GAN), trained jointly using the loss function defined in Eq. (15). The AE is configured with a hidden layer of dimension n, corresponding to the number of system states. Given sufficient training of the hybrid model, the generator in the hybrid model can produce attack measurements whose

residual errors match the theoretical Chi-squared distribution, provided the hyperparameter is chosen as $\lambda_{\text{AE}}^{opt} = \frac{\delta^2 \cdot m}{\sigma^2(n-1)}$.

Proof: Based on Lemma 2 and Lemma 3, the resultant distribution of the LNR value, obtained through training with the hybrid model, is $\frac{1}{1+\lambda_{\text{AE}}} \cdot \text{Gamma}(\frac{m-n}{2}, 2\frac{\delta^2+\sigma^2}{\sigma^2}) + \frac{\lambda_{\text{AE}}}{1+\lambda_{\text{AE}}} \cdot \text{Gamma}(\frac{m-n}{2}, 2\frac{n}{m})$. The density function of this LNR is

$$\frac{e^{-\frac{\text{LNR}}{(2\frac{\delta^2+\sigma^2}{\sigma^2(1+\lambda_{\text{AE}})}+2\frac{n\lambda_{\text{AE}}}{m(1+\lambda_{\text{AE}})})}}}{\Gamma(\frac{m-n}{2})(2\frac{\delta^2+\sigma^2}{\sigma^2(1+\lambda_{\text{AE}})}+2\frac{n\lambda_{\text{AE}}}{m(1+\lambda_{\text{AE}})})} x^{\frac{m-n}{2}-1}, \tag{16}$$

matching the density function of the Chi-squared distribution $\chi^2(m-n)$ when $\lambda_{\text{AE}}^{opt} = \frac{\delta^2 \cdot m}{\sigma^2(n-1)}$. ∎

With the residual error matching the original Chi-squared distribution, Theorem 1 highlights the potential of the hybrid model to bypass the BDD with guarantees. Aside from bypassing the BDD, we want to maximize the impact of the attack. Specifically, we want to incentivize the generator to produce attack measurements $\tilde{\mathbf{z}} = G(z)$ that differ significantly from the corresponding real measurements $\mathbf{z}$. To accomplish this, we incorporate a regularization term based on the $L2$ norm $\|\mathbf{z} - G(z)\|_2^2$:

$$\min_{G} \max_{D \in \mathcal{D}} \quad \mathbb{E}_{\mathbf{z}\sim\mathbb{P}_z} \mathbb{E}_{\tilde{\mathbf{z}}\sim\mathbb{P}_g} \Big[D(\mathbf{z}) - D(\tilde{\mathbf{z}}) + \lambda_{\text{AE}} \cdot \|\tilde{\mathbf{z}} - \text{AE}(\tilde{\mathbf{z}})\|_2^2 - \lambda_{\text{attack}} \cdot \|\mathbf{z} - \tilde{\mathbf{z}}\|_2^2 \Big], \tag{17}$$

where $\tilde{\mathbf{z}} = G(\mathbf{z})$, and λ_{attack} is the hyperparameter controlling the extent of the distance punishment. In Lemma 4, we discuss the effect of this hyperparameter: a larger penalty λ_{attack} leads to a larger attack impact. We note that, since the DC state estimation process can be linearly described by Eq. (2), a larger norm $\|\mathbf{z} - \tilde{\mathbf{z}}\|_2^2$ in the measurement space translates to a larger norm $\|\mathbf{x} - \tilde{\mathbf{x}}\|_2^2$ in the state space. Thus, the regularization term in Eq. (17) tends to make the state estimator to produce erroneous estimated states that deviate more significantly from the actual system states.

Lemma 4: The attack impact, quantified by $\|\mathbf{z} - \tilde{\mathbf{z}}\|_2^2$, has a lower bound $\mathcal{O}(\lambda_{\text{attack}})$, where λ_{attack} is the penalty parameter of $\mathbf{z}$ and $\tilde{\mathbf{z}}$ being too close.

To summarize, our proposed architecture is shown in Fig. 2 with two stages. First, an autoencoder is trained with historical measurement data to minimize the residual error in the state estimator. Second, the GAN is trained with the same data and the two regularization terms: (1) one incentivizes the GAN to produce measurements that will pass the residual error test and (2) another to maximize the attack's impact.

IV. NUMERICAL EXPERIMENTS

This section assesses the efficacy of our proposed FDIA without knowing physical model. Our analysis specifically investigates the residual error distribution to assess its influence on the Bad Data Detector (BDD) bypass performance.

Evaluation Metrics: For BDD, we calculate the *largest normalized residual* (LNR) of attacked data $\tilde{\mathbf{z}}$ following Eq. (3). Then, we compute the rate at which the LNR values fall

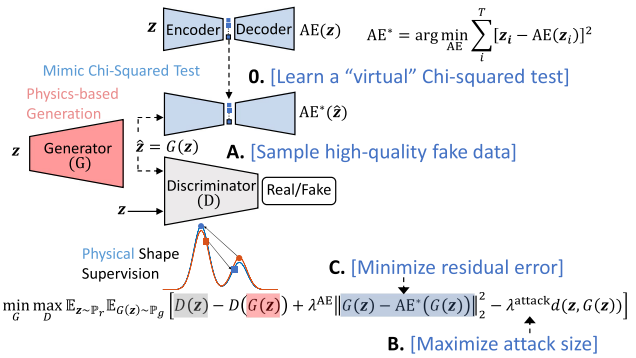


FIGURE 2. Proposed physical-model-free FDIA architecture with an AE-GAN hybrid structure.

below a Chi-squared distribution critical value:

$$\text{succ}_{\text{BDD}} = \mathbb{P}(\text{LNR}(\tilde{\mathbf{z}}) \leq \chi^2_{(m-n),1-\alpha}), \tag{18}$$

where m and n represent the dimensions of the measurement and state vectors, and α is the significance level. To assess how closely the LNR values adhere to the expected Chi-squared distribution $\chi^2_{(m-n)}$, we employ the Earth Mover's Distance (EMD) metric [43]. A lower EMD value signifies better alignment between the generated LNR distribution and the theoretical Chi-squared distribution.

