On Brittleness of Data-Driven Distribution System State Estimation to Targeted Attacks

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

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

ABSTRACT

State estimation techniques that utilize machine learning are gaining popularity in power distribution networks with high penetration of distributed energy resources due to their higher accuracy, faster convergence, and computational efficiency. However, little attention has been paid to their security and robustness, especially to targeted false data injection and evasion attacks. This note aims to investigate if the direction and magnitude of change in the state estimation result can be simultaneously controlled by the attacker, and the kind of access required to perform successful targeted attacks on data-driven state estimation approaches.

CCS CONCEPTS

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

ACM Reference Format:
Afia Afrin and Omid Ardakanian. 2024. On Brittleness of Data-Driven Distribution System State Estimation to Targeted Attacks. In *The 15th ACM International Conference on Future and Sustainable Energy Systems (E-Energy '24), June 4–7, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

Data-driven approaches are considered better alternatives to conventional power system state estimation techniques in terms of accuracy, convergence rate, and computational complexity [6, 31, 33, 34]. However, the incorporation of neural networks into the power system operation necessitates a comprehensive evaluation beyond mere performance metrics. In particular, ensuring overall system security is a key concern in safety-critical applications, such as real-time monitoring and control of power systems. In previous work, electrical model-based state estimation approaches, such as the weighted least squares (WLS) method, have been found vulnerable to various kinds of false data injection attack (FDIA) [8, 17, 22, 37]. In seminal work by Liu et al. [17], it was shown that the attacker with knowledge of the power system structure and configuration can determine the amount of false data that must be introduced to move the state estimates produced by electrical model-based state estimation approaches by a desired amount in a specific direction. With the growing popularity of data-driven distribution system state estimation (DSSE) approaches, it is imperative to understand *if the known vulnerabilities of electrical model-based DSSE approaches extend to the data-driven DSSE approaches, and whether new vulnerabilities may arise that are specific to the data-driven approaches*?

Some efforts have been made in recent years to answer these questions. For example, state estimators that incorporate a neural network were found vulnerable to various FDIAs [15, 16]. Recent work shows that the inherent vulnerability of neural networks to the *evasion attack* – a type of adversarial attack where the neural network input is manipulated once the trained model is deployed – would pose a threat to the power system operation should data-driven DSSE approaches be adopted [2]. Nevertheless, there is no known work that studies how precisely the attacker can control the direction and/or amount of error introduced in the estimated state under different threat models.

In this work, we investigate the following research questions:

- RQ1 (Vulnerability to FDIA): What is the impact of the traditional FDIA [17] on data-driven DSSE?
- RQ2 (Vulnerability to Targeted Adversarial Perturbations): Is it possible to generate a targeted adversarial attack against data-driven DSSE where the attacker can control either the direction or the amount of error being injected?
- RQ3 (Privileges Required for Successful Targeted Attack): What kind of access to distribution network structure and parameters, neural network parameters, and sensor data would be required for the attacker to successfully launch powerful and targeted attacks against data-driven DSSE?

Investigating RQ2 leads us to design a novel targeted evasion attack based on the Fast Gradient Sign Method (FGSM) [10] because, to our knowledge, no targeted evasion attack on data-driven DSSE has been proposed in the literature. This is another contribution of this work besides analyzing vulnerabilities of data-driven DSSE.

2 BACKGROUND AND RELATED WORK

Stealthy FDIAs on WLS-based state estimation were originally introduced in [17]. More recently, the vulnerability of data-driven and electrical model-based distribution system state estimators to various FDIAs has been studied [8, 16, 24, 25, 37]. Among these attack strategies, the FDIA proposed in [17] and some of its extensions [8, 26, 37] are considered targeted attacks as they can control the amount and direction of change in the state estimation result. Yet, none of these attack strategies has been tested against data-driven DSSE. This motivates us to evaluate the efficacy of targeted FDIA on data-driven DSSE approaches.

