Adversarial Attacks on Machine Learning-Based State Estimation in Power Distribution Systems

Afia Afrin
aafrin@ualberta.ca
University of Alberta
Edmonton, AB, Canada

Omid Ardakanian
ardakanian@ualberta.ca
University of Alberta
Edmonton, AB, Canada

ABSTRACT

We examine the robustness of machine learning-based distribution system state estimation (DSSE) techniques to a class of adversarial attacks, known as the black-box evasion attack. In these attacks, the attacker manipulates real-time measurements from sensors installed in the distribution grid by adding carefully crafted perturbations to diminish the accuracy of DSSE. We devise a stealthy attack based on the Fast Gradient Sign Method (FGSM), dubbed Sneaky-FGSM, by analyzing the statistical properties of real-time measurements and adding perturbations accordingly. Using simulation on a standard test distribution system, we show that this attack would remain largely unidentified and the error introduced in the DSSE process could propagate to a voltage control scheme that takes the DSSE result as input. Our result suggests that incorporating machine learning models in DSSE is a double-edged sword and calls for more research to ensure the robustness of these models to adversarial samples.

CCS CONCEPTS

• **Security and privacy**; • **Computer systems organization** → *Embedded and cyber-physical systems*;

KEYWORDS

Adversarial attack, power system operation, bad data detection

ACM Reference Format:
Afia Afrin and Omid Ardakanian. 2023. Adversarial Attacks on Machine Learning-Based State Estimation in Power Distribution Systems. In *The 14th ACM International Conference on Future Energy Systems (e-Energy '23), June 20–23, 2023, Orlando, FL, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3575813.3597352

1 INTRODUCTION

Real-time state estimation plays a pivotal role in realizing the vision of efficient, nimble, and resilient electric power infrastructure as it underpins various monitoring and control applications, from fault detection to Volt/VAR optimization. Historically, state estimation was primarily used in the power transmission system to determine its state, e.g., bus voltages or branch currents, from incomplete or noisy measurements. These measurements can be obtained from the supervisory control and data acquisition (SCADA) system or phasor measurement units (PMUs) installed at specific nodes in the network. But in the past decade, the growing adoption of distributed energy resources (DER) and controllable loads has caused wide fluctuations in voltage and reverse flow in the power distribution system, making it imperative to increase visibility in low-voltage feeders and employ feedback control schemes to maintain its reliable operation. Since real-time state estimation supports these applications, it is anticipated that it will be increasingly incorporated in distribution system operation practices [46].

The state estimation problem can be formulated as a system of nonlinear equations, which is typically solved as a weighted least-squares (WLS) problem [41] in the polar or rectangular coordinate system. However, WLS-based estimators do not yield sufficiently accurate results in the DSSE problem for several reasons. First, unlike the transmission system, real-time measurements are scarce in the distribution system as there is little instrumentation beyond the substation [18]. This results in fewer measurements than unknowns, rendering WLS-based estimators ineffective [69]. Second, a typical distribution system contains numerous unbalanced three-phase lines. These lines are shorter than transmission lines and have a higher r/x ratio. This could lead to ill-conditioned Jacobian and gain matrices, affecting the convergence rate of WLS-based state estimation techniques [7]. Finally, WLS-based state estimation techniques rely on the electrical system model, which encodes the operational structure of the network and parameters of distribution lines and transformers. This model is not available in most distribution systems today [8]. Inspired by the success of machine learning (ML) techniques in approximating complex physics-based model, several attempts have been made to solve DSSE by taking a data-driven approach or a hybrid approach that combines ML models with electrical model-based, static or dynamic state estimation techniques, such as WLS and Kalman filter [24]. In particular, neural networks trained on historical measurements or simulation data have been used to estimate the system state from existing measurements [11, 60, 67, 68], initialize the Gauss-Newton method so it enjoys quadratic convergence to the true latent state of the system [65], or generate pseudo-measurements to compensate the lack of sufficient measurements when solving DSSE using traditional model-based techniques [38]. More recently, physics-aware neural networks [66] have been utilized to increase the accuracy of DSSE by pruning connections in the neural network according to the distribution system model. These studies are unanimous in their conclusion that ML-based state estimators are superior to traditional model-based techniques, which are computationally expensive and often incapable of capturing the nonlinear relationship

e-Energy '23, June 20–23, 2023, Orlando, FL, USA

ACM ISBN 979-8-4007-0032-3/23/06…$15.00
https://doi.org/10.1145/3575813.3597352

between input and output, hence they cannot effectively deal with increased variability and uncertainty in distribution networks.

Despite the vast literature on data-driven and hybrid state estimation techniques, the previous work does not investigate whether these techniques are robust to *adversarial samples* [22] that resemble normal sensor data. This is important because adversarial attacks have been shown to greatly degrade the performance of classification and regression models in other domains [17, 40, 64]. Since DSSE is essentially a regression problem, these attacks can reduce the state estimation accuracy and subsequently the performance of the controller that relies on the DSSE result. For example, the attacker might be able to create power quality issues by misleading the operator into taking actions that exacerbate over or under-voltage problems. Such an attack will be detrimental if it is not detected by the *bad data detection* (BDD) mechanism that is commonly adopted to protect the state estimation process. The real-world application of the newly developed data-driven and hybrid DSSE techniques requires assessing the vulnerability of the underlying machine learning model(s), and developing threat models and mitigation strategies, which are currently missing. This observation serves as the key motivation behind our work.

To investigate the robustness of ML-based state estimation techniques to adversarial samples, we consider a powerful technique, called Stacked ResNetD, from the previous work [11]. It uses an ensemble of dense residual neural networks (ResNet) to map real-time measurements to the system state. This technique is shown to outperform several other electrical model-agnostic state estimation techniques, so we use it as a representative DSSE technique in our study [11]. We propose a *black-box* adversarial attack that uses an arbitrary *surrogate model* trained on historical data – measurements and corresponding states – to add carefully crafted perturbations to the measurements to reduce the accuracy of DSSE. We show that the standard residual-based BDD mechanism fails to flag the modified measurements as bad data in the majority of cases. We then devise an even stealthier version of this attack in which the attacker uses statistical properties of sensor data to selectively apply the perturbations. To demonstrate the damage that could be inflicted, we assess the impact of both attacks on a voltage control scheme that relies on the DSSE result. Our contribution is threefold:

- We present a black-box *evasion attack* against a state-of-the-art DSSE technique that uses an ensemble of residual neural networks to estimate the system state. We then investigate transferability of this attack to two other data-driven state estimators based on a convolutional neural network and the K-nearest neighbors (KNN) algorithm. Using surrogate models that are different from the victim state estimation model, we argue that the attacker neither needs the knowledge of the ML model used in DSSE (as in white-box attacks), nor any information about the distribution system model.
- We devise a stealthier evasion attack, namely *Sneaky-FGSM*, by applying perturbations according to the variance of data generated by the respective sensors. We show that this novel attack can further reduce the accuracy of DSSE at a lower BDD detection rate.
- We conduct a simulation study on an extended version of the IEEE 33-bus test system, in which the IEEE European low-voltage system is used to model the secondary networks and real load data is used to represent the household demands, to investigate how the error introduced in the state estimation process propagates and affects a voltage control scheme that relies on the DSSE output.

To our knowledge, *this is the first work that investigates the adversarial robustness of data-driven state estimators and analyzes the impact of adversarial attacks on the distribution system control process.* Our findings suggest that ML-based DSSE techniques are not presently robust to carefully crafted adversarial examples and more research is warranted to address their vulnerability before they can be incorporated into distribution system operation practices.

2 RELATED WORK

Machine learning-based state estimation techniques garnered attention in recent years as they were shown to be superior to traditional static and dynamic state estimation techniques, such as WLS and Kalman filter [46], especially in distribution networks with high DER penetration. For example, real-time distribution system state estimators based on various deep neural network architectures [6, 67, 68], long short-term memory (LSTM) [5], and KNN [60] were proposed in the literature. An ML-based state estimator that takes advantage of an ensemble of ResNets [11] has been recently shown to outperform several other ML-based techniques, including multilayer perceptron (MLP) and convolutional neural network (CNN). A physics-aware neural network is proposed for DSSE in [66], where the knowledge of the physical system is utilized to prune the dense neural network, reducing overfitting. Several studies use a hybrid approach where an ML model is combined with a traditional approach (such as WLS and least absolute value) [12, 13, 65]. The fundamental concept underlying these hybrid approaches is to leverage the ML model to map available measurements or historical data to the neighborhood of the true latent state. These approximate state values are then used as a starting point for iterative methods, such as the Gauss-Newton method.

2.1 Adversarial Attacks

Adversarial machine learning studies how to fool machine learning models by providing malicious inputs during training or test. The two most common types of adversarial attack algorithms are:

- *Evasion Attacks:* Attacks in this category add carefully crafted perturbations to the benign samples in the test set with the goal of producing erroneous output, thereby reducing the accuracy of the machine learning model during deployment. Popular evasion attacks include *FGSM* [22], *Basic Iterative Method (BIM)* [28], *Projected Gradient Descent (PGD)* [36], *DeepFool* [42], and *Carlini–Wagner Attack (C&W)* [14].
- *Poisoning Attacks:* These attacks affect the model by targeting its availability or integrity. In the former case, the attacker injects malicious data into the training set to corrupt the learned model [43, 44, 63], whereas in the latter case, the adversary creates a *backdoor* into the learning model using poisoning strategies [15].