Dataset Configuration: We evaluate across diverse system configurations. This includes experiments on transmission systems using the IEEE 14-bus, 30-bus, 39-bus, 57-bus, 118-bus, 200-bus network, and the Reliability Test System - Grid Modernization Lab Consortium (RTS-GMLC) system [44], [45]. The experiments also cover distribution grids including the IEEE 8-bus, 123-bus networks [14], [46], along with two representative European systems: a medium voltage network in an urban area (MV urban) and a low voltage network in a suburban area (LV suburban) [47], [48]. The time-series measurements $\mathbf{z}$ are simulated by solving DC power flow equations in MATLAB Power System Simulation Package (MATPOWER) [30]. To generate more authentic data, we incorporate real power profiles into our experiments. Specifically, we utilize the profile provided by Duquesne Light Company in Pittsburgh for our transmission grid experiments. For the distribution grid experiments, we employ the Pecan Street profile. To enhance the richness of our simulations, we introduce variability by scaling the load and generation profiles using randomly selected loading parameters. Additionally, we inject white noise into measurements with a standard deviation set to 0.02 p.u. [49], [50].

Implementing Details: In the linear AE model, the input consists of active power flow measurements on all branches, representing the physical measurements collected from the grid. For example, in the IEEE 14-bus system with 20 branches, the input vector has a dimension of 20. The output layer of the AE mirrors the input, aiming to reconstruct the original measurement vector with minimal reconstruction error. To capture the underlying system behavior, we set the

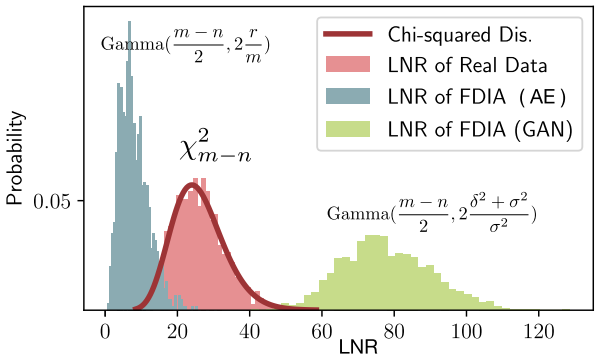


FIGURE 3. Empirical histogram of LNR values using AE/GAN compared to real LNR values in IEEE 14-bus system.

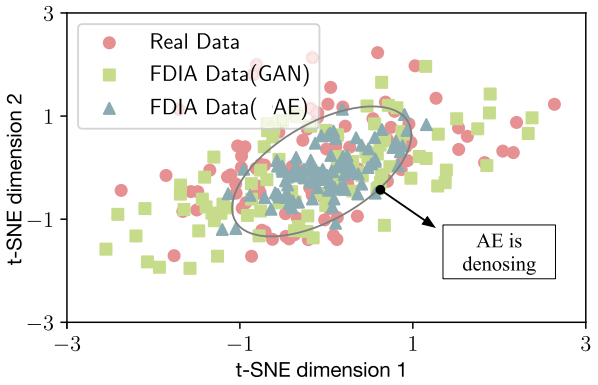


FIGURE 4. FDIA measurements using AE/GAN compared to real measurements in IEEE 14-bus system.

width of the latent layer to match the number of independent system states (n), which corresponds to the number of bus voltage angles excluding the reference bus. For instance, in the IEEE 14-bus system, the latent space dimension is set to 13. For the linear encoder, linear decoder, generator, and discriminator architectures, we employ five fully-connected layers where each layer comprises approximately ten neurons, and the neurons are activated through Rectified Linear Units (ReLU). We set the maximum number of training epochs to 300 for sufficient training. Additionally, for every 5 iteration, we train the generator so that we prioritize training the discriminator to allow for better convergence. For each iteration, we sample 50 mini-batches to compute gradients for advanced searching for parameters. We update these parameters using the Adam optimizer with a learning rate of 2×10^{-4}. After obtaining system measurements from MATLAB 2022b, the remaining calculations for FDIA are implemented using Python 3.8 on a personal computer with an Intel Core i7 processor clocked at 2.2 GHz, and 16 GB of RAM.

Baseline Methods: In the following experiments, we compare our FDIA approach with recent physical-model-free FDIA baselines, including a principal component analysis **PCA** approach to estimate the system Jacobian matrix [18], a low-rank matrix singular value decomposition approximation **SVD** to estimate the system Jacobian matrix [20], and a generative adversarial network-based approach **iAttackGen** to generate new attack measurements [27].

A. VERIFICATION OF RESIDUAL ERROR USING AE AND GAN ALONE

Before delving into our hybrid model combining AE and GAN, this subsection first examines the individual performances of AE and GAN modules to verify Lemma 2 and Lemma 3. To validate the two Lemmas, we illustrate in Fig. 3 the empirical histograms of LNR values obtained from FDIA data employing AE and GAN individually, comparing them to the LNR values of real measurements. The results demonstrate a noticeable deviation of the residual error distribution from the original Chi-squared distribution when using AE or GAN alone. Specifically, we compute the Earth Mover's Distance (EMD) between the calculated distribution in Fig. 3 and the theoretical ones from Lemma 2 and Lemma 3. The resulting low EMD values (see last column of Table 1) support our theoretical claims.

To gain insights into the above formation of LNR distributions in Fig. 3, we plot the FDIA measurements obtained through the individual application of AE and GAN in Fig. 4. Utilizing the t-SNE visualization technique [51], we reduce the m dimensional plots to 2 dimensions. The observations reveal that GAN tends to accurately capture the distribution of real measurements, whereas AE primarily focuses on learning the central, noiseless components of the real measurements. Specifically, GAN, by accurately learning the measurement distribution, results in FDIA data dominated by pure noise that closely follows this distribution, leading to a significantly large LNR. Conversely, AE, concentrating on the noiseless portion of real measurements, yields FDIA data with minimal noise, consequently resulting in a very small LNR.

B. EVALUATION OF ATTACK PERFORMANCE OF OUR FDIA

After evaluating the impact of FDIA when using AE or GAN independently, we now focus on assessing the effects of FDIA using our hybrid model which combines AE and GAN.