In another line of work, machine learning has been employed to generate untargeted adversarial attacks against state estimation techniques [2, 5]. Specifically, deep adversarial networks have been used for the first time in [5] to craft a stealthy *black-box*[1] adversarial

E-Energy '24, June 4–7, 2024, Singapore, Singapore
2024.

[1] In black-box attacks, the attacker trains an arbitrary surrogate model on the input and output of the victim model that performs state estimation using its *query access* to that model (i.e. the ability to run the model to record its output for specific input), and

Table 1: Our contribution with respect to previous work.

DSSE Approach	Reference	Attack Strategy		Attack Nature	
		FDIA	Adversarial	Untargeted	Targeted
Electrical Model Based	[8, 17, 37]	•		•	•
	[25]	•		•	
	[26]	•			•
	[27]	•	•	•	
	[5]		•	•	
Data Driven	[15, 16]	•		•	
	[2, 24]		•	•	
	Our work	•	•		•

attack. The authors used FGSM to construct attack vectors against a WLS-based state estimation technique. The vulnerability of a data-driven DSSE approach has been analyzed in [2] and a stealthier version of the conventional FGSM-based attack has been proposed and shown to be capable of bypassing the residual-based bad data detection (BDD) mechanism more often than the conventional FGSM. Nevertheless, all these attacks are untargeted, meaning that they move the estimated state in an arbitrary direction.

Targeted evasion attacks have been designed to fool classification models by exploiting the geometry of the decision boundary [7, 13]. However, limited exploration has been conducted to date to extend these attacks to a regression model [12, 19]. Specifically, there is no known work that focuses on generating targeted evasion attacks against data-driven DSSE, which is a multivariate regression task. In this work, we propose a targeted evasion attack capable of moving the estimated state of the distribution system in a specific direction. Table 1 shows our contribution with respect to the previous work, especially the studies that focus on data-driven DSSE approaches.

3 TARGETED ATTACK STRATEGIES

To explore RQ1, we launch the targeted FDIA proposed in [17] on a data-driven DSSE technique that takes the vector of real-time measurements at a given time t, denoted as $\mathbf{z_t}$, and estimates the system state at that time, denoted as $\mathbf{x_t}$. Suppose the state variables are the bus voltage phasors denoted as $\mathbf{x} = [\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_b}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_b]$, with b being the number of buses not equipped with a sensor, and $\mathbf{v_i}$ and $\boldsymbol{\theta}_i$ representing the vectors that contain the three-phase voltage magnitudes and phase angles of bus i, respectively. As discussed in [17], to produce an erroneous state, $\mathbf{x'_t} = \mathbf{x_t} + \mathbf{c}$, the attacker needs to specify the error vector, $\mathbf{c}$, and then compute the attack vector as: $\mathbf{a} = \mathbf{Hc}$. Here, $\mathbf{H}$ is the measurement matrix, used in electrical model-based state estimation approaches as well as the conventional residual-based BDD mechanism [3]. This strategy ensures that the attack vector will bypass the residual-based BDD mechanism [17] and allows the attacker to judiciously choose the attack vector to achieve the desired goal. Since this attack forms the basis for other targeted FDIAs (e.g. [8, 26, 37]), we consider it to assess the vulnerability of data-driven DSSE to targeted FDIA.

To address RQ2, we design a targeted evasion attack based on FGSM. The attacker's goal is to construct an adversarial data sample, $\mathbf{z'_t}$, from a clean data sample, $(\mathbf{z_t})$. For a surrogate model used at the attacker's end, $g(\mathbf{z_t}; \theta')$, we define the adversarial loss function as $L\left(g(\mathbf{z_t}; \theta'), \mathbf{x'_t}\right)$ which is the mean squared error (MSE) between the model output, $g(\mathbf{z_t}; \theta')$, and the target state vector $\mathbf{x'_t}$. Here, $\mathbf{x'_t}$ is a vector of size n that has the same phase angle values as the predicted state vector of the surrogate model but the voltage values are replaced by some randomly chosen values from the range (v_{min}, v_{max}), which must be defined by the attacker according to their objective. For example, an attacker aiming to cause overestimation should use a higher target range, e.g. around 1.05pu. Conversely, an attacker aiming to cause under-estimation should use a lower target range, e.g. 0.95pu. It is important to note that instead of choosing a fixed target value, we pick a value from the predefined range uniformly at random to introduce some nondeterministic behavior into the algorithm.

