On Adversarial Robustness of Data-Driven State Estimation Techniques

by

Afia Afrin

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

Abstract

The increasing complexity of electric power grids, owing to the integration of Distributed Energy Resources (DER), electric vehicles, energy storage systems, and advanced metering infrastructure, has stimulated a surge in research on machine learning-based state estimation and control. In this thesis, we examine the robustness of machine learning-based Distribution System State Estimation (DSSE) techniques to a class of adversarial attacks, known as evasion attacks. In these attacks, the attacker manipulates real-time measurements of 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. Moreover, we present Targeted-FGSM, a powerful targeted evasion attack strategy that is capable of misleading the state estimator in a certain direction to cause specific power quality issues. Upon analyzing the covert nature of these attacks, we propose the adoption of the feature attribution-based detection strategy to build robust safeguarding mechanism for DSSE techniques. Our results suggest that incorporating machine learning models in DSSE is a double-edged sword and calls for more research in this area to ensure the robustness of these models to adversarial samples.

Preface

This thesis is an original work by Afia Afrin. The research conducted as part of this thesis is led by Dr. Omid Ardakanian in the Sustainable Computing Lab at the University of Alberta. A portion of the codebase for this work, specifically the OpenDSS simulation for the extended IEEE-33 bus system, is based on the *VoltageRegulation* github repository created by Moosa Moghimi Haji, a former postdoctoral fellow in the Sustainable Computing Lab.

The untargeted attack strategies presented in Chapter 3 and the corresponding simulation results presented in Chapter 4 are based on the following conference paper:

Afia Afrin and Omid Ardakanian, "Adversarial Attacks on Machine Learning-Based State Estimation in Power Distribution Systems", in Proceedings of the 14*th ACM International Conference on Future Energy Systems*, 2023, pp. 446-458.

I was responsible for conducting the literature review, developing the methodology, running experiments, and producing results. The other author is my supervisor who guided the project and edited the manuscript.

The targeted attack strategy presented in Sections 3.4.3 and 4.2.5, and the detection mechanisms discussed in Chapter 5 will be submitted to a journal.

To Touhidul

For believing in me even when I could not.

Acknowledgements

I would like to express my gratitude to my supervisor, Dr. Omid Ardakanian. From the very beginning, when I was struggling to formulate the problem definition, to the very end as I am putting up the thesis, he has been the lightbearer guiding me towards clarity and understanding. He is to this thesis, what trees are to the earth– without his wonderful guidance and unwavering support this thesis would never exist. I would also like to thank my amazing peers from the Sustainable Computing Research Group for their constructive comments and suggestions, and for the wonderful time we spent together in the lab.

I am grateful to the University of Alberta Future Energy Systems and the Canada First Research Excellence Fund (CFRF) for their generous funding, enabling the realization of this research. I would like to extend my gratitude to Google DeepMind for the *DeepMind scholarship*, which has been instrumental in facilitating this work.

Needless to mention, I am indebted to my parents for everything. Their endless sacrifices have been the driving force behind the fulfillment of my dreams.

Far away from home, this little journey of earning a Master's degree would have been a thousand times more stressful if I had not discovered Ashley Wian and her precious little boys, Kobe Wian and Kodi Wian on the internet. My heartfelt gratitude and well wishes go out to this beautiful family for all the joy and warmth they have brought into the world.

To the fantastic five– Labiba, Mumu, Oyshee, Prova, and Shifat– thank you for being here and making this foreign land feel like *home.* Thank you for gifting me with the most colorful two years– *this part of my life, this little* *part, is called 'the Fall'.*

Lastly, and most importantly, I would like to thank the members of the StackOverflow and CrossValidated communities. Without their invaluable contributions and assistance, two years would not have been sufficient to bring this work to completion.

Contents

List of Tables

List of Figures

Acronyms

BDD Bad Data Detection.

DMS Distribution Management System.

D-PMUs Distribution-level Phasor Measurement Units.

DER Distributed Energy Resources.

DSO Distribution System Operators.

DSSE Distribution System State Estimation.

FDIA False Data Injection Attacks.

FGSM Fast Gradient Sign Method.

GLRT Generalized Likelihood Ratio Test.

KLD Kullback–Leibler Divergence.

NAD Neural Attack Detector.

OLTC On-load Tap Changers.

OpenDSS Open Distribution Simulator Software.

PMU Phasor Measurement Units.

PSSE Power System State Estimation.

SCADA Supervisory Control and Data Acquisition.

VVO Volt/VAR Optimization.

WLS Weighted Least-Squares.

Chapter 1

Introduction

On August 14, 2003, a high-voltage transmission line tripped in Northern Ohio due to a tree contact – a typical *fault* that was supposed to trigger an alarm in the control room for immediate attention. Unfortunately, the alarm never went off, and this seemingly isolated incident spiraled into a series of catastrophic events that led to the great blackout of 2003 in North America, affecting 50 million people, causing at least 11 deaths and damages worth around $6 billion [80]. Although a combination of factors, including software bugs, equipment failures, and human error were initially blamed for the event, an investigation launched by the North American Electric Reliability Corporation (NERC) concluded that the blackout could have been confined to a smaller region had operators been aware of the system state [79], underscoring the importance of situational awareness.

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 (PMU) installed at specific nodes in the network. But in the past decade, the growing adoption of 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

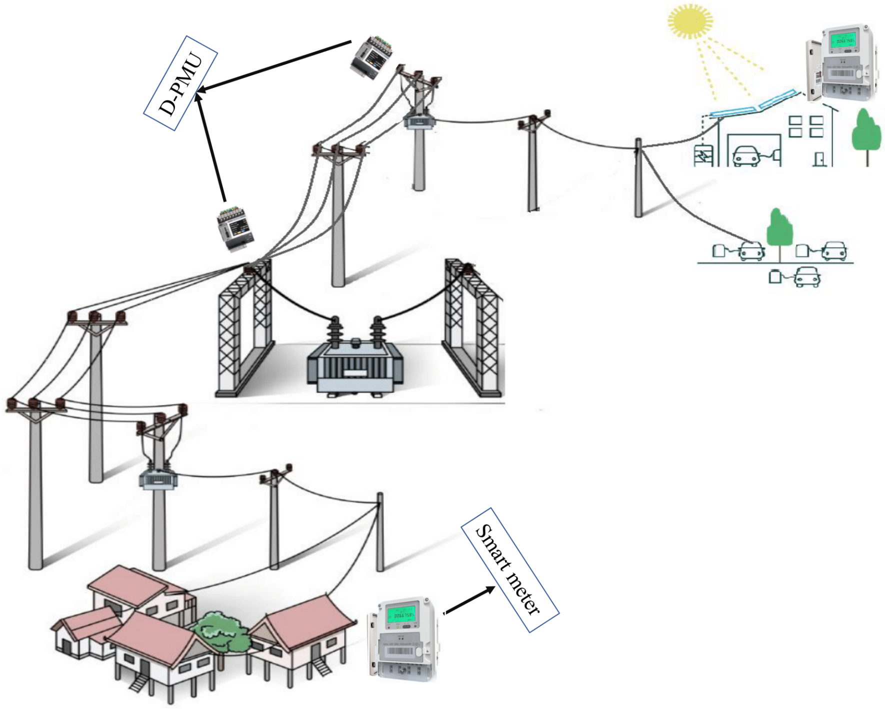


Figure 1.1: Power distribution systems are partially observable with a limited number of measurement devices.

be increasingly incorporated in distribution system operation practices [84].

1.1 State Estimation in Distribution Systems

Power distribution systems are complex networks with a radial operational structure and numerous interconnected components, including substations, step-down transformers, feeders, and loads. Monitoring and control of such systems require a wide range of sensors, reliable communication infrastructure, and data acquisition devices. However, due to factors such as cost, sensor deployment and communication constraints, there is limited observability in distribution systems today [10]. Figure 1.1 shows an example of a partially observable power distribution system equipped with a small number of measurement devices, i.e., Distribution-level Phasor Measurement Units (D-PMUs), in addition to smart meters that are installed at customer premises.

The complexity of distribution systems is growing due to the large-scale integration of DER, increased penetration of electric vehicles, and proliferation of distributed energy storage systems and power electronics. These new components cause bidirectional power flow, wide fluctuations in voltage, and congestion issues, calling for enhanced monitoring and more stringent control of the distribution system through a Distribution Management System (DMS). One of the key components of DMS is *state estimation*, which is defined as the problem of identifying the unobservable parameters, a.k.a. *state variables*, from the available measurements in a power system [63]. Examples of state variables are the voltage magnitude and phase angle of a subset of buses in the distribution network.

We note that the application of state estimation is not limited to power grids. In any large and partially observable distributed system, such as transportation system, communication network, industrial process control, and autonomous vehicles, state estimation plays a critical role in enhancing the situational awareness of the operators by providing an accurate and comprehensive picture of the system's behavior. It enables them to identify potential bottlenecks or congestion issues, and take proactive measures to address them.

1.2 Motivating Data-Driven State Estimation

The state estimation problem can be formulated as a system of nonlinear equations, which is typically solved as a Weighted Least-Squares (WLS) problem [72] 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 [27]. This results in fewer measurements than unknowns, rendering WLS-based estimators ineffective [130]. 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 [3]. 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 [6].

Inspired by the success of ML techniques in approximating complex physics-based models, 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 [40]. In particular, neural networks trained on historical measurements or simulation data have been used to estimate the system state from existing measurements [11], [115], [128], [129], initialize the Gauss-Newton method so it enjoys quadratic convergence to the true latent state of the system [126], or generate pseudo-measurements to compensate for the lack of sufficient measurements when solving DSSE using traditional model-based techniques [67]. More recently, physics-aware neural networks [127] 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 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, previous work does not investigate whether these techniques are robust to *adversarial samples* [38] 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 [23], [71], [125]. 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 *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.

1.3 Problem Statement

In this section, we present 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.

1.3.1 Distribution System State Estimation

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

$$\mathbf{z} = h(\mathbf{x}) + \xi, \tag{1.1}$$

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 [3]:

$$\min_{x} J(\mathbf{x}) = \sum_{i=1}^{m} \left(z_i - h_i\left(\mathbf{x}\right)\right)^2 / R_{ii} \tag{1.2}$$

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 (1.2) in vector/matrix form as follows:

$$\min_{x} \left[\mathbf{z} - h(\mathbf{x})\right]^{\top} \mathbf{R}^{-1} \left[\mathbf{z} - h(\mathbf{x})\right] \tag{1.3}$$

Due to the high computational overhead and possibility of getting stuck in local minima [62], $h(\cdot)$ is often linearized:

$$h(\mathbf{x}) = \mathbf{H}\mathbf{x} \tag{1.4}$$

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 (1.3) and (1.4), we derive the estimated state as:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \left[\mathbf{z} - \mathbf{H}\mathbf{x}\right]^{\top} \mathbf{R}^{-1} \left[\mathbf{z} - \mathbf{H}\mathbf{x}\right] \tag{1.5}$$

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 [106]:

$$\hat{\mathbf{x}} = \left(\mathbf{H}^{\top}\mathbf{W}\mathbf{H}\right)^{-1} \mathbf{H}^{\top}\mathbf{W}\mathbf{z} \tag{1.6}$$

1.3.2 Residual-based Bad Data Detection

As the reliability of estimated states is heavily dependent on the accuracy of measurements, Distribution System Operators (DSO) 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 [3]. From (1.2), the residual can be rewritten as:

$$J(\mathbf{x}) = \sum_{i=1}^{m} \frac{e_i^2}{R_{ii}} = \sum_{i=1}^{m} \left(\frac{e_i}{\sigma_i}\right)^2 \tag{1.7}$$

Notice that Equation (1.7) 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 DSO may either discard the estimated state and replace it with a previous state estimate or try to identify the source of bad data, eliminate the bad measurement(s), and re-estimate the current state.

1.3.3 Voltage Regulation using DSSE Result

The bus voltage is closely related to the load demand in an electrical power distribution system. When the total demand of connected loads increases at a given bus, more power is drawn from the distribution system. This increased demand subsequently causes a drop in the voltage at the distribution bus due to higher power flow through the distribution lines and transformers. On the contrary, when the total load demand decreases, less power is drawn from the distribution system, leading to an increase in the voltage at the distribution bus. Maintaining a stable and appropriate voltage level at the distribution bus is crucial for the reliable and efficient operation of the electrical distribution system. Voltage control equipment, such as transformers and voltage regulators, are used to control and adjust the voltage at the distribution bus to ensure that it remains within acceptable limits and meets the requirements of connected devices and consumers. A voltage limit violation in a power distribution system occurs when the voltage level exceeds or drops below the acceptable 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 (OLTC), are used to quickly respond to voltage fluctuations.

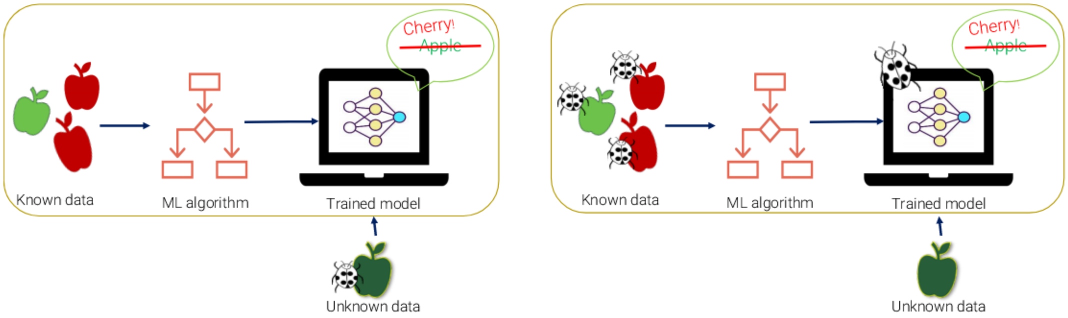
Due to high installation costs, D-PMUs are not currently deployed at each node of a distribution system, despite their ability to provide highly precise and frequent data [95]. Therefore, estimated states from DSSE are often used instead of the measurements when they are missing to detect voltage limit violations [34] and perform Volt/VAR Optimization (VVO) [66]. 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 [86]. The SCADA-controlled VVO is often studied as two independent problems, VAR optimization and Voltage control [86]. 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 [94].

1.3.4 Adversarial Attacks

An adversarial attack is a deliberate and targeted attempt to manipulate the behavior of a machine learning model (e.g. the model used for state estimation) by introducing specially crafted input data, known as adversarial samples, during the model training or inference phase. The two most common types of adversarial attack strategies are:

- *Evasion Attacks:* Attacks in this category add carefully crafted pertur-



(a) Evasion attack

(b) Poisoning attack

Figure 1.2: A pictorial description of evasion and poisoning attack strategies against a classifier.

bations to the benign samples in the test set with the goal of producing erroneous output, thereby reducing the accuracy of the machine learning model post deployment. Popular evasion attacks include *FGSM* [38], *Basic Iterative Method (BIM)* [54], *Projected Gradient Descent (PGD)* [65], *DeepFool* [73], and *Carlini–Wagner Attack (C&W)* [17].

- *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 [76], [77], [120] (a.k.a data poisoning), whereas in the latter case, the adversary creates a *backdoor* into the learning model using poisoning strategies [20] (a.k.a model poisoning).

Figure 1.2 represent a visual overview of the two types of adversarial attacks.

We note that there is no known work that analyzes the vulnerability of data-driven DSSE approaches to adversarial attacks, in particular evasion attacks, and elucidates their effects on the voltage control system. This is the problem we study in this thesis.

1.4 Objectives and Contributions

In this thesis, we aim to address the following research questions.

- While there is a consensus in the research community that data-driven DSSE approaches are superior to conventional static and dynamic state estimation approaches in terms of effectiveness, are they sufficiently reliable and robust to be incorporated into power distribution grids?
- How effective is the conventional BDD mechanism in preventing adversarial attacks on DSSE?
- What impact do adversarial attacks have on control system(s) that rely on the DSSE results?
- Is there a more effective detection mechanism that can replace conventional residual-based BDD to mitigate adversarial attacks?

To answer these questions we investigate the adversarial robustness of two state-of-the-art data-driven DSSE models, namely Stacked ResNetD [11] and Robust K-Nearest Neighbours (R-KNN) [115], that have been shown to outperform several other electrical model-agnostic state estimation techniques. 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. Upon analyzing the untargeted nature of the proposed attack strategies, we devise a targeted adversarial attack which is able to create certain power quality issues. 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. The contribution of this thesis is fourfold:

- We present a black-box *evasion attack* against state-of-the-art data-driven DSSE techniques. Using surrogate models that are different from the victim state estimation model, we argue that the attacker needs neither the knowledge of the ML model used in DSSE (as in white-box attacks), nor any information about the distribution system model. We then 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 acknowledge the untargeted nature of the proposed black-box attack strategies and devise a targeted adversarial attack that is able to mislead the state estimator in a certain direction, thereby inducing certain power quality issues.

- We demonstrate the inefficacy of the conventional residual-based BDD mechanism in detecting adversarial measurements and propose an effective detection-based safeguarding mechanism to protect data-driven state estimators from adversarial attacks. We compare the performance of the proposed detection method with two strong baselines proposed in prior work. Through extensive experiments, we show that our proposed method achieves superior performance in detecting diverse adversarial attacks crafted with different types of surrogate models and varying levels of added noise.

- 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.

Our findings suggest that

Data-driven DSSE techniques are not presently robust to carefully crafted adversarial data, and more research is warranted to address their vulnerabilities and build robust protection strategies before they can be incorporated into distribution system operation practices.

1.5 Outline

The rest of the thesis is organized as follows. Chapter 2 discusses the related work on data-driven DSSE techniques, adversarial attacks, and false data injection attacks on power systems, vulnerability analysis of ML-based power system applications, and safeguarding strategies for state estimators. Chapter 3 presents the methodology used in this thesis– the architecture of the data-driven DSSE strategies, attack formulation, and a brief introduction of the rule-based voltage regulation process. Chapter 4 describes the experimental setup and presents the experimental results on analyzing the adversarial vulnerability of data-driven DSSE approaches. Chapter 5 offers a brief description of the proposed detection strategy, compare its performance with two state-of-the-art baselines and presents the corresponding experimental results. Finally, Chapter 6 concludes the thesis by discussing the limitations of this work and possible future directions.

Chapter 2

Literature Review

In this chapter, we review the previous work concerning data-driven DSSE and its vulnerabilities. Section 2.1 discusses existing data-driven DSSE approaches. Sections 2.2 and 2.3 introduce adversarial attacks and false data injection attacks, respectively. Section 2.4 surveys the literature on robustness analysis of data-driven DSSE approaches and identifies the gaps that we address in this thesis. Finally, Section 2.5 provides a summary of recent efforts to build efficient detection and prevention mechanisms for safeguarding distribution system state estimators.

2.1 Data-Driven DSSE Strategies

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 [84], especially in distribution networks with high DER penetration. For example, real-time distribution system state estimators based on various Deep Neural Network (DNN) architectures [1], [2], [128], [129], and K-nearest neighbors (KNN) [115] were proposed in the literature. An ML-based state estimator that takes advantage of an ensemble of residual neural networks (ResNet) [11] has been recently shown to outperform several other ML-based techniques, including multilayer perceptron (MLP) and convolutional neural network (CNN). A deep learning approach to Bayesian distribution system state estimation for unobservable distribution systems has been proposed in [70]. In that work, a fully connected neural network has been used to learn the parameter of a Bayesian state estimator. In recent work [127], a physics-aware neural network (PAWNN) has been proposed to estimate the state of the distribution system where knowledge of the underlying physical system is used to prune the dense neural network, reducing overfitting. Several studies also employ a hybrid approach in which an ML model is combined with a traditional approach (such as WLS and the least absolute value) [14], [15], [126]. 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.

These data-driven state estimators have been shown to be better alternatives to conventional electric model-based state estimators due to their high accuracy as well as faster and guaranteed convergence. However, the robustness analysis of these models is still an under-explored area of research. We present a detailed discussion on the research gaps present in the literature in Section 2.4.

2.2 Adversarial Attacks

Recall the adversarial attacks introduced in Section 1.3.4. These attacks can be designed using a *white-box* or *black-box* approach. During a white-box attack, the adversary uses the knowledge of the ML model used in the classification or regression task, including its architecture, hyper-parameters, and weights associated with connections, to generate adversarial samples [17], [30], [73]. We refer to this model as the *victim model*. 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 [39], [46], [48]. In the context of black-box attacks, the victim model is often referred to as an *oracle* – an abstract entity that can provide information or answer specific queries [83]. Previous studies have shown that due to the transferability of adversarial samples, it is possible to design black-box attacks by training surrogate models that differ from the victim model [82]. Our work is inspired by this result.

2.2.1 Fast Gradient Sign Method

In this work, we focus primarily on FGSM and its variants. Introduced by Ian Goodfellow et al. in 2014, FGSM is considered to be the very first adversarial attack proposed against neural networks. The key idea behind this attack is to perturb the input data in the direction that maximizes the loss function of the model. By taking a step in the direction of the sign of the gradient, scaled by a small perturbation magnitude, the attack aims to find the adversarial example that is most likely to be misclassified by the model. Concretely, for any input sample $\mathbf{X}$, FGSM generates the corresponding adversarial sample given by:

$$\mathbf{X}' = \mathbf{X} + \epsilon * \operatorname{sign}\left(\nabla_{\mathbf{X}}\left[L(f(\mathbf{X};\theta), y_{true})\right]\right) \tag{2.1}$$

Here, ϵ is a small constant that controls the magnitude of the perturbation, y_{true} is the original label for input $\mathbf{X}$, and $f(\cdot;\theta)$ is a surrogate model utilized by the adversary. In the case of white-box attacks, $f(\cdot;\theta)$ can be the same as the victim model. However, for black-box attacks, the adversary trains an arbitrary model that performs the same task as the victim model. The training data can be collected from publicly available datasets or generated specifically for the purpose of training the surrogate model.

While general adversarial attacks aim to degrade the overall performance of the victim model in any conceivable manner, there exists a more sophisticated type of these attacks, called *targeted adversarial attacks*. The objective of a targeted adversarial attack is to cause the victim model to predict a specific target class or output. Unlike untargeted attacks, which aim to cause any misclassification (in classifier models) or misprediction (in regression models), targeted attacks are designed with a specific goal in mind. These attacks pose significant challenges and raise concerns in various applications. For example, in image recognition, an attacker might seek to manipulate an image so that a classifier identifies it as an object of interest. In autonomous driving, targeted attacks could be used to create misleading road signs or traffic signals that lead self-driving cars to make incorrect decisions, potentially resulting in accidents or other dangerous situations.