The above-listed attack strategies can be designed using a *white-box* or *black-box* approach. During a white-box attack, the adversary uses the knowledge of the *victim* ML model, including its architecture, hyperparameter values, and weights associated with the

connections, to generate adversarial samples [14, 20, 42]. In contrast, in a black-box attack, the adversary has only query access to the victim model and no prior knowledge of the victim model's architecture; therefore, it uses a *surrogate model* to generate adversarial samples [23, 26, 27]. Earlier studies have demonstrated that due to the transferability of adversarial samples, it is possible to devise black-box attacks by training surrogate models that differ from the victim model [45]. Our work is inspired by this result.

2.2 False Data Injection Attacks

False data injection attacks (FDIAs) are a major threat to cyber-physical systems that contain a sense and control loop, such as the smart grid [29, 30]. While both evasion attacks and FDIA manipulate the sensor data, there are fundamental differences between the two in terms of attack formation strategies and threat models. To launch an effective FDIA that bypasses the BDD mechanism, the adversary typically needs to have access to the topology and configuration of the grid or the measurement matrix, in addition to the data-overwrite access [33, 34, 62]. However, only data-overwrite access is sufficient to launch *black-box* adversarial attacks. Furthermore, adversarial samples crafted by models that capture hidden features and trends in data have the property of *transferability*, which allows them to mislead not only a specific target model but also other models even if their architectures differ greatly [45]. To the best of our knowledge, no such evidence regarding transferability of FDIA has been provided in the literature.

2.3 Vulnerability to Attacks in Smart Grid

We now review the related work on analyzing the robustness of ML models that are used in different power system applications. Eklas et al. [25] study the application of machine learning in the smart grid and the emerging security concerns associated with the adoption of this technology. The authors have reviewed recent cyber attacks against electric grid infrastructures that took place around the world and were caused by compromised software, malicious operating systems, or the presence of intruders. To analyze security and vulnerability of learning algorithms used in the power system, Chen et al. [16] propose an evasion attack algorithm that works in a similar manner to FGSM. They examined the efficacy of the proposed attack against a neural network-based power quality disturbance classifier and an RNN-based load forecasting model.

While various ML techniques have been proposed to detect FDIA [48, 54, 58, 61], few papers examined robustness and security issues that arise from the use of machine learning techniques. The impact of two adversarial attacks, namely Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) and Jacobian-based Saliency Map Attack (JSMA), on an MLP-based false data detection technique was analyzed in [48]. Joint adversarial examples and false data injection attacks (AFDIAs) that are able to fool both BDD and neural attack detector (NAD) mechanisms protecting the DC state estimation process have been proposed in [54]. While white-box AFDIAs show promise in bypassing the detection mechanisms, the performance of black-box AFDIAs is subpar. Another recent study uncovers the inefficacy of BDD and NAD mechanisms in DC state estimation in the presence of white-box targeted FDIAs [53].

Turning to data-driven state estimation approaches, ANN-based state estimators have been found vulnerable to FDIAs. For example, optimization techniques based on differential evolution and sequential least-square quadratic programming have been proposed in [31, 32] to construct attack vectors that can fool the BDD mechanism and affect an MLP-based state estimator. More recently, a forward-derivative-based adversarial attack on a neural network-based state estimator is proposed in [51]. However, the authors do not consider the existence of any bad data detection mechanism; thus, it is unclear how effective this attack strategy is when state estimation is safeguarded by the BDD mechanism. We note that all these attacks are white-box, i.e., the attacker is assumed to have the full knowledge of the power grid's structure and model, as well as the architecture and parameters of the ANN used for DSSE, which is a strong assumption in some real-world applications.

The closest work to ours is [10] and [52], where data-driven approaches are used to generate black-box attacks against electrical model-based state estimators. Specifically, a robust linear regression model has been proposed in [52] to generate stealthy attack vectors that can fool the residual-based BDD mechanism integrated with the DC state estimation technique in the black-box setting. For AC state estimation, deep adversarial networks have been used for the first time in [10] to craft a stealthy black-box adversarial attack against power system state estimation. The authors used the vanilla FGSM algorithm to create the attack vectors against an AC-PSSE algorithm that estimates states by solving the WLS optimization. In contrast to these recent studies, we analyze the vulnerability of *data-driven DSSE approaches* to adversarial attacks crafted using surrogate neural networks under the black-box assumption. Moreover, we propose the novel Sneaky-FGSM algorithm, which is able to induce higher measurement noise without getting detected by the conventional BDD mechanism. Lastly, we address a major limitation of the existing literature [10, 51, 54] by analyzing the impact of the proposed attacks on voltage regulation schemes, which is an important control application that relies on the DSSE result.

3 PRELIMINARIES

We give a brief overview of the fundamental concepts that form the foundation of our work. Specifically, we provide the mathematical formulation of DSSE and present a widely used BDD mechanism to protect DSSE. Then, we discuss a rule-based voltage regulation scheme that relies on the DSSE output.

3.1 DSSE Problem Formulation

State estimation is the problem of identifying state variables, e.g., bus voltage magnitudes and phase angles, from the available measurements in a power system [35]. Suppose $h(\cdot)$ is the non-linear function that relates state variables, denoted by vector $\mathbf{x} \in \mathbb{C}^n$ (where $\mathbb{C}$ is the set of complex numbers), to a vector collecting field measurements $\mathbf{z} \in \mathbb{C}^m$. We have

$$\mathbf{z} = h(\mathbf{x}) + \xi, \tag{1}$$

where $\xi \in \mathbb{C}^m$ is the measurement error. Note that $h(\cdot)$ depends on the real-time operational structure and parameters of the distribution system model. To obtain the system state vector of size n from a set of m independent measurements, a WLS estimator minimizes

the following objective function [7]:

$$\min_{x} J(\mathbf{x}) = \sum_{i=1}^{m} \left(z_i - h_i\left(\mathbf{x}\right)\right)^2 / R_{ii} \tag{2}$$

where $\mathbf{R}$ is a diagonal matrix, called the *covariance matrix of measurement errors (ξ)* and given by:

$$\mathbf{R} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_m^2 \end{bmatrix}$$

Here, σ_k^2 is the variance of the k^{th} measurement from the measurement vector $\mathbf{z}$. We can write (2) in vector/matrix form as follows:

$$\min_{x} \left[\mathbf{z} - h(\mathbf{x})\right]^\top \mathbf{R}^{-1} \left[\mathbf{z} - h(\mathbf{x})\right] \tag{3}$$

Due to the high computational overhead and possibility of getting stuck in local minima [33], $h(\cdot)$ is often linearized:

$$h(\mathbf{x}) = \mathbf{H}\mathbf{x} \tag{4}$$

Here, $\mathbf{H}$ is the *measurement matrix* and typically defined as the Jacobian matrix of $h(\cdot)$.

$$\mathbf{H} = \delta h(\mathbf{x}) / \delta \mathbf{x}$$

By combining (3) and (4), we derive the estimated state as:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \left[\mathbf{z} - \mathbf{H}\mathbf{x}\right]^\top \mathbf{R}^{-1} \left[\mathbf{z} - \mathbf{H}\mathbf{x}\right] \tag{5}$$

We note that linearization of $h(\cdot)$ does not work well in distribution grids, so iterative methods, such as Gauss–Newton, can be used instead to estimate the state starting from some initial point.

By adding pseudo-measurements obtained from historical data to field measurements, DSSE is usually solved as an overdetermined problem, where we have fewer states than the measurements, i.e., $n < m$. In this case, the closed-form solution for the maximum likelihood estimate of $\mathbf{x}$ can be derived as follows [57]:

$$\hat{\mathbf{x}} = \left(\mathbf{H}^\top \mathbf{W} \mathbf{H}\right)^{-1} \mathbf{H}^\top \mathbf{W} \mathbf{z} \tag{6}$$

As the reliability of estimated states is heavily dependent on the accuracy of measurements, distribution system operators often deploy a residual-based BDD mechanism to safeguard the state estimation procedure. Specifically, the measurement error, $\mathbf{e}$ is defined as the difference between actual measurements ($\mathbf{z}$) and estimated measurements ($\hat{\mathbf{z}}$), i.e., $\mathbf{z} - \hat{\mathbf{z}}$, where $\hat{\mathbf{z}} = \mathbf{H}\hat{\mathbf{x}}$.

The *chi-square test* is a convenient strategy to identify the presence of bad data [7]. From (2), the residual can be rewritten as:

$$J(\mathbf{x}) = \sum_{i=1}^{m} \frac{e_i^2}{R_{ii}} = \sum_{i=1}^{m} \left(\frac{e_i}{\sigma_i}\right)^2 \tag{7}$$

Notice that Equation (7) is of the form $y = \sum_{i=1}^{d} \chi^2$, which corresponds to the chi-squared distribution with d degrees of freedom. Since it is assumed that $m > n$, at most $(m-n)$ of the measurement residuals will be linearly independent, resulting in $d = m - n$. To detect the presence of bad (measurement) data, $J(x)$ is compared to the critical chi-square value at the degree of freedom d, and a pre-specified level of significance α. If $J(x) < \chi^2_{d,\alpha}$, then the estimated state, i.e., $\hat{\mathbf{x}}$, can be trusted. Otherwise, it is assumed that the measurement contains bad data. Upon detecting bad data, the distribution system operator (DSO) may either discard the estimated state and replace it by a previous state estimate or try to identify the source of bad data, eliminate the bad measurement(s), and re-estimate the current state.

3.2 Voltage Regulation using DSSE Result

A voltage limit violation in a power distribution system occurs when the voltage level exceeds or drops below the limit set by the utility company or some regulatory body. This can happen due to various reasons, such as equipment failure, an increase in load, or a fault on the distribution lines. These violations not only affect the stability of the power grid but also can cause damage to equipment (both at the grid end and consumer end), and power outages. To prevent these calamities, voltage control devices, such as capacitor banks, regulators, and on-load tap changers (OLTCs), are used to quickly respond to voltage fluctuations.