We construct the adversarial sample, $\mathbf{z'_t}$, by moving the original data sample in the opposite direction of the gradient of the adversarial loss function, $L\left(g(\mathbf{z_t}; \theta'), \mathbf{x'_t}\right)$, taken with respect to the input data. Moving the data sample in that direction, which is found using the sign of the gradient, will minimize the loss and eventually move the DSSE result closer to the target state, $\mathbf{x'_t}$. Notice the difference between training the model for data-driven state estimation and crafting adversarial data given the fully trained state estimation model. In the former, we calculate the gradient of the loss with respect to the model parameters, θ, whereas in the latter, we calculate the gradient with respect to the input data, $\mathbf{z_t}$. Algorithm 1 describes the proposed Targeted-FGSM attack. This algorithm can be used in white-box and black-box settings where the difference lies in whether the surrogate model g is identical to the victim DSSE model f, or it is an arbitrary neural network that has sufficient learning capacity. In the black-box setting, the attacker collects a set of measurement-state pairs, i.e. $(\mathbf{z_t}, \mathbf{x_t})$, to train a surrogate model g that mimics the victim DSSE model f. Moreover, the query access to the victim model can be utilized to get the estimated state, $\mathbf{x_t}$, in line 3 of Algorithm 1.

We note that, with some effort, directed perturbations can be calculated using other adversarial attack strategies such as basic iterative method (BIM) and projected gradient descent (PGD) [21]. Yet, we build our targeted evasion attack on FGSM which is easier

Algorithm 1 Targeted-FGSM Attack

1: **Inputs:**
   Surrogate model, $g(\cdot; \theta')$
   Original data sample at timestamp t, $\mathbf{z_t}$
   Predefined voltage range, (v_{min}, v_{max})
2: **Output:**
   Adversarial data sample at timestamp t, $\mathbf{z'_t}$
   ▷ *Initialize the target state vector*
3: $\mathbf{x_t} \leftarrow f(\mathbf{z_t}; \theta)$
4: $\mathbf{x'_t} \leftarrow \mathbf{x_t}$
   ▷ *Let I be the vector denoting indices of voltage magnitudes in x*
5: **for** i in I **do**
6: $\quad \mathbf{x'_t}[i] \leftarrow \text{random}(v_{min}, v_{max})$
7: **end for**
8: $\boldsymbol{\delta}_{\mathbf{z_t}} \leftarrow \nabla_{\mathbf{z_t}} \left[L\left(g(\mathbf{z_t}; \theta'), \mathbf{x'_t}\right)\right]$
9: $\mathbf{z'_t} \leftarrow \mathbf{z_t} - \epsilon \cdot \text{sign}(\boldsymbol{\delta}_{\mathbf{z_t}})$ ▷ *ϵ is a scalar hyperparameter*

uses the surrogate model to craft adversarial data. However, in white-box attacks, the attacker uses the same model as the victim model to craft adversarial data.

to implement and serves as the foundation of PGD and BIM [36]. This is sufficient to address RQ2 as discussed later.

4 EXPERIMENTAL SETUP

Data-Driven DSSE Approaches. We use the *Stacked ResNetD* model described in [2] as our data-driven DSSE approach (i.e. the victim model). This ensemble of deep residual neural networks has been shown to achieve better performance in DSSE than several other deep learning models [6]. The choice is also inspired by the finding of [23, 28] that ensemble models generally have better adversarial robustness. In our implementation of the white-box attacks, we use the Stacked ResNetD model as the surrogate model trained on data-state pairs by the attacker. For black-box attacks, we utilize the 8-layer convolutional neural network (CNN), consisting of three convolution, two pooling, and three dense layers with ReLU activation function, proposed in [6] as the surrogate.

