Received: 17 June 2021 | Revised: 21 September 2021 | Accepted: 1 November 2021

IET Renewable Power Generation

DOI: 10.1049/rpg2.12334

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Adversarial attack and defense methods for neural network based state estimation in smart grid

Jiwei Tian[1] | Buhong Wang[2] | Jing Li[3] | Charalambos Konstantinou[4]

[1] ATC Navigation College, Air Force Engineering University, Xi'an, China

[2] Information and Navigation College, Air Force Engineering University, Xi'an, China

[3] School of Design and Art, Henan University of Technology, Zhengzhou, China

[4] KAUST, Thuwal, Saudi Arabia

Correspondence
Jiwei Tian, ATC Navigation College, Air Force Engineering University, No. 1, Changle East Road, Baqiao District, Xi'an City, Shaanxi Province, China 710043.
Email: tianjiwei2016@163.com

Funding information
National Natural Science Foundation of China, Grant/Award Number: 61902426

Abstract

Deep learning has been recently used in safety-critical cyber-physical systems (CPS) such as the smart grid. The security assessment of such learning-based methods within CPS algorithms, however, is still an open problem. Despite existing research on adversarial attacks against deep learning models, only few works are concerned about safety-critical energy CPS, especially the state estimation routine. This paper investigates security issues of neural network based state estimation in the smart grid. Specifically, the problem of adversarial attacks against neural network based state estimation is analysed and an efficient adversarial attack method is proposed. To thwart this attack, two defense methods based on protection and adversarial training, respectively, are proposed further. The experiments demonstrate that the proposed attack method poses a major threat to neural network based state estimation models. In addition, our results present that defense methods can improve the ability of neural network models to defend against such adversarial attacks.

1 | INTRODUCTION

The smart grid can be seen as the largest internet-of-things (IoT) deployment, since it is based on smart IoT devices and intelligent decision-making algorithms to achieve low-loss, efficient, and environmentally-friendly power control [1–3]. At the same time, the smart grid is also facing increasing threats from cyber-attacks [4–7]. For example, the cyber-attack on Ukraine's power system in 2015 caused hundreds of thousands of electricity customers to lose power. Among the attack threats facing the power grid, cyber-attacks on state estimation have attracted a lot of research attention because estimated variables are direct inputs to critical applications in the other energy management system routines (e.g. contingency analysis, economic dispatch) [8]. Typical cyber-attacks against state estimation include false data injection attacks [9–13], load redistribution attacks [14, 15], data framing attacks [16], and topology attacks [17]. These attacks are all against weighted least squares (WLS)-based state estimation. The optimization-oriented WLS scheme for nonlinear (AC) state estimation requires many iterations, and often fails to converge, due to the increasing system size [18]. In addition, it is impractical for real-time estimation [19].

Recently, neural networks (NNs) have been explored to perform various tasks related to state estimation: detection of topological errors [20], generation of pseudo-measurements [21], and real-time state estimation [18, 19, 22]. Particularly, it has been shown that NNs can provide highly accurate state estimation but have much lower computational burden than conventional WLS methods. However, the security implications of utilizing NNs for state estimation are rarely considered. From the perspective of active defense, their security threats should be studied in depth to design more secure models, so that they can be practically applied in actual systems.

NNs are extremely vulnerable to a form of attack called 'adversarial example' [23–25]. Adversarial examples are inputs that an attacker deliberately crafts to cause the deep learning model to operate as not intended. Although research in the field of adversarial machine learning has been carried out for more than a decade [26], most work has focused on classification problems in the image, speech, and text fields [27–31]. Deep learning is increasingly used in cyber-physical systems (CPS) [32, 33], but its security in CPS is still a new topic [34–36]. In the field of energy CPS, research on machine learning (ML) security has only begun to appear recently [37–39]. The vulnerabilities

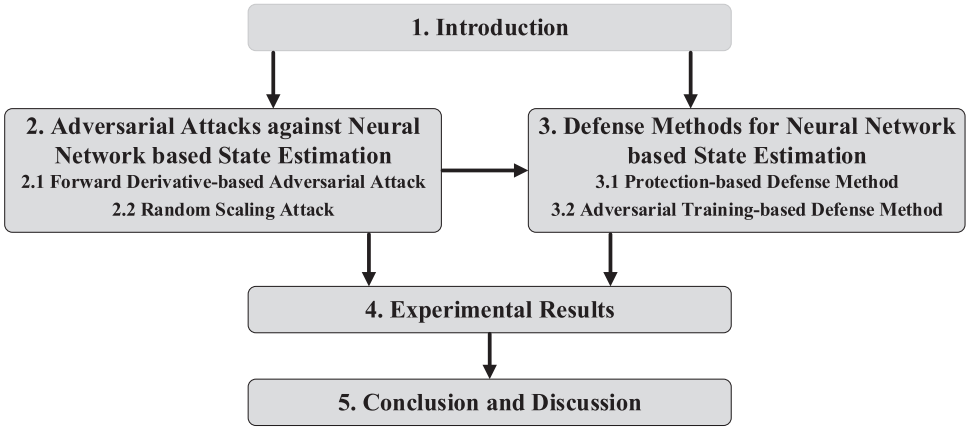

FIGURE 1 The overall flowchart of the paper

of current ML algorithms proposed in power systems are first shown in reference[40], and a recent survey is presented in reference [41]. Concerning NN-based classification tasks in power gird, existing works analyse security issues of grid events classification [42], $N-1$ security classification [43], power quality signal classification [40], [44], energy theft detection [45, 46], non-intrusive load monitoring [47] and false data injection attack detection [48, 49]. In the area of NN-based regression tasks in power grid, research has mainly been focused on security issues of load forecasting. The potential vulnerabilities in load forecasting algorithms by injected perturbations into input temperatures are studied in reference[50]. A domain-specific framework to evaluate security and resilience of load forecasting algorithms is proposed in reference[51]. The experiments showed that NN-based load predictions may suffer worst-case attacks even when only part of the network is compromised [51]. From a defense perspective, adversarial training [52], and game theory [53] are, respectively, adopted to design resilient load forecasting systems. In addition, poisoning attack (implemented during the training phase) on load forecasting is explored in reference[54].

To our best of knowledge, unlike load forecasting, there is little research on security issues of NN-based state estimation. The first study of adversarial attack against NN-based state estimation is performed in reference[55]. However, the iterative optimization algorithms used to carry out such attacks are likely not suitable for large-scale power systems in the real world, and corresponding defense methods have not been considered at all. To address these problems, this paper considers designing efficient adversarial attack crafting methods and exploring corresponding defense strategies. To our knowledge, this is the first study on adversarial attacks against NN-based state estimation from a defense perspective, which calls for more attention on the research about security of NNs in safety-critical applications. The overall flowchart is presented in Figure 1, and the main contributions of this work are summarized as follows:

- We propose an efficient adversarial attack crafting method based on forward derivative. The proposed attack crafting method takes into account multiple factors such as the magnitude of the input elements, the impact of the attack on the multiple regression output, and the number of measurement meters allowed to be controlled. In addition, the attack method is efficient so that corresponding defense methods such as adversarial training can be carried out based on this method.
- We proposed two defense methods to thwart adversarial attacks against state estimation:
  1. *Utilizing traditional security technologies*, the protection-based defense method protects important meters under resource constraints to achieve maximum defense effects.
  2. *Using techniques in the field of adversarial ML*, the adversarial training-based defense method improves the robustness of NN-based state estimation models, thereby reducing the impact of adversarial attacks.

  Both defense methods can improve the ability of NN models to defend against adversarial attacks to a certain extent.

The rest of this paper is organized as follows. In Section 2, the NN-based state estimation and proposed adversarial attack crafting method are introduced. The proposed protection and adversarial training-based defense methods are presented in Section 3. In Section 4, the proposed attack and defense methods are evaluated for various case studies. The conclusion and future work are provided in Section 5.

2 | ADVERSARIAL ATTACKS AGAINST NEURAL NETWORK BASED STATE ESTIMATION

2.1 | Neural network based state estimation

Formally, a NN-based state estimation model represents a function $F : \boldsymbol{X} \rightarrow \boldsymbol{Y}$, where $\boldsymbol{X}$ is an input vector and $\boldsymbol{Y}$ is a regression output vector. Unlike the discrete outputs of classification problems [56], the outputs of regression problems are continuous. Given certain training samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$, the above regression function is computed by minimizing the

expected loss $E_P(R(F(\boldsymbol{x})-\boldsymbol{y}))$. The squared loss function $R(F(\boldsymbol{x})-\boldsymbol{y}) = \|F(\boldsymbol{x})-\boldsymbol{y}\|_2^2$ is often used to measure the difference between $F(\boldsymbol{x})$ and $\boldsymbol{y}$.