2.2.2 Targeted vs. Untargeted Attacks

Consider a classifier $f(\mathbf{x_i})$ that correctly classifies the input sample $\mathbf{x_i}$ into its original class y_i. An adversarial attack algorithm aims to generate an adversarial sample, $\mathbf{x'_i}$, that it is similar to the original sample, $\mathbf{x_i}$, according to some distance metric d_i, but is misclassified as $f(\mathbf{x_i}) \neq y_i$. We can define the two types of attack against a classifier model as follows:

- *Targeted Attack:* All the adversarial samples, $\mathbf{x'_1}, \mathbf{x'_2}, ..., \mathbf{x'_n}$, are generated such that they are misclassified into a pre-determined class y', where $y' \neq y_i$, for $i = 1, 2, ..., n$.
- *Untargeted Attack:* The generated adversarial samples are misclassified as any class except the true class.

Similarly, we can define these two types of attack against a regression model, $g(\mathbf{x_i})$, as follows:

- *Targeted Attack:* All the adversarial samples, $\mathbf{x'_1}, \mathbf{x'_2}, ..., \mathbf{x'_n}$, are generated such that $g(\mathbf{x'_i}) \geq g(\mathbf{x_i}) + t$ or $g(\mathbf{x'_i}) \leq g(\mathbf{x_i}) - t$, where $t > 0$ is the predefined target.
- *Untargeted Attack:* Generated adversarial samples shift the output of the victim model by any amount $t \neq 0$.

Targeted attacks are typically generated by using a customized loss function in the attack algorithm. One of the most popular targeted attack algorithms is the Carlini–Wagner (C&W) attack [17] which aims to find the smallest noise $\delta \in \mathbb{R}^n$ added to an image $\mathbf{X} \in \mathbb{R}^n$ that will change the classification result to a target class t, predefined by the attacker, by optimizing an adversarial loss function that consists of two parts – one for minimizing the perturbation and another for maximizing the adversarial loss for the perturbed input. The iterative least-likely class method, a.k.a iterative target class method (ITCM) [53], is another simple, yet powerful targeted attack generation approach that modifies the untargeted iterative FGSM algorithm to generate targeted attacks.

One major limitation of the above-mentioned targeted attack approaches is the lack of transferability, which hinders the development of black-box targeted attacks. Li et al. [57] identified the two main reasons behind this. First, existing transferable attacks use softmax cross-entropy as loss function which results in vanishing gradient problems in iterative targeted attacks. Second, traditional targeted attack strategies only focus on maximizing the probability of targeted class and ignore whether the adversarial examples are close to the original class. Consequently, in some cases, the targeted adversarial examples neither successfully transfer with the target label nor deceive the victim model effectively. To address these challenges, they introduced two novel concepts: Poincaré distance and triplet loss, for generating transferable targeted attacks. Poincaré distance has been used instead of the cross entropy loss to adapt the size of the gradient. Additionally, the triplet loss ensured that the victim model's output for the adversarial sample not only moved closer to the target label but also moved farther away from the true class label, improving the attack's effectiveness and transferability. In another recent work, replacing cross-entropy loss with logit loss has been found effective in generating efficient transferable targeted attacks that do not suffer from gradient disappearance during the perturbation generation process [132]. A source-independent generative approach to devise transferable targeted attacks against image classifiers has been proposed in [78]. This approach is built on a novel loss function that focuses on matching the distribution-level statistics of perturbed source and target samples.

Most of the targeted attack approaches were originally developed against classifiers, and there has been limited exploration into understanding their effects on regression models. Unlike classification models, where the goal is to misclassify the input, regression models output continuous values, making the attack strategy slightly different. Two targeted attack algorithms against electroencephalogram (EEG) based brain-computer interface (BCI) regression problems have been proposed in [69]. The authors adopted the ideas of C&W and iterative FGSM attacks and modified them to produce two targeted attack strategies that are effective on regression models, namely CW-R and IFGSM-R. A similar approach has been taken to generate targeted attacks on wind power forecasting models by leveraging the white-box PGD attack algorithm to minimize the mismatch between the victim model's prediction and the attacker's target in [43]. To the best of our knowledge, no prior study focused on generating targeted attacks against DSSE, which is essentially a regression problem.

2.3 False Data Injection Attacks in Power Systems

Liu et al. introduced the idea of stealthy false data injection attacks against transmission system state estimation [61]. They showed that an attacker can carry out stealthy FDIA that fool the traditional residual-based BDD mechanism if the attack vector **a** satisfies

$$\mathbf{a} = \mathbf{H}\mathbf{c} \tag{2.2}$$

where **c** is a nonzero vector of the same length as the system state to be estimated.

This result has served as a driving force for researchers to conduct in-depth investigations into the development of FDIA, their impacts on power system operations, and potential protective measures. FDIA can affect a wide range of smart grid functions and applications, including state estimation [28], [61], [87], [123], load forecasting [24], [74], demand response [26], [36], and SCADA system [37], [113]. Given the focus of our research, we only survey the literature on generating FDIA against state estimation in power distribution systems.

FDIA against DSSE: Unlike transmission systems, FDIA at the distribution system level have not been extensively explored yet. A realistic FDIA strategy against the WLS-based distribution system state estimators has been proposed in [28]. The proposed FDIA was generated under a relaxed assumption that the attacker does not have access to the true system states, and therefore, utilizes only information about the local state obtained by approximating the entire system state using a small number of branch flow or bus injection measurements to generate attack vectors. One major limitation of their work is that they used a simplified single-phase feeder model. Later, Zhuang et al. [135] extended that work and investigated the vulnerabilities of a linear DSSE approach in multiphase and unbalanced smart distribution systems.

Unlike targeted adversarial attacks, in the context of FDIA, the term *targeted attack* usually refers to a specific type of security attack aimed at compromising or damaging a particular component, node, or group of nodes within the system. Typically, targeted attacks are launched against *distributed state estimation* in an inter-connected power system where each regional control center performs the state estimation based on the topology and parameters of the region, in addition to the measurements taken in that region. For example, the targeted attacks proposed in [108] are launched against a single compromised control center and are able to affect the outcome of distributed state estimation. Another work by the same authors focuses on detecting and localizing targeted attacks on distributed state estimation in power transmission systems. In this work, we perform state estimation in a centralized manner, and use the term *targeted attack* when we talk about the targeted adversarial attacks described in Section 2.2.

Evasion Attacks vs. FDIA: 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 [61], [62], [119]. 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 [82]. To the best of our knowledge, no such evidence regarding transferability of FDIA has been provided in the literature.

2.4 Vulnerability of ML Models in Power Systems

In recent years machine learning has shown promise in solving a variety of planning and operation problems in the power system. For example, data-driven strategies have been used successfully in renewable generation and residential load forecasting [5], [35], [124], power line outage prediction and localization [31], [41], power system protection and control [19], [121] and state estimation [1], [2], [11], [115]. Despite the growing interest in the integration of ML with planning and operation practices in the power system, research pertaining to the security of ML in this field has only begun to appear recently.

The vulnerabilities of ML algorithms used in the power system are first investigated by Chen et al. in [22] where the authors 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. Eklas et al. [44] 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.

While various ML techniques have been proposed to detect FDIA [91], [102], [109], [117], 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 [91]. 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 [102]. 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 FDIA [101].

Turning to data-driven state estimation approaches, ANN-based state estimators have been found vulnerable to FDIA. For example, optimization techniques based on differential evolution and sequential least-square quadratic programming have been proposed in [59], [60] to construct attack vectors that can fool the BDD mechanism and affect an MLP-based state estimator. However, the iterative optimization algorithms employed to execute these attacks may not be efficient enough to be launched against large-scale power systems in the real world. More recently, a forward-derivative-based adversarial attack on a neural network-based state estimator is proposed in [99]. 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 also note that all these attacks are white-box, i.e., the attacker is assumed to have 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 [9] and [100], 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 [100] 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 [9] to craft a stealthy black-box adversarial attack against power system state estimation. The authors used the vanilla

Table 2.1: Technical contribution and novelty of our work with respect to the related work.

Reference	SE Approach	Attack Strategy	Stealthiness Analysis	Impact on Control System
[59], [60]	AC-Power System State Estimation (PSSE) using MLP	Differential evolution (DE) and sequential least square quadratic programming (SLSQP)	Bypasses BDD under the white-box assumption.	✗
[99]	AC-PSSE using MLP	Forward-derivative-based adversarial attack	Stealthiness analysis is not done.	✗
[102]	DC state estimation	Joint adversarial examples and FDIA	Bypasses BDD and NAD under the white-box assumption. The black-box version performs poorly.	✗
[101]	DC state estimation	Targeted FDIA	Bypasses BDD and NAD under the white-box assumption. The black-box attack is not considered.	✗
[100]	DC state estimation	Data-driven approach based on robust linear regression (RLR)	Bypasses BDD under the **black-box** assumption.	✗
[9]	AC-PSSE with WLS	Vanilla FGSM	Bypasses BDD under the **black-box** assumption.	✗
Our work	Data-driven non-linear AC-DSSE using Stacked ResNetD & R-KNN	Vanilla and Sneaky-FGSM	Bypasses BDD under the **black-box** assumption.	✓
		white-box Targeted-FGSM	Bypasses BDD	

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, in this thesis, we analyze the vulnerability of *data-driven DSSE approaches* to adversarial attacks crafted using surrogate neural networks under the black-box assumption. We propose the novel Sneaky-FGSM algorithm, which is capable of inducing higher measurement noise without being detected by the conventional BDD mechanism. Moreover, we propose the Targeted-FGSM algorithm, which is capable of executing a targeted attack that aims to fulfill a specific adversarial objective by misleading the DSSE mechanism in a certain direction. To the best of our knowledge, this is the first proposed **targeted adversarial attack** strategy that aims to misguide a control operation in a particular direction by inducing error in the DSSE process. Lastly, we address a major limitation of the existing literature [9], [99], [102] by analyzing the impact of the proposed attacks on voltage regulation schemes, which is an important control application that relies on the DSSE result. Table 2.1 provides a concise overview of our contributions and the unique aspects of this work compared to the related work.

2.5 Safeguarding State Estimation

Safeguarding state estimation in power systems is crucial for maintaining reliable and secure operation. The core principles of safeguarding state estimators remain similar between distribution and transmission systems due to the shared objective of ensuring accurate and secure power system operation. In this section, we review the literature on the implementation of various safeguarding mechanisms for state estimation in power systems.

2.5.1 Safeguards Against FDIA

We categorize the existing defense mechanisms against FDIA into two categories as described below.

Protection-based Approaches: Identifying crucial system components, implementing strict access control and secure communication protocols, and maintaining best security practices such as performing regular security assessments, configuring firewalls, updating software, and disabling unnecessary services help prevent adversaries from accessing the system. In [12], Bi et al. demonstrate that it is possible to prevent FDIA against state estimation approaches by protecting carefully selected meter measurements. They propose an arithmetic greedy algorithm that finds the minimum set of protected meter measurements by gradually expanding the set of secure state variables. Considering the high computational complexity of the proposed greedy algorithm, the same authors conduct another study to characterize the optimal protection from a graph theoretical perspective [13]. The optimal state protection problem is mapped into a minimum measured Steiner tree (MMST) problem and two exact solutions are proposed based on the Steiner vertex enumeration algorithm and mixed integer linear programming (MILP). A combination of protection-based and detection-based defense mechanisms for PSSE has been proposed in [123]. As a preventive measure, sensors and meters connected to the largest number of buses have been identified as critical and selected for protection. Besides, temporal- and spatial-based FDIA detection mechanisms have been proposed.

Detection-based Approaches: The problem of bad data detection in power system state estimation has been well-studied for decades. Residual-based bad data detection methods such as chi-square test for detection and largest normalized residual test (LNRT) for localization are widely used in commercial PSSE software and tools [3]. Several other physical model-based approaches that rely on statistical features of measurement data have also been explored in different studies. For example, the authors in [50] argue that the control center can use historical data to maintain and track its belief state of the system and propose a Bayesian formulation of the state estimation problem. The detector presented in [50] uses L_∞ norm on residue errors from the state estimator to detect the statistically unlikely measurements. A

Table 2.2: Detection-based safeguarding mechanisms against FDIA

Approach	**Algorithm**	**Reference**
residual-based	L_∞ detector	[50]
convex optimization	Generalized Likelihood Ratio Test (GLRT)	[51]
checking statistical consistency	robust projection statistics	[131]
error thresholding	topology perturbation method	[52]
solving matrix separation	Fast GoDec	[55]
tracking measurement variations	absolute distance & KLD	[18]
nonparametric density estimation	Kernel density estimation	[21]
supervised, semi-supervised, ensemble learning & online learning	SVE, SVM, SLR, K-NN, Adaboost, & online perceptron (OP)	[81]
supervised learning	SVM, K-NN, ENN, ANN	[89], [118]
supervised learning	CNN	[64], [110]
supervised learning	Recurrent neural network (RNN)	[7], [47], [112], [114]
ensemble learning	Isolation forest	[4]
ensemble learning	ensemble of LR, DBSCAN, & Chebyshev	[133]
deep learning	deep autoencoder	[109]
deep learning	Generative Adversarial Network (GAN)	[58]
supervised & semi-supervised learning	SVM, statistical-based anomaly detection	[32]
supervised learning	attention-based temporal convolutional network (ATCN)	[85]

similar approach based on the Bayesian formulation of SE is considered in [51] where the authors proposed a detector based on the principle of GLRT. The primary drawback of the Bayesian formulation-based approaches lies in their lack of efficacy in identifying falsified measurement data that fits into the historical measurement distribution. This issue is addressed by Chaojun et al. in [18]. They propose a robust FDIA detection mechanism that uses the KLD to track the dynamics of the measurement variations and detect false data injected into the system. This approach has been proven effective in detecting false data samples even if they fit the distribution of historical data. Another effective FDIA detector based on statistical consistency check between two state vectors– one estimated using secure PMU measurements and the other with remaining SCADA and PMU measurements has been proposed in [131]. Besides these physical model-based approaches, data-driven strategies for detecting anomalies and false data have also been explored in the recent literature. Various supervised, semi-supervised, and unsupervised models have been found effective in this prospect. A comprehensive review of various FDIA detection methods leveraging machine learning algorithms has been presented in [90]. Considering the large volume of work in designing and developing FDIA detectors, we present a brief overview of the relevant recent works in detecting FDIA on state estimators in table 2.2.

2.5.2 Safeguards Against Adversarial Attacks

While FDIA against state estimators have been well studied for decades, the damage that could be inflicted by adversarial attacks has been investigated only recently. The existing literature on building safeguarding mechanisms that protect state estimators from adversarial attacks is mostly detection-based. In [102], a joint adversarial and stealthy false data injection attack has been launched against a DC state estimation model that is protected by two detection-based safeguarding mechanisms, namely a conventional BDD and a NAD, which is a simple Fully Connected Neural Network (FCNN) trained to classify bad measurement data. Both BDD and NAD models have been found vulnerable to white-box state-perturbation-based FDIA (S-FDIA). Sayghe et al. [91] conduct a similar study on a WLS-based DC state estimation model protected by an MLP-based classifier trained to distinguish bad measurement samples from good ones. Similar to the results obtained in [102], the MLP-based detection approach has also been found vulnerable to white-box adversarial attacks. While the aforementioned studies focus on white-box attacks, another recent work by Bhattacharjee et al. [9] confirms the ineffectiveness of conventional residual-based BDD mechanisms in safeguarding the WLS-based AC-PSSE approach against black-box adversarial attacks.

To the best of our knowledge, only two safeguarding mechanisms in the literature have been proven to be effective in protecting data-driven state estimators against adversarial attacks. These include a meter protection-based strategy and an adversarial training-based strategy proposed by Tian et al. in [99]. In the meter protection-based strategy, a forward derivative-based approach has been undertaken to rank the importance of the meters and then a subset of important meters have been selected for enhanced protective measures such as encryption, authentication, and access control. This approach resulted in a promising performance in protecting the state estimator. However, as mentioned in the original work, such protection methods might consume a lot of defense resources, especially in large-scale and complex power grids. For the adversarial training-based defense, the data-driven state estimator is retrained using a mixture of benign and adversarial data samples for enhanced robustness. This approach is budget-friendly and easy to incorporate into large-scale systems. However, one major drawback of this defense strategy is that adversarial training leads to a reduction in model performance on benign data, creating a trade-off between robustness and general performance [99]. We note that, both of these safeguarding techniques are protection-based, aiming to thwart adversarial attacks. However, no preventive measure is entirely foolproof and therefore, in addition to integrating preventive measures, it is equally important to devise robust and effective detection mechanisms to ensure comprehensive system security.

Based on this literature survey, it is safe to say that there is a lack of existing research that presents a viable safeguarding mechanism to effectively defend state estimation methodologies against adversarial attacks. Thus, *designing effective and robust safeguarding strategies against adversarial attacks is still an open research area that requires further investigation.* This motivates us to look for an effective safeguarding strategy that would be able to detect adversarial attacks on distribution system state estimators. Moreover, the inefficacy of traditional bad data detection strategies, as reported by previous work, prompts us to approach this problem from a different perspective. Instead of remodeling the age-old power system security mechanisms for modern data-driven DSSE approaches, we aim to explore the possibilities within the machine learning domain and find a viable solution that could be tailored to our advantage.

We note that adversarial examples are hard to detect due to their subtle and imperceptible nature, closely resembling normal data but containing slight perturbations crafted to mislead machine learning models [16], [103]. A growing body of research aims to understand the existence of adversarial examples [33], [38], [97], but a complete understanding of the underlying reasons remains elusive primarily due to the intricate functional structures of deep neural networks, making it challenging to derive precise mathematical descriptions. Recently, *feature attribution* has been found successful in tackling the black-box nature of neural networks [56], [92] and improving transparency and fairness of machine learning models [25], [88]. Based on this observation, Yang et al. [122] introduced an effective method for detecting adversarial attacks by thresholding a scale estimate of feature attribution scores. Their proposed *ML-LOO* detector has demonstrated superiority over state-of-the-art detection methods in its ability to differentiate adversarial images from popular attack methods across a variety of real data sets. Motivated by its success in safeguarding image classifiers against adversarial attacks, we adopt the *feature attribution*-based detection approach for protecting data-driven state estimators. More details regarding this experiment can be found in Chapter 5.

Chapter 3

Methodology

This chapter begins by describing the threat model where we state the assumption about the goal, capability, and resources available to potential adversaries. We then discuss the implementation of DSSE and BDD techniques, present the mathematical formulation of the proposed black-box evasion attack strategies and white-box targeted attack strategy, and provide more details about the voltage control scheme that relies on the DSSE results.

3.1 Threat Model

We analyze the effectiveness of untargeted black-box and targeted white-box FGSM attacks on two state-of-the-art data-driven DSSE models that have been proposed in prior work, namely Stacked ResNetD [11], and R-KNN [115]. For the untargeted attacks, 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) query access to the victim model[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

[1] The query access is basically equivalent to the read access to the model output for user specified inputs. As discussed in Section 2.2, having query access to the victim model is a fundamental assumption for black-box attacks.

gaining access to the DSO's authorized user account. For example, during the 2015 Ukraine power outage, the hackers managed to infiltrate the control system by sending phishing emails to the operators [116]. Once inside, they executed multiple unauthorized commands, resulting in power disruption to over $225,000$ customers lasting between 1 and 6 hours. The PMU networks and utility data centers have been found vulnerable to cyber attacks in several recent studies [105], [111], indicating a high risk of the presence of such adversaries, lending credence to this threat model.

The goal of a successful attack is to distort the measurements in a way that significantly alters the state estimation results and at the same time remains undetected. Hence, in the second phase of experimentation, we investigate the stealthiness of the proposed attack to the traditional residual-based BDD mechanism. Upon analyzing the results from BDD, we propose a new attack strategy, namely *Sneaky-FGSM*, which is able to induce noise in a stealthier fashion, fooling the BDD strategy more often than the vanilla FGSM attack. In light of the observation that both vanilla and Sneaky-FGSM attack methods are generated in an untargeted manner, resulting in a random impact on the output of the victim model, we develop a white-box targeted attack strategy, called *Targeted-FGSM*. The objective of this attack is to induce specific power quality issues, such as under-voltage or over-voltage problems, in the distribution buses by misleading the state estimator in a certain direction.

3.2 DSSE and BDD Techniques

Neural networks, being universal function approximators, can precisely approximate the state estimation module. 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 two state-of-the-art data-driven models as our victim models– (a) an ensemble learning model, namely *Stacked ResNetD*, which has been proposed in [11] and shown to outperform several other deep neural networks, and (b) a robust K-nearest neighbors approach, namely *R-KNN*, which has been proposed in [115] and

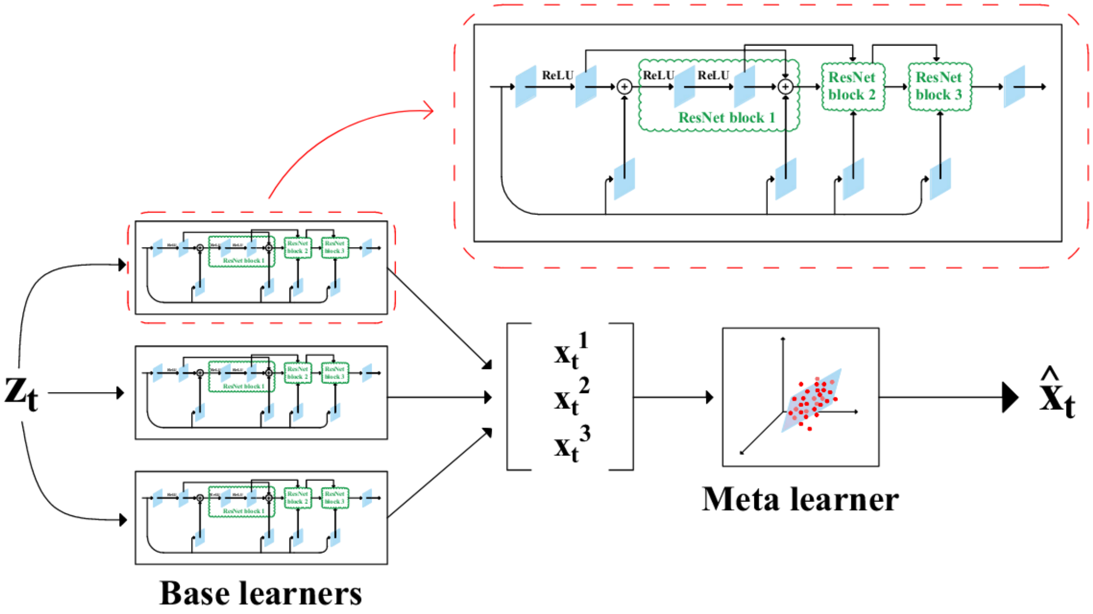


Figure 3.1: Stacked ResNetD architecture [11]

shown to outperform the traditional WLS-based model in terms of both accuracy and speed. We discuss the basic architectures of these two models below.