Test Distribution System. Following [2], we use the 33-bus system [4] and the IEEE European low voltage test feeder [1] as the primary and secondary distribution networks respectively. To represent the system loads, we use the Multifamily Residential Electricity Dataset (MFRED) [18], which contains daily load profiles of 390 US apartments with 15 minutes resolution over 12 months. Once the system is built, we run power flow analysis in OpenDSS [9] to generate training and test datasets for the data-driven DSSE approach. Note that the training dataset can be generated in a similar fashion in the real world, i.e., by solving the power flow equations to obtain the system states using historical load and generation data [32].

Data Preparation & Simulation. We assume that all load buses within the secondary distribution network are equipped with smart meters that have 15-minute resolution. Aggregating these data from all load buses yields the real and reactive power consumption at the primary bus, which will be treated as pseudo-measurements. Moreover, six primary buses are considered to be equipped with distribution level phasor measurement units (D-PMUs).[2] Thus, the measurement vector, $\mathbf{z_t}$, contains three-phase real and reactive power consumption at each of the primary load buses, and three-phase voltage magnitudes of the six buses. We have 27 buses that are not monitored by D-PMUs. The three-phase voltage magnitudes and phase angles of these buses comprise the system state.

We use the OpenDSS simulation results for the first half of every month to train the victim model. With a dataset resolution of 15 minutes, we amass a total of 17, 280 training samples. For the test dataset, we randomly select load data from three consecutive days of each month and obtain corresponding OpenDSS simulation results. Consequently, we generate 3, 456 instances of test samples, organized into 12 groups of 288 consecutive measurements. These groups are evenly distributed throughout the year, covering three consecutive days from each month, with 96 samples per day. The remaining samples, comprising 12 days in the latter half of every month, are employed to train the CNN surrogate model for the black-box version of the proposed Targeted-FGSM attack.

[2]We install six D-PMUs since this level of observability led to reasonable state estimation performance in [11]. Note that determining the optimal placement of measurement devices, such as D-PMUs, is not within the scope of our study, so we just adopted a reasonable sensor placement strategy.

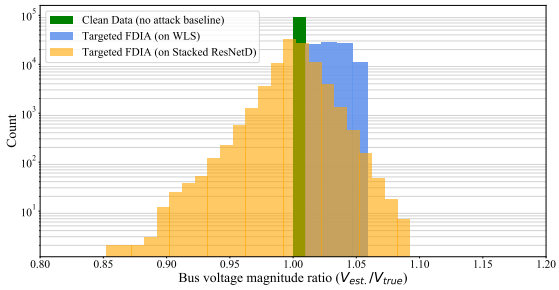

Figure 1: Distribution of bus voltage magnitude ratio over all unobserved buses under targeted FDIA attack on two DSSE models. Note the y-axis is in logarithmic scale.

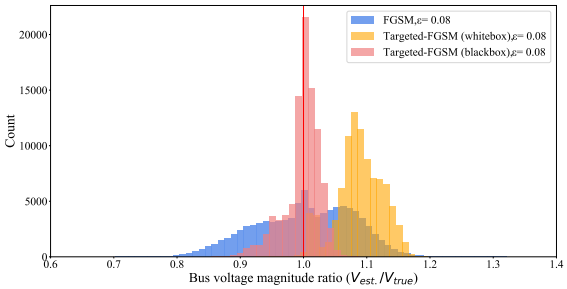

Figure 2: Distribution of bus voltage magnitude ratio over all unobserved buses under FGSM and Targeted-FGSM attack (both white-box and black-box versions). An ideal estimation should reside near the vertical line drawn at 1.0.