1) QUALITY OF CREATED SAMPLES

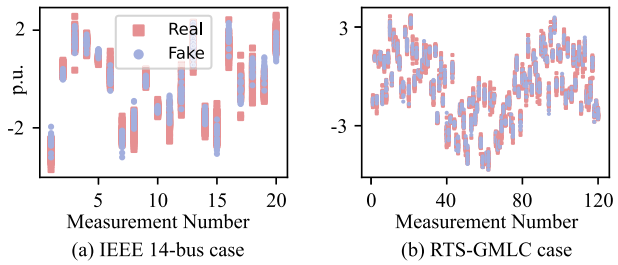
Fig. 5 depicts the real measurements (red) alongside the fake measurements (blue) generated by our FDIA. While the fake measurements visually resemble the real data, they do not completely overlap. This divergence from the original dataset is anticipated, as the attack regularization term incentivizes GAN to produce measurements that reside within the boundary of the original data distribution.

2) PERTURBATION OF SYSTEM STATES

Fig. 6 shows the real system states (red) and the states produced by the fake measurements (blue). Notably, the fake states exhibit a greater degree of dispersion compared to

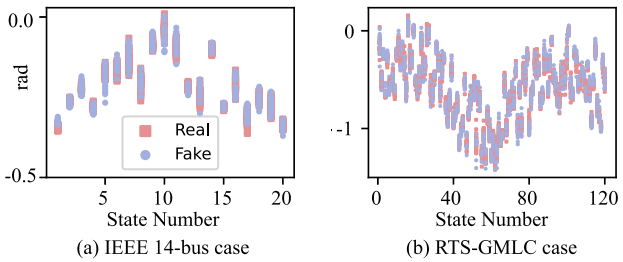
TABLE 1. Performance comparison on various systems. Significance level $\alpha = 3\%$.

physical model			Succ_{BDD} (%) ↑				Earth-Moving Distance (EMD) ↓			
	m	n	PCA [18]	SVD [20]	iAttackGen [27]	our FDIA	PCA [18]	SVD [20]	iAttackGen [27]	our FDIA
IEEE 14-bus	20	14	96.2	95.9	95.1	97.6	2.93	1.84	2.53	0.34
IEEE 30-bus	41	30	94.7	95.3	96.7	97.2	4.33	4.57	5.04	1.12
IEEE 39-bus	46	39	94.3	94.4	94.7	97.1	3.17	3.56	2.89	0.89
IEEE 57-bus	80	57	94.5	95.5	95.4	97.7	13.8	12.4	17.3	1.42
IEEE 118-bus	186	118	94.5	96.6	95.2	98.3	36.1	33.6	29.5	4.67
IEEE 200-bus	245	200	95.3	96.0	98.0	95.3	23.6	18.4	19.0	3.23
RTS-GMLC	120	73	93.6	95.7	95.5	98.5	21.1	18.4	24.3	2.57
IEEE 8-bus	14	8	96.4	96.0	95.3	98.2	2.56	1.96	2.28	0.45
IEEE 123-bus	244	123	93.8	94.3	97.7	94.2	18.1	12.6	17.2	3.51
MV urban	46	36	96.2	96.3	96.4	97.4	1.48	2.77	2.11	1.04
LV suburban	226	115	93.8	94.3	97.9	94.7	9.57	4.49	13.9	3.12



(a) IEEE 14-bus case (b) RTS-GMLC case

FIGURE 5. FDIA measurements compared to real measurements.



(a) IEEE 14-bus case (b) RTS-GMLC case

FIGURE 6. FDIA states compared to real states.

the real states. This divergence aligns with the attacker's objective of manipulating the system state estimation process by solely tampering with the measurements.

3) RESIDUAL ERROR DISTRIBUTION

To validate the assertion that our FDIA can accurately recover the Chi-squared distribution with a carefully chosen hyperparameter λ_{AE}, as per Theorem 1, we present empirical histograms of LNR values in different systems using our FDIA approach in Fig. 7. The findings demonstrate our FDIA model successfully reproduces the original Chi-squared distribution, presenting a challenge for defenders in distinguishing between real data and attack data.

Fig. 7 shows that the LNR values of our FDIA model closely adhere to the Chi-squared distribution χ^2_{m-n}, suggesting a high likelihood of bypassing the BDD. To provide quantitative validation, we calculate the success rate of passing BDD using Eq. (18) and present the results in Table 1.

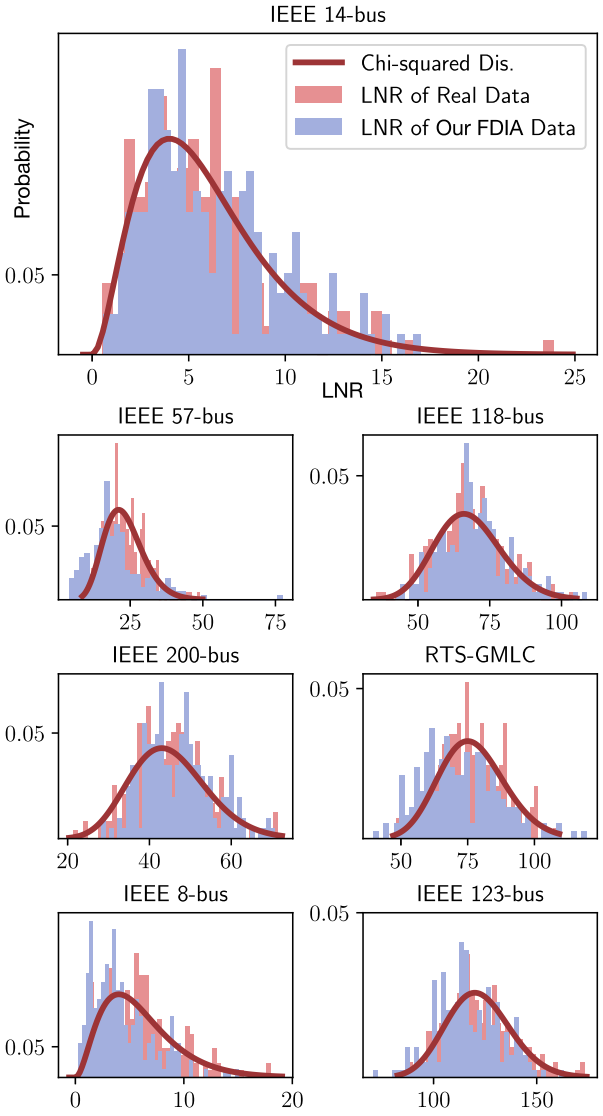


FIGURE 7. The empirical histogram of LNR values. $\alpha = 3\%$.