2.2 | Attack model and objective

In our attack model, we propose a threat model as follows:

- We assume that the attackers have knowledge of the NN-based state estimation model parameters, and use this knowledge to carry out adversarial attacks. In practice, the related knowledge can be obtained either because of insider threats [57], or because attackers can use various methods, such as eavesdropping on network traffic and destroying database systems [58].
- We assume that the attackers have some resources to maliciously control some energy-CPS meters. This can be achieved through a variety of cyber-attacks such as man-in-the-middle attack, Trojan-horse attack, firmware modification attacks [59], etc.

Although the above assumptions may not always represent the practical cases, they enable us to explore the robustness and vulnerabilities of the NN-based state estimation under the worst-case scenario. In addition, the worst-case scenario can guide, in a more structural way in terms of probable attack impacts, the design of defense strategies.

2.3 | Forward derivative-based adversarial attack

A forward derivative-based adversarial attack method is proposed for power systems state estimation. The forward derivative is defined as the Jacobian matrix of F:

$$J_F(\boldsymbol{X}) = \frac{\partial F(\boldsymbol{X})}{\partial \boldsymbol{X}} = \left[\frac{\partial F_j(\boldsymbol{X})}{\partial \boldsymbol{X}_i}\right]_{i\in 1\cdots M,\ j\in 1\cdots N}, \tag{1}$$

where M and N represent the dimensions of the input and output, respectively.

The gradient of the forward derivative calculation is similar to the gradient of the backpropagation calculation, but there are two important differences: we directly use the derivative of the network instead of its cost function, and we differentiate in regard to the input vector rather than model parameters. The forward derivative approach is chosen because it allows us to find the input components that cause model outputs to change significantly [56].

To derive the forward derivative in (1), the recursive process is started from the first hidden layer of the network.

$$\frac{\partial \boldsymbol{L}_k(\boldsymbol{X})}{\partial \boldsymbol{X}_i} = \left[\frac{f_{k,q}\left(\boldsymbol{W}_{k,q}\cdot \boldsymbol{L}_{k-1} + b_{k,q}\right)}{\boldsymbol{X}_i}\right]_{q\in 1\cdots m_k}, \tag{2}$$

where $\boldsymbol{L}_k$, $f_{k,q}$ represent the output vector and the qth activation function of kth hidden layer ($k \in 1\cdots n+1$), respectively. The connections between the kth layer and the previous layer are represented in $\boldsymbol{W}_{k,q}$. $b_{k,q}$ is the bias for neuron q of layer k. Based on the chain rule, a series of formulas can be written for $k \geq 2$:

$$\begin{aligned}\frac{\partial \boldsymbol{L}_k(\boldsymbol{X})}{\partial \boldsymbol{X}_i}|_{q\in 1\cdots m_k} = &\left(\boldsymbol{W}_{k,q}\cdot \frac{\partial \boldsymbol{L}_{k-1}}{\partial \boldsymbol{X}_i}\right)\times \\ &\frac{\partial f_{k,q}}{\partial \boldsymbol{X}_i}\left(\boldsymbol{W}_{k,q}\cdot \boldsymbol{L}_{k-1} + b_{k,q}\right).\end{aligned} \tag{3}$$

Then $\frac{\partial \boldsymbol{L}_n}{\partial \boldsymbol{X}_i}$ can be expressed and the jth output neuron computes:

$$F_j(\boldsymbol{X}) = f_{n+1,\,j}\left(\boldsymbol{W}_{n+1,j}\cdot \boldsymbol{L}_n + b_{n+1,j}\right). \tag{4}$$

Thus, $J_F[i,j](\boldsymbol{X})$ can be obtained based on the chain rule:

$$\begin{aligned}\frac{\partial F_j(\boldsymbol{X})}{\partial \boldsymbol{X}_i} = &\left(\boldsymbol{W}_{n+1,j}\cdot \frac{\partial \boldsymbol{L}_n}{\partial \boldsymbol{X}_i}\right)\times \\ &\frac{\partial f_{n+1,j}}{\partial \boldsymbol{X}_i}\left(\boldsymbol{W}_{n+1,j}\cdot \boldsymbol{L}_n + b_{n+1,j}\right).\end{aligned} \tag{5}$$

Based on the derived forward derivative, we can add some perturbation leading to significant changes in regression outputs. Unlike only one regression output in load forecasting [50–53], there are usually multiple outputs in regression for state estimation. To compare the cumulative effect of each input disturbance on all outputs, we compute the row sum of the absolute matrix of $J_F(\boldsymbol{X})$,

$$J_{F\Sigma}(\boldsymbol{X}) = \sum_{j=1}^{N} abs(J_F(\boldsymbol{X})(:,j)). \tag{6}$$

The $M\times 1$ vector $J_{F\Sigma}(\boldsymbol{X})$ can reflect the impact of each input disturbance on the overall output to some extent. Among various adversarial attack methods developed in the image domain so far, the well-known and popular method Fast Gradient Sign Method (FGSM) adds the constant magnitude perturbation to the original sample to maximize the loss function [24]. *Unlike the pixels of images in the range of (0,1), the range of the grid measurements used for state estimation is much larger and each meter (value in the input vector) also has a different range of variation.* Based on these conditions, it is more reasonable to express the deviation as a ratio rather than a fixed value. Therefore, we compute

$$J_{F\Sigma i}(\boldsymbol{X}) = J_{F\Sigma}(\boldsymbol{X})\cdot |\boldsymbol{X}|, \tag{7}$$

where $|\boldsymbol{X}|$ represents the absolute vector of $M\times 1$ input.

Considering the resource constraints of attackers, the number of meters allowed to be perturbed is also limited. The larger the element (meter) in $J_{F\Sigma i}(\boldsymbol{X})$, the greater the impact on overall

ALGORITHM 1 Forward derivative-based adversarial attack crafting for NN-based state estimation

Input: $\boldsymbol{x}$: An original input measurement vector;F: The NN-based state estimation model;γ: The ratio of meters attackers can perturb;θ: The scaling ratio made to measurements.

Output: $\boldsymbol{x}'$, The adversarial example.

Initialization: $\boldsymbol{x}' \leftarrow \boldsymbol{x}$.

Calculate the forward derivative w.r.t $\boldsymbol{x}$:$J_F(\boldsymbol{x}) = \frac{\partial F(\boldsymbol{x})}{\partial \boldsymbol{x}} = [\frac{\partial F_j(\boldsymbol{x})}{\partial \boldsymbol{x}_i}]$.

Calculate the metric $J_{F\mathfrak{I}}(\boldsymbol{x})$:$J_{F\mathfrak{I}}(\boldsymbol{x}) = \sum_{j=1}^{N} abs(J_F(\boldsymbol{x})(:, j))$.

Calculate the metric $J_{F\mathfrak{I}i}(\boldsymbol{x})$:$J_{F\mathfrak{I}i}(\boldsymbol{x}) = J_{F\mathfrak{I}}(\boldsymbol{x}) \cdot |\boldsymbol{x}|$.

Find set Υ: the largest γ ratio elements of $J_{F\mathfrak{I}i}(\boldsymbol{x})$

for i in Υ **do**

  $\boldsymbol{x}'_i = \boldsymbol{x}_i \times \theta$

end for

returnThe generated adversarial sample $\boldsymbol{x}'$.

ALGORITHM 2 Random scaling attack crafting for NN-based state estimation

Input: $\boldsymbol{x}$: An original input measurement vector;γ: The ratio of meters attackers can perturb;θ: The scaling ratio made to measurements.

Output: $\boldsymbol{x}'$, The adversarial example.

Initialization: $\boldsymbol{x}' \leftarrow \boldsymbol{x}$.

Find set Υ:Randomly choose γ ratio of meters

for i in Υ **do**

  $\boldsymbol{x}'_i = \boldsymbol{x}_i \times \theta$

end for

returnThe generated adversarial sample $\boldsymbol{x}'$.

regression output. Sophisticated attackers will always choose meters with greater influence to carry out adversarial attacks in order to achieve a high impact on the state estimation with low effort. The method proposed for generating adversarial samples is summarized in Algorithm 1. Note that this method is very efficient since the iterative optimization process is not required (in image domain, C&W [60] and PGD [61] are iterative attack methods and time consuming).