Stacked ResNetD [11]: 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, that is, 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 3.1 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. Previous work has shown that ensemble learning models have enhanced adversarial robustness [93], [104]. This observation together with the strong performance of Stacked ResNetD motivates our choice of this victim model.

R-KNN [115]: This approach utilizes similar measurements from historical data to estimate the state given a measurement sample, $\mathbf{z}_t$, at any time step t. Given a specific $\mathbf{z}_t$, the state vector, $\mathbf{x}_t$, is considered a random variable that follows a normal distribution with unknown hyper-parameters $\mathbf{q}$ and $\mathbf{\Sigma}_d$: $\mathcal{N}\left(\mathbf{q}^T\mathbf{z}_t, \mathbf{\Sigma}_d\right)$.

From the historical dataset, K training points closest[2] to $\mathbf{z}_t$ are selected. Then, ridge regression is used to estimate the unknown hyper-parameter $\hat{\mathbf{q}}$.

$$\hat{\mathbf{q}} = \arg\min_{\mathbf{q}} \sum_{i=1}^{K} \left(\mathbf{x}_i - \mathbf{q}^{\mathbf{T}}\mathbf{z_i}\right)^2 + 2\gamma\|\mathbf{q}\|^2$$

Here, $2\gamma\|\mathbf{q}\|^2$ is the regularization term. The unknown hyper-parameter $\mathbf{\Sigma}_d$ has been absorbed in the penalty constant γ.

Assuming that the historical data obtained from the K-nearest neighbors is stored in matrices,

$$\mathbf{Z}_{mat} = (\mathbf{z_1}, \mathbf{z_2}, ..., \mathbf{z_K})$$
$$\mathbf{X}_{mat} = (\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_K}),$$

the closed-form solution is given by:

$$\hat{\mathbf{q}} = \left(\mathbf{Z}_{mat}\mathbf{Z}_{mat}^T + 2\gamma\mathbf{I}\right)^{-1}\mathbf{Z}_{mat}\mathbf{X}_{mat} \tag{3.1}$$

Once the hyper-parameter $\hat{\mathbf{q}}$ is estimated using Equation 3.1, it is then used for Bayesian inference to generate the current state estimate, $\hat{\mathbf{x}}_t$,

$$\hat{\mathbf{x}_t} = \hat{\mathbf{q}}^T\mathbf{z}_t$$

To enhance the robustness of the state estimation model, a two-stage filtering process is applied to the historical data. In the first stage, the historical measurement data is carefully cleaned by filtering out data points with large measurement residuals.[3] This step helps to get rid of any bad historical measurement data resulting from various factors such as telecommunication errors,

[2] $L2$-norm has been used as the distance metric.

[3] As described in Section 1.3.2, the measurement residual is defined as the sum of squared differences between the original measurement ($\mathbf{z}$) and the estimated measurement ($\hat{\mathbf{z}}$).

incorrect topology information, equipment failure, finite accuracy, infrequent instrument calibration and measurement scaling procedure at the control center. In the second stage, all the historical state estimates undergo an outlier detection algorithm to mitigate the impact of any potentially malicious data that could have been injected by an adversary. The authors proposed an outlier detection algorithm that assesses the local density of a data point (i.e. state vectors) compared to its K-neighboring points. If the density is significantly lower, the data point is flagged as an outlier and excluded from the process. This thesis focuses solely on evasion attacks, where the adversary injects adversarial examples exclusively during runtime, and the training dataset (i.e. the historical data used in the R-KNN model) is assumed to be clean and free of malicious data. Therefore, we safely omit the robustness enhancement steps and just implement the core R-KNN method.

Residual-based BDD: Bad data detection is a classic problem inherent in the original formulation of state estimation. 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 [106]. In this work, we implement the residual-based BDD mechanism described in Section 1.3.2 with a level of significance α=0.05, and degrees of freedom d=48.

3.3 Voltage Regulation Scheme

We implement the rule-based voltage regulation scheme that relies on DSSE and was previously described in Section 1.3.3. In this scheme, the 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*

Algorithm 1 Vanilla FGSM Attack

```
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 $t$, $(\mathbf{z_t}, \mathbf{x_t})$
Output:
    Adversarial measurement sample at timestamp $t$, $\mathbf{z'_t}$
Initialize:
    $\theta_0$ with small random values
    Surrogate model, $f_\theta$ with appropriate loss function, $L$
▷ Training the surrogate model $f$, parameterized by $\theta$
for $j = 0, 1, \cdots, maxIter$ do
    $\boldsymbol{\theta_{j+1}} \leftarrow \boldsymbol{\theta_j} - \alpha \nabla_{\boldsymbol{\theta_j}} \left[\frac{1}{N_{train}} \sum_i L\left(f(\mathbf{z_i^{train}}; \boldsymbol{\theta}_j), \mathbf{x_i^{train}}\right)\right]$
end for                                                    ▷ $\alpha$ is the learning rate
▷ Calculating gradient of the loss w.r.t. the input, $\mathbf{z_t}$
$\boldsymbol{\delta}_{\mathbf{z_t}} = \nabla_{\mathbf{z_t}} \left[L\left(f(\mathbf{z_t}; \boldsymbol{\theta}), \mathbf{x_t}\right)\right]$
$\mathbf{z'_t} = \mathbf{z_t} + \epsilon \cdot \text{sign}(\boldsymbol{\delta}_{\mathbf{z_t}})$          ▷ $\epsilon$ is a hyper-parameter (scalar)
return $\mathbf{z'_t}$                                      ▷ Return the adversarial sample
```

position accordingly" [86]. 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 4.2.4.

3.4 Attack Strategies

Now, we describe the three gradient-based adversarial attack strategies, namely (a) vanilla FGSM, (b) Sneaky-FGSM, and (c) Targeted-FGSM, that will be used in the next chapter. Note that the last two attack strategies are novel and are developed in this thesis.

3.4.1 Vanilla FGSM

We employ the FGSM attack presented in section 2.2 in a black-box setting under the hypothesis that the adversary, being unaware of the victim model's architecture, can choose any suitable neural network as the surrogate model. To test this hypothesis, we use four different surrogate models and investigate their effectiveness against the two victim models. The initial two surrogate models used in the study are variations of a MLP architecture. One version consists of five dense layers with ReLU activation functions, while the other version incorporates tanh activation functions. Additionally, a CNN model proposed in [11], comprising three convolutional layers and three dense layers with ReLU activation functions, has been employed as a surrogate. Another variant of this CNN model, utilizing tanh activation functions, has also been utilized as a surrogate model.

Our aim is to investigate how 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 1.3.4, 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 [134], 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 and Targeted-FGSM discussed later.

3.4.2 Sneaky-FGSM

From Equation (1.7), 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.[4]

Typically, measurement variances calculated from the historical data are stored in the data center and are updated periodically. Depending on the level of access that the adversary has to the data center, they might be able to read the stored variance data. Even in a stricter case when the adversary does not have access to the variance data, on any day D, an adversary with access to the measurement data can easily 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 the 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 version of the attack.

In this experiment, we use the household power consumption dataset (described later in Section 4.1), 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 hyper-parameter 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

[4] 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 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 measurement sample at timestamp $t$, $\mathbf{z'_t}$
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: $\mathbf{z'_t} = \mathbf{z_t} + \epsilon \cdot \mathbf{S}$                ▷ $\epsilon$ is a hyper-parameter (scalar)
9: return $\mathbf{z'_t}$
```

adding noise to Q measurements and perhaps other measurements that are intrinsically low variance.

In Line 6 of Algorithm 2, we define a binary vector, **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 7, we modify the perturbation vector obtained from vanilla FGSM (i.e., $\text{sign}(\boldsymbol{\delta}_{\mathbf{z_i}})$) by calculating its Hadamard (element-wise) product with **select**.

3.4.3 Whitebox Targeted FGSM

Vanilla FGSM and Sneaky-FGSM, despite being successful in misleading the victim state estimator model, causing increased wear and tear of voltage regulation equipment, and occasional under-voltage or over-voltage incidents[5], share a common limitation. Specifically, both of them generate adversarial samples in an untargeted manner without a specific objective in mind. This means, while the adversarial samples are capable of misleading the victim state estimator model, there is no promise on the *direction* in which the vic-

[5] Corresponding experimental results are presented in the following chapter.

Algorithm 3 Targeted-FGSM Attack