The results indicate that our FDIA model achieves a higher rate (around 97%) of bypassing BDD to baseline

models (around 95%). This improvement is likely attributed to the model's capability of accurately replicating the Chi-squared distribution, as formally proven in Theorem 1. Meanwhile, we observe that although other baselines also demonstrate relatively high rates of bypassing BDD, they exhibit a significantly higher Earth Mover's Distance (EMD) metric compared to our FDIA model. It implies that these baseline attacks are more easily detectable when scrutinizing the distribution of LNR values. In contrast, our FDIA model achieves a remarkably small EMD, indicating close alignment of its LNR values with the exact Chi-squared distribution.

C. SENSITIVITY ANALYSIS TO HIDDEN STATE DIMENSION

A key design choice in our FDIA model is the hidden state dimension within the autoencoder (AE) module. We hypothesize that setting this dimension equal to the number of free system states is essential for achieving accurate measurement reconstruction and replicating the behavior of the state estimation process. To validate this hypothesis, we conduct an experiment using the IEEE 14-bus test case. We analyze the residual error and the reconstruction error while varying the AE's latent dimension from 1 to 20. It is important to note that, in DC power flow models, the system states are the voltage phase angles of all buses except the reference bus, whose angle is fixed and non-free. Therefore, the number of free system states in the chosen system is 13.

Lemma 2 states that the residual error of attack measurements processed by the AE is proportional to the hidden dimension size. To validate this relationship, Fig. 8 (upper half) plots the residual errors of attack measurements from AE for various hidden dimension sizes. The blue curve represents the averaged residual error, and the shaded area shows the range of residual errors. The plot demonstrates that the residual errors of the attack measurements exhibit a trend that roughly aligns with a linear relationship to the size of the hidden dimension. Furthermore, as the hidden dimension size approaches 20 (which corresponds to the number of system measurements in this case study), the residual error approaches the level observed for real measurements (horizontal pink line).

While reducing the hidden layer dimension in the AE model can lower the residual error—thereby improving stealthiness—it may also lead to a higher reconstruction error. This is because a smaller latent space lacks the capacity to capture the full complexity of the measurement data. The lower half of Fig. 8 illustrates the reconstruction error of a well-trained AE model across various hidden dimensions. We observe that the reconstruction error drops sharply and approaches zero once the latent dimension exceeds 13, which corresponds to the number of independent system states in this case study. This observation aligns with the principle that system states represent the minimal set of variables required to reconstruct all measurements [32]. These results indicate that setting the AE's hidden layer dimension equal to the number of system states achieves a desirable balance:

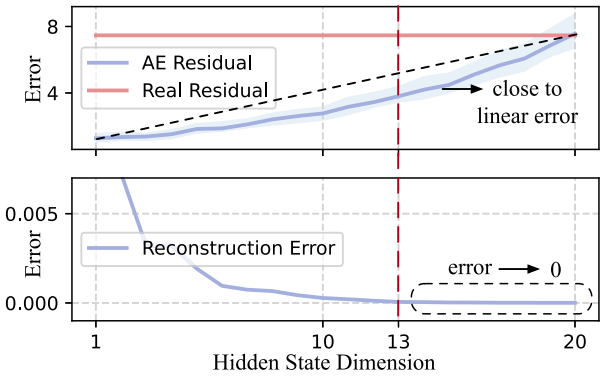


FIGURE 8. Residual error and reconstruction error against various hidden layer size for the IEEE 14-bus test case.

it minimizes residual error while preserving accurate reconstruction of the original measurements.

1) DISCUSSION ON DATA AVAILABILITY

The feasibility of training our AE-GAN model depends on access to historical measurement data. While real-time topology and parameters are typically protected, archived measurements are often available for operational or regulatory purposes [52], and may be obtained via cyber intrusions, insider threats, or public platforms [53]. When data is limited, training—especially for GANs—becomes more challenging due to overfitting risk and reduced diversity. To mitigate this, data augmentation [54] and transfer learning [55] can improve generalization. We conduct a sensitivity analysis on the IEEE 57-bus system by varying training sizes from 500 to 5000 samples. Results show the BDD bypass rate stabilizes around 95% after 1750 samples, suggesting moderate data is sufficient. This reflects the model's data efficiency, as it learns the residual structure rather than the full physical model. Future work will explore few-shot and meta-learning techniques to further reduce data requirements.

D. SENSITIVITY ANALYSIS TO HYPER-PARAMETERS

1) SENSITIVITY TO HYPER-PARAMETER λ_{AE}

We examine the effect of λ_{AE} by plotting the success rate of passing BDD (succ_{BDD}) and earth-moving distance (EMD) for various λ_{AE} in Fig. 9. Here, we denote the theoretically optimal hyperparameter from Theorem 1 as λ_{AE}^{opt} and compare it with $0.5 \times \lambda_{AE}^{opt}$ and $2 \times \lambda_{AE}^{opt}$. The findings demonstrate that choosing $\lambda_{AE} = 0.5 \times \lambda_{AE}^{opt}$, smaller than the optimal value, leads to the predominance of the GAN model within the hybrid framework. This positioning shifts the residual error distribution to the right of the true Chi-squared distribution, resulting in a decreased succBDD and an increased EMD. Conversely, opting for $2 \times \lambda_{AE}^{opt}$, larger than the optimal value, favors the DAE model dominance, causing the residual error distribution to shift to the left of the true Chi-squared distribution. This leads to a higher succBDD but a larger EMD. Notably, neither scenario proves as advantageous as

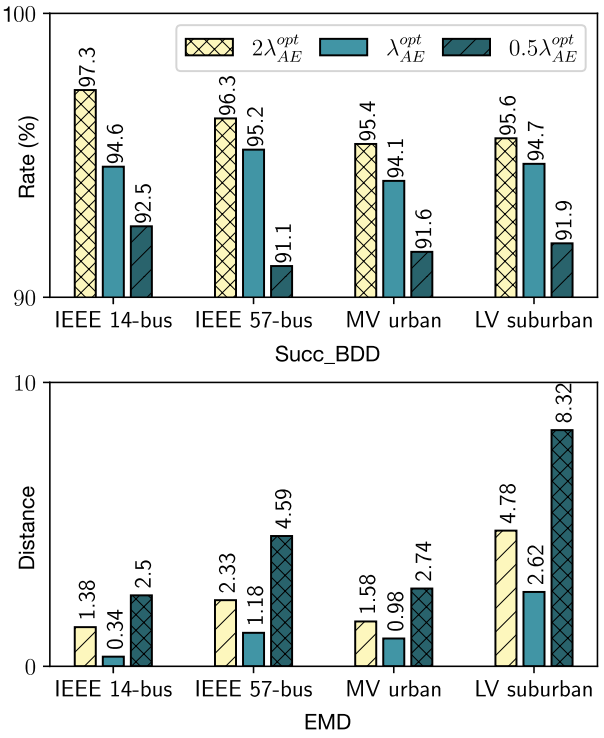


FIGURE 9. Performance against different λ_{AE}. Upper: The successful rate of passing BDD. Lower: The earth-moving distance between LNR values of fake/true data. $\alpha = 3\%$.