2.4 | Random scaling attack

In random scaling attack (RSA) [52], the attacker randomly selects a proportion γ of all meters and modifies the selected true measurements via a constant scaling factor θ as in Algorithm 1. The RSA method for generating adversarial samples is given in Algorithm 2.

For the above two attack methods, the scaling ratio θ and attack ratio γ will determine the final attack effect. We analyse the related influences in the experiments (Section 4). Note that although the generated adversarial examples by the above two attack methods might raise detection flags within power system operation routines, a smaller number of attack meters and a lower attack ratio can still lead towards implementing stealthy attacks (the threshold-based detection method cannot detect attacks as long as the residual does not exceed the threshold). Furthermore, as mentioned earlier, the proposed attack methods are mainly for exploring vulnerabilities and worst-case scenarios of the model, which can provide guidance for the corresponding defense strategies.

ALGORITHM 3 Calculating the overall importance order of meters

Input: $\boldsymbol{X}$: A set of input measurement vectors (number,N).F: The NN-based state estimation model.

Output: $\boldsymbol{S}$, The overall importance order of meters.

for $\boldsymbol{x}$ in $\boldsymbol{X}$ **do**

  Calculate the forward derivative w.r.t $\boldsymbol{x}$:$J_F(\boldsymbol{x}) = \frac{\partial F(\boldsymbol{x})}{\partial \boldsymbol{x}} = [\frac{\partial F_j(\boldsymbol{x})}{\partial \boldsymbol{x}_i}]$.

  Calculate the metric $J_{F\mathfrak{I}}(\boldsymbol{x})$:$J_{F\mathfrak{I}}(\boldsymbol{x}) = \sum_{j=1}^{N} abs(J_F(\boldsymbol{x})(:, j))$.

  Calculate the metric $J_{F\mathfrak{I}i}(\boldsymbol{x})$:$J_{F\mathfrak{I}i}(\boldsymbol{x}) = J_{F\mathfrak{I}}(\boldsymbol{x}) \cdot |\boldsymbol{x}|$.

  Sort the elements in $J_{F\mathfrak{I}i}(\boldsymbol{x})$ and record the order of each elements:$J_{F\mathfrak{I}is}(\boldsymbol{x}) = argsort(J_{F\mathfrak{I}i}(\boldsymbol{x}))$

end for

Calculate the overall importance order:$\boldsymbol{S} = argsort(\frac{1}{N}\sum_{j=1}^{N} J_{F\mathfrak{I}is}(\boldsymbol{x}))$.

return The overall importance order of meters $\boldsymbol{S}$.

3 | DEFENSE METHODS FOR NEURAL NETWORK BASED STATE ESTIMATION

3.1 | Protection-based defense method

For false data injection attacks against WLS state estimation, protection strategies based on carefully chosen meters are explored in [62]. Similarly as in the case of attack generation, this implies that different sensors also have widely varying importance in defense measures. This phenomenon still exists in NN-based state estimation (see Section 4 for the details).

In fact, in Algorithm 1, the proposed adversarial attack method ranks the importance of the meters. In addition, during our experiments, the significance of some meters remains relatively consistent in different measurement samples. Therefore, we can record the importance ranking of each sensor in different input samples and calculate the order of overall importance of each meter. Based on the overall importance order, we can choose the more vital meters to protect under resource constraints[1]. The related methodologies and techniques for protecting meters include data encryption, authentication, access control, and auditing. The method for calculating the overall importance order is given in Algorithm 3. Note that although different meters can have a varying degree of importance over

[1] In reality, large-scale power grids have a large number of sensors, and defense resources are always limited. Therefore, it is unrealistic to protect all sensors.

ALGORITHM 4 Defending NN-based state estimation using adversarial training

Input: D_{tr}: A training dataset of clean samples.N_{iter}: The number of iterations used for training.

Output: A robust NN-based state estimation model F_{Θ}, where Θ is a set of learnable parameters of the model F.**Training the model F_{Θ} using a training dataset D_{tr}:**

for 1, 2, ... , N_{iter} **do**

  Optimize Θ by training F_{Θ} using D_{tr}.

end for

Generate a set of adversarial samples $\tilde{\mathcal{X}}$ using Algorithm 1.**Retraining the model F_{Θ} using the set of samples $\tilde{\mathcal{X}}$:**

for 1, 2, ... , N_{iter} **do**

  Update Θ by training F_{Θ} using $\tilde{\mathcal{X}}$.

end for

returnA robust NN-based state estimation model F_{Θ}.

time (power systems are dynamic systems), the overall importance order derived by Algorithm 3 is global and comprehensive (based on a set of input measurements vectors). Therefore, the derived importance order reflects the importance of each meter from the overall perspective.

3.2 | Adversarial training-based defense method

Adversarial training is used to improve robustness of a NN model to adversarial attacks [24, 63]. In this method, a large number of adversarial examples generated are used to retrain the model. The basic requirement of this method is to use the strongest possible attack crafting algorithm to generate as many adversarial examples as possible. In references[24] and [64], the results show that adversarial training improves robustness of NNs to adversarial attacks. Adversarial training can provide regularization for NNs and improve their classification accuracy [24]. Here, we also adopt adversarial training as our defense method to improve robustness of NN-based state estimation in power systems. Our proposed adversarial training method is summarized in Algorithm 4.

4 | EXPERIMENTAL RESULTS

We conducted experiments against NN-based state estimation algorithms proposed in reference[18]. The corresponding dataset, code, and well-trained feed-forward neural networks (FNN) for the IEEE 118-bus benchmark system are publicly available, which makes it easy to evaluate our attack and defense methods. Specifically, real load data from the 2012 Global Energy Forecasting Competition[2] were used to generate the datasets, where the load series were sub-sampled for size reduction by a factor of 2 for the IEEE 118-bus system [18]. Subsequently, the resultant load instances were normalized to match the scale of power demands in the simulated system. The MATPOWER toolbox [65] was used to solve the AC power flow equations to obtain the ground-truth voltages and measurements. The dataset includes 18,528 voltage-measurement pairs, with 14,822 pairs employed for training and 3706 kept for testing. Each voltage-measurement pair consists of 490 dimensional measurement input and 236 dimensional state output. Three trained NN-based state estimation models are: 6-layer FNN, 8-layer FNN and Prox-linear Net, respectively[3]. In the experiments, we consider two NN-based state estimation models: 8-layer FNN and Prox-linear Net (in the following parts of the section, we analyse the results of 8-layer FNN first and then analyse the results of Prox-linear Net). The normalized root mean-square error (RMSE) is adopted as the evaluation metric (the RMSE of normal test data of 8-layer FNN and Prox-linear Net are 0.02 and 0.0003[4], respectively).

4.1 | Impact analysis of adversarial attacks

The performance of the forward derivative-based adversarial attack (FAA) and RSA are both affected by two factors: attack ratio γ and scaling ratio θ. In order to better study the influence of related factors, we carry out considerable experiments. The chosen percentage of measurement points (total of 490) being compromised γ is listed in Table 1. For each γ, we explore the attack performance under different scaling ratio θ, which is also listed in Table 1. Each experiment runs on all training measurement instances. The main results are summarized in Figure 2. Figure 2a shows that as the attack ratio γ and the scaling ratio θ increase, the average RMSE increases accordingly. This is in line with our intuitive idea that the larger the perturbation, the greater the error in the regression results. However, as pointed out in reference[23], the influence of randomly generated noise (even large) on classification tasks is much smaller than that of the specific perturbation, which can be referred to as an *adversarial example*. This observation still exists in the NN-based state estimation tasks: under the same attack parameter configuration, the RMSE of RSA (Figure 2b) is much smaller than that of FAA (Figure 2a), implying the RSA has much less impact than the forward derivative-based adversarial attack. Therefore, our proposed FAA method searches meters that have a large impact on the regression results first and then applies perturbations. In addition, from Figure 2a, attackers can cause a non-negligible impact on state estimation results even the attack ratio γ is small (the RMSE of clean data is only 0.02). The similar results of Prox-linear Net are provided in Figure 3, which implies that the state estimation model based on different NN structures is still

[2] https://www.kaggle.com/c/global-energy-forecasting-competition-2012-load-forecasting/data

[3] The specific model architecture and parameters can be referred to in [18]. Due to space constraints, they are omitted here.

[4] The well-trained 8-layer FNN and Prox-linear Net models are publicly available and we evaluated the RMSEs based on the provided test data in reference [18]. Although the Prox-linear Net is better than the 8-layer FNN, we considered both the models to evaluate the proposed attack and defense methods.