```
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 $t$, $(\mathbf{z_t}, \mathbf{x_t})$
    The victim model, $g(\mathbf{z}; \theta)$
    Pre-defined range, $v_{min}, v_{max}$
Output:
    Adversarial measurement at timestamp $t$, $\mathbf{z'_t}$
$\mathbf{x'_t} = \mathbf{x_t}$.copy()
▷ Replace the voltage values of $\mathbf{x'_t}$ with random values selected from the range $v_{min}, v_{max}$
for $i = 0, 1, \cdots, n/2$ do
    $\mathbf{x'_t}[i] = \text{random}(v_{min}, v_{max})$
end for
$\boldsymbol{\delta}_{\mathbf{z_t}} = \nabla_{\mathbf{z_t}} \left[L\left(g(\mathbf{z_t}; \theta), \mathbf{x'_t}\right)\right]$
$\mathbf{z'_t} = \mathbf{z_t} - \epsilon \cdot \text{sign}(\boldsymbol{\delta}_{\mathbf{z_t}})$        ▷ $\epsilon$ is a hyper-parameter (scalar)
return $\mathbf{z'_t}$        ▷ Return the adversarial sample
```

tim model's output moves (e.g., the estimated bus voltage may be higher or lower than its true value). To address this limitation, we introduce a goal-oriented attack strategy, namely *Targeted-FGSM*, that aims to create system instability at a higher rate compared to the untargeted methods at the expense of a stricter assumption: the adversary possesses complete knowledge of the victim model's architecture, denoted $g(\mathbf{z}; \theta)$.

The goal of our proposed targeted attack strategy is to mislead the victim model so that its estimated voltage values are shifted upwards. Consequently, the adversary increases the likelihood of under-voltage incidents occurring by misleading the voltage violation detection mechanism in two ways: (a) producing a false negative result– concealing the true state during under-voltage situations (i.e. estimating a voltage within the acceptable range), and (b) producing a false positive result– generating an over-estimation that surpasses the upper safe-range threshold, even when the bus voltage is actually within the safe range. An elaborate discussion on the impact of adversarial attacks on the rule-based voltage regulation scheme is presented in Section 4.1.3.

Similar to some of the recent work [17], [53], we design the Targeted-FGSM algorithm by modifying the surrogate model's loss function. Suppose that we are trying to construct an adversarial measurement data sample, $\mathbf{z'_t}$, from a clean data pair, $(\mathbf{z_t}, \mathbf{x_t})$. For a victim model $g(\mathbf{z_t}; \theta)$, we define the adversarial loss function for the white-box Targeted-FGSM attack as $L\left(g(\mathbf{z_t}; \theta), \mathbf{x'_t}\right)$. Here, $\mathbf{x'_t}$ is a vector of size n which has the same phase angle values as $\mathbf{x_t}$ but the voltage values are replaced by some randomly chosen values from the voltage range (v_{min}, v_{max}), which is pre-defined by the adversary based on their objective. An adversary who aims to trigger under-voltage occurrences should define (v_{min}, v_{max}) from a higher range e.g. around the upper safe-voltage threshold. Conversely, an adversary who aims to trigger over-voltage incidents should define (v_{min}, v_{max}) around the lower safe-voltage threshold. It is important to note that using a fixed value for each element of $\mathbf{x'_t}$ instead of randomly selecting them from the range (v_{min}, v_{max}) is also a viable option. However, we prefer the random selection approach to affirm that $\mathbf{x'_t}$ is not a hyper-parameter requiring fine-tuning for the targeted attack strategy to be effective. By opting for random values of elements of $\mathbf{x'_t}$, we contend that the success of the proposed Targeted-FGSM algorithm, i.e., its ability to push the estimated states in a specific direction, primarily relies on the perturbation factor, ϵ, rather than the choice of $\mathbf{x'_t}$.

Once the adversarial loss function is defined, we construct the adversarial sample, $\mathbf{z'_t}$ using the gradient descent formula to minimize the adversarial loss function, $L\left(g(\mathbf{z_t}; \theta), \mathbf{x'_t}\right)$, with respect to the input sample, $\mathbf{z_t}$. Algorithm 3 describes the proposed Targeted-FGSM attack.

There are two key restrictions inherent in the proposed Targeted-FGSM algorithm. First, similar to some of the gradient-based targeted attacks discussed in Section 2.2, our proposed Targeted-FGSM also lacks transferability, hence it cannot be used in the black-box setting. Second, the effectiveness of this attack highly depends on the perturbation factor, ϵ. An ϵ value that is too small will not be able to push the estimated state enough to produce a false positive or false negative while detecting voltage violation. On the other hand, if the value of ϵ is excessively large, it will result in a notable increase in measurement noise that will be immediately detected by the safeguard system tied to the state estimator.

In this chapter, we established the groundwork for this thesis by explaining the approach and methodologies employed in our study. In the next chapter, we will delve into the practical implementation of these methods and showcase the experimental results obtained for vulnerability analysis of data-driven DSSE approaches when faced with the three adversarial attack strategies introduced earlier in this chapter.

Chapter 4

Adversarial Vulnerability of the DSSE Models

This chapter provides an overview of the simulation environment and the experiments carried out to analyze adversarial vulnerability of the data-driven DSSE methods described in Section 3.2. This includes a detailed description of the test case, data preparation techniques, and evaluation criteria, followed by an in-depth presentation of the experimental results.

4.1 Experimental Setup

4.1.1 Test Case

Our test system is structurally similar to the customized IEEE 33-bus test system presented in [40]. Specifically, we use the 33-bus system [8] as the primary distribution network and the IEEE European low voltage test feeder [45] 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 4.1 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) [68], 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 aggrega-

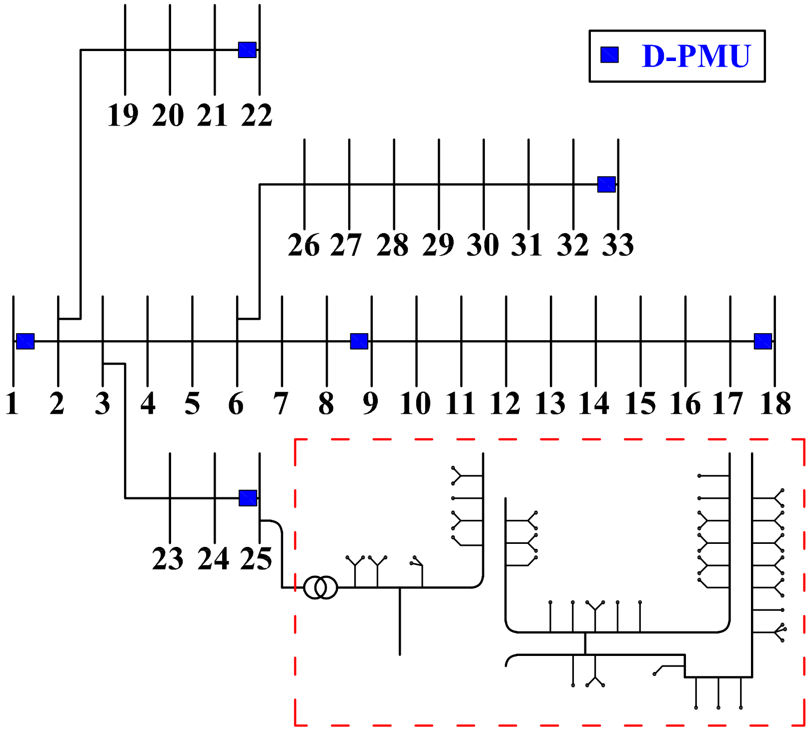


Figure 4.1: Single-line diagram of the customized IEEE 33-bus test system. Node 25 shows the IEEE European low-voltage system, which is connected to each of the primary nodes.

tion standard for publishing utility data in the State of New York [98]. 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 deviations $0.01, 0.02, \cdots, 0.1$ 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 [8]. More specifically, we randomly select buildings and connect them to each 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 datasheet. Finally, we run the AC power flow analysis using the Open Distribution Simulator Software (OpenDSS) [29] 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 [126].

4.1.2 Data Preparation and 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 [84] 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-PMUs 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 of buses equipped with D-PMUs. We install six D-PMUs since this level of observability led to reasonable state estimation performance in [40]. Figure 4.1 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, for the first 15 days of every month for a year). 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 × 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.

4.1.3 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 data-driven state estimators we use the root-mean-square error (RMSE) defined as:

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

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 limit 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) [107].

Impact on the voltage regulation scheme 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 to analyze the impact of the proposed adversarial attacks on the rule-based 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.

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 or a false positive in the opposite direction that erroneously detects an under-voltage occurrence instead of the true over-voltage state). In this case, the affected bus experiences an over-voltage problem, but since it is not accurately detected, the voltage control scheme either does not take any remedial action (in the case of a false negative) or takes the opposite action that makes the situation worse (in the case of false positive). Thus, the over-voltage situation persists; **(b)** when the bus voltage is within the specified range but an under-voltage occurrence 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.

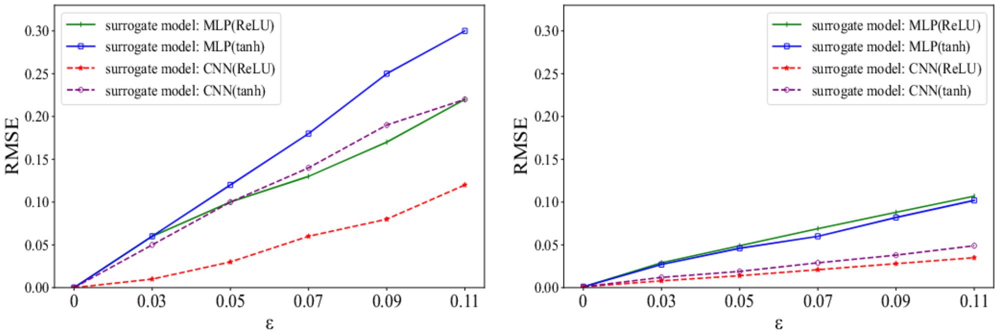
Scenario 3 (Increased under-voltage incidents) This is the exact opposite of the previous scenario and may occur in two different ways: **(a)** when the bus voltage is below the lower threshold and it does not get detected (i.e., a false negative or a misdirected false positive that erroneously detects an over-voltage occurrence instead of the true under-voltage state); **(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.

4.2 Experimental Results

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

4.2.1 Effectiveness of Black-box Attacks

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 3.4, the choice of the surrogate model rests exclusively with the attacker. To analyze the impact of black-box attacks on the state estimator's performance, we employ four different surrogate models, namely MLP and CNN with tanh and ReLU activations, 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 error is directly proportional to the



(a) Victim: Stacked ResNetD

(b) Victim: R-KNN

Figure 4.2: Increase in the state estimation error under the black-box vanilla FGSM attack.

amount of noise added to the dataset.

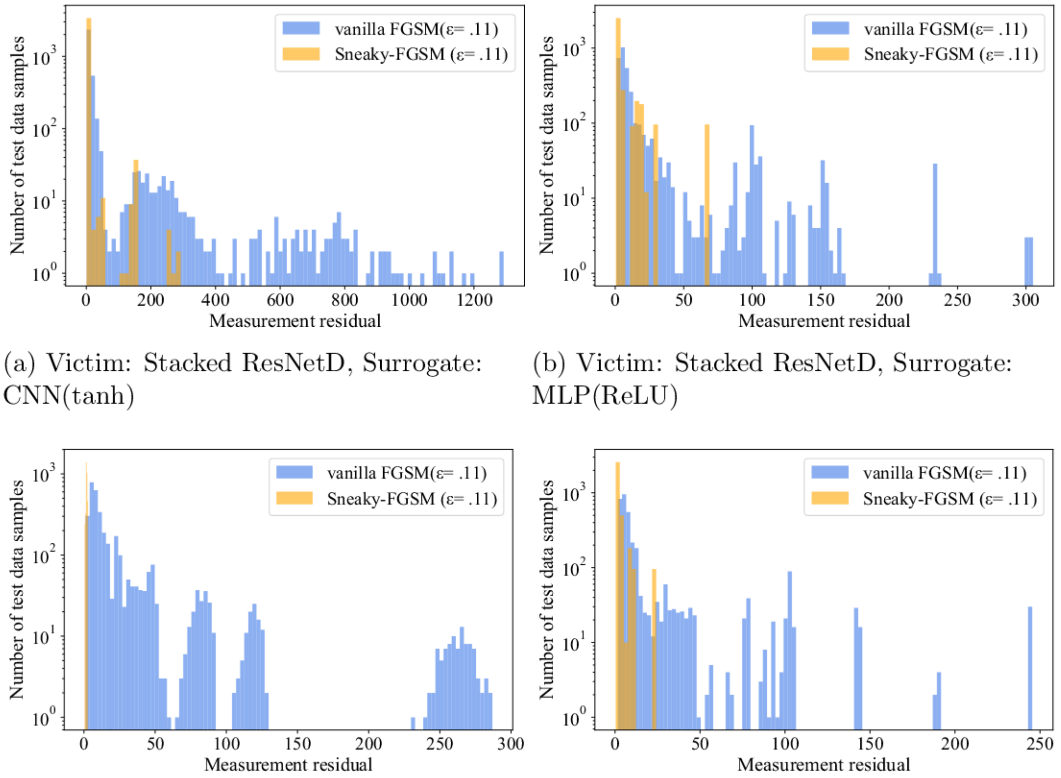
Figure 4.2 presents the impact of vanilla FGSM attack on the victim models. It can be readily seen that regardless of the choice of the surrogate model, the performance of both of the victim models is adversely impacted. This result highlights the vulnerability of data-driven DSSE approaches to black-box attacks which are more practical and likely than the conventional FDIA since the attacker does not necessarily need to know the victim model's architecture to generate effective attacks.

An increased amount of added noise, ϵ, increases the estimation error raising the question, *how much measurement error can an adversary induce without getting detected by the system operator?* As discussed in Section 3.1, we conduct a stealthiness analysis of the proposed attacks to address this question. The results are presented below.

4.2.2 Stealthiness Analysis

We introduced the idea of Sneaky-FGSM, a stealthier attack strategy specifically designed to fool the residual-based BDD mechanism, in Section 3.4.2. Now, we discuss the experimental results analyzing the stealthiness of Sneaky-FGSM compared to vanilla FGSM.

As discussed in Section 3.4.2, Sneaky-FGSM aims to perturb the measurements in such a way that the overall measurement residual, $J(\mathbf{x})$ is lower



(a) Victim: Stacked ResNetD, Surrogate: CNN(tanh)

(b) Victim: Stacked ResNetD, Surrogate: MLP(ReLU)

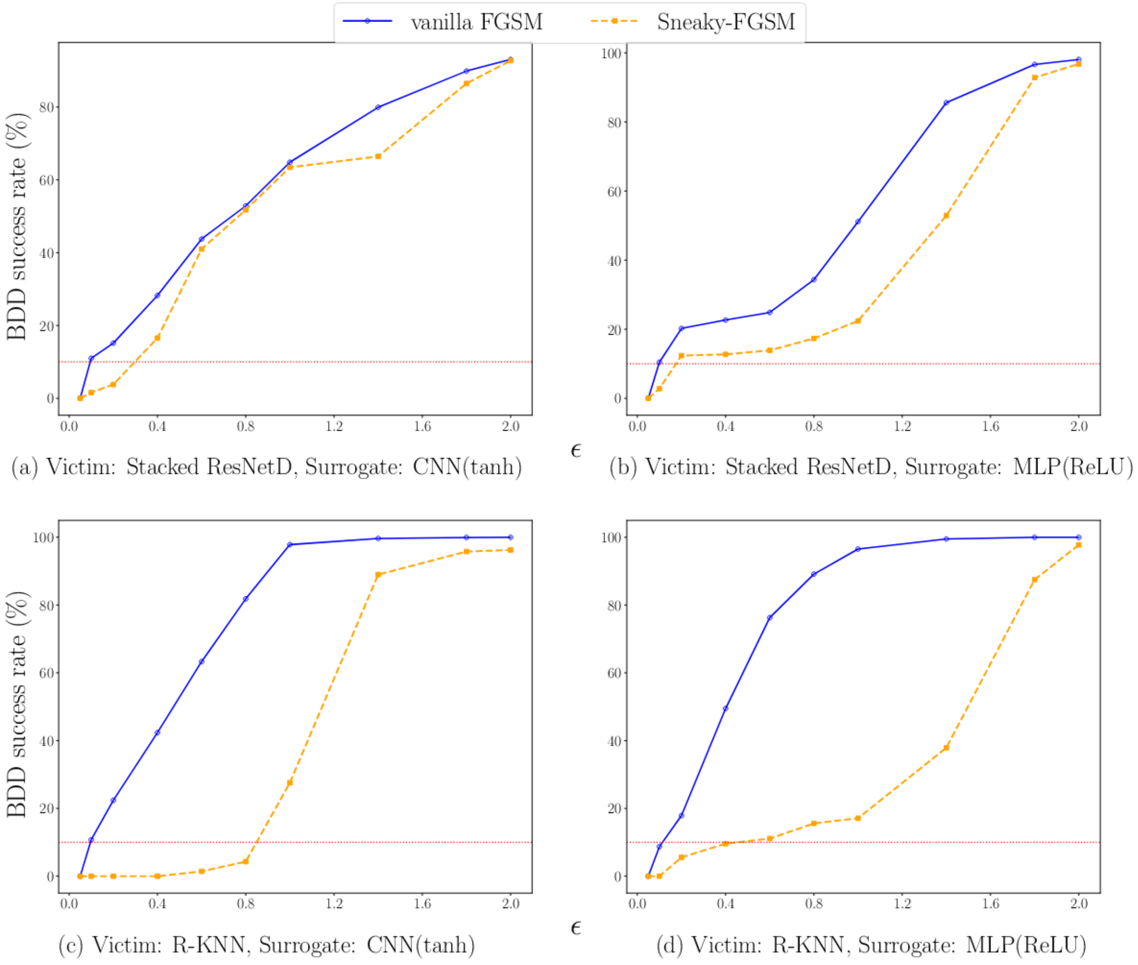
(c) Victim: R-KNN, Surrogate: CNN(tanh)

(d) Victim: R-KNN, Surrogate: MLP(ReLU)

Figure 4.3: Frequency plot of the measurement residual ($J(\mathbf{x})$) under vanilla and Sneaky-FGSM attacks. The plots placed side-by-side are for the same victim model. The plots aligned vertically are for the same surrogate model. Note the y-axis is in logarithmic scale.

than that of vanilla FGSM. We craft both vanilla FGSM and Sneaky-FGSM with the same perturbation factor (ϵ) and compare the corresponding measurement residuals. Figure 4.3 shows the experimental results. While adding the same amount of adversarial perturbation as the vanilla attack strategy, Sneaky-FGSM accounts for lower residual, resulting in fewer detections by the conventional residual-based BDD mechanism.

Next, we craft both vanilla FGSM and Sneaky-FGSM attacks with varying the perturbation factor ϵ, and test the efficacy of the residual-based 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 (N_{total}=3456).



(a) Victim: Stacked ResNetD, Surrogate: CNN(tanh)

(b) Victim: Stacked ResNetD, Surrogate: MLP(ReLU)

(c) Victim: R-KNN, Surrogate: CNN(tanh)

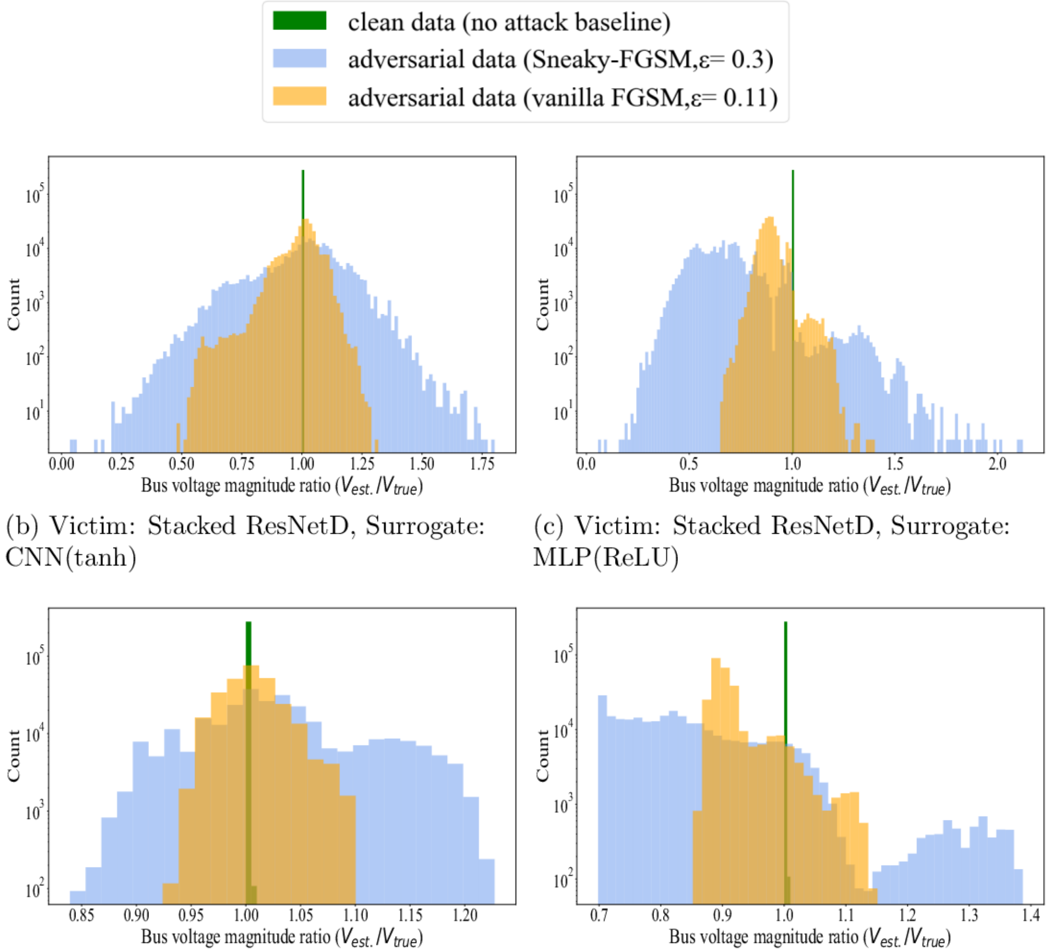
(d) Victim: R-KNN, Surrogate: MLP(ReLU)

Figure 4.4: Stealthiness of vanilla and Sneaky-FGSM attacks. The horizontal line is drawn at 10% detection rate.

Figure 4.4 compares the stealthiness of vanilla FGSM and Sneaky-FGSM attacks crafted with different surrogates and launched against different victim models. As discussed in the above section, any suitable surrogate model can be used to generate adversarial perturbation. To underscore the persistence of Sneaky-FGSM's stealthiness, we employ two distinct surrogate models, namely CNN(tanh) and MLP(ReLU), to create Sneaky-FGSM attacks against each of the two victim models. The intent is to demonstrate that Sneaky-FGSM retains its covert nature regardless of the surrogate model chosen.

As can be seen from Figure 4.4, 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 arises 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?



(b) Victim: Stacked ResNetD, Surrogate: CNN(tanh)

(c) Victim: Stacked ResNetD, Surrogate: MLP(ReLU)

(d) Victim: R-KNN, Surrogate: CNN(tanh)

(e) Victim: R-KNN, Surrogate: MLP(ReLU)

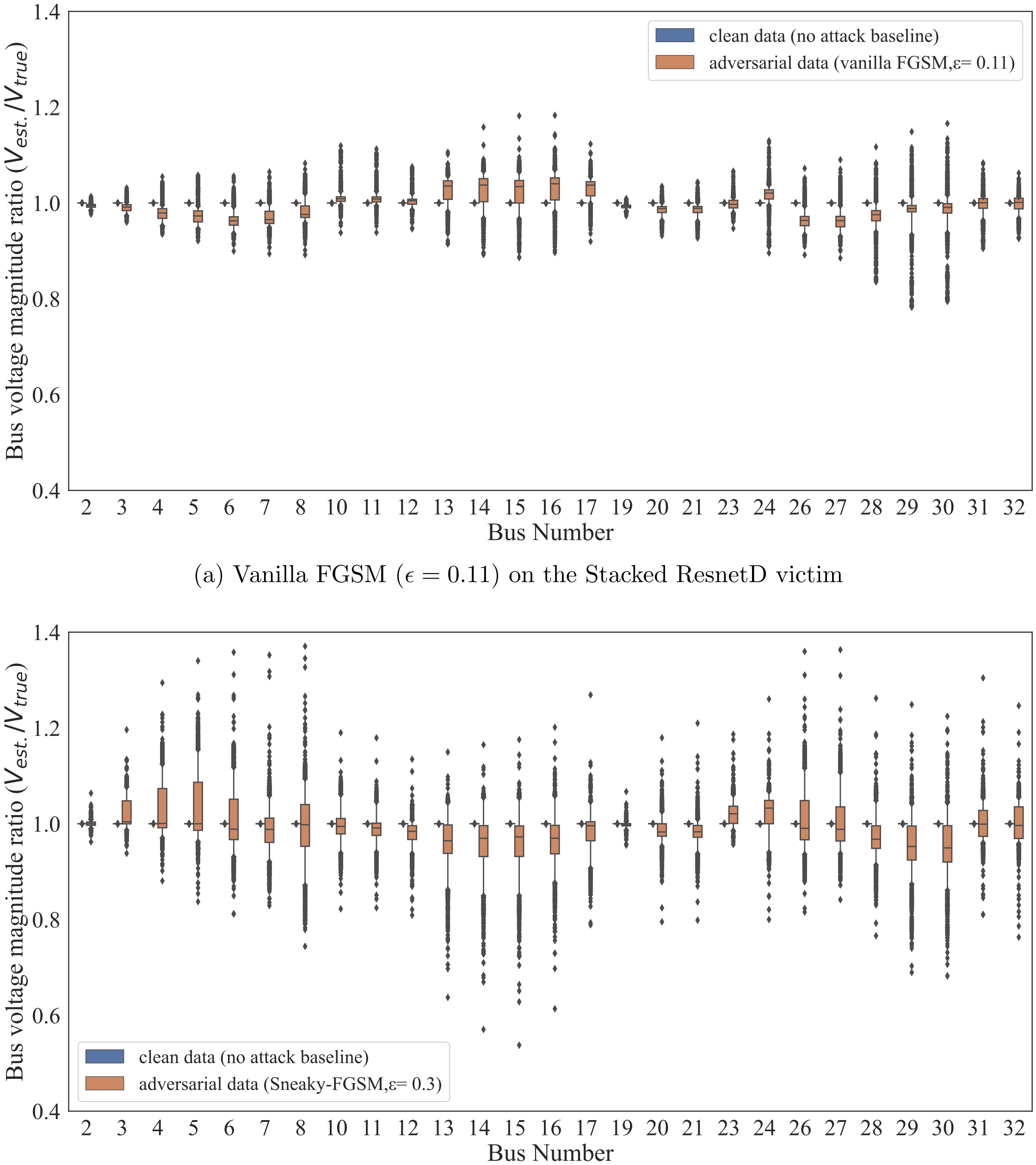
Figure 4.5: Distribution of bus voltage magnitude ratio over all unobserved buses. The ϵ values are chosen such that the average BDD detection rate does not exceed 10%. Note the y-axis is in logarithmic scale.

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, i.e. at least 90% success rate. Observe that Sneaky-FGSM is capable of bypassing BDD with higher ϵ values than that of vanilla FGSM (Figure 4.4). In other words, by utilizing the Sneaky-FGSM algorithm, it is possible to add more perturbation without being detected while keeping BDD detection rate the same.

Figure 4.5 shows the distribution of two ratios, $\frac{|V_{est.}|}{|V_{true}|}$ and $\frac{|V_{est.}^{adv}|}{|V_{true}|}$, which helps compare the effect of vanilla FGSM ($\epsilon = 0.11$) and Sneaky-FGSM ($\epsilon = 0.3$) attacks on the (victim) state estimation models. Here, $|V_{true}|$ is the true bus voltage magnitude, $|V_{est.}|$ and $|V_{est.}^{adv}|$ are the estimated voltage magnitudes under normal conditions and under adversarial attack, respectively. As expected, the original DSSE 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 4.6 shows a more detailed comparison of the vanilla FGSM and Sneaky-FGSM attacks, generated using ϵ values such that the average BDD detection rate is not more than 10%, 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. For brevity, we only present the box and whisker plots for two surrogate models.

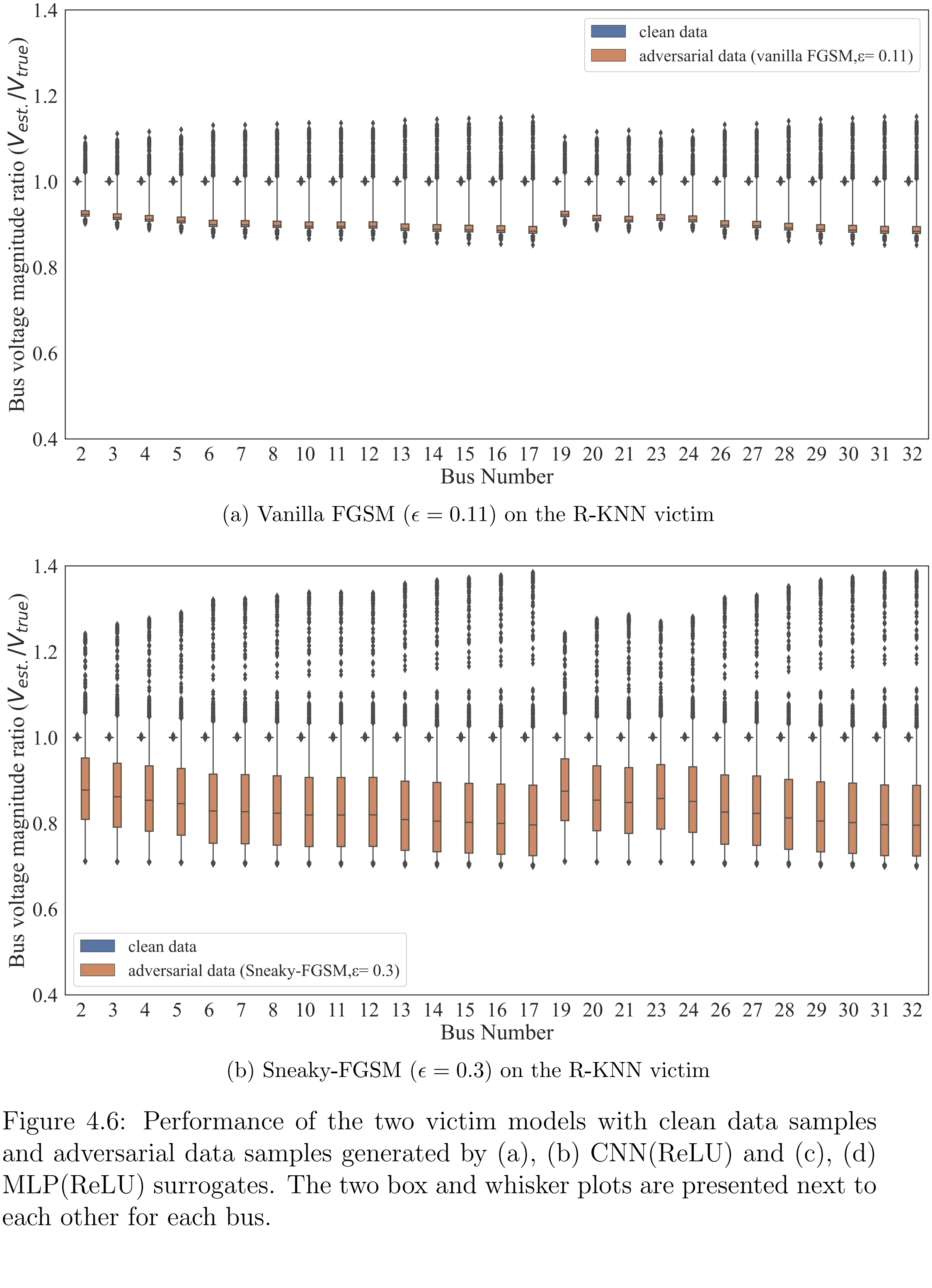
4.2.3 Impact on Voltage Limit Violation Detection

It is essential for system operators to determine the real-time state of the power distribution system to ensure its reliable operation. One important application of DSSE is detecting voltage limit violations in the network. 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.



(a) Vanilla FGSM ($\epsilon = 0.11$) on the Stacked ResnetD victim

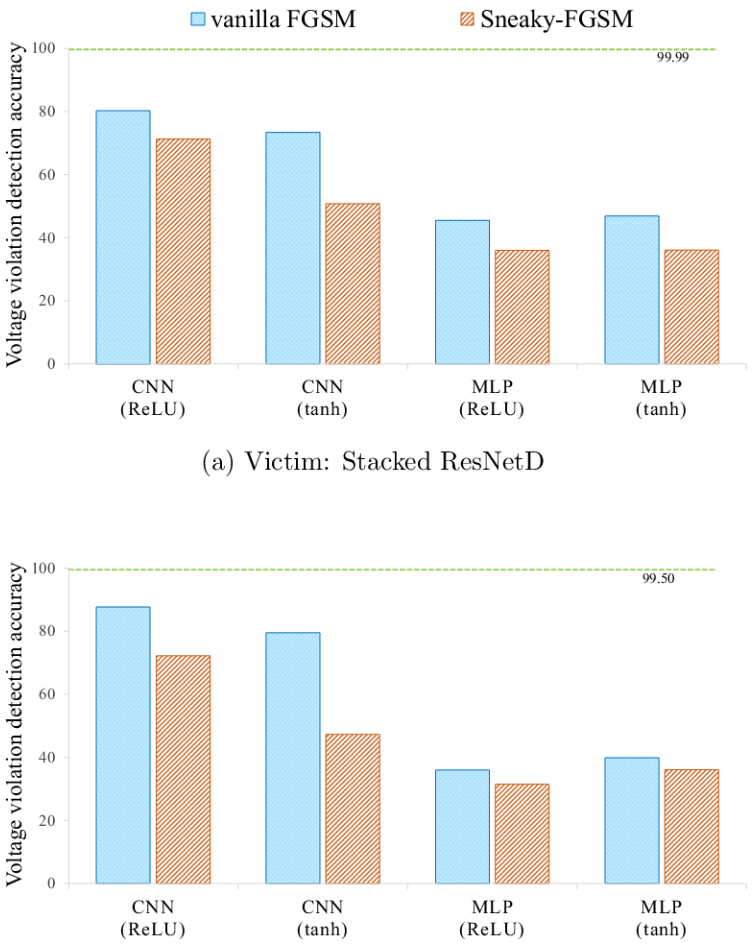
(b) Sneaky-FGSM ($\epsilon = 0.3$) on the Stacked ResnetD victim



(a) Vanilla FGSM ($\epsilon = 0.11$) on the R-KNN victim

(b) Sneaky-FGSM ($\epsilon = 0.3$) on the R-KNN victim

Figure 4.6: Performance of the two victim models with clean data samples and adversarial data samples generated by (a), (b) CNN(ReLU) and (c), (d) MLP(ReLU) surrogates. The two box and whisker plots are presented next to each other for each bus.



(a) Victim: Stacked ResNetD