Due to high installation costs, distribution-level PMUs (D-PMUs) are not currently deployed at each node of a distribution system, despite their ability to provide highly precise and frequent data [50]. Therefore, estimated states from DSSE are often used instead of the measurements when they are missing to detect voltage limit violations [21] and perform Volt/VAR optimization (VVO) [37]. In this context, an adversarial attack launched against the data-driven state estimator would eventually impact these control decisions.

The most prevalent VVO approach is the *SCADA-controlled VVO*, which is a rule-based strategy where voltage and VAR control devices, such as voltage regulators and capacitor banks, are controlled based on some pre-defined set of rules [47]. The SCADA-controlled VVO is often studied as two independent problems, VAR optimization and Voltage control [47]. For this study, we focus on the voltage control part of the SCADA-controlled VVO mechanism which aims to maintain acceptable voltage levels at all points along the distribution feeder under all load conditions by controlling tap changers and/or voltage regulators [3].

4 THREAT MODEL

We choose the Stacked ResNetD state estimator as the victim model, which is a strong baseline among electrical model-agnostic state estimation techniques. We conduct our experiments under the assumption that the attacker has no knowledge of the architecture of the victim model, hence it is a black-box attack. Nevertheless, the attacker is assumed to have (a) read and write access to the real-time measurement of all sensors, $\mathbf{z}$, and (b) read access to the victim model's output, $\mathbf{x}$, which can be paired with the corresponding measurement to construct the training dataset, $\left\{\left(\mathbf{z_i^{train}}, \mathbf{x_i^{train}}\right)\right\}$, for the surrogate model f, described in Algorithm 1. Given these assumptions, the primary attack point would be the utility data center where the state estimation (victim) model is run and sensor data are stored. The attacker can be an insider (e.g., a malicious operator), or an intruder hacking into the server, using compromised software installed on the server that hosts the victim model, or gaining access to the DSO's authorized user account. The PMU networks and utility data centers have been found vulnerable to cyber attacks in several recent studies [56, 59], indicating a high risk of the presence of such adversaries, lending credence to this threat model.

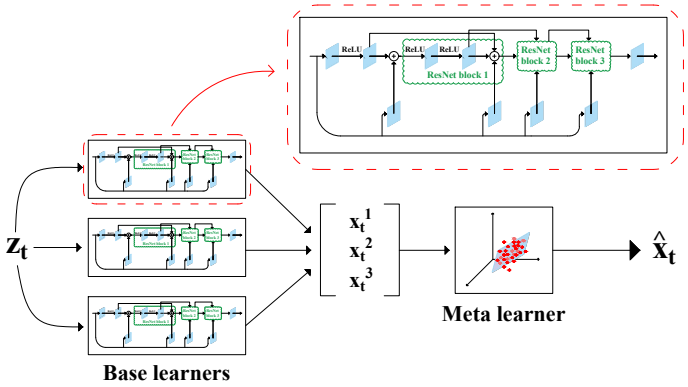


Figure 1: Stacked ResNetD architecture

The goal of an attacker is to distort the measurements in a way that significantly alters the state estimation result and at the same time remains undetected. Hence, in the second phase of experimentation, we investigate the stealthiness of the proposed attack to the conventional BDD mechanism. Upon analyzing the results from the BDD mechanism, we propose a new attack strategy, namely *Sneaky-FGSM*, which is able to induce noise in a stealthier fashion.

5 METHODOLOGY

We describe the implementation of DSSE and BDD, present the proposed black-box evasion attack strategy, and provide more details about the voltage control scheme that relies on the DSSE result.

5.1 DSSE and BDD Techniques

While it is possible to train a variety of ML models and incorporate them in DSSE, instead of introducing yet another architecture and identifying its vulnerabilities, we use a state-of-the-art ensemble learning model, namely *Stacked ResNetD*, which has been proposed in [11] and shown to outperform several other deep neural networks, as our victim model. Previous work has shown that ensemble learning models have enhanced adversarial robustness [49, 55]. This together with the strong performance of Stacked ResNetD motivates our choice of the victim model.

The term Stacked ResNetD refers to an ensemble learning model consisting of B base learners and a meta-learner. In this work, we choose the value of B by trial and error, i.e., empirically evaluating DSSE for different values of B on our test network and choosing the one that yields the lowest error. We employ three base learners (B=3), each of which is a 13-layer dense ResNet model, trained to estimate states from the measurements. Figure 1 shows the architecture of the Stacked ResNetD model. The output states produced by the base learners are combined together and fed into the meta-learner. A multivariate linear regressor (MLR) is employed as the meta-learner. Finally, we use historical measurement-state pairs, $(\mathbf{z}, \mathbf{x})$, to train the ensemble model.

Detecting bad measurements is extremely valuable for the state estimation procedure and is typically implemented as a residual-based method. The intuition behind the residual-based BDD approach is that the residual, $J(x)$, determined after the state estimator algorithm converges, will be minimal if the measurement set contains no bad data [57]. We implement the residual-based BDD mechanism described in Section 3.1 with level of significance α=0.05, and degrees of freedom d=48.

5.2 Attack Strategy

Vanilla FGSM. We hypothesize that the adversary, being unaware of the victim model's architecture, can choose any suitable neural network as the surrogate model to produce the black-box attack. To test this hypothesis, we use two surrogate models and investigate their effectiveness against a fixed victim model (i.e., Stacked ResNetD) that is different from them. The first surrogate model is an MLP that consists of 5 dense layers. The second one is the CNN model proposed in [11] consisting of 3 convolutional layers and 3 dense layers. For both models, we use ReLU as the activation function and mean absolute error (MAE) as the loss function. Note that from the vast ocean of ML models, the adversary is free to choose any suitable surrogate, meaning that there are hundreds of possible ways to affect the data-driven DSSE approaches.

Our aim is to investigate how black-box evasion attacks could mislead distribution network control systems by affecting the data-driven state estimation process. While most evasion attacks are designed to fool classifiers, FGSM and its iterative versions, in particular, BIM and PGD introduced in Section 2.1, can be applied against regression models as well. Since FGSM is the foundation of the other two attacks and all of these three attack strategies work in a similar manner [70], we choose this as our primary attack strategy. For the rest of this paper, we refer to the standard black-box FGSM, presented in Algorithm 1, as *vanilla FGSM* to distinguish it from the novel Sneaky-FGSM discussed below.

Sneaky-FGSM. From Equation (7), it can be seen that the tolerance of the residual-based BDD mechanism is determined by the variance of measurement data. Thus, intuitively, perturbing the measurements that do not show much variance increases the chance of being detected by the BDD mechanism. Based on this insight, we formulate the novel Sneaky-FGSM attack strategy, which improves the vanilla FGSM attack by perturbing only the measurements with high variance to increase the stealthiness of the attack. The proposed Sneaky-FGSM approach is presented in Algorithm 2.

Power system measurement data exhibits seasonality and temporal variation. Thus, the data used in the variance calculation step plays an important role in correctly detecting bad measurements. For example, taking into account measurements collected over one year would result in higher variance (hence a less stringent BDD process) than considering measurements collected over a week for this calculation. In this study, we use the daily variance of measurements, i.e., we calculate the variance of a batch of data generated over 24 hours, while implementing the BDD mechanism.[1]

On any day D, an adversary with access to the measurement data can readily estimate the daily variance of each of the m measurements, $\{\sigma_k^2\}_{k=1}^m$, by calculating the daily variance using the measurements from the previous day, $D-1$, or using a batch of latest data samples. These estimates will be used by the adversary to identify which measurements have an exceptionally low variance, and therefore, should not be perturbed in the stealthy attack.

[1] In practice, the daily variance data can be estimated using historical measurements from the same day in prior year(s) or the previous day.

Algorithm 1 Vanilla FGSM Attack

1: **Inputs:**
Training data, $\left\{\left(\mathbf{z_i^{train}}, \mathbf{x_i^{train}}\right)\right\}_{i=1,\cdots,N_{train}}$
Maximum training iteration, $maxIter$
Clean data sample at timestamp t, $(\mathbf{z_t}, \mathbf{x_t})$
2: **Output:**
Adversarial sample at timestamp t, $(\mathbf{z'_t}, \mathbf{x'_t})$
3: **Initialize:**
θ_0 with small random values
Surrogate model, f_θ with appropriate loss function, L
▷ *Training the surrogate model f, parameterized by θ*
4: **for** $j = 0, 1, \cdots, maxIter$ **do**
5: $\boldsymbol{\theta_{j+1}} \leftarrow \boldsymbol{\theta_j} - \alpha \nabla_{\boldsymbol{\theta_j}} \left[\frac{1}{N_{train}} \sum_i L\left(f(\mathbf{z_i^{train}}; \boldsymbol{\theta}_j), \mathbf{x_i^{train}}\right)\right]$
6: **end for** ▷ *α is the learning rate*
▷ *Calculating gradient of the loss w.r.t. the input, $\mathbf{z_t}$*
7: $\boldsymbol{\delta}_{\mathbf{z_t}} = \nabla_{\mathbf{z_t}} \left[L\left(f(\mathbf{z_t}; \boldsymbol{\theta}), \mathbf{x_t}\right)\right]$
8: $\mathbf{z'_t} = \mathbf{z_t} + \epsilon \cdot \text{sign}(\boldsymbol{\delta}_{\mathbf{z_t}})$ ▷ *ϵ is a hyperparameter (scalar)*
9: return $(\mathbf{z'_t}, \mathbf{x'_t})$ ▷ *Return the adversarial sample*

In this experiment, we use the household power consumption dataset (described later in Section 6.1), in which reactive power consumption (Q) exhibits exceptionally low variance (less than 1). In light of this, we design the first version of Sneaky-FGSM by perturbing all measurements except the Q measurements. We found that using this attack it is possible to fool the BDD mechanism more frequently than the vanilla FGSM; however, perturbing the Q measurements in addition to the other measurements would increase the BDD detection rate. This successful attempt led to a more general version of the proposed Sneaky-FGSM, where we do not perturb a particular measurement $\mathbf{z}_t[k]$ if its variance, $\sigma_t^2[k]$, is lower than a pre-defined threshold value. The threshold value that is being used to determine whether a variance value is 'low' or not, is a hyperparameter that is tuned according to the attacker's intent. Using a higher threshold value will produce a stealthier but less effective attack and vice versa. In this experiment, we define the thresholds for power consumption measurements as follows: $\upsilon_1 = \cdots = \upsilon_m = 1$ to avoid adding noise to Q measurements and perhaps other measurements that are intrinsically low variance.