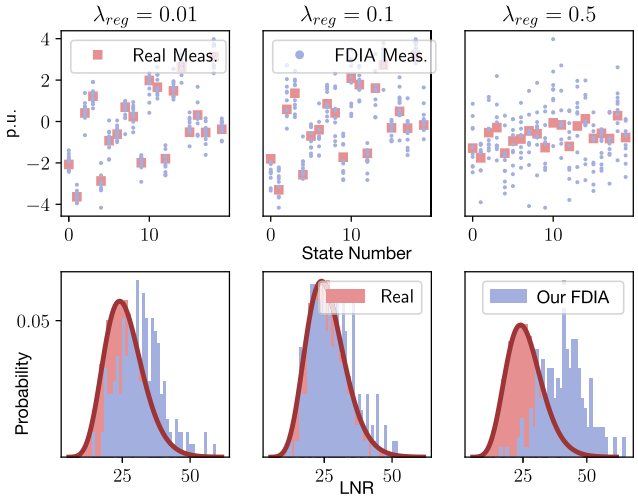


FIGURE 10. Measurements and residual error distribution of our FDIA model under various hyperparameters λ_{attack}. $\alpha = 3\%$.

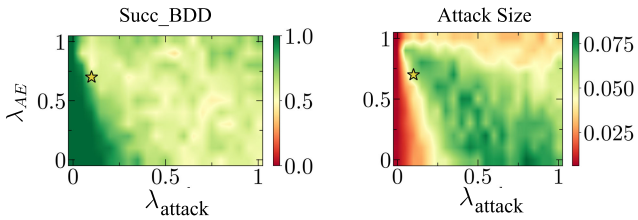


FIGURE 11. The success rate of bypassing BDD and the attack size w.r.t. the hyper-parameters λ_{AE} and λ_{attack} in IEEE 14-bus system. The optimal choice is $\lambda_{AE}^{*} = 0.7$ and $\lambda_{attack}^{*} = 0.1$.

selecting the optimal hyperparameter λ_{AE}^{opt}, which achieves a perfect recovery of the Chi-squared distribution.

2) SENSITIVITY TO HYPER-PARAMETER λ_{attack}

We also assess the effects of λ_{attack}, which governs the extent of attack size regularization. In Fig. 10, we depict the relationship between fake measurements and real measurements across various selections of the hyperparameter λ_{attack}. Additionally, we visualize the corresponding residual error for each scenario. As we vary the hyperparameter λ_{attack}, which penalizes the distance between FDIA and real measurements, interesting trends emerge. With a small value like $\lambda_{attack} = 0.01$, we observe that the generated FDIA data clusters closely around the real data. As we increase λ_{attack} to values like 0.1 and 0.5, the fake measurements progressively deviate from the real ones, indicating a more successful attack. However, a notable observation is made at $\lambda_{attack} = 0.5$, where the significant increase starts to compromise the Chi-squared distribution of the residual error, potentially leading to failure at the distribution detector.

3) TRADE-OFF BETWEEN HYPERPARAMETERS λ_{AE} AND λ_{attack}

An effective attack achieves two key objectives: (1) bypassing the BDD and (2) creating a significant attack vector. Our prior sensitivity analysis revealed that the hyperparameters λ_{AE} and λ_{attack} influence these objectives in opposing ways. To illustrate this trade-off, Fig. 11 presents the successful rate of passing BDD (succ_{BDD}) and the expected squared norm of the attack size $\mathbb{E}\|\tilde{\mathbf{z}} - \mathbf{z}\|_2^2$ across various combinations of λ_{AE} and λ_{attack}. Indicated by the results, the optimal weights are $\lambda_{AE}^{*} = 0.7$ and $\lambda_{attack}^{*} = 0.1$, where λ_{AE}^{*} is actually very close to the theoretically optimal value in Theorem 1.

4) SENSITIVITY ANALYSIS TO MEASUREMENT COVERAGE

In real-world scenarios, complete historical data from all grid measurements may not always be available, and certain measurements might be inherently protected or immune to manipulation. This practical constraint necessitates an analysis of our FDIA's performance under varying degrees of measurement coverage. Denoting the number of measurements immune to attack as $m^{\text{NoAttack}} \in [0, m]$, we conduct a sensitivity analysis to evaluate our FDIA across a range of m^{NoAttack} values. To do so, we randomly select and remove m^{NoAttack} measurements and apply our FDIA process. As shown in Fig. 12, the success rate of bypassing BDD drops with an increasing number of unattacked measurements. This decline is unsurprising, given the diminished attack information resulting from the reduced measurement coverage, thereby presenting greater challenges for effective FDIA.

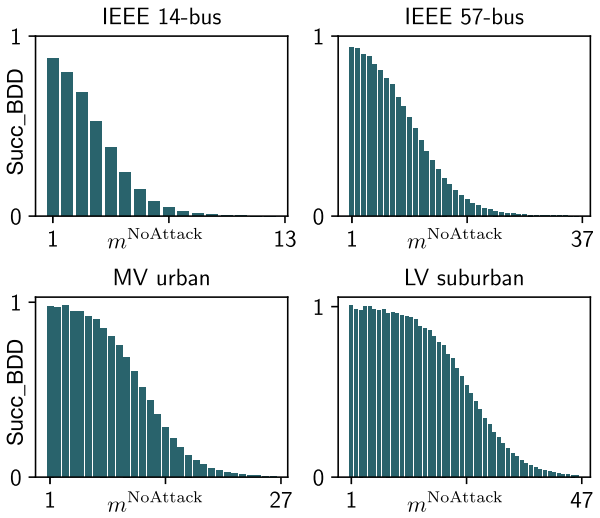


FIGURE 12. **Successful rate of bypassing BDD against m^{NoAttack}.**

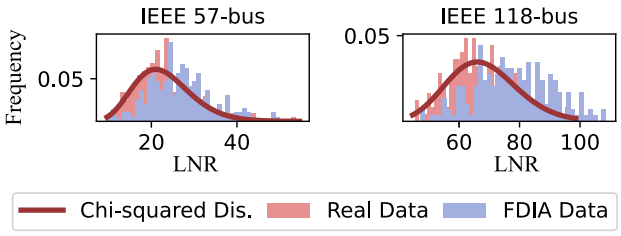


FIGURE 13. **Empirical histograms of LNR of attack measurements in IEEE 57-bus and 118-bus AC systems compared to the theoretical Chi-squared distribution.**