5 VULNERABILITY TO TARGETED ATTACKS

Without loss of generality, we consider an adversary who intends to push the voltage levels in the state estimation vector in an upward direction. Unless otherwise stated, we set the hyper-parameters of the Targeted-FGSM attack as follows: v_{min}=1.03, v_{max}=1.04, ϵ=0.08.

Figure 1 shows the experimental results for RQ1 in terms of the distribution of the ratio of the estimated bus voltage magnitude (either under normal conditions or in the presence of an attacker) to its true voltage magnitude, for all the unobserved buses.

As found in [17], targeted FDIA pushes the estimation of the WLS-based DSSE model in the desired direction and by the specific amount defined in the error vector, $\mathbf{c}$. However, it does not work like a targeted attack on the Stacked ResNetD model. It can be seen that it moves the state estimate in both directions and the resulting state $\mathbf{x}'$ has arbitrary errors injected instead of the specified error vector, $\mathbf{c}$. This is expected because the Stacked ResNetD model has no information regarding the $\mathbf{H}$ matrix and transforms the input measurement into a state vector using some non-linear transformation that takes place inside the complex neural network architecture. Thus, the traditional targeted FDIA that works against electrical model-based DSSE approaches does not retain its targeted property when used against the Stacked ResNetD model.

Next we evaluate the performance of the proposed Targeted-FGSM attack (Algorithm 1) on the Stacked ResNetD model. Figure 2 compares the white-box FGSM from [2] and the targeted evasion attack (both white-box and black-box versions) that we proposed in this work. As we observe, FGSM, due to its untargeted nature,

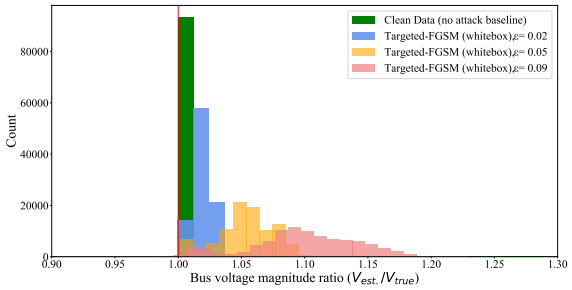

Figure 3: Distribution of bus voltage magnitude ratio over all unobserved buses under Targeted-FGSM attack crafted with different ϵ values (v_{min} and v_{max} are kept fixed at 1.03 and 1.04, respectively).

moves the estimates in both directions while white-box Targeted-FGSM attack maintains a majority of cases with a ratio greater than 1, pushing the estimations in an upward direction. However, similar to FGSM and targeted FDIA, the black-box Targeted-FGSM attack moves the estimated states in both directions, indicating its failure in fulfilling the adversarial objective. It is important to note that by adjusting the perturbation factor hyperparameter (ϵ) of the Targeted-FGSM attack one could further skew the distribution to the right at the cost of increasing the chance of the attack being detected. We discuss more about this in the following section.

6 DISCUSSION

6.1 Required Resources and Privileges

We now address RQ3. As shown in Section 5, the proposed Targeted-FGSM attack is effective only in white-box setting, where the attacker has (a) complete knowledge of the data-driven DSSE model, $f(\cdot;\theta)$, referred to as the *victim model*, and (b) read and write access to the sensor measurements, $\mathbf{z}$. Due to the first assumption, the *surrogate model* used by the attacker can be identical to the victim model, making it a white-box attack. It is worth noting that, in the machine learning domain, the lack of transferability is a well-known shortcoming of the targeted evasion attacks [14].

The primary attack point is the utility data center where field measurements are stored, and the data-driven DSSE model is stored and executed eventually. The attacker can be a malicious system operator, or an outsider gaining unauthorized access to the server using compromised software or via a compromised user account. PMU networks and utility data centers have been found vulnerable to such attacks in recent studies [29, 30], lending credence to this threat model. That said, one could argue that the requirement for white-box access is too strict, reducing the chances of this targeted evasion attack taking place in real life.