In Line 6, we define a binary vector, $\mathbf{select} \in \{0, 1\}^m$, which holds 0 at index k if the variance of the k^{th} measurement of the data sample $\mathbf{z_t}$ is below the predefined threshold (i.e., $\sigma_t^2[k] < \upsilon_k$), and 1 otherwise. Finally, in Line 9, we modify the perturbation vector obtained from vanilla FGSM (i.e., $\text{sign}(\boldsymbol{\delta}_{\mathbf{z_t}})$) by calculating its Hadamard (element-wise) product with **select**.

5.3 Voltage Regulation under Adversarial Attack

We implement the rule-based voltage regulation strategy that relies on DSSE and was previously outlined in Section 3.2. In this scheme, control rules are generally determined based on operational constraints. An example VAR optimization rule can be– *"switch on the capacitor bank, if the power factor is less than* 0.95*"* and an example of the voltage control rule can be– *"if voltage at bus n drops below or goes above the pre-defined setpoint, change the OLTC tap position accordingly"* [47]. In this study, we implement the rule-based voltage control strategy by installing an OLTC and setting up a voltage control rule similar to the example we gave for the voltage control rule. Detailed analysis of this experimentation is presented in Section 7.3.

Algorithm 2 Sneaky-FGSM Attack

1: **Inputs:**
Training data, $\left\{\left(\mathbf{z_i^{train}}, \mathbf{x_i^{train}}\right)\right\}_{i=1,\cdots,N_{train}}$
Maximum training iteration, $maxIter$
Clean data sample at timestamp t, $(\mathbf{z_t}, \mathbf{x_t})$
2: **Output:**
Adversarial sample at timestamp t, $(\mathbf{z'_t}, \mathbf{x'_t})$
3: Train the surrogate model f parameterized by $\boldsymbol{\theta}$, following the steps described in Algorithm 1 (Line 4 to 6).
4: Define the minimum thresholds, $[\upsilon_1, \upsilon_2, ..., \upsilon_m]$, for the variance of measurements
5: Define the vector, $select$, of size m as follows:
$$\mathbf{select}[k] = \begin{cases} 0 & \text{if } \sigma_t^2[k] < \upsilon_k \\ 1 & \text{otherwise} \end{cases}$$
▷ *Calculating gradient of the loss w.r.t. the input, $\mathbf{z_t}$*
6: $\boldsymbol{\delta}_{\mathbf{z_t}} = \nabla_{\mathbf{z_t}} \left[L\left(f(\mathbf{z_t}; \boldsymbol{\theta}), \mathbf{x_t}\right)\right]$
7: $\mathbf{S} = \mathbf{select} \odot \text{sign}(\boldsymbol{\delta}_{\mathbf{z_t}})$
8: $\mathbf{z'_t} = \mathbf{z_t} + \epsilon \cdot \mathbf{S}$ ▷ *ϵ is a hyperparameter (scalar)*
9: return $(\mathbf{z'_t}, \mathbf{x'_t})$

6 EXPERIMENTAL SETUP

This section describes the experimental setting that we used to study the adversarial vulnerability of the Stacked ResNetD state estimator.

6.1 Test Case

Our test system is structurally similar to the customized IEEE 33-bus test system presented in [24]. Specifically, we use the 33-bus system [9] as the primary distribution network and the IEEE European low voltage test feeder [2] to model the secondary networks. We assume each of the primary buses, except the first one, is connected to a low-voltage feeder, representing the secondary network. Figure 2 shows one of the low-voltage feeders that originates from Bus 25. Other low-voltage feeders are not depicted in this figure.

Each secondary feeder supplies 55 single-phase loads. To represent these loads, we adopt the Multifamily Residential Electricity Dataset (MFRED) [39], which contains daily load profile of 390 US apartments with 15-minute resolution over a 12 month period (January 2019 to December 2019). The loads are grouped into 26 apartment groups as per the recommended data aggregation standard for publishing utility data in the State of New York [1]. Thus, each of the apartment groups contains the average real and reactive power consumption of 15 apartments.

To simulate a real-world setting, we add Gaussian noise with standard deviation of 1%, 2%, $\cdots$, 10% to each of the 26 household load data to generate 286 distinct apartment load data including the original 26 households. This way, 500 hypothetical buildings are created, each containing 1 to 10 apartments chosen randomly from the 286 apartments. We determine the suitable aggregation level at each low-voltage bus using the network data provided in [9]. More specifically, we randomly select buildings and connect them to each

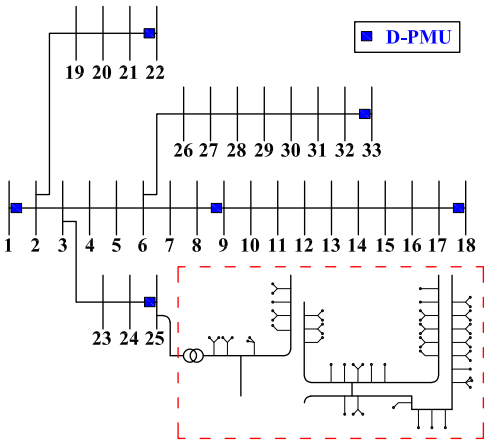


Figure 2: Single-line diagram of the customized IEEE 33-bus test system. Node 25 shows the IEEE European low-voltage system, which is connected to each of the primary nodes.

secondary bus until the loads in the low-voltage network under each primary bus add up to the load given in the 33-bus system data sheet. Finally, we run the AC power flow analysis using the Open Distribution Simulator Software (OpenDSS) [19] to generate the training and test datasets for the ML models. We note that the training dataset can be generated in a similar fashion in the real-world setting, i.e., by solving the power flow equations to obtain the system states using historical load and generation data [65].

6.2 Data Preparation & Simulation

At a given time t, the input to the state estimator is the real-time measurements collected by vector $\mathbf{z_t}$, and the output is the system state, $\mathbf{x_t}$. In defining measurements and states, we use the conventional approach [46] where the state variables are the bus voltage phasors, denoted by $x = [\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_b}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_b]$, with b being the number of buses that do not have D-PMU installed. Here, $\mathbf{v_i}$ and $\boldsymbol{\theta}_i$ represent the vectors containing the three-phase voltage magnitudes and phase angles of bus i, respectively. Any combination of redundant network data (i.e., bus voltage phasors, real and reactive power consumption, branch flows) can be considered as the measurement for the DSSE process. For this study, we assume all load buses in the secondary distribution network are equipped with smart meters providing real and reactive power consumption data every 15 minutes. We aggregate the smart meter data from all load buses in a secondary network, without accounting for losses, to produce the real and reactive power consumption at the primary bus, which are treated as pseudo-measurements. Thus, the measurement vector contains three-phase real and reactive power consumption at each of the primary load buses, and three-phase voltage magnitudes and phase angles of buses equipped with D-PMUs. We install six D-PMUs since this level of observability led to reasonable state estimation performance in [24]. Figure 2 shows the placement of the D-PMUs that collect the voltage phasor measurement data. One D-PMU is installed at the substation (Bus 1). The remaining D-PMUs are installed at the end of the primary feeders and one in the middle of the longest feeder to ensure system-wide observability. Note that determining the optimal placement of measurement devices, such as D-PMUs, is outside the scope of this work, so we just tried one reasonable sensor placement strategy.

Treating the first bus as the slack bus, we have 32 load buses in our primary distribution system. Therefore, we have $32 \times 3 \times 2$ pseudo-measurements for real and reactive power consumption at these buses: $(\mathbf{P}, \mathbf{Q})$. From the buses equipped with a D-PMU, we have 6×3 voltage magnitude measurements. Thus, the input measurement vector is of size 210×1. Excluding the D-PMU-installed buses, we have 27 buses that comprise the system state; thus, the state vector is of size 162×1.

We consider the OpenDSS simulation results obtained for the first half of every month to train the victim model. Since the dataset has 15−minute resolution, we have a total of 17280 training samples (i.e., 96 instances from each day). To form the test dataset, we randomly choose the load data from three consecutive days of each month and obtain the corresponding OpenDSS simulation result. Thus, we generate 3456 instances of test samples, grouped in 12 groups of 288 consecutive measurements (i.e., 3 consecutive days from each month $\times$ 96 samples from each day) that are evenly distributed over the one-year time period. The remaining samples, pertaining to 12 days in the second half of every month, are used to train the surrogate model.

6.3 Evaluation Criteria

We use the following measures to evaluate the performance of the ML-based DSSE technique and the voltage control scheme under normal operating conditions and in the presence of the black-box evasion attack.