TABLE 1 Parameter setting for adversarial attacks

Attack number (attack ratio γ)									
1(< 1%)	2(< 1%)	3(< 1%)	4(1%)	24(5%)	49(10%)	73(15%)	98(20%)	122(25%)	147(30%)
Scaling ratio θ									
0.5	0.6	0.7	0.8	0.9	1.1	1.2	1.3	1.4	1.5

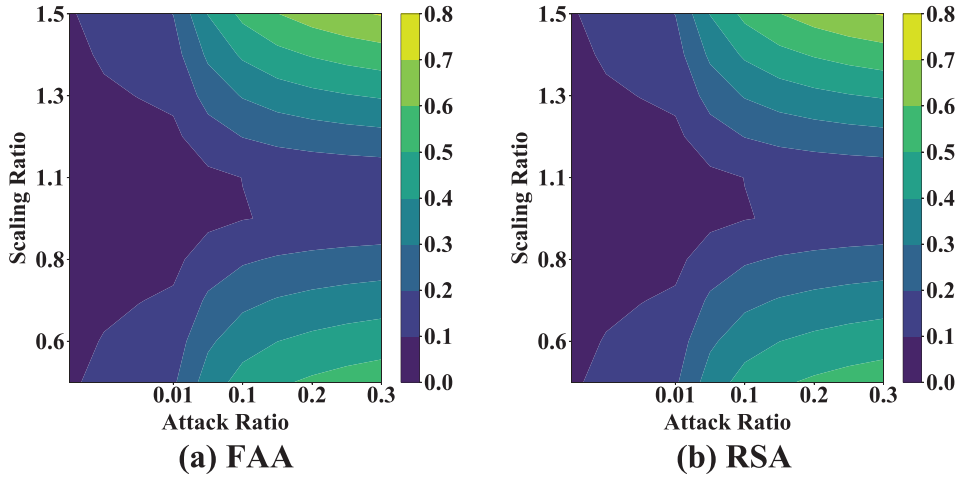

FIGURE 2 RMSE of (a) FAA and (b) RSA for varying values of attack ratio γ and scaling ratio θ (8-layer FNN)

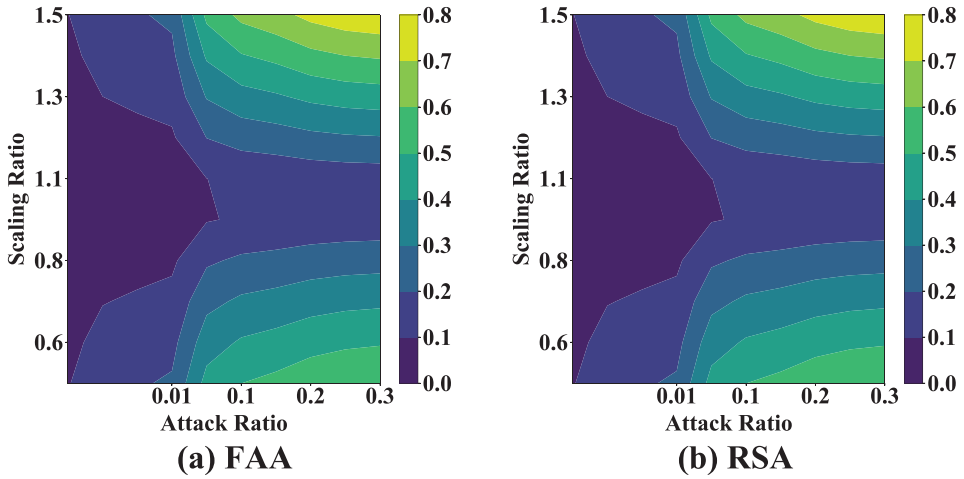

FIGURE 3 RMSE of (a) FAA and (b) RSA for varying values of attack ratio γ and scaling ratio θ (Prox-linear Net)

vulnerable to adversarial attacks. Besides, as shown in Figure 4, most of the RMSEs of 8-layer FNN are bigger than those of Prox-linear Net, which is consistent with the original model performance.

4.2 | Evaluation of protection-based defense

The protection-based defense method focuses on the fundamental elimination of the possibility of injecting bad data based on security measures such as encryption, authentication and access control. Due to the limited resources from the defense perspective, power grid utilities cannot protect against all meters. As shown in Figure 5, based on the same scaling ratio, the impact of attacking different meters will be significantly different. At this point, how to choose important meters for defense is crucial.

Based on Algorithm 3, defenders can derive the overall importance order of meters $\boldsymbol{S}$. For 8-layer FNN, the overall importance order of meters $\boldsymbol{S}$ is shown in Figure 6. In Figure 6, the meters are ranked based on importance from 1 to 490. The smaller the meter's sequence value, the more important the

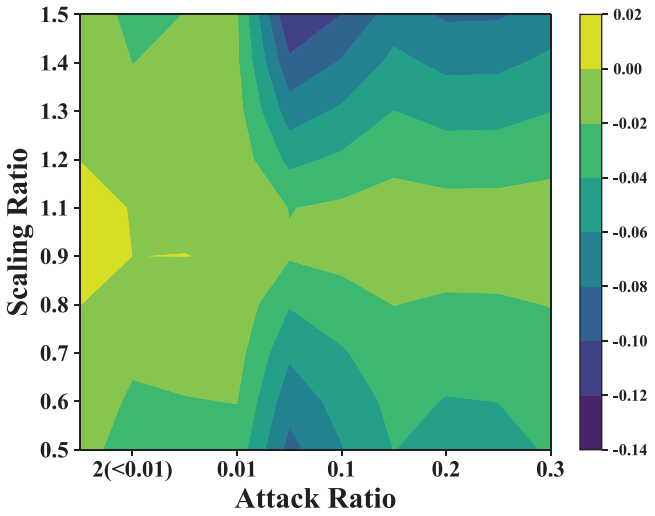

FIGURE 4 Difference of RMSE between Prox-linear Net and 8-layer FNN for varying values of attack ratio γ and scaling ratio θ (negative values mean that RMSEs of 8-layer FNN are bigger than those of Prox-linear Net)

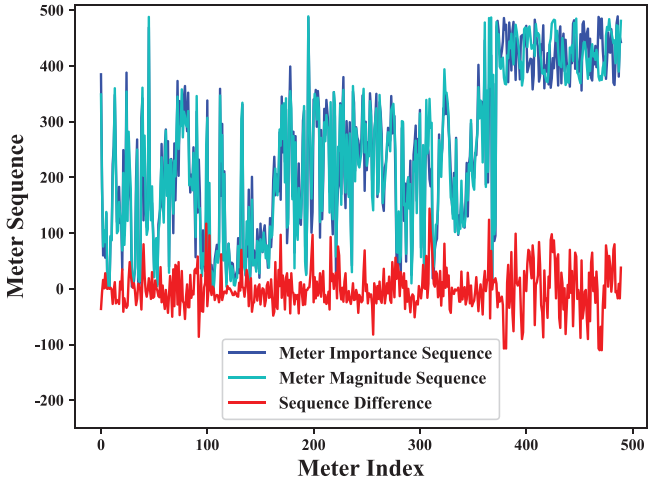

FIGURE 6 The overall importance order of meters (8-layer FNN)

sensor is. Besides, we also rank the meters based on magnitude from 1 to 490 and calculate difference between two ranks. In Figure 6, although there is a degree of relationship between importance and magnitude (the curve of '*Sequence Difference*' fluctuates up and down with 0 as the center), it is not reasonable to consider magnitude only (some values of 'Sequence Difference' are large). The Algorithm 3 considering both magnitude and gradient provides a more reasonable way to rank the overall importance order of meters, and choose related meters to protect, accordingly. The similar results of Prox-linear Net are provided in Figure 7. In addition, we compare the importance orders of 8-layer FNN and Prox-linear Net. The result is shown in Figure 8. The overall importance orders of 8-layer FNN and Prox-linear Net are quite different. This indicates that *in different NN-based state estimation models, the meters that require priority in terms of protection are likely to be different, even if their selection does not lead to better model performance when there is no attack*. The reason may be that models with different structures learn different representations, causing the same meter to have different effects in the model. Therefore, the protection strategy is for a specific NN model. In other words, for the deployed NN-based state estimation model, the specific protection strategy is chosen to achieve the maximum defense effect.