6.2 Hyperparameter Sensitivity of the Targeted Evasion Attack

Algorithm 1 contains three hyperparameters, namely v_{min}, v_{max}, and ϵ. We tested the proposed Targeted-FGSM attack with various combinations of these hyperparameters to understand their importance for generating a successful targeted attack. Our analysis reveals an interesting observation. While testing different hyperparameter combinations, we found that the impact of ϵ supersedes that of the other two hyperparameters, v_{min} and v_{max}, which define the target range. Typically, distribution systems are operated in a way that bus voltages are maintained in a specific range (for example, between 0.95 and 1.05 p.u). From the standpoint of the attacker, opting for a target range close to the upper (lower) pre-defined value is sufficient to trigger targeted behavior of the algorithm, resulting in overestimation (underestimation). However, the effectiveness of the attack, which is measured as how much it can push the estimated states in a certain direction, depends on ϵ primarily. Figure 3 shows the sensitivity to ϵ. Observe that v_{min} and v_{max} are kept fixed at 1.03 and 1.04 – yielding quite a small range! In spite of that, we are able to push the estimated states to around 1.15p.u. just by tweaking ϵ. This implies that while the target range, $[v_{min}, v_{max}]$, defines the attack *direction*, ϵ determines the size of the *leap* that we are taking in that particular direction. By choosing higher ϵ values, it is possible to generate attack vectors that are more effective indeed. However, this comes at the cost of an increased chance of getting detected by the DSSE safeguard mechanism. We defer designing a hyperparameter selection strategy for a known safeguard mechanism to future work and just note that striking a good trade-off between effectiveness and stealthiness of the attack would be of utmost importance to the attacker.

7 CONCLUSION AND FUTURE WORK

We explored the vulnerability of a data-driven DSSE approach to the traditional targeted FDIA that was found successful in deceiving the electrical model-based DSSE in a predicted way. Upon observing the failure of this targeted FDIA against the data-driven DSSE, we designed a targeted adversarial attack, namely Targeted-FGSM, to ensure over-estimation of the system states. In practice, such an attack can cause consistent power quality (under-voltage) issues in the distribution network.

Our proposed Targeted-FGSM attack approach, despite being quite successful in achieving its goal, suffers from several key limitations. We believe these limitations illuminate the path toward further research. The limitations along with future research directions are listed below:

- There are a number of recent studies that aim to develop targeted adversarial attack strategies against image classifiers that are *transferable* across different victim models [14, 20, 35]. Extending these attack strategies to multi-variate regression problems can help us understand the impact of these attack strategies on various data-driven smart grid applications including DSSE.
- Evasion attacks are significantly different from the conventional FDIAs not only in how they create the attack vector but also in the way they affect the overall system. This calls for analyzing the effectiveness of conventional DSSE safeguard mechanisms (such as BDD) against these attacks and designing better safeguard mechanisms.
- As discussed in Section 6.2, the effectiveness and stealthiness of the proposed targeted evasion attack strategy highly depend on the perturbation factor, ϵ. Automatically tuning this hyper-parameter to get an acceptable balance between effectiveness and stealthiness of the attack presents an intriguing future work direction.

REFERENCES

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

[2] Afia Afrin and Omid Ardakanian. 2023. Adversarial Attacks on Machine Learning-Based State Estimation in Power Distribution Systems. In *Proc. 14th ACM Int. Conf. Future Energy Syst.* 446–458.

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

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

[5] Arnab Bhattacharjee, Sukumar Mishra, and Ashu Verma. 2022. Deep Adversary based Stealthy False Data Injection Attacks against AC state estimation. In *14th Asia-Pacific Power Energy Eng. Conf.* IEEE, 1–7.

[6] Narayan Bhusal, Raj Mani Shukla, Mukesh Gautam, Mohammed Benidris, and Shamik Sengupta. 2021. Deep ensemble learning-based approach to real-time power system state estimation. *Int J. Electr. Power Energy Syst.* 129 (2021), 106806.