State estimation accuracy. To evaluate the performance of the Stacked ResNetD state estimator we use the root-mean-square error (RMSE) defined as:

$$RMSE = \sqrt{\frac{1}{T \cdot n} \sum_{t=1}^{T} \sum_{i=1}^{n} \left(x_i^t - \hat{x}_i^t \right)^2}$$

Here, n is the total number of estimated states, T is the total number of test samples, and x_i^t and $\hat{x}_i^t$ represent actual and predicted states, respectively.

Voltage violation detection accuracy. Detecting voltage limit violations is the first step of voltage regulation, which is crucial to ensure the reliable operation of the distribution system. To analyze the impact of the black-box FGSM attack on the ability to detect voltage limit violations using the estimated state, we set the acceptable voltage range as ±5% of the nominal voltage level. We remark that the optimal acceptable range varies from system to system. We followed the range specified for the Range A service voltage in the American national standard for utilization voltage regulation (ANSI C84.1) [4].

Impact on the voltage regulation scheme. Controlling voltage control devices based on an inaccurate state estimation result may lead to one of the three unfavorable scenarios described below. We use the number of unnecessary tap change operations and the amount of voltage limit violations (including both over or under-voltage incidents) at the selected bus as our performance measures.

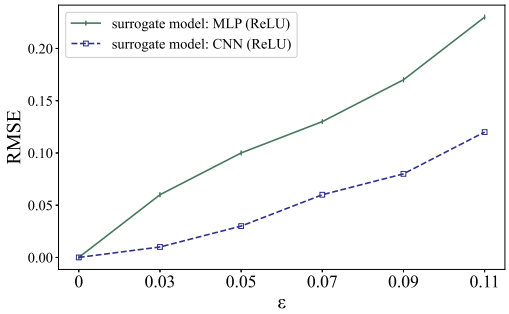


Figure 3: Increase in the state estimation error under the black-box vanilla FGSM attack.

Scenario 1 (Increased tap operations). This occurs when there is a false positive: even though the bus voltage is within the specified range, unnecessary voltage control operations (such as OLTC tap changes) are performed due to the error in the state estimation result. This increases wear and tear on voltage regulation devices, reducing their lifetime..

Scenario 2 (Increased over-voltage incidents). It may occur in two different ways: **(a)** when the bus voltage is above the upper threshold but it does not get detected because of the erroneous state estimation result (i.e., a false negative). In this case, the affected bus experiences an over-voltage problem, but since it is not detected, the voltage control scheme does not take any remedial action. Thus, the over-voltage situation persists; **(b)** when the bus voltage is within the specified range but under-voltage is detected (i.e., a false positive). In this case, the controller sends a command to increase the OLTC tap position. Due to this unnecessary tap change operation, the voltage level increases in that bus and possibly other buses downstream of the OLTC. This may lead to an over-voltage problem, degrading the power quality.

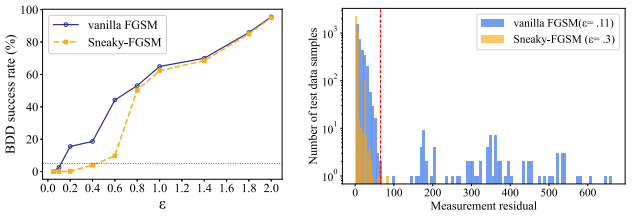
Scenario 3 (Increased under-voltage incidents). This is the exact opposite of the previous scenario and may occur in two different ways: **(a)** when the bus voltage is below the lower threshold and it does not get detected (i.e., a false negative); **(b)** when the bus voltage is within the specified range but over-voltage is detected (i.e., a false positive). In this case, an unnecessary tap change operation is performed to lower the tap setting. This may lead to an under-voltage problem, degrading the power quality.

7 EXPERIMENTAL RESULTS

We present the simulation results and evaluate the effectiveness and stealthiness of the proposed attacks, and analyze the impact of these attacks on a rule-based voltage regulation scheme.

7.1 Black-box FGSM against the Stacked ResNetD model

In the first phase of experimentation, we design an attacker who constructs adversarial data samples using the vanilla FGSM presented in Algorithm 1 and modifies the measurements ($\mathbf{z}$) accordingly. As discussed in Section 5.2, we employ two different surrogate models, namely MLP and CNN, to generate the adversarial data samples. These adversarial samples, when fed to the victim state estimator model, increase the state estimation error. The induced estimation



(a) Stealthiness of vanilla and sneaky FGSM attacks crafted using a CNN surrogate. The horizontal line is drawn at 5% detection rate.

(b) Frequency plot of the measurement residual ($J(\mathbf{x})$) under vanilla and sneaky FGSM attacks. The vertical line at $J(x) = 65.17$ represents the critical chi-square value being used as the threshold for detecting bad data. Note the y-axis is logarithmic scale.

Figure 4: Performance of the residual-based BDD mechanism under Vanilla and Sneaky FGSM.

error is directly proportional to the amount of noise added to the dataset. Figure 3 presents the impact of vanilla FGSM attack on the Stacked ResNetD state estimator. It can be readily seen that both surrogate models are successful in misleading the state estimator. However, the effectiveness of vanilla FGSM depends greatly on the choice of the surrogate model.

As discussed in Section 5.2, the choice of the surrogate model rests exclusively with the attacker. We use the CNN surrogate for the rest of the experiments and later compare the two surrogate models in Section 7.4. Note that we need to tune the hyperparameter, ϵ, according to the choice of the surrogate.

Next, we study whether the conventional (residual-based) BDD can identify adversarial samples. We craft both vanilla FGSM and Sneaky-FGSM attacks with different perturbation factors, ϵ, and test the efficacy of the BDD mechanism. We quantify BDD efficacy by its *success rate*, which is defined as $\frac{N_c}{N_{total}}$, where N_c is the number of bad data samples that get detected by BDD and N_{total} is the total number of bad data samples used. In this experiment, we used 3456 data samples (i.e., N_{total}=3456).

Figure 4a compares the stealthiness of vanilla FGSM and Sneaky-FGSM attacks. For smaller perturbations ($\epsilon \leq 0.7$), the proposed Sneaky-FGSM manages to bypass BDD more often. However, as the amount of perturbation (ϵ) increases, the measurement residual, $J(x)$, starts to exceed the critical chi-square value, resulting in a level of stealthiness that is comparable with vanilla FGSM. We conclude that *the attacker should carefully tweak ϵ to maximize the impact of the attack while bypassing the BDD mechanism with high probability.* This observation gives rise to an interesting research question: how much can the attacker affect a control application that relies on the state estimation result by launching black-box adversarial attacks while remaining undetected? To address this question, we analyze the impact of vanilla FGSM and Sneaky-FGSM attacks that are able to bypass BDD with a high success rate. The distribution of measurement residuals ($J(x)$) under vanilla FGSM with $\epsilon = 0.11$ and Sneaky-FGSM with $\epsilon = 0.3$ is shown in Figure 4b. The chosen ϵ values ensure that the corresponding attacks can bypass the BDD mechanism with at least 95% success rate. Observe that Sneaky-FGSM is capable of bypassing BDD with a higher ϵ value ($\epsilon = 0.3$) than that of vanilla FGSM ($\epsilon = 0.11$). In other words, by utilizing the Sneaky-FGSM algorithm, it is possible to add more

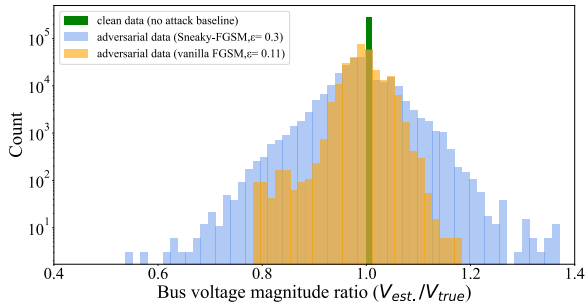


Figure 5: Distribution of bus voltage magnitude ratio over all unobserved buses. Note the y-axis is logarithmic scale.

perturbation without being detected (approximately three times more in our test case), while keeping BDD detection rate the same. We stick with these ϵ values in the following sections.

Figure 5 shows the distribution of two ratios, $\frac{|V_{est.}|}{|V_{true}|}$ and $\frac{|V^{adv}_{est.}|}{|V_{true}|}$, which helps compare the effect of vanilla FGSM and Sneaky-FGSM attacks on the (victim) state estimation model. Here, $|V_{true}|$ is the true bus voltage magnitude, $|V_{est.}|$ and $|V^{adv}_{est.}|$ are the estimated voltage magnitudes under normal condition and under adversarial attack, respectively. As expected, the original Stacked ResNetD model keeps the ratio very close to 1. However, the vanilla FGSM attack causes the number of outliers to increase significantly and with the Sneaky-FGSM attack, the induced estimation error is even higher. Figure 6 shows a more detailed side-by-side comparison of the vanilla FGSM and Sneaky-FGSM attacks by presenting the box and whisker plot of bus voltage magnitude ratios at each of the unobserved buses, with outliers marked at 5^{th} and 95^{th} percentiles.