E. EXTENSION TO AC POWER SYSTEMS

Although our method is developed under the DC power flow model for rigorous theoretical guarantees, we evaluate its performance on AC systems using the IEEE 57-bus and 118-bus test cases. Fig. 13 compares the empirical LNR distributions of the generated attack measurements with the theoretical Chi-squared distribution used in BDD. Results show that our method generalizes reasonably well: the attack residuals closely match the expected distribution, indicating a high likelihood of bypassing BDD even under nonlinear AC conditions. However, the fit is slightly less precise compared to the DC case. This degradation stems from the nonlinear nature of AC systems, which complicates the measurement manifold and limits the ability of a standard autoencoder to fully capture it. Since our theoretical guarantees rely on linear models, this gap is expected. These findings motivate future work on developing specialized or physics-informed autoencoders better suited for nonlinear AC systems.

F. ROBUSTNESS UNDER SYSTEM CONTINGENCIES

In practical power systems, the Assumption 1 of steady-state operation and stationary data distributions can be challenged by real-world contingencies. These contingencies may temporarily or permanently shift the system behavior, raising concerns about the robustness and adaptability of data-driven attack models. In this section, we examine two representative scenarios that deviate from steady-state operation and evaluate the performance of our proposed AE-GAN model under each: (1) power-related fluctuations that preserve the grid topology, and (2) structural changes that alter the system topology and trigger updates to the state estimator and BDD mechanism.

For power-related fluctuations, the system experiences short-term deviations such as demand spikes or generation adjustments, which do not change the network topology. Since the Jacobian matrix $\mathbf{H}$ remains fixed, the BDD test remains valid. While these fluctuations shift the operating point, they still follow the same physical model and thus lie near the boundary of the original data distribution assumed in our training. To validate robustness under such conditions, we simulate a scenario in the IEEE 57-bus system where 10% of the time window includes synthetic deviations in the state variables. We compute new measurements from the unchanged $\mathbf{H}$, and use the entire dataset (including perturbed data) for training. As shown in Fig. 14 (left), the residuals of the generated false measurements remain well aligned with the Chi-squared distribution, indicating that our method remains stealthy and effective under moderate load fluctuations.

For topology-altering contingencies, such as line outages or reconfiguration events, the system undergoes a permanent structural change that alters its physical model. In response, the operator updates the state estimator to reflect the new topology, resulting in a modified Jacobian matrix $\mathbf{H}$ and a corresponding update to the BDD mechanism. To simulate such a scenario, we remove the transmission line between bus 25 and bus 30 in the IEEE 57-bus system and generate DC power flow data under both the original and modified topologies. We concatenate the two datasets to form a composite measurement distribution with two distinct modes. To ensure compatibility with the updated BDD logic, we apply MATLAB's built-in change-point detection algorithm `findchangepts` to isolate the post-contingency data, which aligns with the updated system configuration. The AE-GAN model is then retrained on this post-change data. As shown in Fig. 14 (right), the residuals of the generated attacks continue to follow the Chi-squared distribution closely, demonstrating that, with proper preprocessing, the proposed method remains stealthy even under topology changes.

The above experiment regarding topology-altering events also provides insights into the performance of our approach under a class of defense mechanisms known as Moving Target Defense (MTD), which intentionally and periodically reconfigures the grid topology to disrupt attack strategies. While our experiment reflects a single MTD-induced change, it shows that with access to post-reconfiguration data, change-point detection enables the attacker to adapt to the new topology and retrain the model to maintain stealthiness. However, we also observe that if MTD schemes are applied

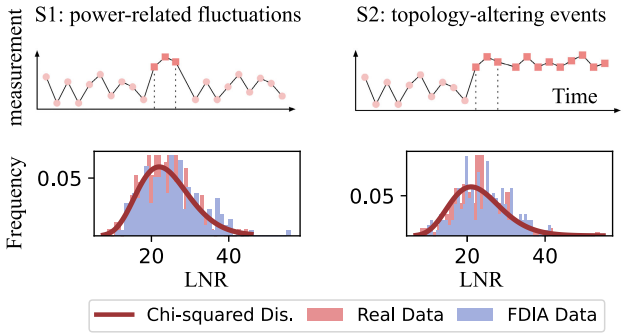


FIGURE 14. Empirical LNR histograms of attack measurements under two scenarios: (left) power-related fluctuations, (right) topology-altering events, in IEEE 57-bus system, compared to the theoretical Chi-squared distribution.

at high frequency, the attacker may lack sufficient data under each topology to retrain effectively, thereby degrading the feasibility of data-driven FDIAs. Fortunately, in practical power systems, MTD strategies are implemented at moderate timescales due to operational, physical, and economic constraints. As such, our method remains applicable in many realistic MTD settings, and we highlight adaptation to rapid or unpredictable MTD as an important direction for future work.

V. CONCLUSION

This paper presents a novel physical-model-free False Data Injection Attack (FDIA) framework that reliably bypasses Bad Data Detection (BDD) without requiring any physical model, such as grid topology or line parameters. The proposed method introduces a principled hybrid architecture combining an autoencoder (AE) and a generative adversarial network (GAN). The AE module reduces residual errors by mimicking the state estimation process in terms of denoising and projecting noisy measurements onto a physically meaningful manifold. It is achieved by using a latent space dimension equal to the number of system states to denoise measurements and replicate the residual error characteristics of real data. The GAN module then explores the measurement manifold to induce significant deviations in the estimated states, thereby maximizing the impact of the attack. Theorem 1 formally demonstrates that the residual errors of the generated false measurements follow the same theoretical Chi-squared distribution as true measurements, ensuring a statistically equivalent likelihood of bypassing the BDD. The effectiveness of the proposed method is validated across 11 representative grid systems, including both transmission and distribution networks, using real-world power profiles. The results show consistently high BDD evasion success rates and low Earth Mover's Distance (EMD) values, outperforming existing model-free baselines and confirming the ability of our method to replicate the residual distribution with high fidelity. These findings reveal a critical vulnerability in modern power grids, where attackers without system knowledge can still launch highly effective and stealthy FDIAs. Future work will extend this framework to more realistic AC power flow models and explore advanced detection and defense mechanisms to counteract such data-driven cyber threats.