Based on derived overall importance order of meters, defenders can choose a number of the most important meters to protect. The chosen percentage of meters being protected is listed

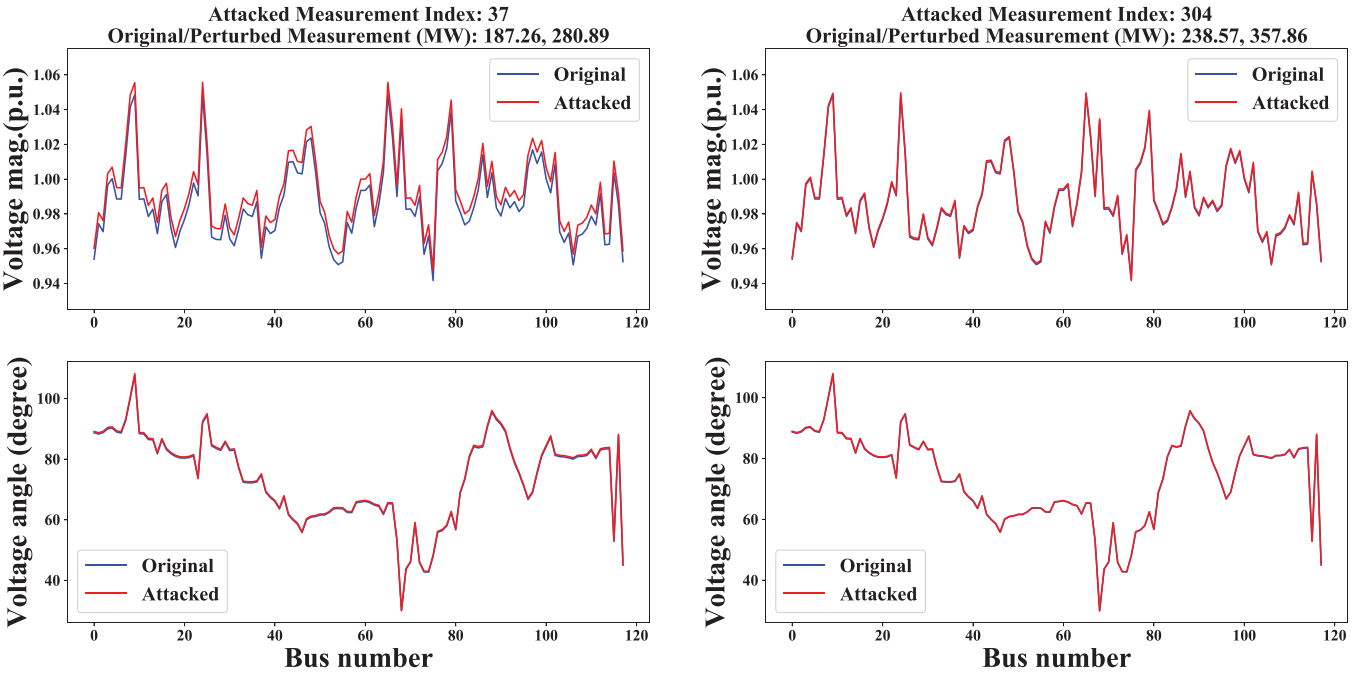

FIGURE 5 Comparison of attacking two different meters respectively (8-layer FNN, scaling ratio: 1.5)

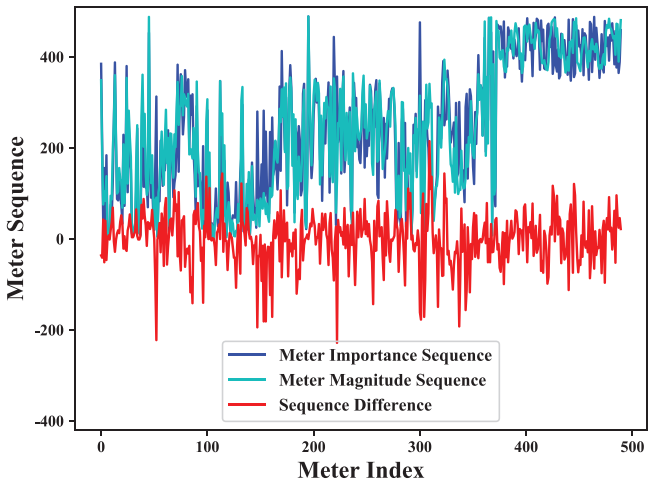

FIGURE 7 The overall importance order of meters (Prox-linear Net)

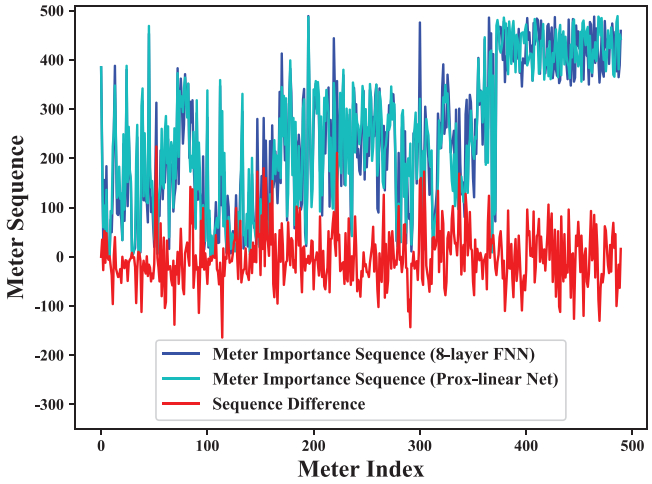

FIGURE 8 The overall importance order of meters (8-layer FNN vs. Prox-linear Net)

TABLE 2 Parameter setting for protection-based defense

Protect number (protect ratio)					
1 (< 1%)	2 (< 1%)	3 (< 1%)	4 (1%)	24 (5%)	49 (10%)

in Table 2. Then, for each protection configuration settings, we conduct FAAs based on Table 1. The only difference is that chosen protected meters cannot be perturbed. The main results are summarized in Figure 9. Figures 9 and Figure 2a show that by protecting a small number of important meters, the impact of FAAs on the NN-based model will be greatly reduced (e.g. by protecting only 10% of meters, the worst-case RMSE is reduced from 0.65 to 0.2). The similar results for Prox-linear Net are provided in Figure 10 (note that the protected meters of 8-layer FNN and Prox-linear Net may be different, see Figure 11).

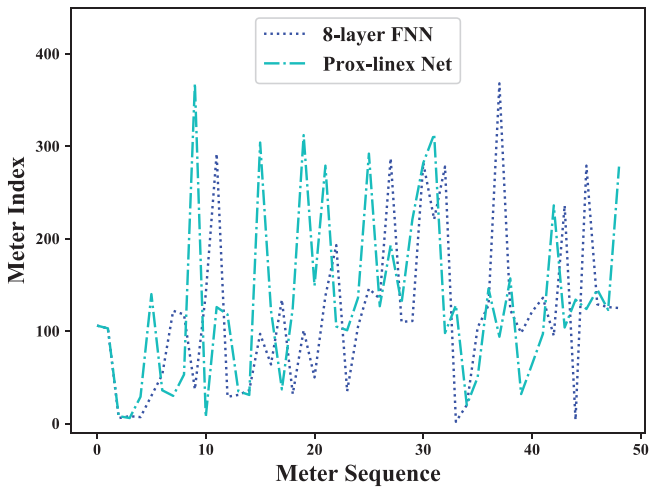

FIGURE 9 RMSE (FAA) under different protection settings (8-layer FNN): (a) 1 (< 1%), (b) 2 (< 1%), (c) 3 (< 1%), (d) 4 (1%), (e) 24 (5%), (f) 49 (10%)

TABLE 3 Parameter setting for generating adversarial examples for adversarial training

Attack number (attack ratio γ)					
4 (1%)		24 (5%)		49 (10%)	
Scaling ratio θ					
0.5	0.7	0.9	1.1	1.3	1.5

4.3 | Evaluation of adversarial training-based defense

The adversarial training-based defense method focuses on improving the robustness of NN-based state estimation models. In order to better evaluate the effect of adversarial training, we carry out substantial experiments: for each test, we retrain the model using generated adversarial examples based on a fixed γ and a fixed θ as in Table 3. Then, based on the trained model we calculate RMSE against normal test data and RMSE of generated new adversarial examples based on the same γ and θ. The results of 8-layer FNN are shown in Figure 12. Figure 12 shows that based on adversarial training, the effect of FAAs based on the same γ and θ have been significantly reduced. However, Figure 12c shows that adversarial training affects estimation performance on normal samples, which also exists in image domain [25]. The corresponding similar results of Prox-linear Net are provided in Figure 13. However, we can observed that compared with the 8-layer FNN model, adversarial training of the Prox-linear Net achieves better performance in both estimation against normal samples and adversarial examples. This indication might be representative of the advantages of this specific NN model. In addition, from the perspective of the defender, we can set a threshold about RMSE of normal samples, and then adversarial training is conducted based on this threshold to improve the robustness of the model.