7.2 Impact on Voltage Limit Violation Detection

We use the voltage phasor magnitudes obtained from the OpenDSS simulation results using the test dataset to identify the true voltage limit violation incidents during the simulation period. Since we have 3456 test data instances and 27 load buses that are not equipped with D-PMUs in our test system, there is a total of $3456 \times 27 = 93312$ instances where voltage magnitude violation may occur. We define a binary vector, VLV, of size 93312, as follows

$$\text{VLV}[i] = \begin{cases} 1 & \text{if voltage limit violation occurs at instance } i \\ 0 & \text{otherwise} \end{cases}$$

In a similar manner, we obtain (a) VLV_{clean}– a binary vector representing the detection of voltage violation incidents from the estimated states when clean test data samples are fed to the Stacked ResNetD model, (b) VLV_{FGSM}– a binary vector representing the detection of voltage violation incidents from the estimated states when adversarial test data samples generated by vanilla FGSM are fed to the Stacked ResNetD model, and (c) $\text{VLV}_{sneakyFGSM}$– a binary vector representing the detection of voltage violation incidents from the estimated states when adversarial test data samples generated by Sneaky-FGSM are fed to the Stacked ResNetD model.

Each of these three binary vectors is then compared to the true detection vector (VLV) to analyze the impact of the proposed attacks on voltage violation detection accuracy. Figure 7 shows the final outcome of the experiment. As we can see, the Stacked ResNetD estimator captures the true state of the system and detects all the voltage limit violations correctly. However, under the proposed black-box attacks, we encounter some false positives and some false negatives. Inaccurately detecting voltage limit violations may mislead the voltage regulation process and result in poor management of voltage-control tools, power quality degradation, and even worse, catastrophic operational failures such as persistent over-voltage or under-voltage problems at load buses causing equipment damage. We illustrate these scenarios in the next section.

Attack Type	#Unnecessary Tap Changes Initiated	#Voltage Violation Occurrences Caused by the Attack
None	0	0
Vanilla FGSM	13	0
Sneaky-FGSM	74	23

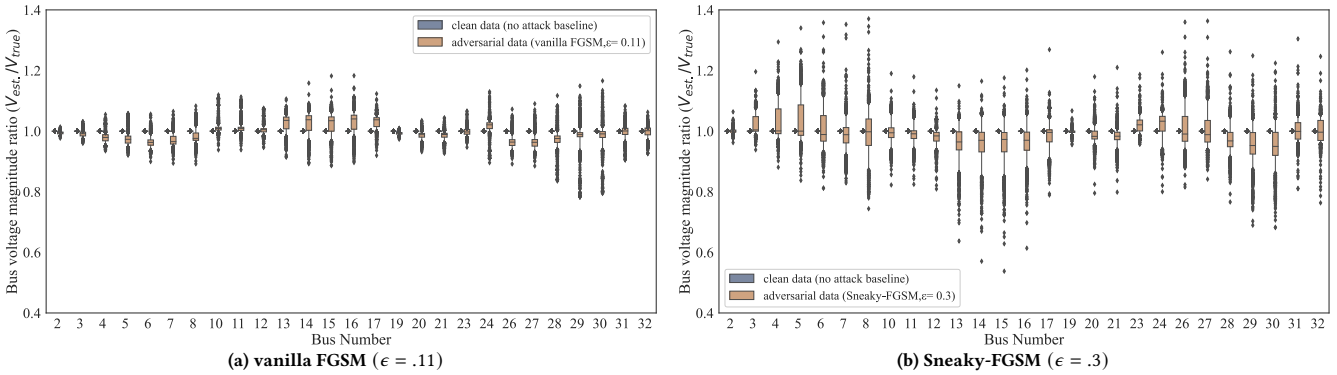
Table 1: Impact of adversarial attacks on the rule-based voltage regulation process.

7.3 Impact on the Voltage Control Scheme

To maximize the system observability with a small number of measurement devices, we instrument all of the head-ends of the primary feeders of our test system with D-PMUs as depicted in Figure 2. Hence, voltage regulators close to the endpoints can be controlled using the D-PMU measurements. However, we must rely on the estimated states to apply the VVO mechanism at intermediate buses, which may experience over-voltage or under-voltage issues due to changes in load during peak and off-peak hours. Until this phase of our experiment, we ran the simulation without installing any voltage regulator. As the simulation results suggest, a number of intermediate nodes experience the under-voltage problem during peak hours. We observe that the closest node near the substation bus that is affected by this issue is bus 6, and the problem persists as we travel further along the feeder. To address this, we utilize the *RegControl* object from the OpenDSS simulator to install a transformer with OLTC at line $5-6$ and set the corresponding control rule as *"If the voltage at bus* 6 *violates the limits, change the tap setting accordingly"*.[2] To investigate how the proposed Sneaky-FGSM attack affects the voltage control scheme, we simulate the BDD-integrated DSSE-based voltage regulation process using 24-hour load data (from 6:15am to 6:15am of the next day) in three different settings: a) in the absence of an attacker; b) under the vanilla FGSM attack; and c) under the Sneaky-FGSM attack. For the last two settings, we initiate the attacks starting from the second hour. During this simulation, if a particular measurement is flagged as 'bad' data, we replace the corresponding state estimate with the latest state estimate that was computed using a 'good' measurement.

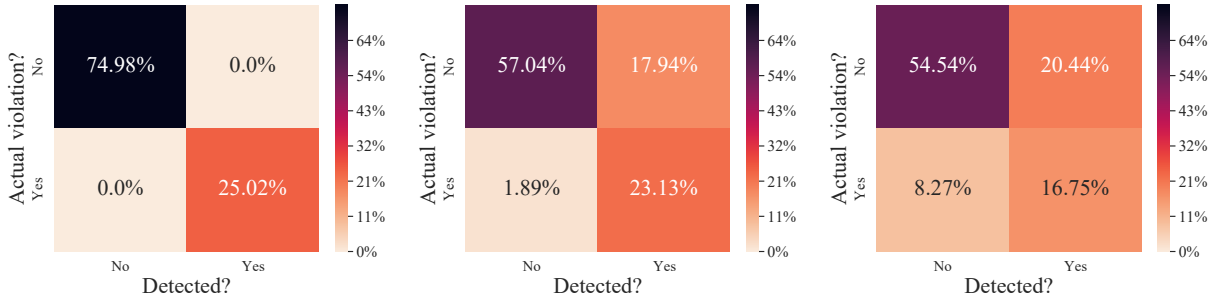
Figure 8 shows the simulation results. As we observe, in the absence of the attacker, the voltage control scheme correctly identifies the violation that took place at 6:45*am*, initializes the command to increase the tap setting, and brings the voltage to the specified range. However, when the attacker is present, the violation detection mechanism often fails, resulting in unnecessary tap operations

[2]Depending on the amount of the violation, one or more tap change actions may be performed.



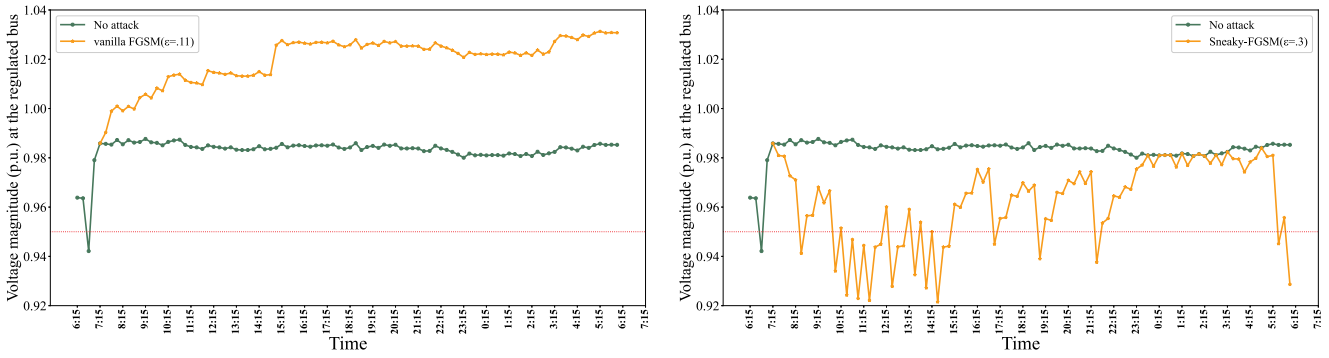
(a) vanilla FGSM (ϵ = .11) **(b) Sneaky-FGSM (ϵ = .3)**

Figure 6: Performance of the Stacked ResNetD model with clean data samples and adversarial data samples generated by the CNN surrogate. The two bar and whisker plots are presented next to each other for each bus.



(a) Stacked ResNetD state estimator **(b) Stacked ResNetD state estimator, under vanilla FGSM (ϵ = .11) attack** **(c) Stacked ResNetD state estimator, under Sneaky-FGSM (ϵ = .3) attack**

Figure 7: Confusion matrix: detecting voltage violation at load buses using the estimated states. Chosen ϵ values ensure that the attacker bypasses BDD mechanism with at least 95% success rate.



(a) Vanilla FGSM (ϵ=0.11) **(b) Sneaky-FGSM (ϵ=0.3)**

Figure 8: Impact of adversarial attacks on the rule-based voltage control mechanism. The green curve shows the voltage profile at the regulated bus under normal conditions. The orange curve shows the same voltage profile when the attacker is present. The system was operating normally with one tap change to resolve the under-voltage problem at 6:45am; the attack was then launched, starting at 7:15am, causing unnecessary tap operations throughout the day.

as well as voltage fluctuations and occasional under-voltage problems at the regulated bus. These issues are more pronounced under the Sneaky-FGSM attack. Table 1 shows the impact of these attacks in terms of the number of unnecessary tap changes and voltage limit violations during a day-long simulation.