(b) Victim: R-KNN

Figure 4.7: Reduction in voltage limit violation detection accuracy under adversarial attacks crafted using different surrogate models. The horizontal line in green shows the victim model's accuracy on benign data samples.

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

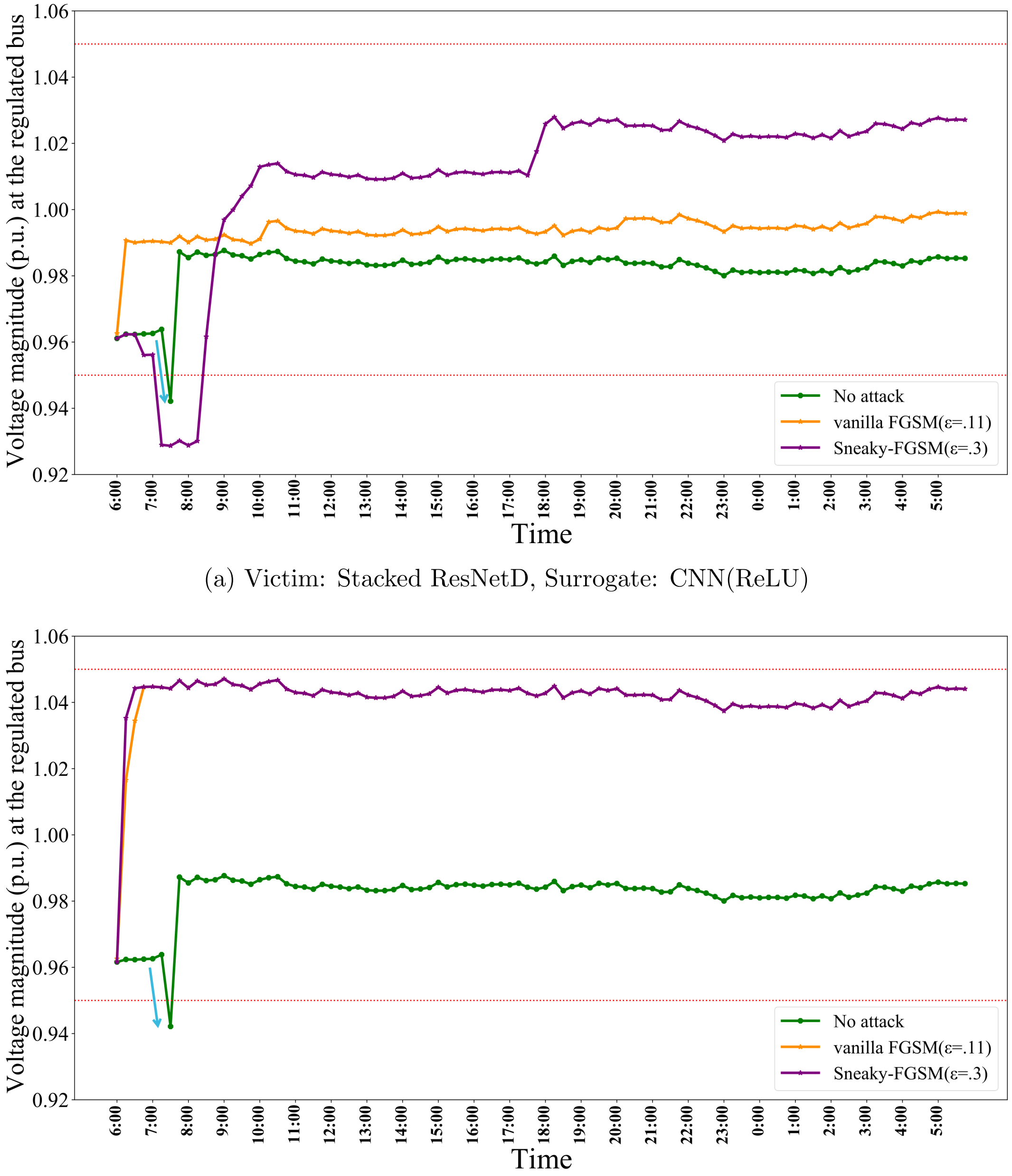
Then each of these three binary vectors is compared to the true detection vector (VLV) to analyze the impact of the proposed attacks on the accuracy of voltage limit violation detection. Figure 4.7 shows the final outcome of the experiment. As we observe, although both of the victim models exhibit high accuracy when tested in non-adversarial cases, they experience a substantial drop in accuracy under adversarial settings.

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.

4.2.4 Impact on the Voltage Control Scheme

To maximize the system observability with a small number of measurement devices, we instrument all the head-ends of the primary feeders of our test system with D-PMUs as depicted in Figure 4.1. 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 use the *RegControl* object from the OpenDSS simulator to install a transformer with OLTC on Line $5-6$ and set the corresponding control rule as *"If the voltage at Bus* 6 *violates the limits, change the tap setting accordingly"*.[1] To

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



(a) Victim: Stacked ResNetD, Surrogate: CNN(ReLU)

(b) Victim: R-KNN, Surrogate: MLP(ReLU)

Figure 4.8: Impact of adversarial attacks on the rule-based voltage control mechanism. All of the major spikes in the voltage profile are due to multiple concurrent tap change operations except the one pointed with the cyan arrow, which occurred due to a sudden increase in load demand.

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 6:00am to 6:00am of the next day) in three different settings: a) in the absence of an attacker; b) under the vanilla FGSM

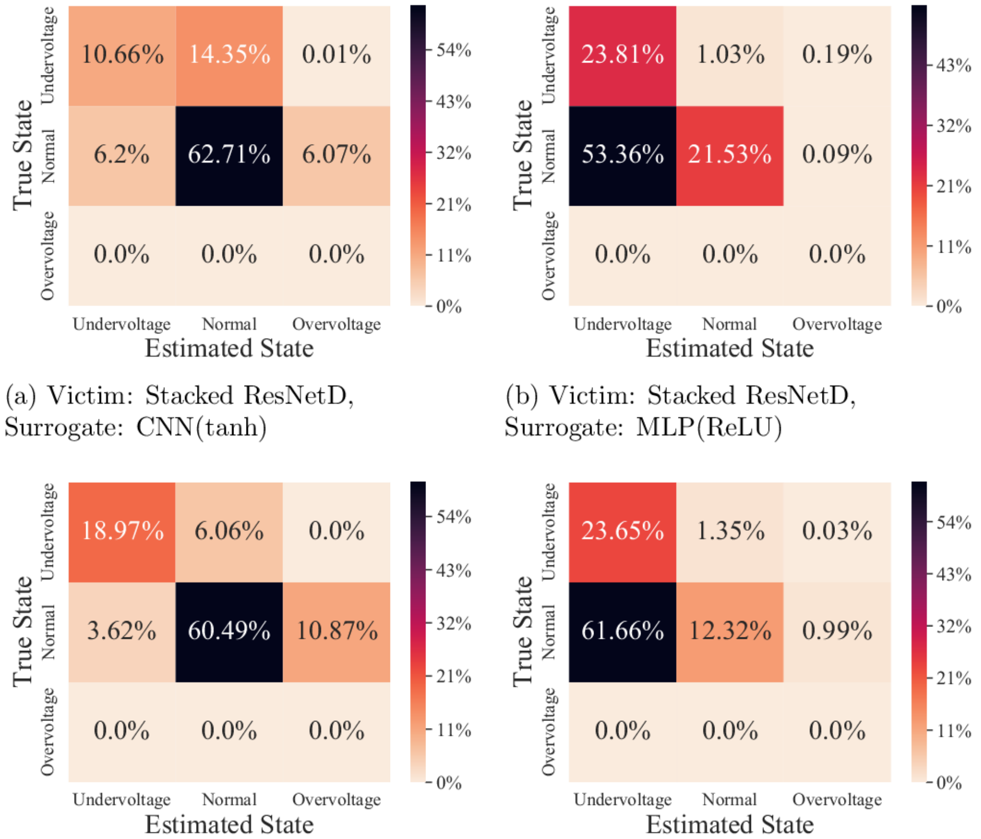
Table 4.1: Impact of adversarial attacks on the rule-based voltage regulation process during a day-long simulation.

Victim	Surrogate	#Unnecessary Tap Changes Initiated		#Voltage Limit Violations	
		Vanilla FGSM	Sneaky-FGSM	vanilla FGSM	Sneaky-FGSM
Stacked ResNetD	CNN(ReLU)	4	11	0	5
	CNN(tanh)	1	11	0	23
	MLP(ReLU)	96	96	0	0
	MLP(tanh)	96	96	0	0
R-KNN	CNN(ReLU)	6	21	1	88
	CNN(tanh)	10	56	83	85
	MLP(ReLU)	68	96	0	0
	MLP(tanh)	66	96	0	0

attack; and c) under the Sneaky-FGSM attack. 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 4.8 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 7:30am, 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 as well as voltage fluctuations and occasional under-voltage problems at the regulated bus. These issues might get more pronounced under the Sneaky-FGSM attack. Table 4.1 shows the impact of these attacks in terms of the number of unnecessary tap changes and voltage limit violations during a day-long simulation.

As we observe from Table 4.1, the main limitation of the proposed black-box attacks is their untargeted nature. While it is guaranteed that the proposed attack strategies will mislead the voltage detection mechanism and impact the control system, the magnitude of the impact varies across victim models and surrogate models being used. In some cases, we see a significant rise in the number of unnecessary tap change operations and no voltage limit viola-



(a) Victim: Stacked ResNetD, Surrogate: CNN(tanh)

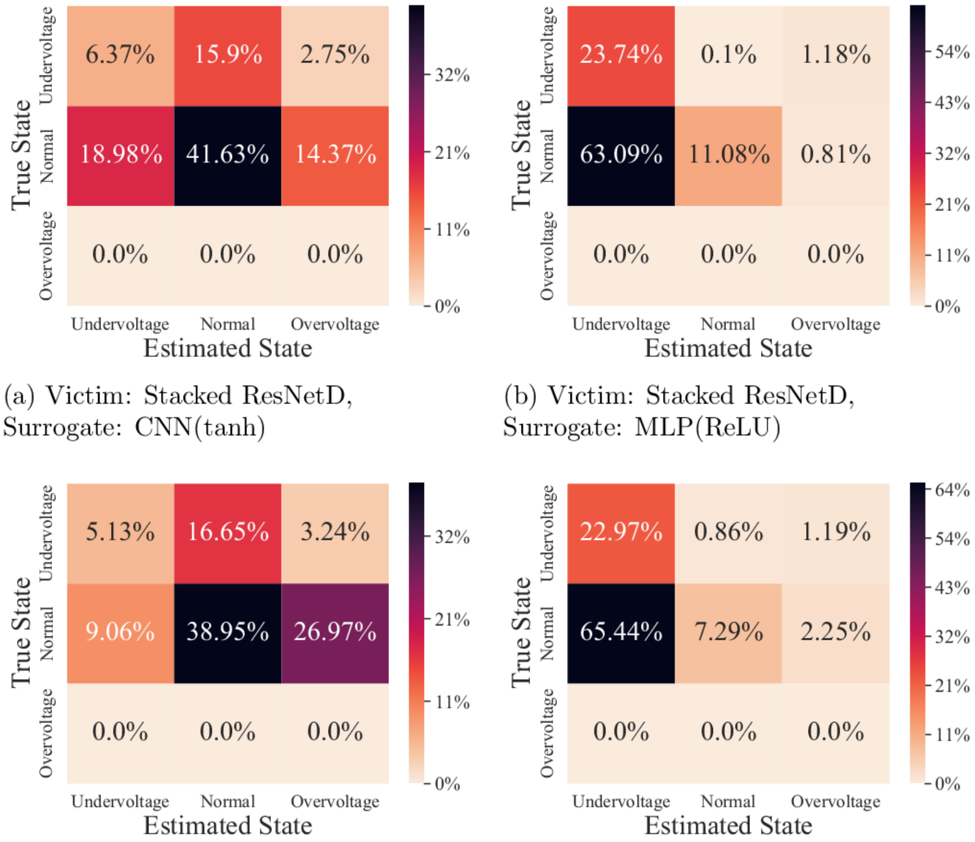
(b) Victim: Stacked ResNetD, Surrogate: MLP(ReLU)

(c) Victim: R-KNN, Surrogate: CNN(tanh)

(d) Victim: R-KNN, Surrogate: MLP(ReLU)

Figure 4.9: Untargeted nature of the vanilla FGSM attack ($\epsilon = 0.11$).

tion. In other cases, we observe both increment in tap change operations and voltage limit violations. Figure 4.8a shows the voltage profile of the regulated bus, presenting a visual representation of the simulation result (for brevity, we present only two plots among the eight possible victim-surrogate combinations presented in Table 4.1). We observe vanilla FGSM attack initiates four unnecessary tap-up operations and moves the voltage profile in an upward direction. On the other hand, Sneaky-FGSM initiates a combination of tap-up and tap-down operations, creating some under-voltage incidents. Moreover, the effects of these attacks on the control system vary based on the victim model's architecture and the surrogate model being used at the attacker's end. For example, in Figure 4.8b, an MLP-based surrogate has been used to generate adversarial measurements against the R-KNN-based state estimator.



(a) Victim: Stacked ResNetD, Surrogate: CNN(tanh)

(b) Victim: Stacked ResNetD, Surrogate: MLP(ReLU)

(c) Victim: R-KNN, Surrogate: CNN(tanh)

(d) Victim: R-KNN, Surrogate: MLP(ReLU)

Figure 4.10: Untargeted nature of the Sneaky-FGSM attack ($\epsilon = 0.3$).

As can be seen from Figure 4.6, regardless of the attack algorithm used, this particular surrogate model tends to push the predicted states by the R-KNN victim significantly below the lower-voltage threshold (i.e. 0.95p.u. in our test case); therefore, multiple unnecessary tap-up operations take place at 6:00am, causing a sudden fluctuation in the voltage profile. The misprediction continues and unnecessary tap-up command is initiated at every instance during the 24-hours simulation period. However, no fluctuation in the voltage profile is observed afterward, because the OLTC reaches its maximum tap position[2] after the first hour. That said, the bus voltage profile would have entered the over-voltage state if the system supported further tap-up operations.

[2] In OpenDSS, the RegControl objects are limited by a maximum of 16 tap operations in one direction.

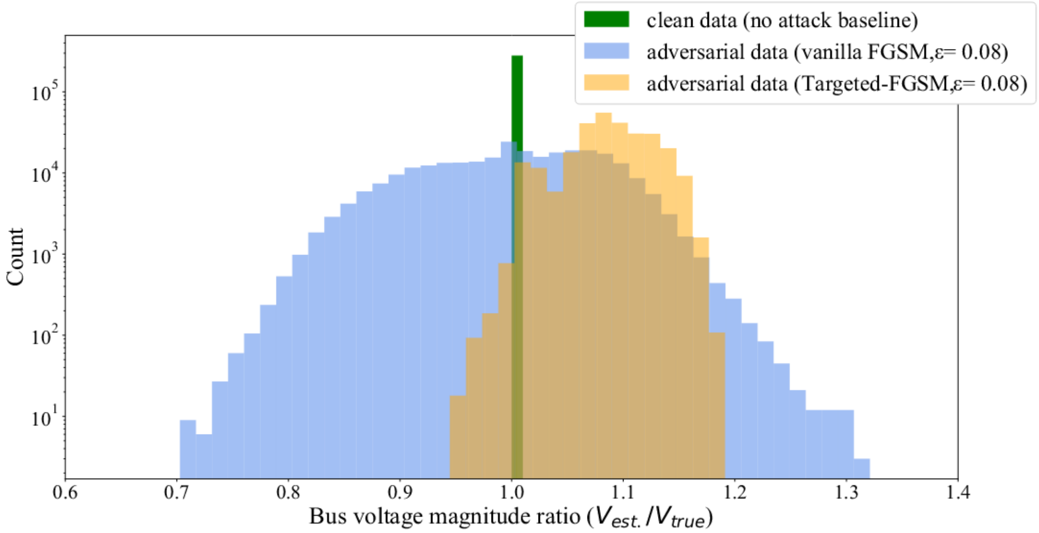
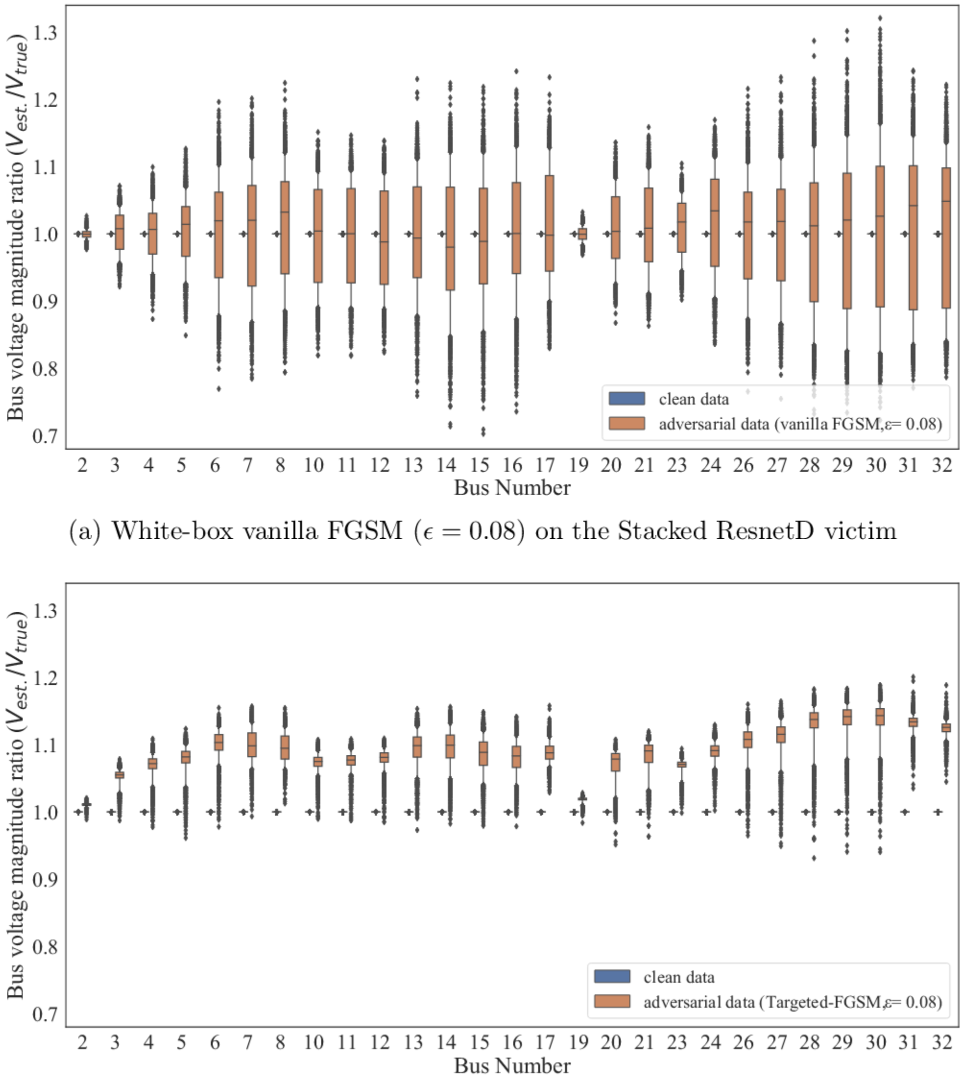


Figure 4.11: Distribution of bus voltage magnitude ratio over all unobserved buses under vanilla FGSM ($\epsilon = 0.08$) and Targeted-FGSM ($\epsilon = 0.08$) attacks. Note the y-axis is in logarithmic scale.

Figures 4.9 and 4.10 present a further elaborated explanation of the untargeted nature of vanilla FGSM and Sneaky-FGSM attacks. We evaluate two victim models under the proposed black-box attacks and compare the actual states with the estimated states for all the unobserved buses using confusion matrices. The confusion matrices placed side by side show the impact of different surrogate models on the same victim model. The confusion matrices aligned vertically show the impact of the same surrogate models on different victims. The type of estimation error varies in different settings based on how the victim model at the DSOs' end and the surrogate model at the adversary's end capture the inherent features of the input data.

4.2.5 Targeted Attack on DSSE

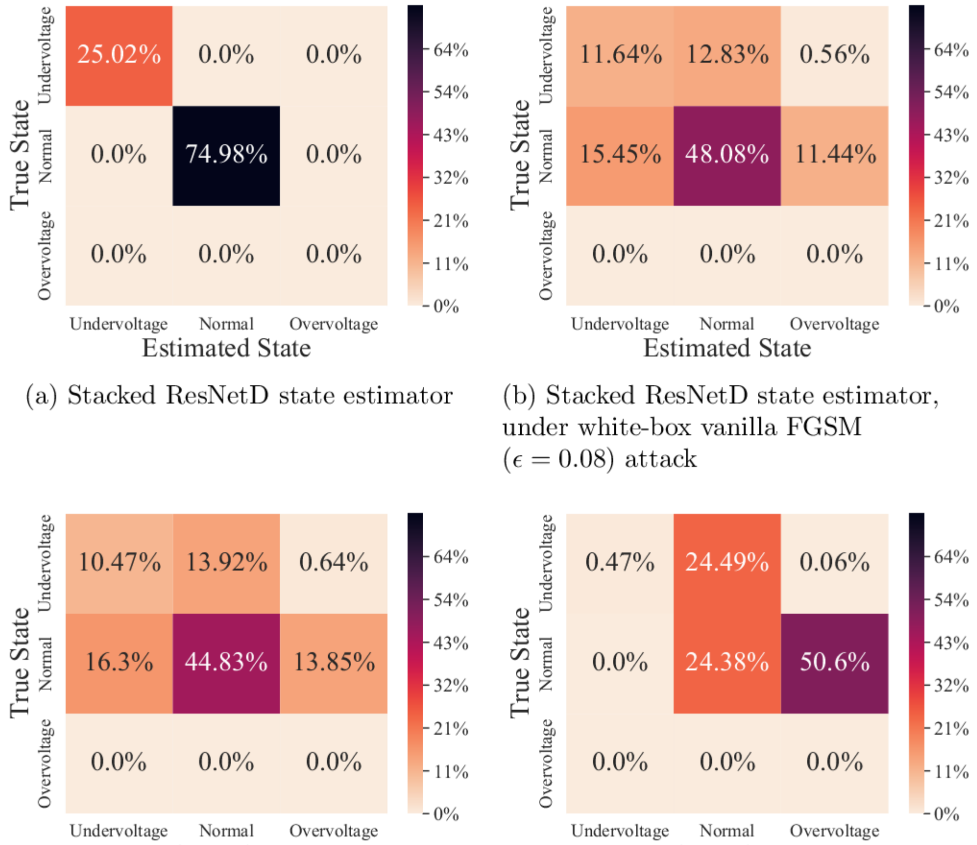
As discussed in the previous section, despite being successful in misleading the state estimator and affecting the voltage regulation scheme, vanilla and Sneaky-FGSM attacks do not guarantee any specific power quality issues (i.e. guaranteed under-voltage or over-voltage occurrences) due to their untargeted nature. In Section 3.4.3, we introduced Targeted-FGSM, a targeted adversarial attack capable of misleading the state estimator in a certain direction, thereby



(a) White-box vanilla FGSM ($\epsilon = 0.08$) on the Stacked ResnetD victim

(b) White-box Targeted-FGSM ($\epsilon = 0.08$) on the Stacked ResnetD victim

Figure 4.12: Performance of the two victim models with clean data samples and adversarial data samples generated by (a) white-box vanilla FGSM and (b) white-box Targeted-FGSM surrogates. 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 white-box vanilla FGSM ($\epsilon = 0.08$) attack

(c) Stacked ResNetD state estimator, under white-box Sneaky-FGSM ($\epsilon = 0.08$) attack

(d) Stacked ResNetD state estimator, under Targeted-FGSM ($\epsilon = 0.08$) attack

Figure 4.13: Impact of untargeted and targeted white-box attacks on the voltage limit violation detection mechanism.

inducing certain power quality issues, e.g. under-voltage or over-voltage incidents. In this section, we present the corresponding experimental results.

Our proposed Targeted-FGSM is a white-box gradient-based attack that utilizes the gradient of the victim model's loss function with respect to the input sample to generate stealthy adversarial data. Since the R-KNN approach does not have a loss function, it is not possible to perform a gradient-based attack against the R-KNN victim. Therefore, we perform the Targeted-FGSM attack only on the Stacked-ResNetD model and compare its performance with a white-box attack with the same perturbation factor, ϵ, on the same victim.

As described in Section 3.4.3, we consider an adversary who intends to create more under-voltage events. Thus, the adversary needs to push the voltage levels in the estimated state vector in an upward direction. We define $v_{min} = 1.03$ and $v_{max} = 1.04$ in Algorithm 3.

Figure 4.11 compares white-box vanilla FGSM ($\epsilon = 0.08$) and Targeted-FGSM ($\epsilon = 0.08$) on the Stacked ResNetD victim model in terms of the distribution of two ratios, $\frac{|V_{est.}|}{|V_{true}|}$ and $\frac{|V_{est.}^{adv}|}{|V_{true}|}$. Here, $|V_{true}|$ is the true bus voltage magnitude, $|V_{est.}|$ and $|V_{est.}^{adv}|$ are the estimated voltage magnitudes under normal conditions and under adversarial attack, respectively. It can be seen that vanilla FGSM moves the estimates in both directions while Targeted-FGSM maintains a majority of cases with a ratio greater than 1. A detailed per-bus comparison using box and whisker plots is shown in Figure 4.12, which confirms the targeted nature of this attack.

Figure 4.13 presents a quantitative comparison between the untargeted and targeted attacks. As we observe in Figure 4.13a, the Stacked ResNetD model, under normal conditions, is able to accurately estimate the true system state. Figures 4.13b and 4.13c illustrate the impact of white-box vanilla FGSM and Sneaky-FGSM, respectively. We notice a rise in false positives and false negatives due to these attacks, although the estimate can shift in either direction. Figure 4.13d shows the impact of the Targeted-FGSM attack. It is evident that the estimated states are now being pushed upward, fulfilling the adversary's objective.

We conduct a day-long simulation and present the bus voltage profile in Figure 4.14. As it can be seen, the proposed Targeted-FGSM attack misleads the state estimator to predict over-voltage even when the bus voltage is within the acceptable range. As a result, multiple unnecessary tap-down operations are triggered, causing a persistent under-voltage issue in the regulated bus. On the other hand, with the white-box vanilla-FGSM, we obtain a mixture of tap-up and tap-down operations due to its untargeted nature.

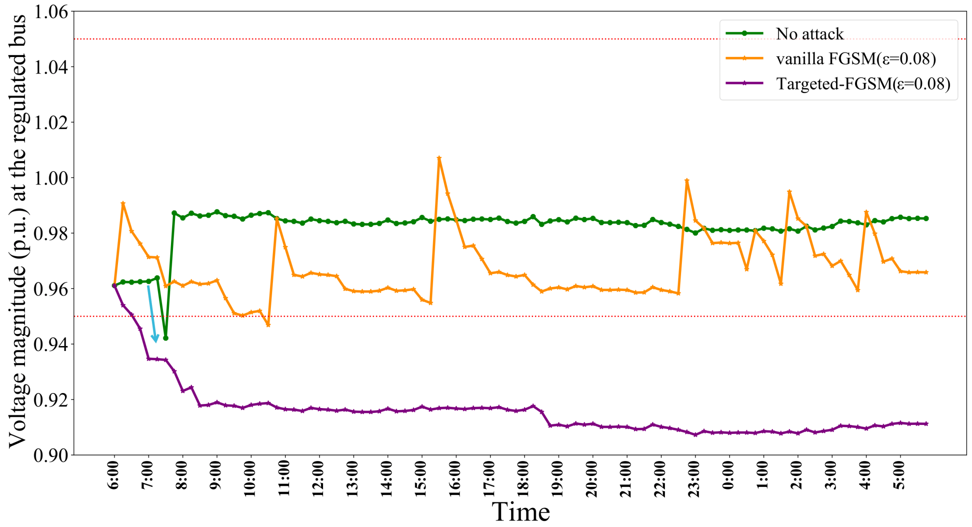


Figure 4.14: Impact of the white-box vanilla FGSM and Targeted-FGSM attacks on the rule-based voltage control mechanism. All of the major spikes in the voltage profile are due to multiple concurrent tap change operations except the one pointed with the cyan arrow, which occurred due to a sudden increase in load demand.

4.3 Discussion

We have presented three different adversarial attack algorithms, each designed with distinct adversarial objectives while sharing a common intention of deceiving data-driven DSSE techniques and consequently affecting the control operations. Now we present a brief comparison of the adversarial attack algorithms explored in this chapter.

In Table 4.2, we compare the proposed attack strategies in terms of (a) effectiveness (i.e. their ability to affect the DSSE technique and the control operation), (b) stealthiness (i.e. their ability to fool the residual-based BDD mechanism), (c) their ability to achieve a specific adversarial goal, and (d) their ability to work in a black-box setting. We consider the proposed Targeted-FGSM algorithm *highly* effective due to its ability to achieve specific adversarial objectives. However, in terms of stealthiness, Sneaky-FGSM stands out among the three. Note, the stealthiness comparison is solely based on the residual-based BDD mechanism. In the next chapter, we propose an effective detection-based safeguarding mechanism that is able to achieve high accuracy in detecting various types of adversarial attacks, including the proposed Sneaky-FGSM attack.

Table 4.2: Comparison of the three proposed adversarial attack strategies.

Attack Strategy	Effectiveness	Stealthiness (w.r.t. BDD)	Satisfies Specific Adversarial Goal?	Works in Black-box Setting?
Vanilla FGSM	Low	Medium	✗	✓
Sneaky-FGSM	Medium	High	✗	✓
Targeted-FGSM	High	Medium	✓	✗

In this chapter, we analyzed the adversarial vulnerabilities of data-driven DSSE approaches from the perspective of an adversary and summarized the experimental results. In the next chapter, we look at the problem from the perspective of the system operator and seek protective measures to mitigate these attacks on state estimation.

Chapter 5

Detecting Adversarial Attacks

In the previous chapter, we have explored the adversarial vulnerabilities associated with data-driven state estimators and analyzed the impact of different types of adversarial attacks on the control system that relies on the estimated states. In this chapter, we propose an effective detection-based defense mechanism that can be used as a safeguard for distribution system state estimators.

5.1 Feature Attribution-based Detection Strategy

Feature attribution refers to the process of quantifying the contribution of individual features of an input sample in a machine learning model's decision-making process. Feature attribution techniques help in understanding how much each feature influences the model's output and gaining insights into why a specific prediction was made. In [122], a *leave-one-out (LOO)*-based feature attribution method has been used to detect adversarial samples generated against a C-class classifier, $f(\mathbf{z};\theta) : \mathbb{R}^d \to [0,1]^C$, that maps an input sample $\mathbf{z}$ of dimension d to a probability vector $f(\mathbf{z})$ of dimension C. The proposed feature attribution method maps an input sample $\mathbf{z} \in \mathbb{R}^d$ to an attribution vector $\phi(\mathbf{z}) \in \mathbb{R}^d$, such that the i^{th} element of $\phi(\mathbf{z})$ is the contribution of feature i in the prediction of $f(\mathbf{z};\theta)$. Each element of the feature attribution vector for input sample $\mathbf{z}$ is computed as

$$\phi(\mathbf{z})_i = f(\mathbf{z};\theta)_c - f(\mathbf{z}_i;\theta)_c, \text{where, } c = \arg\max_{j \in C} f(\mathbf{z};\theta)_j \tag{5.1}$$

Here, $\mathbf{z}_i$ is a masked example generated by masking the i^{th} feature of $\mathbf{z}$ by 0, and $f(\mathbf{z};\theta)_c$ denotes the c^{th} element of the probability vector returned by the classifier $f(\mathbf{z};\theta)$.

Then, the interquartile range (IQR) of $\phi(\mathbf{z})$, defined as the difference between the 75^{th} percentile and the 25^{th} percentile among all entries of $\phi(\mathbf{z}) \in \mathbb{R}^d$, is used to calculate statistical dispersion in the feature attribution matrix:

$$\mathrm{IQR}(\phi(\mathbf{z})) = Q_{\phi(\mathbf{z})}(0.75) - Q_{\phi(\mathbf{z})}(0.25) \tag{5.2}$$

where

$$Q_{\phi(\mathbf{z})}(p) = \min\{\beta : \frac{\#\{i : \phi(\mathbf{z})_i < \beta\}}{d} \geq p\}.$$