REFERENCES

[1] S. Mantravadi, R. Schnyder, C. Møller, and T. D. Brunoe, "Securing IT/OT links for low power IIoT devices: Design considerations for industry 4.0," *IEEE Access*, vol. 8, pp. 200305–200321, 2020.

[2] J. Styczynski and N. Beach-Westmoreland. (2019). *When The Lights Went Out: A Comprehensive Review Of The 2015 Attacks On Ukrainian Critical Infrastructure*. [Online]. Available: https://www.boozallen.com/s/insight/thought-leadership/lessons-from-ukranians-energy-grid-cyber-attack.html

[3] S. Tatum. (2018). *U.S. Accuses Russia of Cyberattacks on Power Grid*. CNN. [Online]. Available: https://www.cnn.com/2018/03/15/politics/dhs-fbi-russia-power-grid/index.html

[4] C. Strategic and I. Studies. (Jun. 2024). *Significant Cyber Incidents*. CSIS. [Online]. Available: https://www.csis.org/programs/strategic-technologies-program/significant-cyber-incidents

[5] G. C. Rasner, *Cybersecurity and Third-party Risk: Third Party Threat Hunting*. Hoboken, NJ, USA: Wiley, 2021.

[6] M. Mohammadpourfard, Y. Weng, and M. Tajdinian, "Benchmark of machine learning algorithms on capturing future distribution network anomalies," *IET Gener., Transmiss. Distrib.*, vol. 13, no. 8, pp. 1441–1455, Apr. 2019.

[7] T. Hong and A. Hofmann, "Data integrity attacks against outage management systems," *IEEE Trans. Eng. Manag.*, vol. 69, no. 3, pp. 765–772, Jun. 2022.

[8] K. Paul, "Modified grey wolf optimization approach for power system transmission line congestion management based on the influence of solar photovoltaic system," *Int. J. Energy Environ. Eng.*, vol. 13, no. 2, pp. 751–767, Jun. 2022.

[9] M. A. Rahman, E. Al-Shaer, and R. Kavasseri, "Impact analysis of topology poisoning attacks on economic operation of the smart power grid," in *Proc. IEEE 34th Int. Conf. Distrib. Comput. Syst.*, Jun. 2014, pp. 649–659.

[10] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Trans. Inf. Syst. Secur.*, vol. 14, no. 1, pp. 1–33, May 2011.

[11] G. Hug and J. A. Giampapa, "Vulnerability assessment of AC state estimation with respect to false data injection cyber-attacks," *IEEE Trans. Smart Grid*, vol. 3, no. 3, pp. 1362–1370, Sep. 2012.

[12] P. M. Anderson, C. F. Henville, R. Rifaat, B. Johnson, and S. Meliopoulos, *Power System Protection*. Hoboken, NJ, USA: Wiley, 2022.

[13] M. A. Basit, S. Dilshad, R. Badar, and S. M. S. ur Rehman, "Limitations, challenges, and solution approaches in grid-connected renewable energy systems," *Int. J. Energy Res.*, vol. 44, no. 6, pp. 4132–4162, May 2020.

[14] Y. Liao, C. Xiao, and Y. Weng, "Quickest line outage detection with low false alarm rate and no prior outage knowledge," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Jul. 2022, pp. 1–5.

[15] N. Costilla-Enriquez, Y. Weng, and B. Zhang, "Combining Newton-raphson and stochastic gradient descent for power flow analysis," *IEEE Trans. Power Syst.*, vol. 36, no. 1, pp. 514–517, Jan. 2021.

[16] M. Jin, J. Lavaei, and K. H. Johansson, "Power grid AC-based state estimation: Vulnerability analysis against cyber attacks," *IEEE Trans. Autom. Control*, vol. 64, no. 5, pp. 1784–1799, May 2019.

[17] M. Esmalifalak, H. Nguyen, R. Zheng, L. Xie, L. Song, and Z. Han, "A stealthy attack against electricity market using independent component analysis," *IEEE Syst. J.*, vol. 12, no. 1, pp. 297–307, Mar. 2018.

[18] Z.-H. Yu and W.-L. Chin, "Blind false data injection attack using PCA approximation method in smart grid," *IEEE Trans. Smart Grid*, vol. 6, no. 3, pp. 1219–1226, May 2015.

[19] A. Anwar and A. N. Mahmood, "Stealthy and blind false injection attacks on SCADA EMS in the presence of gross errors," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Jul. 2016, pp. 1–5.

[20] H. Yang and Z. Wang, "A false data injection attack approach without knowledge of system parameters considering measurement noise," *IEEE Internet Things J.*, vol. 11, no. 1, pp. 1452–1464, Jan. 2024.

[21] H. Yang, X. He, Z. Wang, R. C. Qiu, and Q. Ai, "Blind false data injection attacks against state estimation based on matrix reconstruction," *IEEE Trans. Smart Grid*, vol. 13, no. 4, pp. 3174–3187, Jul. 2022.