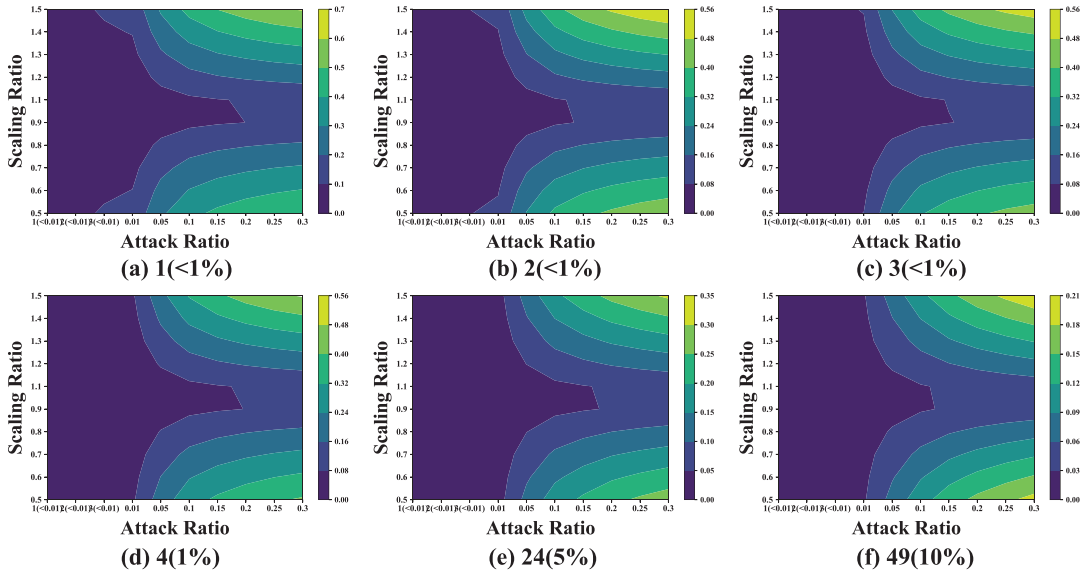

FIGURE 10 RMSE (FAA) under different protection settings (Prox-linear Net): (a) 1 (< 1%), (b) 2 (< 1%), (c) 3 (< 1%), (d) 4 (1%), (e) 24 (5%), (f) 49 (10%)

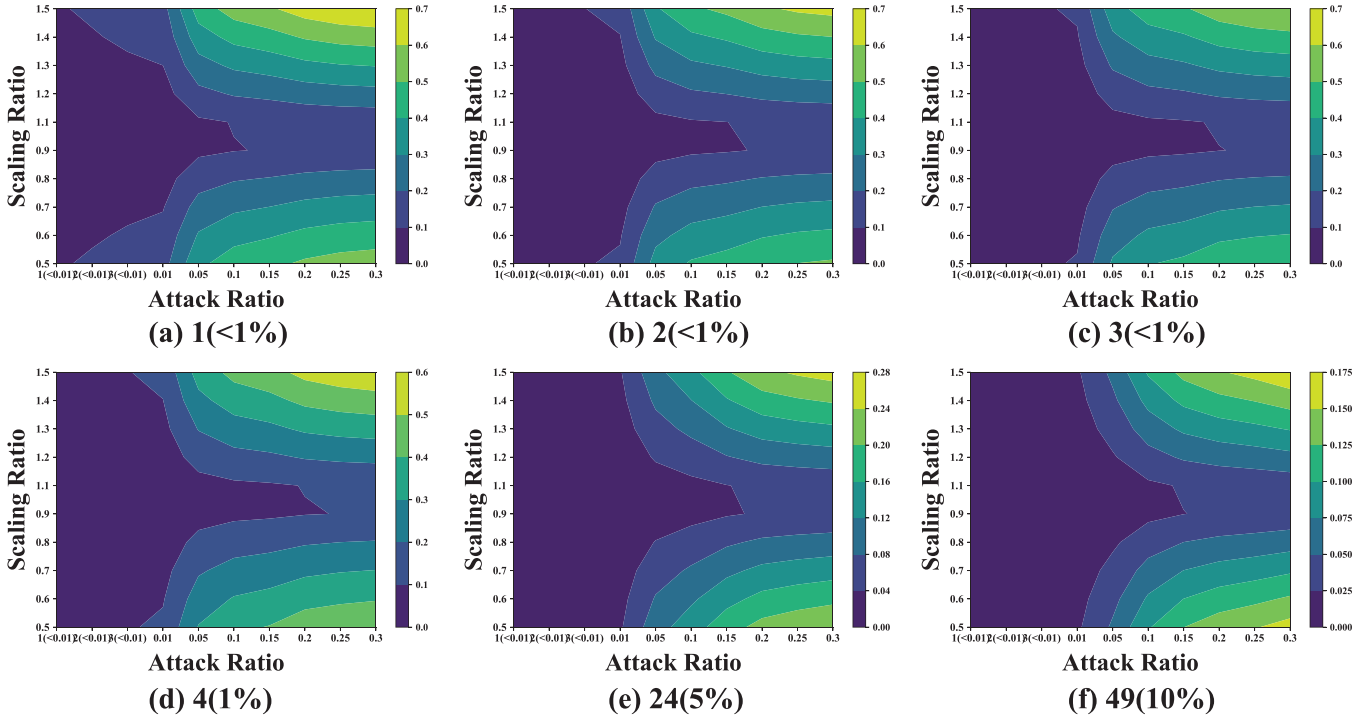

FIGURE 11 The indices of the 49 (10%) most important measurement points (meters) (8-layer FNN vs. Prox-linear Net)

4.4 | Comparison of protection-based and adversarial training-based defense

In our aforementioned experiments, the protection-based defense can greatly reduce the impact of adversarial attacks like FAAs. However, such protection method might consume a lot of defense resources, especially in large-scale and complex power grids. The adversarial training method does not require defensive resources. Nevertheless, while improving the robustness of the NN model against adversarial examples, the performance of the model on normal samples will be reduced.

5 | CONCLUSION AND DISCUSSION

Here, we propose a forward derivative-based method to attack NN-based state estimation algorithms. To counter the threat

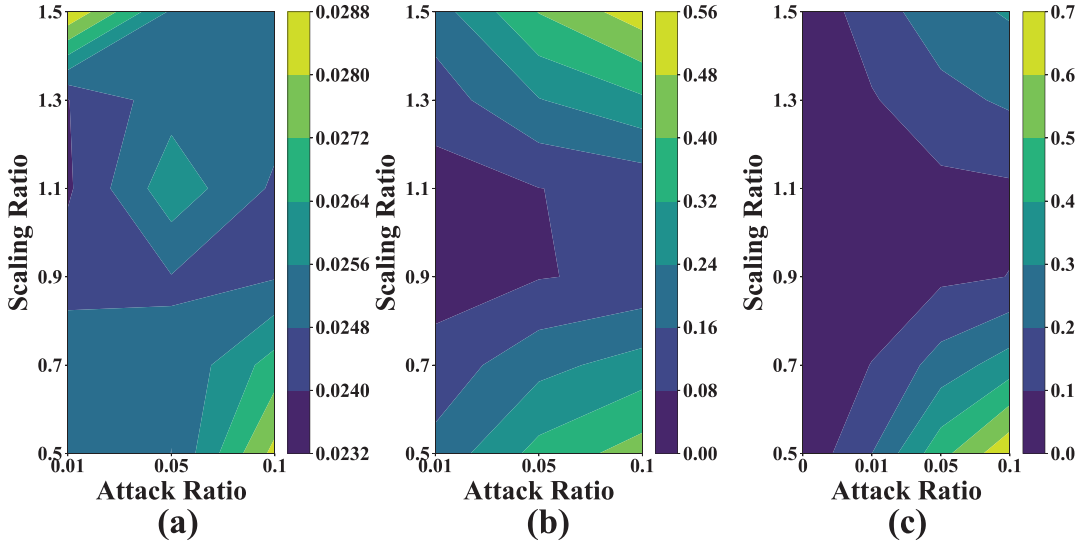

FIGURE 12 Performance of adversarial training under different settings (8-layer FNN): (a) RMSE of generated adversarial examples against retrained model based on the fixed pair of γ and θ; (b) RMSE of generated adversarial examples against original model based on the fixed pair of γ and θ; (c) RMSE of normal test samples against retrained model based on the fixed pair of γ and θ