As shown in [122], for an adversarially perturbed data, $\mathbf{z}'$, $\mathrm{IQR}(\phi(\mathbf{z}'))$ is larger than that of the corresponding benign data. An intuitive explanation behind this behavior lies in the working principle of algorithms generating adversarial perturbation. These perturbations can alter the importance and influence of different features in the data, causing the model to make incorrect predictions. As a result, the dispersion measure in feature attribution scores for adversarial data tends to exhibit comparatively higher values. Therefore, adversarial samples can be distinguished from benign ones either by thresholding the IQR of feature attribution maps or by fitting a logistic regression model to the IQR values. This approach achieved superior performance in detecting adversarial image samples generated against multiclass classifiers [122], motivating us to adopt this detection strategy in safeguarding the data-driven state estimators.

In Equation (5.1), the prediction score for the class label that has the highest value has been used to calculate the feature attribution value. However, for regression tasks there is no ‘class label’, and therefore, the formulation for obtaining $\phi(\mathbf{z})$ must be different. To adopt the same detection method in regression tasks, we define each element, $\phi(\mathbf{z})_i$, of the feature attribution matrix as the difference between the original prediction of the DSSE model on the true measurement data $\mathbf{z}$ and the prediction when the i^{th} feature of $\mathbf{z}$ is masked by 0.

$$\phi(\mathbf{z})_i = f(\mathbf{z}) - f(\mathbf{z}_i) \tag{5.3}$$

Then, we calculate IQR $(\phi(\mathbf{z}))$ using Equation (5.2) and fit a logistic regression model to the IQR values calculated for benign and adversarial training samples. Since it is not realistic for the system operator to generate the exact same adversarial samples as the attacker, we train the logistic regression model using a particular set of adversarial samples, generated using the vanilla FGSM ($\epsilon = 0.1$) algorithm crafted with the CNN(ReLU) surrogate. Once the logistic regression model is trained, we deploy it for testing.

Since all three detection mechanisms evaluated in this thesis are binary classifiers, we use the receiver operating characteristics (ROC) curves to evaluate the performance of the detection mechanisms. The ROC curve plots the classifiers' false positive rates against their true positive rates across different thresholds. Note, for the LOO-based detection algorithm, the threshold that varies throughout the ROC curve is the cutoff value of the logistic regression model.

5.2 Baseline Detection Methods

We adopt the feature attribution-based adversarial attack detector described in the previous section as a DSSE safeguarding mechanism and compare its performance with two state-of-the-art detection-based SE safeguarding strategies proposed in prior works, namely the neural attack detector (NAD) [102] and the KL Divergence-based FDIA detector [18]. In this section, we present a brief overview of the working principle of the two baseline detection strategies.

Neural Attack Detection (NAD): Various types of neural networks, including CNN [64], [110], RNN [7], [112], [114], MLP [96], [102], and DNN [42] have been proven effective in detecting stealthy FDIA that are able to bypass traditional BDD mechanisms. Inspired by these results, we implement the NAD strategy proposed in [102] as one of the baselines for detection models. We use the fully connected neural network (FCNN), comprised of six hidden layers with ReLU activation functions, originally proposed as a safeguarding mechanism for the IEEE 30-bus system in [102]. According to [102], once trained, this model can detect traditional stealthy FDIA (as described in Equation (2.2)) with 99.6% accuracy.

The NAD model is trained using a mixture of benign and adversarial data samples at the operator's end. Similar to the LOO-based detection mechanism described above, we train the NAD model using a particular set of adversarial samples generated using the CNN(ReLU) surrogate model and vanilla FGSM algorithm with $\epsilon = 0.1$. Then, we test the efficacy and generality of this detection strategy using a variety of adversarial samples generated using four different surrogates and varying ϵ values.

The NAD model calculates the probability estimates for both benign and adversarial classifications of any given input sample, $\mathbf{z}$. We consider the 'positive' class to represent adversarial samples. By comparing the probability estimate of the positive class with the actual class labels at different thresholds, we generate the ROC curves.

KLD-based FDIA Detection Strategy: The Kullback–Leibler divergence (KLD) has been utilized in several previous works to detect FDIA [18], [49], [75]. Chaojun et al. [18] proposed an effective FDIA detection strategy in AC state estimation where at any time step k, the KLD between two distributions p_k and q is compared against a predefined threshold, τ, to detect if the measurement sample z_k contains bad data. The detection mechanism can be described as follows:

$$\text{isBad=}\begin{cases}1, & \text{if } D_{KL}\left(p_k \| q\right) > \tau \\ 0, & \text{otherwise}\end{cases} \tag{5.4}$$

Before delving into the details, it is important to clarify how the term 'distribution of measurement variations' has been used in the context of the KLD-based detection method. We need to detect bad measurement data in an online fashion. At any time step k, we obtain the measurement vector $\mathbf{z_k}$ and decide if it contains bad measurement. For this, we calculate the difference between $\mathbf{z_k}$ and $\mathbf{z_{k-1}}$; this difference is considered as the distribution of measurement variations at time step k, p_k. This might be confusing at first because only one data point is defined as a distribution. However, the basic idea is based on the observation that power systems behave as quasi-static systems, where the system state changes constantly but slowly. Thus, when all the measurements are normalized in the per unit system, most of the measurement variations, i.e. each of the elements of p_k, should be small and close to zero, despite the fact that they are collected from different types of sensors that are placed in different locations. In other words, regardless of the type of the measurement (be it active load, reactive load, or bus voltage), the difference between two consecutive measurements obtained from any sensor can be treated as a random variable, namely *measurement variation*, which takes arbitrary values. If no bad data is injected in $\mathbf{z_k}$, then the distribution of measurement variation at time step k, i.e. p_k, should be similar to the distribution of average measurement variations obtained from the historical data. We note that this is indeed a strict assumption and might be one of the reasons behind the subpar performance of this detection strategy.

In Equation (5.4), q is the distribution of average measurement variations obtained from the historical data, p_k is the distribution of measurement variations between the current time step and the previous time step, and $D_{KL}\left(p_k \| q\right)$ is the Kullback–Leibler divergence between p_k and q which is given by:

$$D_{KL}\left(p_k \| q\right) = \sum_x p_k(x) ln \frac{p_k(x)}{q(x)}$$

The accuracy of this detection strategy greatly depends on the selection of τ. We outline a process to select the optimal value of τ below. The experimental results showing the performance of this baseline detector are presented in the next section.

Selecting the value of τ: As described in the original work [18], maximum KLD from historical data with 99% confidence level is considered the optimal value for τ. To find out the optimal τ value, we consider the training data obtained from the first 11 months (January to November) to calculate the distribution of average historical measurement variation, q. Then, the training samples collected in December are used to calculate the distribution of measurement variations, p_{k-dec}, for each of the benign data samples. Figure 5.1

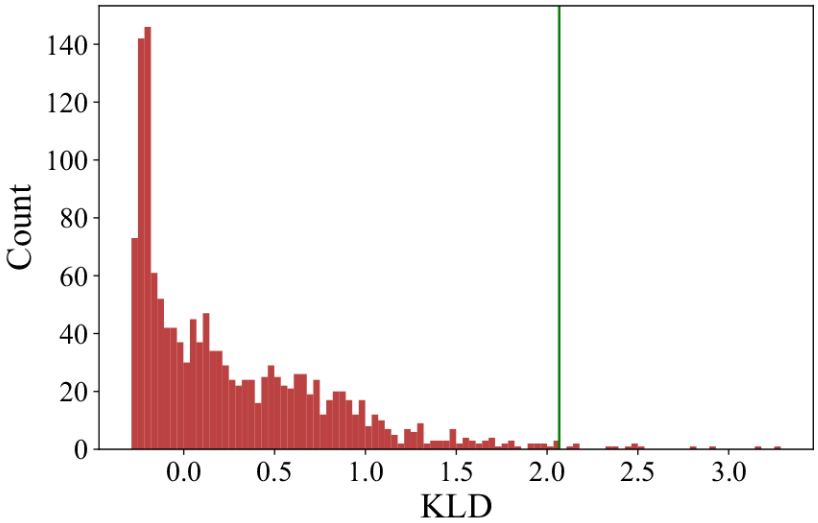


Figure 5.1: Histogram of KLD on December 1 to 15 (benign data). The vertical line in green shows the optimal value for τ.

shows the histogram of $D_{KL}\left(p_{k-dec} \| q\right)$. As we observe, the maximum KLD for 99% confidence level is 2.067. Therefore, from our historical data, $\tau = 2.067$ is the optimal threshold value.

We test the efficacy of this detection strategy on the test measurement dataset, $\mathbf{Z_{test}}$, comprised of 3456 benign and 3456 adversarial measurement samples. We derive the distribution of measurement variations $p_k = \mathbf{z}_k - \mathbf{z}_{k-1}$ from each of the measurement samples $\mathbf{z}_k$ in $\mathbf{Z_{test}}$ and calculate $D_{KL}\left(p_k \| q\right)$. However, instead of sticking into a fixed threshold, we compare the detection mechanisms using ROC curves and area under the curve (AUC). For the KLD-based detector, τ is the discrimination threshold that has been used to plot the ROC curves.

5.3 Comparing Adversarial Data Detection Techniques

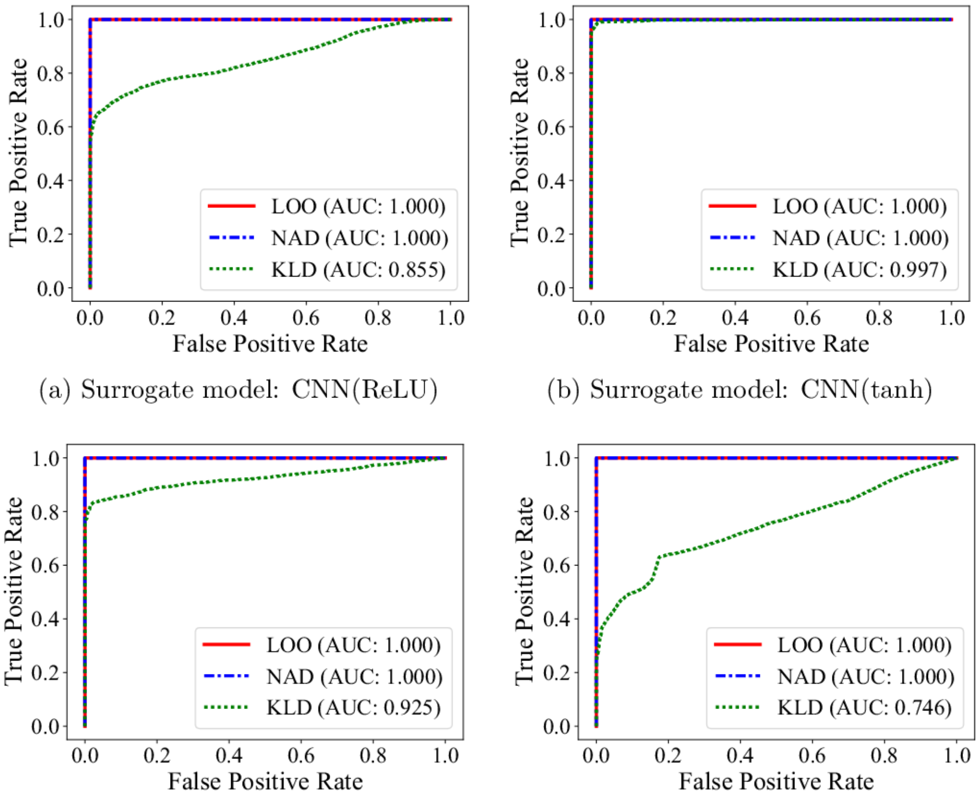
We deploy the three detection methods described above and analyze their efficacy in detecting adversarial measurements. For each of the proposed attack strategies, we generate 12 batches of adversarial measurements using four different surrogates and three distinct ϵ values for each of the surrogates. Similar to the previous chapter, we generate the white-box Targeted-FGSM attack against the Stacked ResNetD victim model. However, while analyzing the performance of these detection strategies on black-box attacks, defining a victim model is not necessary. This is because unlike the residual-based BDD mechanism (described in Section 1.3.2) that uses the states estimated by the victim model to calculate the measurement residual, these detection strategies directly use the sensor measurements, $\mathbf{z}_k$, collected at time step k and classify it as benign or adversarial. Since we do not need to define a victim model to generate adversarial samples in a black-box setting, the performance of the detection strategies on the untargeted black-box attacks remains consistent across different victim models.

Let the true positive rate (TPR) be the proportion of adversarial measurement samples that are correctly classified and the false positive rate (FPR) be the proportion of benign measurements that are misclassified as adversarial. We use the area under the curve (AUC) of the receiver operating characteristic (ROC) curves as the performance evaluation metric for this experiment. Table 5.1 summarizes the experimental results. The ROC curves of the detection methods under vanilla FGSM ($\epsilon = 0.15$), Sneaky-FGSM ($\epsilon = 0.15$), and Targeted-FGSM ($\epsilon = 0.15$) are depicted in Figures 5.2, 5.3, and 5.4, respectively.

Our proposed LOO detector outperforms the other two methods in detecting various types of adversarial attacks generated with different perturbation factors, ϵ. While NAD shows comparable performance in detecting the vanilla FGSM attack, its effectiveness decreases when it is tested with Sneaky-FGSM and Targeted-FGSM attacks crafted with lower values of ϵ. This limitation stems from the fact that the NAD model is trained on adversarial data generated solely by the vanilla FGSM algorithm, highlighting the inherent limitations of neural attack detectors as they are constrained by the data they are trained on, which implies they cannot be generalized. On the other hand, the KLD-based detection strategy has subpar performance (similar to the residual-based BDD mechanism) and fails to offer robust protection against adversarial attacks. This result reveals an interesting observation: *unlike traditional*

Table 5.1: Performance of detection methods against different adversarial attack strategies crafted with different surrogate models. Note: Stacked ResNetD model is used as the victim model for the Targeted-FGSM attack.

Attack	Surrogate Model	ϵ	AUC		
			NAD	KLD	LOO
vanilla FGSM	CNN(ReLU)	0.05	0.999	0.782	0.999
		0.15	1.0	0.855	1.0
		0.30	1.0	0.874	1.0
	CNN(tanh)	0.05	0.999	0.952	1.0
		0.15	0.999	0.997	1.0
		0.30	0.999	0.999	1.0
	MLP(ReLU)	0.05	0.996	0.884	1.0
		0.15	0.999	0.925	1.0
		0.30	1.0	0.929	1.0
	MLP(tanh)	0.05	0.997	0.636	1.0
		0.15	0.999	0.746	1.0
		0.30	1.0	0.771	1.0
Sneaky-FGSM	CNN(ReLU)	0.05	0.687	0.591	0.889
		0.15	0.771	0.635	0.940
		0.30	0.887	0.653	0.946
	CNN(tanh)	0.05	0.888	0.757	0.992
		0.15	0.967	0.844	0.999
		0.30	0.982	0.859	0.999
	MLP(ReLU)	0.05	0.878	0.513	0.892
		0.15	0.973	0.518	0.944
		0.30	0.982	0.519	0.957
	MLP(tanh)	0.05	0.910	0.508	0.960
		0.15	0.987	0.511	0.998
		0.30	0.999	0.511	0.999
Targeted-FGSM	White-box	0.05	0.721	0.938	1.0
		0.15	0.965	0.995	1.0
		0.30	0.997	0.997	1.0



(a) Surrogate model: CNN(ReLU)

(b) Surrogate model: CNN(tanh)

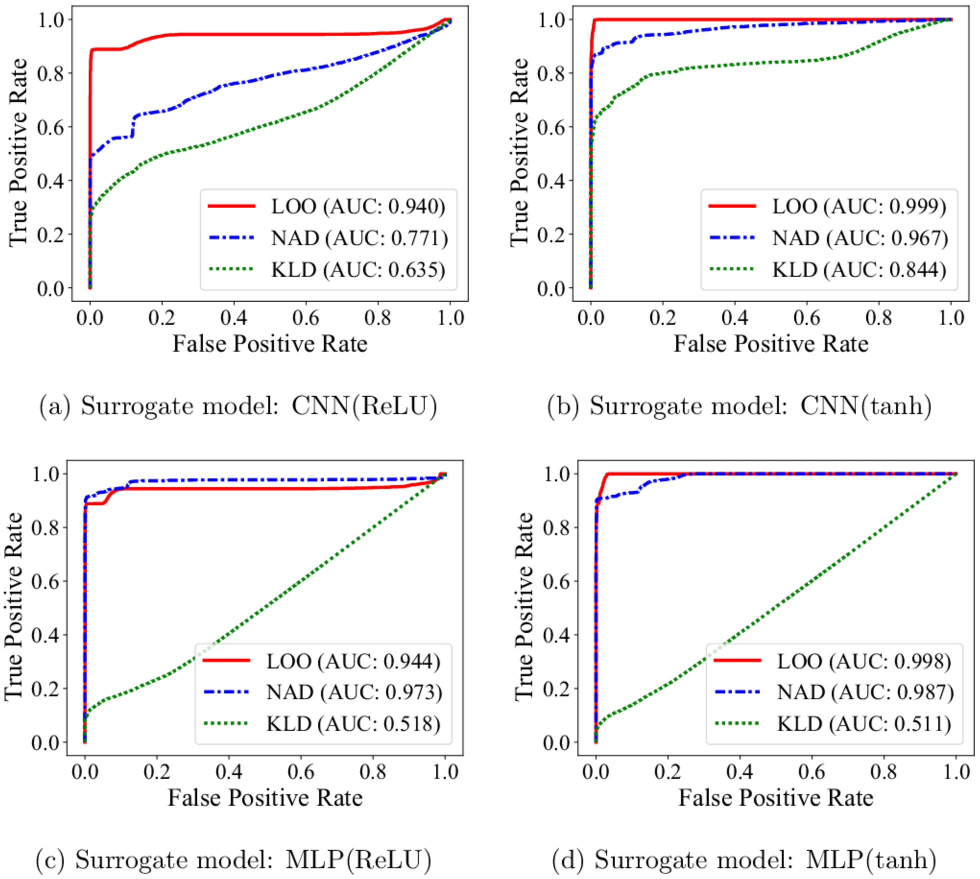
(c) Surrogate model: MLP(ReLU)

(d) Surrogate model: MLP(tanh)

Figure 5.2: ROC curves of detection methods under vanilla FGSM ($\epsilon = 0.15$) attack.

FDIA, surface-level analysis of measurement samples is insufficient to detect adversarial attacks. We delve further into this observation in the following.

Discussion. Adversarial samples are crafted using optimization techniques to find small perturbations that maximally confuse the victim model's decision boundaries while keeping the adversarial sample as close as possible to the original one. As a result, these samples can be indistinguishable from legitimate data in terms of statistical properties, such as mean, variance, and distribution. Traditional statistical methods rely on the assumption that data follows a specific statistical pattern. However, adversarial examples violate these assumptions, rendering such methods ineffective for detection. As a result, despite being effective in detecting FDIA, conventional approaches such as residual-based BDD (described in Section 1.3.2) and the KLD-based detec-



(a) Surrogate model: CNN(ReLU)

(b) Surrogate model: CNN(tanh)

(c) Surrogate model: MLP(ReLU)

(d) Surrogate model: MLP(tanh)

Figure 5.3: ROC curves of detection methods under Sneaky-FGSM ($\epsilon = 0.15$) attack.

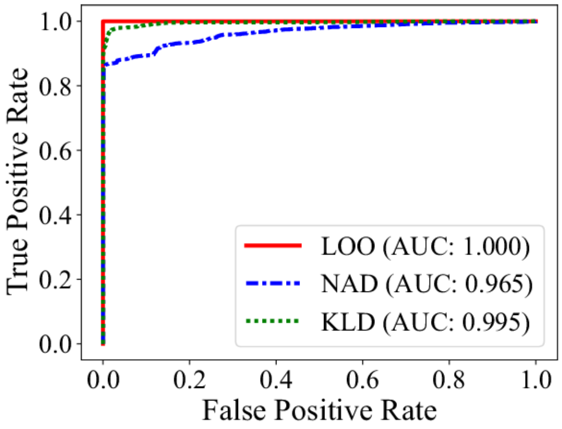


Figure 5.4: ROC curves of detection methods under Targeted-FGSM ($\epsilon = 0.15$) attack.

tion strategy cannot offer robust protection against adversarial attacks. This is evident in Figure 5.5 which shows the histogram of KLDs for the benign and adversarial test samples, $D\left(p_{benign}\|q\right)$ and $D\left(p_{adv}\|q\right)$, respectively. Here, q represents the distribution of measurement variation obtained from the historical data. It can be seen that there is a significant overlap between the two histograms, making it impossible to find a good threshold (τ) value that would achieve high TPR and low FPR.

Adversarial samples, despite being indistinguishable from normal data at the surface level, leave traces on hidden features used by neural networks for classification or prediction purposes. This means the key to identifying adversarial samples lies in examining how the neural network's decision-making process in being affected by these samples. This is where feature attribution comes into play, as they offer profound insight into the model's decision-making process by highlighting the significance of particular features in its predictions. Figure 5.6 compares the histogram of average[1] dispersion measures of feature attributions for benign and adversarial samples from the test dataset. We observe that adversarially perturbed samples have a larger average dispersion in feature attribution compared to the benign samples and there is no overlap between the two histograms. Thus, it is easier to distinguish adversarial samples from benign ones either by thresholding the dispersion measures (when the dispersion measures are scalars) or by fitting a simple classifier, such as a logistic regression model (when the dispersion measures are vectors). This explains the superior performance of the proposed LOO-based detector.