7.4 Transferability of Adversarial Attacks

Given the vulnerability of the Stacked ResNetD state estimation model to the proposed adversarial attack, it is worth exploring the following questions: (a) how would the effectiveness of the

Victim Model	Attack Type	Surrogate Model	DSSE Error (RMSE)		Volt Violation Detection Acc (%)	
			Vanilla FGSM	Sneaky-FGSM	Vanilla FGSM	Sneaky-FGSM
CNN [11] ($\sim 3 \times 10^{-3}$, 97.16%)	White-box	CNN (ReLU)	**0.04**	**0.05**	78.93	**62.38**
	Black-box	CNN (tanh)	0.04	0.04	76.89	74.86
	Black-box	MLP (ReLU)	0.03	0.05	81.13	70.94
	Black-box	MLP (tanh)	0.03	0.05	**74.40**	71.66
Stacked ResNetD [11] ($\sim 3 \times 10^{-5}$, 99.99%)	White-box	Stacked ResNetD	**0.38**	**1.02**	53.23	39.63
	Black-box	CNN (ReLU)	0.12	0.22	80.17	71.29
	Black-box	CNN (tanh)	0.23	0.70	73.38	50.75
	Black-box	MLP (ReLU)	0.23	0.66	**45.52**	**36.00**
	Black-box	MLP (tanh)	0.31	0.54	46.87	36.13
KNN [60] ($\sim 8 \times 10^{-4}$, 99.49%)	Black-box	CNN (ReLU)	0.03	0.09	87.58	72.21
	Black-box	CNN (tanh)	0.05	0.15	79.45	47.32
	Black-box	MLP (ReLU)	**0.11**	**0.23**	**35.99**	**31.45**
	Black-box	MLP (tanh)	0.10	0.21	39.94	31.69

Table 2: Evaluating three victim models under vanilla FGSM ($\epsilon = 0.11$) and Sneaky-FGSM ($\epsilon = 0.3$) attacks crafted using the same surrogate model as the victim model (white-box) and other surrogate models (black-box). The victim model's performance, i.e. the DSSE RMSE and detection accuracy of voltage limit violations, on original data is reported in brackets in the first column.

proposed attacks change if a different surrogate model was used for state estimation? (b) are the proposed attacks effective against other data-driven state estimators? To address these questions, we employ two other data-driven DSSE techniques, namely a CNN with the same architecture as the surrogate model described in Section 5.2 and the KNN-based DSSE approach proposed in [60], and examine the efficacy of the proposed adversarial attacks in these cases. Additionally, we use four neural networks (CNN and MLP with ReLU and tanh activation functions) as the surrogate model. Table 2 summarizes the result. It can be seen that regardless of the choice of the surrogate model and whether it is identical to the victim model (as in the white-box attack) or not, the DSSE error increases to a great extent (especially in the case of Sneaky-FGSM), causing a significant drop (up to 64%) in the detection accuracy of voltage limit violations. We attribute this to the fact that feed-forward neural networks can represent a wide variety of functions and even approximate neural networks that have different architectures. While it is impossible to predict which surrogate model would be the best choice for the attacker (under the black-box assumption), our result suggests that by training either the CNN or MLP surrogate model, the attacker could cause sufficiently high error in the state estimation process to foil voltage regulation.

We wish to emphasize that higher error in the state estimation process does not always imply lower voltage violation detection accuracy. This is because the predicted state (i.e., the voltage magnitude) can be far from the true state, yet both may be below/above the minimum/maximum threshold. This implies that the attacker needs to take into consideration not only the error induced in DSSE but also the direction of the induced error. We defer the analysis of targeted evasion attacks to future work.

8 CONCLUSION

In this paper, we present a comprehensive analysis of (a) the security and robustness of data-driven DSSE techniques, (b) the effectiveness of conventional BDD mechanism against black-box adversarial (evasion) attacks, and (c) the deleterious impact of these types of attacks on distribution system control and operation practices that rely on the DSSE result. Our analysis suggests that in general, data-driven DSSE processes are vulnerable to stealthy and effective adversarial attacks that can fool the BDD mechanism with at least 95% success rate, this is while the attacker does not need to have any prior knowledge of the distribution system model or the ML model used for state estimation. This makes these types of attacks more practical and likely than conventional FDIA, in which the attacker is assumed to have some prior knowledge of the system model to launch an effective attack. We also propose a novel Sneaky-FGSM attack that outwits the BDD mechanism more frequently than the vanilla FGSM, and is capable of wreaking greater havoc in the control system.

This study opens the path to a number of interesting research directions in the area of smart grid cyber-security. One major limitation of the proposed attacks is that they solely care about inducing error in the estimation process while bypassing the BDD mechanism, without aiming for a specific target. In future work, we aim to analyze the effects of targeted evasion attacks, where the attacker pushes the state estimates in a certain direction causing only certain power quality issues (e.g., just over-voltage incidents), and model poisoning attacks on ML-based DSSE approaches, and undertake a comparative analysis of different security measures that can be taken to prevent and recover from adversarial attacks. We will also consider other threat models, e.g., when the adversary has read and write access to real-time measurements of specific sensors only. We plan to investigate whether adversarial samples can be added during model training to design more resilient ML-based state estimators. Furthermore, we intend to explore redesigning the conventional BDD mechanism such that it can detect adversarial samples effectively and reliably.

ACKNOWLEDGMENTS

This research was supported in part by funding from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2019-04349), and the Canada First Research Excellence Fund as part of the University of Alberta's Future Energy Systems research initiative (CFREF-2015-00001).

REFERENCES

[1] [n. d.]. The Aggregated Challenges of Regulating Energy Usage Data. Retrieved January 23, 2023 from https://eq-research.com/blog/the-aggregated-challenges-of-regulating-energy-usage-data/

[2] [n. d.]. IEEE PES distribution systems analysis subcommittee, radial test feeders. Retrieved January 23, 2023 from https://cmte.ieee.org/pes-testfeeders/resources/

[3] [n. d.]. Smart Distribution Systems for a Low Carbon Energy Future Workshop, CIRED workshop, 6 June 2011, Frankfurt, Germany. Retrieved January 31, 2023 from https://grouper.ieee.org/groups/td/dist/da/doc/2011%20CIRED%20Panel%20Tutorial%20binder_AH.pdf

[4] [n. d.]. Voltage tolerance boundary. Retrieved January 31, 2023 from https://www.pge.com/includes/docs/pdfs/mybusiness/customerservice/energystatus/powerquality/voltage_tolerance.pdf

[5] Faya Safirra Adi, Yee Jin Lee, and Hwachang Song. 2020. State estimation for dc microgrids using modified long short-term memory networks. *Applied Sciences* 10, 9 (2020), 3028.

[6] Fiaz Ahmad, Muhammad Tariq, and Ajmal Farooq. 2019. A novel ANN-based distribution network state estimator. *International Journal of Electrical Power & Energy Systems* 107 (2019), 200–212.

[7] Mukhtar Ahmad. 2013. *Power system state estimation.* Artech house.

[8] Omid Ardakanian, Vincent W. S. Wong, Roel Dobbe, Steven H. Low, Alexandra von Meier, Claire J. Tomlin, and Ye Yuan. 2019. On Identification of Distribution Grids. *IEEE Transactions on Control of Network Systems* 6, 3 (2019), 950–960.

[9] Mesut E Baran and Felix F Wu. 1989. Network reconfiguration in distribution systems for loss reduction and load balancing. *IEEE Power Engineering Review* 9, 4 (1989), 101–102.

[10] Arnab Bhattacharjee, Sukumar Mishra, and Ashu Verma. 2022. Deep Adversary based Stealthy False Data Injection Attacks against AC state estimation. In *2022 IEEE PES 14th Asia-Pacific Power and Energy Engineering Conference (APPEEC)*. IEEE, 1–7.

[11] Narayan Bhusal, Raj Mani Shukla, Mukesh Gautam, Mohammed Benidris, and Shamik Sengupta. 2021. Deep ensemble learning-based approach to real-time power system state estimation. *International Journal of Electrical Power & Energy Systems* 129 (2021), 106806.

[12] Arturo S Bretas, Aquiles Rossoni, Rodrigo D Trevizan, and Newton G Bretas. 2020. Distribution networks nontechnical power loss estimation: A hybrid data-driven physics model-based framework. *Electric Power Systems Research* 186 (2020), 106397.

[13] Zhiyuan Cao, Yubo Wang, Chi-Cheng Chu, and Rajit Gadh. 2020. Scalable distribution systems state estimation using long short-term memory networks as surrogates. *IEEE Access* 8 (2020), 23359–23368.

[14] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*. IEEE, 39–57.

[15] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. 2017. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526* (2017).

[16] Yize Chen, Yushi Tan, and Deepjyoti Deka. 2018. Is machine learning in power systems vulnerable?. In *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. IEEE, 1–6.

[17] Yize Chen, Yushi Tan, and Baosen Zhang. 2019. Exploiting vulnerabilities of load forecasting through adversarial attacks. In *Proceedings of the tenth ACM international conference on future energy systems*. 1–11.

[18] Kaveh Dehghanpour, Zhaoyu Wang, Jianhui Wang, Yuxuan Yuan, and Fankun Bu. 2018. A survey on state estimation techniques and challenges in smart distribution systems. *IEEE Transactions on Smart Grid* 10, 2 (2018), 2312–2322.

[19] Roger C Dugan and Thomas E McDermott. 2011. An open source platform for collaborating on smart grid research. In *2011 IEEE power and energy society general meeting*. IEEE, 1–7.

[20] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-Box Adversarial Examples for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 31–36.

[21] Maria Fotopoulou, Stefanos Petridis, Ioannis Karachalios, and Dimitrios Rakopoulos. 2022. A Review on Distribution System State Estimation Algorithms. *Applied Sciences* 12, 21 (2022), 11073.