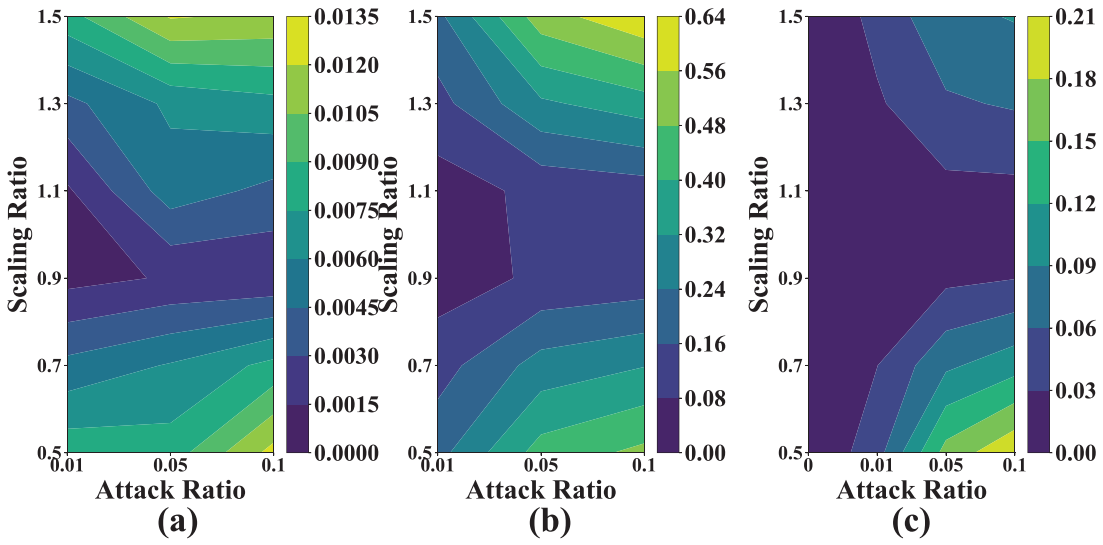

FIGURE 13 Performance of adversarial training under different settings (Prox-linear Net): (a) RMSE of generated adversarial examples against retrained model based on the fixed pair of γ and θ; (b) RMSE of generated adversarial examples against original model based on the fixed pair of γ and θ; (c) RMSE of normal test samples against retrained model based on the fixed pair of γ and θ

of this attack, we also propose two defense methods based on protection and adversarial training, respectively. Both defense methods can improve the ability of the NN model to defend against adversarial attacks to a certain extent. There may be some possible limitations of our proposed methods. In detail, the attack we proposed did not consider the existence of physical limitations or bad data detection. Therefore, in the future, we will explore more practical and smarter adversarial attacks. Through a more comprehensive exploration of possible attacks, we can design more reasonable defense methods while improving the performance of the NN model on benign samples.

ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (No. 61902426).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Jiwei Tian https://orcid.org/0000-0002-1485-7465
Charalambos Konstantinou https://orcid.org/0000-0002-3825-3930

REFERENCES

1. Wang, Z., Liu, Y., Ma, Z., Liu, X., Ma, J.: Lipsg: Lightweight privacy-preserving Q-learning-based energy management for the IoT-enabled smart grid. IEEE Internet Things J. 7(5), 3935–3947 (2020)
2. Rana, M.M., Xiang, W., Wang, E.: IoT-based state estimation for microgrids. IEEE Internet Things J. 5(2), 1345–1346 (2018)
3. Xu, W., Li, J., Dehghani, M., GhasemiGarpachi, M.: Blockchain-based secure energy policy and management of renewable-based smart microgrids. Sustain. Cities Soc. 72, 103010 (2021)
4. Zhang, Y., Krishnan, V.V.G., Pi, J., Kaur, K., Srivastava, A., Hahn, A., Suresh, S.: Cyber physical security analytics for transactive energy systems. IEEE Trans. Smart Grid 11(2), 931–941 (2020)
5. McLaughlin, S., Konstantinou, C., Wang, X., Davi, L., Sadeghi, A., Maniatakos, M., Karri, R.: The cybersecurity landscape in industrial control systems. Proc. IEEE 104(5), 1039–1057 (2016)
6. Tian, J., Wang, B., Li, T., Shang, F., Cao, K.: Coordinated cyber-physical attacks considering DoS attacks in power systems. Int. J. Robust Nonlinear Control 30(11), 4345–4358 (2020)
7. Tian, J., Wang, B., Li, X.: Data-driven and low-sparsity false data injection attacks in smart grid[J]. Secur. Commun. Netw. 2018, Art. no. 8045909 (2018). https://doi.org/10.1155/2018/8045909.
8. Liu, X., Hu, Y., Konstantinou, C., et al.: CHIMERA: A Hybrid Estimation Approach to Limit the Effects of False Data Injection Attacks[J]. arXiv preprint arXiv:2103.13568, (2021)
9. Konstantinou, C., Maniatakos, M.: A case study on implementing false data injection attacks against nonlinear state estimation. In: Proceedings of the 2nd ACM Workshop on Cyber-Physical Systems Security and Privacy, pp. 81–92. CPS-SPC '16, New York (2016)
10. Anubi, O.M., Konstantinou, C.: Enhanced resilient state estimation using data-driven auxiliary models. IEEE Trans. Ind. Inf. 16(1), 639–647 (2020)
11. Dehghani, M., Kavousi-Fard, A., Dabbaghjamanesh, M., Avatefipour, O.: Deep learning based method for false data injection attack detection in ac smart islands. IET Gener. Transm. Distrib. 14(24), 5756–5765 (2020)
12. Tian, J., Wang, B., Li, T., Shang, F., Cao, K., Guo, R.: TOTAL: Optimal protection strategy against perfect and imperfect false data injection attacks on power grid cyber–physical systems. IEEE Internet Things J. 8(2), 1001–1015 (2021)
13. Konstantinou, C., Anubi, O.M.: Resilient cyber-physical energy systems using prior information based on Gaussian process. IEEE Trans. Ind. Inf. 1–1 (2021)
14. Xiang, Y., Ding, Z., Zhang, Y., Wang, L.: Power system reliability evaluation considering load redistribution attacks. IEEE Trans. Smart Grid 8(2), 889–901 (2016)
15. Ospina, J., Liu, X., Konstantinou, C., Dvorkin, Y.: On the feasibility of load-changing attacks in power systems during the COVID-19 pandemic. IEEE Access 9, 2545–2563 (2021)
16. Kim, J., Tong, L., Thomas, R.J.: Data framing attack on state estimation. IEEE J. Sel. Areas Commun. 32(7), 1460–1470 (2014)
17. Zhang, J., Sankar, L.: Physical system consequences of unobservable state-and-topology cyber-physical attacks. IEEE Trans. Smart Grid 7(4)
18. Zhang, L., Wang, G., Giannakis, G.B.: Real-time power system state estimation and forecasting via deep unrolled neural networks. IEEE Trans. Signal Process. 67(15), 4069–4077 (2019)
19. Zamzam, A.S., Sidiropoulos, N.D.: Physics-aware neural networks for distribution system state estimation. IEEE Trans. Power Syst. 35(6), 4347–4356 (2020)
20. Krstulovic, J., Miranda, V., Simões Costa, A.J.A., Pereira, J.: Towards an auto-associative topology state estimator. IEEE Trans. Power Syst. 28(3), 3311–3318 (2013)
21. Miranda, V., Krstulovic, J., Keko, H., Moreira, C., Pereira, J.: Reconstructing missing data in state estimation with autoencoders. IEEE Trans. Power Syst. 27(2), 604–611 (2012)
22. Mestav, K.R., Luengo-Rozas, J., Tong, L.: Bayesian state estimation for unobservable distribution systems via deep learning. IEEE Trans. Power Syst. 34(6), 4910–4920 (2019)
23. Szegedy, C., Zaremba, W., Sutskever, I., et al.: Intriguing properties of neural networks[J]. arXiv preprint arXiv:1312.6199, (2013).
24. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples, arXiv:1412.6572
25. Moosavi-Dezfooli, S.-M., Fawzi, A., Frossard, P.: Deepfool: A simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2574–2582 (2016)
26. Biggio, B., Roli, F.: Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognit. 84, 317–331 (2018)
27. Yuan, X., He, P., Zhu, Q., Li, X.: Adversarial examples: Attacks and defenses for deep learning. IEEE Trans. Neural Networks Learn. Syst. 30(9), 2805–2824 (2019)
28. Wang, D., Li, C., Wen, S., et al.: Daedalus: Breaking nonmaximum suppression in object detection via adversarial examples[J]. IEEE Trans. Cybern. (2021). https://doi.org/10.1109/TCYB.2020.3041481.
29. Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T., Xu, W.: Dolphinattack: Inaudible voice commands. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 103–117 (2017)
30. Li, J., Ji, R., Liu, H., Hong, X., Gao, Y., Tian, Q.: Universal perturbation attack against image retrieval. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4899–4908 (2019)
31. Zheng, T., Chen, C., Ren, K.: Distributionally adversarial attack. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, pp. 2253–2260 (2019)
32. Fei, X., Shah, N., Verba, N., Chao, K.-M., Sanchez-Anguix, V., Lewandowski, J., James, A., Usman, Z.: CPS data streams analytics based on machine learning for cloud and fog computing: A survey. Future Gener. Comput. Syst. 90, 435–450 (2019)
33. Tian, J., Wang, B., Guo, R., Wang, Z., Cao, K., Wang, X.: Adversarial attacks and defenses for deep learning-based unmanned aerial vehicles. IEEE Internet Things J. 1–1 (2021). http://doi.org/10.1109/JIOT.2021.3111024
34. Ren, K., Wang, Q., Wang, C., Qin, Z., Lin, X.: The security of autonomous driving: Threats, defenses, and future directions. Proc. IEEE 108(2), 357–372 (2019)
35. Jiang, W., Li, H., Liu, S., Luo, X., Lu, R.: Poisoning and evasion attacks against deep learning algorithms in autonomous vehicles. IEEE Trans. Veh. Technol. 69(4), 4439–4449 (2020)
36. Konstantinou, C., Maniatakos, M., Saqib, F., Hu, S., Plusquellic, J., Jin, Y.: Cyber-physical systems: A security perspective. In: 2015 20th IEEE European Test Symposium (ETS), pp. 1–8 (2015)
37. Bor, M.C., Marnerides, A.K., Molineux, A., Wattam, S., Roedig, U.: Adversarial machine learning in smart energy systems. In: Proceedings of the Tenth ACM International Conference on Future Energy Systems, pp. 413–415 (2019)
38. Zografopoulos, I., Ospina, J., Liu, X., Konstantinou, C.: Cyber-physical energy systems security: Threat modeling, risk assessment, resources, metrics, and case studies. IEEE Access 9, 29775–29818 (2021)
39. Mohammadpourfard, M., Khalili, A., Genc, I., Konstantinou, C.: Cyber-resilient smart cities: Detection of malicious attacks in smart grids. Sustain. Cities Soc. 75, 103116 (2021)
40. Chen, Y., Tan, Y., Deka, D.: Is machine learning in power systems vulnerable? In: 2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), pp. 1–6. IEEE, New Jersey (2018)
41. Sayghe, A., Hu, Y., Zografopoulos, I., Liu, X., Dutta, R.G., Jin, Y., Konstantinou, C.: Survey of machine learning methods for detecting false data injection attacks in power systems. IET Smart Grid 3(5), 581–595 (2020)
42. Niazazari, I., Livani, H.: Attack on grid event cause analysis: An adversarial machine learning approach[C]//2020 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT). IEEE. 1–5 (2020)
43. Venzke, A., Chatzivasileiadis, S.: Verification of neural network behaviour: Formal guarantees for power system applications[J]. IEEE Trans. Smart Grid 12(1), 383–397 (2020)