[1] For each measurement $\mathbf{z}$, $\phi(\mathbf{z})$ is a matrix of size 210×162 and $\mathrm{IQR}(\phi(\mathbf{z}))$ is a vector of length 162. To plot the histogram, we take the average of the vector $\mathrm{IQR}(\phi(\mathbf{z}))$.

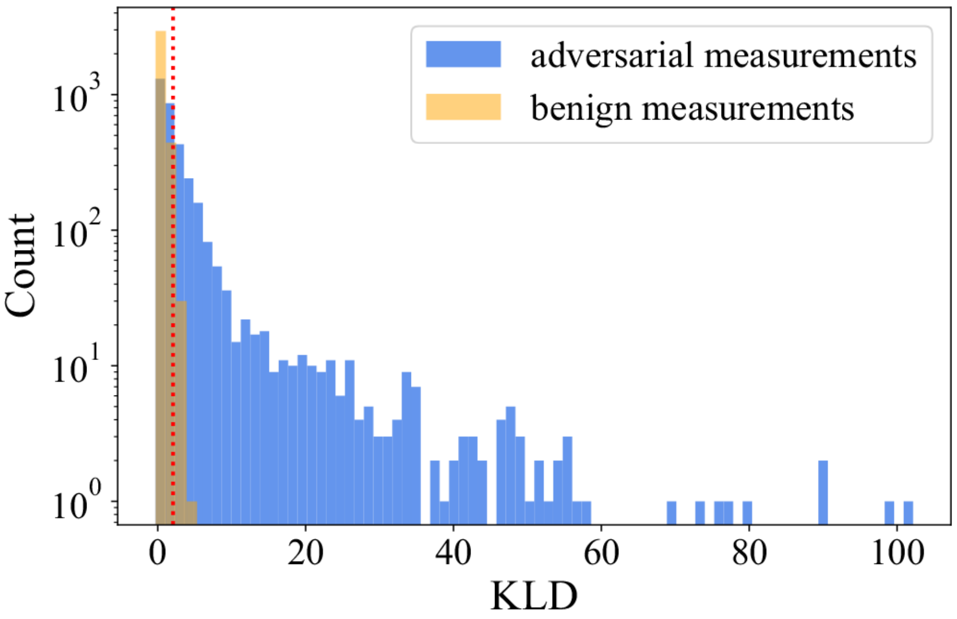


Figure 5.5: Histogram of KLD for the benign and adversarial test samples crafted with vanilla FGSM ($\epsilon = 0.15$) using the MLP(tanh) surrogate. The red vertical line shows the threshold (τ) value obtained from the historical measurement data. Note the y-axis is in logarithmic scale.

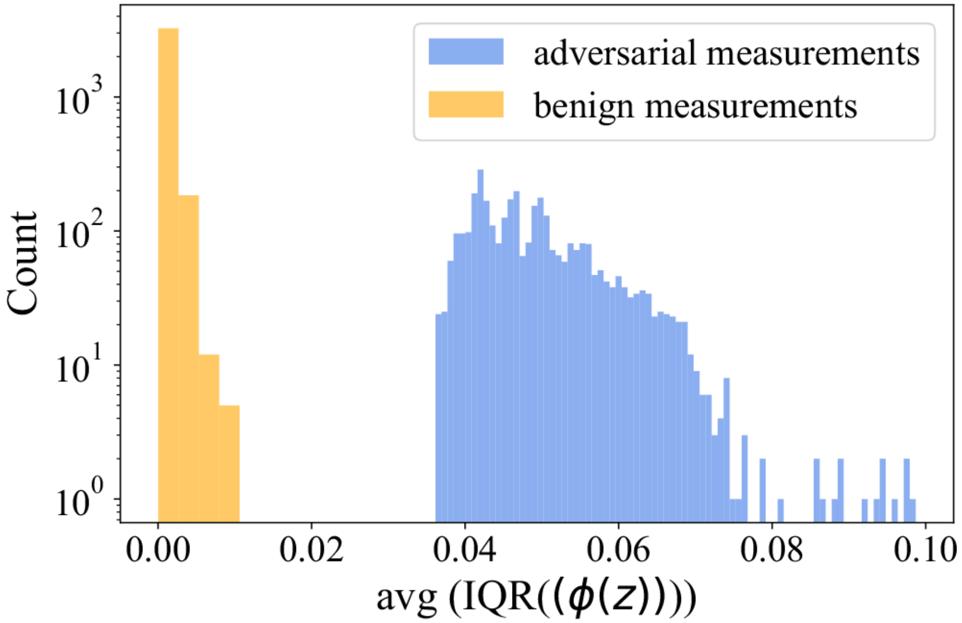


Figure 5.6: Histogram of average dispersion measures of feature attributions for benign samples and adversarial samples crafted with vanilla FGSM ($\epsilon = 0.15$) using the MLP(tanh) surrogate. Note the y-axis is in logarithmic scale.

Chapter 6

Conclusion

Data-driven DSSE approaches offer distinct advantages over traditional methods that rely on assumptions and simplified models. This is because they have the ability to learn from historical data, capturing intricate patterns and non-linear relationships in the system, resulting in improved efficiency and accuracy. Nonetheless, it is crucial to acknowledge the vulnerabilities inherent in these models before integrating them with monitoring and control systems of critical infrastructure, such as the power distribution system. This thesis presents a comprehensive analysis of the security and robustness of data-driven DSSE techniques in the presence of adversarial evasion attacks from two different perspectives: from an adversary's viewpoint (Chapter 4) and from the DSO's viewpoint (Chapter 5).

From the adversary's perspective, we design effective and stealthy evasion attacks that are able to induce error in the state estimation process, thereby exerting deleterious impact on distribution system control and operation practices that rely on the state estimates. We show that data-driven DSSE processes are vulnerable to stealthy and effective black-box adversarial attacks that can fool the BDD mechanism with at least 90% 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, while inflicting comparable or potentially greater harm to the control system. Through comprehensive experimentation involving different combinations of victim and surrogate models, we demonstrate that our proposed Sneaky-FGSM retains its covert nature regardless of the surrogate models being used by the adversary.

After analyzing the proposed black-box attacks that are untargeted, we introduce a targeted, white-box adversarial attack strategy, namely Targeted-FGSM, that creates specific power quality issues by misleading the state estimator in a certain direction. Using real household load data, we conduct simulations on a 33-bus test distribution system as the primary network and the IEEE European low-voltage test feeder as the secondary distribution system connected to each primary node, validating the experimental results.

From the DSO's perspective, our goal is to design effective and robust safeguards for the data-driven state estimators. Based on our analysis of the rationale behind the ineffectiveness of traditional statistical feature-based BDD strategies in detecting adversarial attacks, we propose a highly effective detection-based safeguarding mechanism, namely LOO, that utilizes feature attribution scores to distinguish between benign and adversarial measurements. We corroborate that LOO achieves superior performance compared to two other state-of-the-art detection strategies.

6.1 Limitations and Future Work

We discuss the limitations of this work and possible future research directions below.

- As discussed in Section 3.4.3, our proposed Targeted-FGSM algorithm lacks transferability, hindering the development of black-box targeted attacks. A recent line of work aims to develop targeted adversarial attack strategies that are transferable across different victim models [57], [78], [132]. Exploring the impact of these attack strategies on various smart grid applications can be an interesting future research direction.

- Another drawback of the proposed Targeted-FGSM algorithm is its dependence on the perturbation factor, ϵ, to ensure the desired targeted effect. Exploring alternative ideas from automatically tuning this hyperparameter to overcoming this dependency presents an intriguing research direction.

- In this thesis, we have analyzed the vulnerabilities of data-driven DSSE approaches to evasion attacks only. In future, we aim to conduct similar studies on model poisoning and data poisoning attacks that could take place during the training period. One interesting idea is analyzing the impacts of *backdoor attacks*, where the adversary implants a backdoor into the ML model during training and later, during its deployment, exploits that to achieve adversarial goals.

- Our proposed safeguarding mechanism, LOO, is a detection-based approach. Designing protective measures and developing robust data-driven state estimators present promising future research direction.

6.2 Applicability

Data-driven DSSE techniques, being able to capture the non-linearity in the complex distribution systems, offer better alternatives to the conventional state estimation techniques in terms of accuracy, efficiency, and convergence rate. In this thesis, we analyze the adversarial robustness of data-driven DSSE strategies, uncover their vulnerabilities to adversarial attacks, and demonstrate the potential impacts of such attacks on the control systems that rely on state estimation. Our experimental findings highlight that intelligent models, despite being highly efficient, might introduce novel security risks. Moreover, we illustrate the reasoning behind the sub-optimal performance of conventional bad data detection strategies in detecting adversarial samples and based on the analysis, we propose an effective adversarial attack detector.

Accurate state estimation supports optimized grid operation, better management of power flows, improved power quality, and timely detection of contingencies and faults, which allows faster response to prevent the occurrence of blackouts and minimize service disruptions. We believe this thesis would prove beneficial to the research community in providing insight into potential security risks and reliability concerns that might arise due to the integration of data-driven techniques with monitoring and control tools developed for critical infrastructure such as the power grid and shedding light on the importance of designing robust DSSE approaches as well as effective safeguard mechanisms.

References

[1] F. S. Adi, Y. J. Lee, and H. Song, “State estimation for dc microgrids using modified long short-term memory networks,” *Applied Sciences*, vol. 10, no. 9, p. 3028, 2020.

[2] F. Ahmad, M. Tariq, and A. Farooq, “A novel ann-based distribution network state estimator,” *International Journal of Electrical Power & Energy Systems*, vol. 107, pp. 200–212, 2019.

[3] M. Ahmad, *Power system state estimation.* Artech house, 2013.

[4] S. Ahmed, Y. Lee, S.-H. Hyun, and I. Koo, “Unsupervised machine learning-based detection of covert data integrity assault in smart grid networks utilizing isolation forest,” *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 10, pp. 2765–2777, 2019.

[5] G. Alkhayat and R. Mehmood, “A review and taxonomy of wind and solar energy forecasting methods based on deep learning,” *Energy and AI*, vol. 4, p. 100 060, 2021.

[6] O. Ardakanian, V. W. S. Wong, R. Dobbe, *et al.*, “On identification of distribution grids,” *IEEE Transactions on Control of Network Systems*, vol. 6, no. 3, pp. 950–960, 2019.

[7] A. Ayad, H. E. Z. Farag, A. Youssef, and E. F. El-Saadany, “Detection of false data injection attacks in smart grids using recurrent neural networks,” in *2018 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, 2018, pp. 1–5. DOI: `10.1109/ISGT.2018.8403355`.

[8] M. E. Baran and F. F. Wu, “Network reconfiguration in distribution systems for loss reduction and load balancing,” *IEEE Power Engineering Review*, vol. 9, no. 4, pp. 101–102, 1989.

[9] A. Bhattacharjee, S. Mishra, and A. Verma, “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, 2022, pp. 1–7.

[10] S. Bhela, V. Kekatos, and S. Veeramachaneni, “Enhancing observability in distribution grids using smart meter data,” *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 5953–5961, 2017.

[11] N. Bhusal, R. M. Shukla, M. Gautam, M. Benidris, and S. Sengupta, "Deep ensemble learning-based approach to real-time power system state estimation," *International Journal of Electrical Power & Energy Systems*, vol. 129, p. 106 806, 2021.

[12] S. Bi and Y. J. Zhang, "Defending mechanisms against false-data injection attacks in the power system state estimation," in *2011 IEEE GLOBECOM Workshops (GC Wkshps)*, IEEE, 2011, pp. 1162–1167.

[13] S. Bi and Y. J. Zhang, "Graphical methods for defense against false-data injection attacks on power system state estimation," *IEEE Transactions on Smart Grid*, vol. 5, no. 3, pp. 1216–1227, 2014.

[14] A. S. Bretas, A. Rossoni, R. D. Trevizan, and N. G. Bretas, "Distribution networks nontechnical power loss estimation: A hybrid data-driven physics model-based framework," *Electric Power Systems Research*, vol. 186, p. 106 397, 2020.

[15] Z. Cao, Y. Wang, C.-C. Chu, and R. Gadh, "Scalable distribution systems state estimation using long short-term memory networks as surrogates," *IEEE Access*, vol. 8, pp. 23 359–23 368, 2020.

[16] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, pp. 3–14.

[17] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy*, IEEE, 2017, pp. 39–57.

[18] G. Chaojun, P. Jirutitijaroen, and M. Motani, "Detecting false data injection attacks in ac state estimation," *IEEE Transactions on Smart Grid*, vol. 6, no. 5, pp. 2476–2483, 2015.

[19] X. Chen, G. Qu, Y. Tang, S. Low, and N. Li, "Reinforcement learning for decision-making and control in power systems: Tutorial, review, and vision," *arXiv*, 2021.

[20] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv preprint arXiv:1712.05526*, 2017.

[21] Y. Chen, S. Huang, F. Liu, Z. Wang, and X. Sun, "Evaluation of reinforcement learning-based false data injection attack to automatic voltage control," *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 2158–2169, 2018.

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

[23] Y. Chen, Y. Tan, and B. Zhang, "Exploiting vulnerabilities of load forecasting through adversarial attacks," in *Proceedings of the tenth ACM international conference on future energy systems*, 2019, pp. 1–11.

[24] M. Cui, J. Wang, and M. Yue, "Machine learning-based anomaly detection for load forecasting under cyberattacks," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5724–5734, 2019.

[25] A. Datta, S. Sen, and Y. Zick, "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems," in *2016 IEEE symposium on security and privacy (SP)*, IEEE, 2016, pp. 598–617.

[26] T. Dayaratne, C. Rudolph, A. Liebman, and M. Salehi, "We can pay less: Coordinated false data injection attack against residential demand response in smart grids," in *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*, 2021, pp. 41–52.

[27] K. Dehghanpour, Z. Wang, J. Wang, Y. Yuan, and F. Bu, "A survey on state estimation techniques and challenges in smart distribution systems," *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 2312–2322, 2018.

[28] R. Deng, P. Zhuang, and H. Liang, "False data injection attacks against state estimation in power distribution systems," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2871–2881, 2018.

[29] R. C. Dugan and T. E. McDermott, "An open source platform for collaborating on smart grid research," in *2011 IEEE power and energy society general meeting*, IEEE, 2011, pp. 1–7.

[30] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou, "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)*, 2018, pp. 31–36.

[31] R. Eskandarpour and A. Khodaei, "Machine learning based power grid outage prediction in response to extreme events," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 3315–3316, 2016.

[32] M. Esmalifalak, L. Liu, N. Nguyen, R. Zheng, and Z. Han, "Detecting stealthy false data injection using machine learning in smart grid," *IEEE Systems Journal*, vol. 11, no. 3, pp. 1644–1652, 2014.

[33] A. Fawzi, O. Fawzi, and P. Frossard, "Analysis of classifiers' robustness to adversarial perturbations," *Machine learning*, vol. 107, no. 3, pp. 481–508, 2018.

[34] M. Fotopoulou, S. Petridis, I. Karachalios, and D. Rakopoulos, "A review on distribution system state estimation algorithms," *Applied Sciences*, vol. 12, no. 21, p. 11 073, 2022.

[35] A. Gensler, J. Henze, B. Sick, and N. Raabe, “Deep learning for solar power forecasting—an approach using autoencoder and lstm neural networks,” in *2016 IEEE international conference on systems, man, and cybernetics (SMC)*, IEEE, 2016, pp. 002 858–002 865.

[36] P. A. Giglou and S. N. Ravadanegh, “Defending against false data injection attack on demand response program: A bi-level strategy,” *Sustainable Energy, Grids and Networks*, vol. 27, p. 100 506, 2021.

[37] S. Gönen, H. H. Sayan, E. N. Yılmaz, F. Üstünsoy, and G. Karacayılmaz, “False data injection attacks and the insider threat in smart systems,” *Computers & Security*, vol. 97, p. 101 955, 2020.

[38] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.

[39] C. Guo, J. Gardner, Y. You, A. G. Wilson, and K. Weinberger, “Simple black-box adversarial attacks,” in *International Conference on Machine Learning*, PMLR, 2019, pp. 2484–2493.

[40] M. M. Haji and O. Ardakanian, “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, 2019, pp. 1–6.

[41] J. He and M. X. Cheng, “Machine learning methods for power line outage identification,” *The Electricity Journal*, vol. 34, no. 1, p. 106 885, 2021.

[42] Y. He, G. J. Mendis, and J. Wei, “Real-time detection of false data injection attacks in smart grid: A deep learning-based intelligent mechanism,” *IEEE Transactions on Smart Grid*, vol. 8, no. 5, pp. 2505–2516, 2017.

[43] R. Heinrich, C. Scholz, S. Vogt, and M. Lehna, “Targeted adversarial attacks on wind power forecasts,” *arXiv preprint arXiv:2303.16633*, 2023.

[44] E. Hossain, I. Khan, F. Un-Noor, S. S. Sikander, and M. S. H. Sunny, “Application of big data and machine learning in smart grid, and associated security concerns: A review,” *IEEE Access*, vol. 7, pp. 13 960–13 988, 2019.

[45] *Ieee pes distribution systems analysis subcommittee, radial test feeders.* [Online]. Available: `https://cmte.ieee.org/pes-testfeeders/resources/`.

[46] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, “Black-box adversarial attacks with limited queries and information,” in *International conference on machine learning*, PMLR, 2018, pp. 2137–2146.

[47] J. James, Y. Hou, and V. O. Li, “Online false data injection attack detection with wavelet transform and deep neural networks,” *IEEE Transactions on Industrial Informatics*, vol. 14, no. 7, pp. 3271–3280, 2018.

[48] L. Jiang, X. Ma, S. Chen, J. Bailey, and Y.-G. Jiang, “Black-box adversarial attacks on video recognition models,” in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 864–872.

[49] K. Khanna, S. K. Singh, B. K. Panigrahi, R. Bose, and A. Joshi, “On detecting false data injection with limited network information using transformation based statistical techniques,” in *2017 IEEE Power & Energy Society General Meeting*, IEEE, 2017, pp. 1–5.

[50] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, “Limiting false data attacks on power system state estimation,” in *2010 44th Annual Conference on Information Sciences and Systems (CISS)*, IEEE, 2010, pp. 1–6.

[51] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, “Malicious data attacks on the smart grid,” *IEEE Transactions on Smart Grid*, vol. 2, no. 4, pp. 645–658, 2011.