[7] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *IEEE Symp. Secur. Priv.* IEEE, 39–57.

[8] Ruilong Deng, Peng Zhuang, and Hao Liang. 2018. False data injection attacks against state estimation in power distribution systems. *IEEE Transactions on Smart Grid* 10, 3 (2018), 2871–2881.

[9] Roger C Dugan and Thomas E McDermott. 2011. An open source platform for collaborating on smart grid research. In *2011 IEEE Power and Energy Society General Meeting*. IEEE, 1–7.

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

[11] Moosa Moghimi Haji and Omid Ardakanian. 2019. Practical Considerations in the Design of Distribution State Estimation Techniques. In *2019 IEEE Int. Conf. Commun., Control, Comput. Techn. Smart Grids*. IEEE, 1–6.

[12] René Heinrich, Christoph Scholz, Stephan Vogt, and Malte Lehna. 2023. Targeted Adversarial Attacks on Wind Power Forecasts. *arXiv preprint arXiv:2303.16633* (2023).

[13] Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. 2016. Adversarial examples in the physical world.

[14] Maosen Li, Cheng Deng, Tengjiao Li, Junchi Yan, Xinbo Gao, and Heng Huang. 2020. Towards transferable targeted attack. In *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recog.* 641–649.

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

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

[17] Yao Liu, Peng Ning, and Michael K Reiter. 2011. False data injection attacks against state estimation in electric power grids. *ACM Trans. Inf. Syst. Secur.* 14, 1 (2011), 1–33.

[18] Christoph Johannes Meinrenken et al. 2020. MFRED (public file, 15/15 aggregate version): 10 second interval real and reactive power in 390 US apartments of varying size and vintage. https://doi.org/10.7910/DVN/X9MIDJ

[19] Lubin Meng, Chin-Teng Lin, Tzyy-Ping Jung, and Dongrui Wu. 2019. White-box target attack for EEG-based BCI regression problems. In *26th International Conference on Neural Information Processing*. Springer, 476–488.

[20] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. 2021. On generating transferable targeted perturbations. In *Proc. IEEE/CVF Int. Conf. Computer Vision*. 7708–7717.

[21] Yao Qin, Nicholas Frosst, Sara Sabour, Colin Raffel, Garrison Cottrell, and Geoffrey Hinton. 2019. Detecting and diagnosing adversarial images with class-conditional capsule reconstructions. *arXiv preprint arXiv:1907.02957* (2019).

[22] Md Ashfaqur Rahman and Hamed Mohsenian-Rad. 2013. False data injection attacks against nonlinear state estimation in smart power grids. In *2013 IEEE Power & Energy Society General Meeting*. IEEE, 1–5.

[23] Rui Shu, Tianpei Xia, Laurie Williams, and Tim Menzies. 2022. Omni: Automated ensemble with unexpected models against adversarial evasion attack. *Empir. Softw. Eng.* 27 (2022), 1–32.

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

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

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

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

[28] Florian Tramr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. 2018. Ensemble adversarial training: Attacks and defenses. In *Int. Conf. Learn. Representations*, Vol. 1. 2.

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

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

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

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

[33] Ahmed Samir Zamzam and Nicholas D. Sidiropoulos. 2020. Physics-Aware Neural Networks for Distribution System State Estimation. *IEEE Transactions on Power Systems* 35, 6 (2020), 4347–4356.

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

[35] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. 2021. On success and simplicity: A second look at transferable targeted attacks. *Advances in Neural Information Processing Systems* 34 (2021), 6115–6128.

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

[37] Peng Zhuang, Ruilong Deng, and Hao Liang. 2019. False data injection attacks against state estimation in multiphase and unbalanced smart distribution systems. *IEEE Transactions on Smart Grid* 10, 6 (2019), 6000–6013.