[22] S. Lakshminarayana, A. Kammoun, M. Debbah, and H. V. Poor, "Data-driven false data injection attacks against power grids: A random matrix approach," *IEEE Trans. Smart Grid*, vol. 12, no. 1, pp. 635–646, Jan. 2021.
[23] A. S. Musleh, G. Chen, Z. Y. Dong, C. Wang, and S. Chen, "Attack detection in automatic generation control systems using LSTM-based stacked autoencoders," *IEEE Trans. Ind. Informat.*, vol. 19, no. 1, pp. 153–165, Jan. 2023.
[24] N. Costilla-Enriquez and Y. Weng, "Attack power system state estimation by implicitly learning the underlying models," *IEEE Trans. Smart Grid*, vol. 14, no. 1, pp. 649–662, Jan. 2023.
[25] S. Ahmadian, H. Malki, and Z. Han, "Cyber attacks on smart energy grids using generative adverserial networks," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2018, pp. 942–946.
[26] M. Mohammadpourfard, F. Ghanaatpishe, M. Mohammadi, S. Lakshminarayana, and M. Pechenizkiy, "Generation of false data injection attacks using conditional generative adversarial networks," in *Proc. IEEE PES ISGT Innov. Smart Grid Technol. (ISGT) Europe*, Delft, The Netherlands, 2020, pp. 41–45. [Online]. Available: https://ieee-isgt-europe.org/
[27] M. H. Shahriar, A. A. Khalil, M. A. Rahman, M. H. Manshaei, and D. Chen, "IAttackGen: Generative synthesis of false data injection attacks in cyber-physical systems," in *Proc. IEEE Conf. Commun. Netw. Secur. (CNS)*, Oct. 2021, pp. 200–208.
[28] N. Costilla-Enriquez and Y. Weng, "Exposing cyber-physical system weaknesses by implicitly learning their underlying models," in *Proc. Asian Conf. Mach. Learn.*, Aug. 2021, pp. 1333–1348.
[29] A. Cooper, A. Bretas, and S. Meyn, "Anomaly detection in power system state estimation: Review and new directions," *Energies*, vol. 16, no. 18, p. 6678, Sep. 2023.
[30] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, "MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Trans. Power Syst.*, vol. 26, no. 1, pp. 12–19, Feb. 2011.
[31] Q. Yang, J. Yang, W. Yu, D. An, N. Zhang, and W. Zhao, "On false data-injection attacks against power system state estimation: Modeling and countermeasures," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 3, pp. 717–729, Mar. 2014.
[32] A. Abur and A. G. Exposito, *Power System State Estimation: Theory and Implementation*. Boca Raton, FL, USA: CRC Press, 2004.
[33] B. M. Horowitz and K. M. Pierce, "The integration of diversely redundant designs, dynamic system models, and state estimation technology to the cyber security of physical systems," *Syst. Eng.*, vol. 16, no. 4, pp. 401–412, Dec. 2013.
[34] Z. Wang, H. He, Z. Wan, and Y. Sun, "Detection of false data injection attacks in AC state estimation using phasor measurements," *IEEE Trans. Smart Grid*, early access, Feb. 10, 2020, doi: 10.1109/TSG.2020.2972781.
[35] S. Bolognani, N. Bof, D. Michelotti, R. Muraro, and L. Schenato, "Identification of power distribution network topology via voltage correlation analysis," in *Proc. 52nd IEEE Conf. Decis. Control*, Dec. 2013, pp. 1659–1664.
[36] D. Deka, S. Backhaus, and M. Chertkov, "Structure learning in power distribution networks," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 3, pp. 1061–1074, Sep. 2018.
[37] J. Zhang and L. Sankar, "Physical system consequences of unobservable state-and-topology cyber-physical attacks," *IEEE Trans. Smart Grid*, vol. 7, no. 4, pp. 2016–2025, Jul. 2016.
[38] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning (Adaptive Computation and Machine Learning Series)*. Cambridge, MA, USA: MIT Press, 2016.
[39] P. Baldi and K. Hornik, "Neural networks and principal component analysis: Learning from examples without local minima," *Neural Netw.*, vol. 2, no. 1, pp. 53–58, Jan. 1989.
[40] I. J. Goodfellow et al., "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 27, 2014, pp. 2672–2680.
[41] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, Aug. 2017, pp. 214–223.
[42] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1125–1134.
[43] M. N. Dinh, C. Trung Vo, and D. Abramson, "Tracking scientific simulation using online time-series modelling," in *Proc. 20th IEEE/ACM Int. Symp. Cluster, Cloud Internet Comput. (CCGRID)*, May 2020, pp. 202–211.
[44] C. Barrows et al., "The IEEE reliability test system: A proposed 2019 update," *IEEE Trans. Power Syst.*, vol. 35, no. 1, pp. 119–127, Jan. 2020.
[45] T. Nguyen, S. Wang, M. Alhazmi, M. Nazemi, A. Estebsari, and P. Dehghanian, "Electric power grid resilience to cyber adversaries: State of the art," *IEEE Access*, vol. 8, pp. 87592–87608, 2020.
[46] W. H. Kersting, "Radial distribution test feeders," *IEEE Trans. Power Syst.*, vol. 6, no. 3, pp. 975–985, Aug. 1991.
[47] C. Mateo et al., "European representative electricity distribution networks," *Int. J. Electr. Power Energy Syst.*, vol. 99, pp. 273–280, Jul. 2018.
[48] C. Xiao, Y. Liao, and Y. Weng, "Distribution grid line outage identification with unknown pattern and performance guarantee," *IEEE Trans. Power Syst.*, vol. 39, no. 2, pp. 3987–3999, Mar. 2024.
[49] M. Shahriar, I. Habiballah, and H. Hussein, "Optimization of phasor measurement unit (PMU) placement in supervisory control and data acquisition (SCADA)-based power system for better state-estimation performance," *Energies*, vol. 11, no. 3, p. 570, Mar. 2018.
[50] C. Xiao, Y. Liao, and Y. Weng, "Privacy-preserving line outage detection in distribution grids: An efficient approach with uncompromised performance," *IEEE Trans. Power Syst.*, vol. 40, no. 1, pp. 866–878, Jan. 2025.
[51] L. Van der Maaten and G. E. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, Jan. 2008.
[52] Z. Liu, Q. Wang, Y. Ye, and Y. Tang, "A GAN-based data injection attack method on data-driven strategies in power systems," *IEEE Trans. Smart Grid*, vol. 13, no. 4, pp. 3203–3213, Jul. 2022.
[53] T. Krause, R. Ernst, B. Klaer, I. Hacker, and M. Henze, "Cybersecurity in power grids: Challenges and opportunities," *Sensors*, vol. 21, no. 18, p. 6225, 2021.
[54] N. Tran, V. Tran, N. Nguyen, T. Nguyen, and N. Cheung, "On data augmentation for GAN training," *IEEE Trans. Image Process.*, vol. 30, pp. 1882–1897, 2021.
[55] W. He, J. Chen, Y. Zhou, X. Liu, B. Chen, and B. Guo, "An intelligent machinery fault diagnosis method based on GAN and transfer learning under variable working conditions," *Sensors*, vol. 22, no. 23, p. 9175, Nov. 2022.