[52] K. Kuntz, M. Smith, K. Wedeward, and M. Collins, “Detecting, locating, & quantifying false data injections utilizing grid topology through optimized d-facts device placement,” in *2014 North American Power Symposium (NAPS)*, IEEE, 2014, pp. 1–6.

[53] A. Kurakin, I. Goodfellow, S. Bengio, *et al.*, *Adversarial examples in the physical world*, 2016.

[54] A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial machine learning at scale,” *arXiv preprint arXiv:1611.01236*, 2016.

[55] B. Li, T. Ding, C. Huang, J. Zhao, Y. Yang, and Y. Chen, “Detecting false data injection attacks against power system state estimation with fast go-decomposition approach,” *IEEE Transactions on Industrial Informatics*, vol. 15, no. 5, pp. 2892–2904, 2019. DOI: 10.1109/TII.2018.2875529.

[56] J. Li, W. Monroe, and D. Jurafsky, “Understanding neural networks through representation erasure,” *arXiv preprint arXiv:1612.08220*, 2016.

[57] M. Li, C. Deng, T. Li, J. Yan, X. Gao, and H. Huang, “Towards transferable targeted attack,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 641–649.

[58] Y. Li, Y. Wang, and S. Hu, “Online generative adversary network based measurement recovery in false data injection attacks: A cyber-physical approach,” *IEEE Transactions on Industrial Informatics*, vol. 16, no. 3, pp. 2031–2043, 2019.

[59] T. Liu and T. Shu, “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, 2019, pp. 365–379.

[60] T. Liu and T. Shu, “On the security of ANN-based ac state estimation in smart grid,” *Computers & Security*, vol. 105, p. 102 265, 2021.

[61] Y. Liu, P. Ning, and M. K. Reiter, “False data injection attacks against state estimation in electric power grids,” *ACM Transactions on Information and System Security (TISSEC)*, vol. 14, no. 1, pp. 1–33, 2011.

[62] Y. Liu, O. Ardakanian, I. Nikolaidis, and H. Liang, “False data injection attacks on smart grid voltage regulation with stochastic communication model,” *IEEE Transactions on Industrial Informatics*, pp. 1–11, 2022.

[63] C. Lu, J. Teng, and W.-H. Liu, “Distribution system state estimation,” *IEEE Transactions on Power systems*, vol. 10, no. 1, pp. 229–240, 1995.

[64] M. Lu, L. Wang, Z. Cao, Y. Zhao, and X. Sui, “False data injection attacks detection on power systems with convolutional neural network,” in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1633, 2020, p. 012 134.

[65] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” *arXiv preprint arXiv:1706.06083*, 2017.

[66] A. Majumdar, Y. P. Agalgaonkar, B. C. Pal, and R. Gottschalg, “Centralized Volt–Var optimization strategy considering malicious attack on distributed energy resources control,” *IEEE Transactions on Sustainable Energy*, vol. 9, no. 1, pp. 148–156, 2017.

[67] E. Manitsas, R. Singh, B. C. Pal, and G. Strbac, “Distribution system state estimation using an artificial neural network approach for pseudo measurement modeling,” *IEEE Transactions on Power Systems*, vol. 27, no. 4, pp. 1888–1896, 2012.

[68] C. J. Meinrenken, N. Rauschkolb, S. Abrol, *et al.*, *MFRED (public file, 15/15 aggregate version): 10 second interval real and reactive power in 390 US apartments of varying size and vintage*, version V1, 2020. DOI: 10.7910/DVN/X9MIDJ. [Online]. Available: https://doi.org/10.7910/DVN/X9MIDJ.

[69] L. Meng, C.-T. Lin, T.-P. Jung, and D. Wu, “White-box target attack for eeg-based bci regression problems,” in *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part I 26*, Springer, 2019, pp. 476–488.

[70] K. R. Mestav, J. Luengo-Rozas, and L. Tong, "Bayesian state estimation for unobservable distribution systems via deep learning," *IEEE Transactions on Power Systems*, vol. 34, no. 6, pp. 4910–4920, 2019.

[71] G. R. Mode and K. A. Hoque, "Adversarial examples in deep learning for multivariate time series regression," in *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, IEEE, 2020, pp. 1–10.

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

[73] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2574–2582.

[74] A. Moradzadeh, M. Mohammadpourfard, C. Konstantinou, I. Genc, T. Kim, and B. Mohammadi-Ivatloo, "Electric load forecasting under false data injection attacks using deep learning," *Energy Reports*, vol. 8, pp. 9933–9945, 2022.

[75] R. Moslemi, A. Mesbahi, and J. M. Velni, "A fast, decentralized covariance selection-based approach to detect cyber attacks in smart grids," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 4930–4941, 2017.

[76] L. Muñoz-González, B. Biggio, A. Demontis, *et al.*, "Towards poisoning of deep learning algorithms with back-gradient optimization," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, pp. 27–38.

[77] L. Muñoz-González, B. Pfitzner, M. Russo, J. Carnerero-Cano, and E. C. Lupu, "Poisoning attacks with generative adversarial nets," *arXiv preprint arXiv:1906.07773*, 2019.

[78] M. Naseer, S. Khan, M. Hayat, F. S. Khan, and F. Porikli, "On generating transferable targeted perturbations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7708–7717.

[79] *NERC final blackout report.* [Online]. Available: `https://www.nerc.com/pa/rrm/ea/August%2014%202003%20Blackout%20Investigation%20DL/NERC_Final_Blackout_Report_07_13_04.pdf`.

[80] C. Owens, *Biggest Blackouts In History: Northeastern USA & Canada 2003.* [Online]. Available: `https://www.theblackoutreport.co.uk/2019/10/10/northeastern-blackout-usa-canada-2003/`.

[81] M. Ozay, I. Esnaola, F. T. Y. Vural, S. R. Kulkarni, and H. V. Poor, "Machine learning methods for attack detection in the smart grid," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 8, pp. 1773–1786, 2015.

[82] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: From phenomena to black-box attacks using adversarial samples," *arXiv preprint arXiv:1605.07277*, 2016.

[83] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017, pp. 506–519.

[84] A. Primadianto and C.-N. Lu, "A review on distribution system state estimation," *IEEE Transactions on Power Systems*, vol. 32, no. 5, pp. 3875–3883, 2016.

[85] Y. Raghuvamsi and K. Teeparthi, "Detection and reconstruction of measurements against false data injection and dos attacks in distribution system state estimation: A deep learning approach," *Measurement*, vol. 210, p. 112 565, 2023.

[86] S. Rahimi, M. Marinelli, and F. Silvestro, "Evaluation of requirements for volt/var control and optimization function in distribution management systems," in *2012 IEEE International Energy Conference and Exhibition (ENERGYCON)*, IEEE, 2012, pp. 331–336.

[87] M. A. Rahman and H. Mohsenian-Rad, "False data injection attacks against nonlinear state estimation in smart power grids," in *2013 IEEE Power & Energy Society General Meeting*, IEEE, 2013, pp. 1–5.

[88] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[89] J. Sakhnini, H. Karimipour, and A. Dehghantanha, "Smart grid cyber attacks detection using supervised learning and heuristic feature selection," in *2019 IEEE 7th international conference on smart energy grid engineering (SEGE)*, IEEE, 2019, pp. 108–112.

[90] A. Sayghe, Y. Hu, I. Zografopoulos, *et al.*, "Survey of machine learning methods for detecting false data injection attacks in power systems," *IET Smart Grid*, vol. 3, no. 5, pp. 581–595, 2020.

[91] A. Sayghe, J. Zhao, and C. Konstantinou, "Evasion attacks with adversarial deep learning against power system state estimation," in *2020 IEEE Power & Energy Society General Meeting (PESGM)*, IEEE, 2020, pp. 1–5.

[92] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International conference on machine learning*, PMLR, 2017, pp. 3145–3153.

[93] R. Shu, T. Xia, L. Williams, and T. Menzies, "Omni: Automated ensemble with unexpected models against adversarial evasion attack," *Empirical Software Engineering*, vol. 27, pp. 1–32, 2022.

[94] *Smart distribution systems for a low carbon energy future workshop, cired workshop, 6 june 2011, frankfurt, germany.* [Online]. Available: `https://grouper.ieee.org/groups/td/dist/da/doc/2011%20CIRED%20Panel%20Tutorial%20binder_AH.pdf`.

[95] M. Starke, D. Rizy, and M. Young, "Synchrophasor technologies and their deployment in the recovery act smart grid programs," *Report US Department of Energy*, 2013.

[96] A. Tabakhpour and M. M. Abdelaziz, "Neural network model for false data detection in power system state estimation," in *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, IEEE, 2019, pp. 1–5.

[97] T. Tanay and L. Griffin, "A boundary tilting persepective on the phenomenon of adversarial examples," *arXiv preprint arXiv:1608.07690*, 2016.

[98] *The aggregated challenges of regulating energy usage data.* [Online]. Available: `https://eq-research.com/blog/the-aggregated-challenges-of-regulating-energy-usage-data/`.

[99] J. Tian, B. Wang, J. Li, and C. Konstantinou, "Adversarial attack and defense methods for neural network based state estimation in smart grid," *IET Renewable Power Generation*, vol. 16, no. 16, pp. 3507–3518, 2022.

[100] J. Tian, B. Wang, J. Li, and C. Konstantinou, "Datadriven false data injection attacks against cyber-physical power systems," *Computers & Security*, vol. 121, p. 102 836, 2022.

[101] J. Tian, B. Wang, J. Li, Z. Wang, B. Ma, and M. Ozay, "Exploring targeted and stealthy false data injection attacks via adversarial machine learning," *IEEE Internet of Things Journal*, vol. 9, no. 15, pp. 14 116–14 125, 2022.

[102] J. Tian, B. Wang, Z. Wang, K. Cao, J. Li, and M. Ozay, "Joint adversarial example and false data injection attacks for state estimation in power systems," *IEEE Transactions on Cybernetics*, vol. 52, no. 12, pp. 13 699–13 713, 2021.

[103] F. Tramer, "Detecting adversarial examples is (nearly) as hard as classifying them," in *International Conference on Machine Learning*, PMLR, 2022, pp. 21 692–21 702.

[104] F. Tramr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *International Conference on Learning Representations*, vol. 1, 2018, p. 2.

[105] C. Tu, X. He, X. Liu, and P. Li, "Cyber-attacks in pmu-based power network and countermeasures," *IEEE Access*, vol. 6, pp. 65 594–65 603, 2018.

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

[107] *Voltage tolerance boundary.* [Online]. Available: `https://www.pge.com/includes/docs/pdfs/mybusiness/customerservice/energystatus/powerquality/voltage_tolerance.pdf`.

[108] O. Vuković and G. Dán, "On the security of distributed power system state estimation under targeted attacks," in *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, 2013, pp. 666–672.

[109] J. Wang, D. Shi, Y. Li, J. Chen, H. Ding, and X. Duan, "Distributed framework for detecting pmu data manipulation attacks with deep autoencoders," *IEEE Transactions on smart grid*, vol. 10, no. 4, pp. 4401–4410, 2018.

[110] S. Wang, S. Bi, and Y.-J. A. Zhang, "Locational detection of the false data injection attack in a smart grid: A multilabel classification approach," *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8218–8227, 2020.

[111] X. Wang, D. Shi, J. Wang, Z. Yu, and Z. Wang, "Online identification and data recovery for pmu data manipulation attack," *IEEE Transactions on Smart Grid*, vol. 10, no. 6, pp. 5889–5898, 2019.

[112] Y. Wang, D. Chen, C. Zhang, X. Chen, B. Huang, and X. Cheng, "Wide and recurrent neural networks for detection of false data injection in smart grids," in *Wireless Algorithms, Systems, and Applications: 14th International Conference, WASA 2019, Honolulu, HI, USA, June 24–26, 2019, Proceedings 14*, Springer, 2019, pp. 335–345.

[113] Y. Wang, Z. Xu, J. Zhang, L. Xu, H. Wang, and G. Gu, "Srid: State relation based intrusion detection for false data injection attacks in scada," in *Computer Security-ESORICS 2014: 19th European Symposium on Research in Computer Security, Wroclaw, Poland, September 7-11, 2014. Proceedings, Part II 19*, Springer, 2014, pp. 401–418.

[114] Y. Wang, W. Shi, Q. Jin, and J. Ma, "An accurate false data detection in smart grid based on residual recurrent neural network and adaptive threshold," in *2019 IEEE International Conference on Energy Internet (ICEI)*, IEEE, 2019, pp. 499–504.

[115] Y. Weng, R. Negi, C. Faloutsos, and M. D. Ilić, "Robust data-driven state estimation for smart grid," *IEEE Transactions on Smart Grid*, vol. 8, no. 4, pp. 1956–1967, 2016.

[116] D. E. Whitehead, K. Owens, D. Gammel, and J. Smith, "Ukraine cyber-induced power outage: Analysis and practical mitigation strategies," in *2017 70th Annual Conference for Protective Relay Engineers (CPRE)*, IEEE, 2017, pp. 1–8.

[117] J. Yan, B. Tang, and H. He, "Detection of false data attacks in smart grid with supervised learning," in *2016 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2016, pp. 1395–1402.

[118] J. Yan, B. Tang, and H. He, "Detection of false data attacks in smart grid with supervised learning," in *2016 International Joint Conference on Neural Networks (IJCNN)*, 2016, pp. 1395–1402. DOI: `10.1109/IJCNN.2016.7727361`.

[119] W. Yan *et al.*, "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, 2021, pp. 108–114.

[120] C. Yang, Q. Wu, H. Li, and Y. Chen, "Generative poisoning attack method against neural networks," *arXiv preprint arXiv:1703.01340*, 2017.

[121] H. Yang, X. Liu, D. Zhang, T. Chen, C. Li, and W. Huang, "Machine learning for power system protection and control," *The Electricity Journal*, vol. 34, no. 1, p. 106 881, 2021.

[122] P. Yang, J. Chen, C.-J. Hsieh, J.-L. Wang, and M. Jordan, "Ml-loo: Detecting adversarial examples with feature attribution," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 6639–6647.

[123] 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 Transactions on Parallel and Distributed Systems*, vol. 25, no. 3, pp. 717–729, 2013.

[124] B. Yildiz, J. I. Bilbao, and A. B. Sproul, "A review and analysis of regression and machine learning models on commercial building electricity load forecasting," *Renewable and Sustainable Energy Reviews*, vol. 73, pp. 1104–1122, 2017.

[125] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Transactions on neural networks and learning systems*, vol. 30, no. 9, pp. 2805–2824, 2019.

[126] A. S. Zamzam, X. Fu, and N. D. Sidiropoulos, "Data-driven learning-based optimization for distribution system state estimation," *IEEE Transactions on Power Systems*, vol. 34, no. 6, pp. 4796–4805, 2019.

[127] A. S. Zamzam and N. D. Sidiropoulos, "Physics-aware neural networks for distribution system state estimation," *IEEE Transactions on Power Systems*, vol. 35, no. 6, pp. 4347–4356, 2020. DOI: 10.1109/TPWRS.2020.2988352.

[128] L. Zhang, G. Wang, and G. B. Giannakis, "Power system state forecasting via deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 8092–8096.

[129] L. Zhang, G. Wang, and G. B. Giannakis, "Real-time power system state estimation and forecasting via deep unrolled neural networks," *IEEE Transactions on Signal Processing*, vol. 67, no. 15, pp. 4069–4077, 2019.

[130] Y. Zhang, A. Bernstein, A. Schmitt, and R. Yang, "State estimation in low-observable distribution systems using matrix completion," National Renewable Energy Lab.(NREL), Golden, CO (United States), Tech. Rep., 2019.

[131] J. Zhao, G. Zhang, and R. A. Jabr, "Robust detection of cyber attacks on state estimators using phasor measurements," *IEEE Transactions on Power Systems*, vol. 32, no. 3, pp. 2468–2470, 2016.

[132] Z. Zhao, Z. Liu, and M. Larson, "On success and simplicity: A second look at transferable targeted attacks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 6115–6128, 2021.

[133] M. Zhou, Y. Wang, A. K. Srivastava, Y. Wu, and P. Banerjee, "Ensemble-based algorithm for synchrophasor data anomaly detection," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2979–2988, 2018.

[134] M. Zhou and V. M. Patel, "On trace of pgd-like adversarial attacks," *arXiv preprint arXiv:2205.09586*, 2022.

[135] P. Zhuang, R. Deng, and H. Liang, "False data injection attacks against state estimation in multiphase and unbalanced smart distribution systems," *IEEE Transactions on Smart Grid*, vol. 10, no. 6, pp. 6000–6013, 2019.