[22] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[23] Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. 2019. Simple black-box adversarial attacks. In *International Conference on Machine Learning*. PMLR, 2484–2493.

[24] Moosa Moghimi Haji and Omid Ardakanian. 2019. Practical Considerations in the Design of Distribution State Estimation Techniques. In *2019 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. IEEE, 1–6.

[25] Eklas Hossain, Imtiaj Khan, Fuad Un-Noor, Sarder Shazali Sikander, and Md Samiul Haque Sunny. 2019. Application of big data and machine learning in smart grid, and associated security concerns: A review. *IEEE Access* 7 (2019), 13960–13988.

[26] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. 2018. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*. PMLR, 2137–2146.

[27] Linxi Jiang, Xingjun Ma, Shaoxiang Chen, James Bailey, and Yu-Gang Jiang. 2019. Black-box adversarial attacks on video recognition models. In *Proceedings of the 27th ACM International Conference on Multimedia*. 864–872.

[28] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236* (2016).

[29] Subhash Lakshminarayana, Jabir Shabbir Karachiwala, Teo Zhan Teng, Rui Tan, and David K. Y. Yau. 2019. Performance and Resilience of Cyber-Physical Control Systems With Reactive Attack Mitigation. *IEEE Transactions on Smart Grid* 10, 6 (2019), 6640–6654.

[30] Gaoqi Liang, Junhua Zhao, Fengji Luo, Steven R Weller, and Zhao Yang Dong. 2016. A review of false data injection attacks against modern power systems. *IEEE Transactions on Smart Grid* 8, 4 (2016), 1630–1638.

[31] Tian Liu and Tao Shu. 2019. Adversarial false data injection attack against nonlinear ac state estimation with ANN in smart grid. In *15th EAI International Conference on Security and Privacy in Communication Networks (SecureComm)*. Springer, 365–379.

[32] Tian Liu and Tao Shu. 2021. On the security of ANN-based AC state estimation in smart grid. *Computers & Security* 105 (2021), 102265.

[33] Yuan Liu, Omid Ardakanian, Ioanis Nikolaidis, and Hao Liang. 2022. False Data Injection Attacks on Smart Grid Voltage Regulation with Stochastic Communication Model. *IEEE Transactions on Industrial Informatics* (2022), 1–11.

[34] Yao Liu, Peng Ning, and Michael K Reiter. 2011. False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security (TISSEC)* 14, 1 (2011), 1–33.

[35] CN Lu, JH Teng, and W-HE Liu. 1995. Distribution system state estimation. *IEEE Transactions on Power systems* 10, 1 (1995), 229–240.

[36] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[37] Ankur Majumdar, Yashodhan P Agalgaonkar, Bikash C Pal, and Ralph Gottschalg. 2017. Centralized Volt–Var optimization strategy considering malicious attack on distributed energy resources control. *IEEE Transactions on Sustainable Energy* 9, 1 (2017), 148–156.

[38] Efthymios Manitsas, Ravindra Singh, Bikash C. Pal, and Goran Strbac. 2012. Distribution System State Estimation Using an Artificial Neural Network Approach for Pseudo Measurement Modeling. *IEEE Transactions on Power Systems* 27, 4 (2012), 1888–1896.

[39] Christoph Johannes Meinrenken, Noah Rauschkolb, Sanjmeet Abrol, Tuhin Chakrabarty, Victor C Decalf, Christopher Hidey, Kathleen McKeown, Ali Mehmani, Vijay Modi, and Patricia J Culligan. 2020. MFRED (public file, 15/15 aggregate version): 10 second interval real and reactive power in 390 US apartments of varying size and vintage. https://doi.org/10.7910/DVN/X9MIDJ

[40] Gautam Raj Mode and Khaza Anuarul Hoque. 2020. Adversarial examples in deep learning for multivariate time series regression. In *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*. IEEE, 1–10.

[41] Alcir Monticelli. 2012. *State estimation in electric power systems: a generalized approach.* Springer Science & Business Media.

[42] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2574–2582.

[43] Luis Muñoz-González, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C Lupu, and Fabio Roli. 2017. Towards poisoning of deep learning algorithms with back-gradient optimization. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*. 27–38.

[44] Luis Muñoz-González, Bjarne Pfitzner, Matteo Russo, Javier Carnerero-Cano, and Emil C Lupu. 2019. Poisoning attacks with generative adversarial nets. *arXiv preprint arXiv:1906.07773* (2019).

[45] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. 2016. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277* (2016).

[46] Anggoro Primadianto and Chan-Nan Lu. 2016. A review on distribution system state estimation. *IEEE Transactions on Power Systems* 32, 5 (2016), 3875–3883.

[47] Saaed Rahimi, Mattia Marinelli, and Federico Silvestro. 2012. Evaluation of requirements for Volt/Var control and optimization function in distribution management systems. In *2012 IEEE International Energy Conference and Exhibition (ENERGYCON)*. IEEE, 331–336.

[48] Ali Sayghe, Junbo Zhao, and Charalambos Konstantinou. 2020. Evasion attacks with adversarial deep learning against power system state estimation. In *2020 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, 1–5.

[49] Rui Shu, Tianpei Xia, Laurie Williams, and Tim Menzies. 2022. Omni: automated ensemble with unexpected models against adversarial evasion attack. *Empirical*

Software Engineering 27 (2022), 1–32.

[50] MR Starke, DT Rizy, and MA Young. 2013. Synchrophasor technologies and their deployment in the recovery act smart grid programs. *Report US Department of Energy* (2013).

[51] Jiwei Tian, Buhong Wang, Jing Li, and Charalambos Konstantinou. 2022. Adversarial attack and defense methods for neural network based state estimation in smart grid. *IET Renewable Power Generation* 16, 16 (2022), 3507–3518.

[52] Jiwei Tian, Buhong Wang, Jing Li, and Charalambos Konstantinou. 2022. Datadriven false data injection attacks against cyber-physical power systems. *Computers & Security* 121 (2022), 102836.

[53] Jiwei Tian, Buhong Wang, Jing Li, Zhen Wang, Bowen Ma, and Mete Ozay. 2022. Exploring targeted and stealthy false data injection attacks via adversarial machine learning. *IEEE Internet of Things Journal* 9, 15 (2022), 14116–14125.

[54] Jiwei Tian, Buhong Wang, Zhen Wang, Kunrui Cao, Jing Li, and Mete Ozay. 2021. Joint adversarial example and false data injection attacks for state estimation in power systems. *IEEE Transactions on Cybernetics* 52, 12 (2021), 13699–13713.

[55] Florian Tramr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. 2018. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, Vol. 1. 2.

[56] Chunming Tu, Xi He, Xuan Liu, and Peng Li. 2018. Cyber-attacks in PMU-based power network and countermeasures. *IEEE Access* 6 (2018), 65594–65603.

[57] P Venkatesh, BV Manikandan, S Charles Raja, and A Srinivasan. 2012. *Electrical power systems: analysis, security and deregulation.* PHI Learning Pvt. Ltd.

[58] Jingyu Wang, Dongyuan Shi, Yinhong Li, Jinfu Chen, Hongfa Ding, and Xianzhong Duan. 2018. Distributed framework for detecting PMU data manipulation attacks with deep autoencoders. *IEEE Transactions on smart grid* 10, 4 (2018), 4401–4410.

[59] Xinan Wang, Di Shi, Jianhui Wang, Zhe Yu, and Zhiwei Wang. 2019. Online identification and data recovery for PMU data manipulation attack. *IEEE Transactions on Smart Grid* 10, 6 (2019), 5889–5898.

[60] Yang Weng, Rohit Negi, Christos Faloutsos, and Marija D Ilić. 2016. Robust data-driven state estimation for smart grid. *IEEE Transactions on Smart Grid* 8, 4 (2016), 1956–1967.

[61] Jun Yan, Bo Tang, and Haibo He. 2016. Detection of false data attacks in smart grid with supervised learning. In *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1395–1402.

[62] Weili Yan et al. 2021. A Stealthier False Data Injection Attack against the Power Grid. In *IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. IEEE, 108–114.

[63] Chaofei Yang, Qing Wu, Hai Li, and Yiran Chen. 2017. Generative poisoning attack method against neural networks. *arXiv preprint arXiv:1703.01340* (2017).

[64] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. 2019. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on neural networks and learning systems* 30, 9 (2019), 2805–2824.

[65] Ahmed S Zamzam, Xiao Fu, and Nicholas D Sidiropoulos. 2019. Data-driven learning-based optimization for distribution system state estimation. *IEEE Transactions on Power Systems* 34, 6 (2019), 4796–4805.

[66] Ahmed Samir Zamzam and Nicholas D. Sidiropoulos. 2020. Physics-Aware Neural Networks for Distribution System State Estimation. *IEEE Transactions on Power Systems* 35, 6 (2020), 4347–4356. https://doi.org/10.1109/TPWRS.2020.2988352

[67] Liang Zhang, Gang Wang, and Georgios B Giannakis. 2019. Power system state forecasting via deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8092–8096.

[68] Liang Zhang, Gang Wang, and Georgios B Giannakis. 2019. Real-time power system state estimation and forecasting via deep unrolled neural networks. *IEEE Transactions on Signal Processing* 67, 15 (2019), 4069–4077.

[69] Yingchen Zhang, Andrey Bernstein, Andreas Schmitt, and Rui Yang. 2019. *State estimation in low-observable distribution systems using matrix completion.* Technical Report. National Renewable Energy Lab.(NREL), Golden, CO (United States).

[70] Mo Zhou and Vishal M Patel. 2022. On Trace of PGD-Like Adversarial Attacks. *arXiv preprint arXiv:2205.09586* (2022).