44. Tian, J., Li, T., Shang, F., Cao, K., Li, J., Ozay, M.: Adaptive normalized attacks for learning adversarial attacks and defenses in power systems. In: 2019 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), pp. 1–6 (2019)
45. Li, J., Yang, Y., Sun, J.S.: Searchfromfree: Adversarial measurements for machine learning-based energy theft detection. In: 2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), pp. 1–6. IEEE, New Jersey (2020)
46. Li, J., Yang, Y., Sun, J.S.: Exploiting vulnerabilities of deep learning-based energy theft detection in ami through adversarial attacks[J]. arXiv preprint arXiv:2010.09212, (2020)
47. Wang, J., Srikantha, P.: Stealthy black-box attacks on deep learning non-intrusive load monitoring models. IEEE Trans. Smart Grid 12(4), 3479–3492 (2021)
48. Sayghe, A., Zhao, J., Konstantinou, C.: Evasion attacks with adversarial deep learning against power system state estimation. In: 2020 IEEE Power & Energy Society General Meeting (GM 2020), pp. 1–5. IEEE, New Jersey (2020)
49. Sayghe, A., Anubi, O.M., Konstantinou, C.: Adversarial examples on power systems state estimation. In: 2020 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), pp. 1–5. IEEE, New Jersey (2020)
50. Chen, Y., Tan, Y., Zhang, B.: Exploiting vulnerabilities of load forecasting through adversarial attacks. In: Proceedings of the Tenth ACM International Conference on Future Energy Systems, pp. 1–11 (2019)
51. Zhou, X., Li, Y., Barreto, C.A., Li, J., Volgyesi, P., Neema, H., Koutsoukos, X.: Evaluating resilience of grid load predictions under stealthy adversarial attacks. In: 2019 Resilience Week (RWS), Vol. 1, pp. 206–212. IEEE, New Jersey (2019)
52. Tang, Z., Jiao, J., Zhang, P., Yue, M., Chen, C., Yan, J.: Enabling cyberattack-resilient load forecasting through adversarial machine learning. In: 2019 IEEE Power & Energy Society General Meeting (PESGM), pp. 1–5. IEEE, New Jersey (2019)
53. Barreto, C., Koutsoukos, X.: Design of load forecast systems resilient against cyber-attacks. In: International Conference on Decision and Game Theory for Security, pp. 1–20. Springer, Berlin (2019)
54. Liang, Y., He, D., Chen, D.: Poisoning attack on load forecasting. In: 2019 IEEE Innovative Smart Grid Technologies-Asia (ISGT Asia), pp. 1230–1235. IEEE, New Jersey (2019)
55. Liu, T., Shu, T.: Adversarial false data injection attack against nonlinear ac state estimation with ann in smart grid. In: International Conference on Security and Privacy in Communication Systems, pp. 365–379. Springer, Berlin (2019)
56. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: 2016 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 372–387. IEEE, New Jersey (2016)
57. Baracaldo, N., Joshi, J.: An adaptive risk management and access control framework to mitigate insider threats. Comput. Secur. 39, 237–254 (2013)
58. Konstantinou, C., Maniatakos, M.: Security Analysis of Smart Grid. pp. 451–487. [Online]. Available: (2017). https://digitallibrary.theiet.org/content/books/10.1049/pbpo095e_ch15
59. Konstantinou, C., Maniatakos, M.: Impact of firmware modification attacks on power systems field devices. In: 2015 IEEE International Conference on Smart Grid Communications (SmartGridComm), pp. 283–288 (2015)
60. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 39–57. IEEE, New Jersey (2017)
61. Madry, A., Makelov, A., Schmidt, L., et al.: Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint arXiv:1706.06083 (2017)
62. Khanna, K., Panigrahi, B.K., Joshi, A.: Priority-based protection against the malicious data injection attacks on state estimation[J]. IEEE Syst. J. 14(2), 1945–1952 (2019)
63. Tramèr, F., Kurakin, A., Papernot, N., et al.: Ensemble adversarial training: Attacks and defenses[J]. arXiv preprint arXiv:1705.07204 (2017)
64. Huang, R., Xu, B., Schuurmans, D., et al.: Learning with a strong adversary[J]. arXiv preprint arXiv:1511.03034 (2015)
65. Zimmerman, R.D., Murillo-Sánchez, C.E., Thomas, R.J.: MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education. IEEE Trans. Power Syst. 26(1), 12–19 (2010)

How to cite this article: Tian, J., Wang, B., Li, J., Konstantinou, C.: Adversarial attack and defense methods for neural network based state estimation in smart grid. IET Renew. Power Gener. 16, 3507–3518 (2022). https://doi.org/10.1049/rpg2.12334