EasyChair Preprint

№ 15370

A Power System Data-Driven State Estimation Adversarial Attack Method Based on Conditional Generative Adversarial Network

Shiyi Hou, Jing Zhang, Lei Zhu and Qi Wang

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

November 5, 2024

A Power System Data-driven State Estimation Adversarial Attack Method Based on Conditional Generative Adversarial Network

1st Shiyi Hou
School of Electrical Engineering
Southeast University
Nanjing, China
houshiyi02@163.com

2nd Jing Zhang
School of Electrical Engineering
Southeast University
Nanjing, China
220232906@seu.edu.cn

3rd Lei Zhu
School of Electrical Engineering
Southeast University
Nanjing, China
230199104@seu.edu.cn

4rd Qi Wang
School of Electrical Engineering
Southeast University
Nanjing, China
wangqi@seu.edu.cn

***Abstract*—Data-driven methods based on artificial intelligence technology exhibit the ability to process data quickly and accurately, thus currently being widely applied in the field of power system state estimation (SE). However, recent studies have found that processing data during the training phase of data-driven algorithms can mislead the operational results, demonstrating security risks in data-driven methods. In light of this, to reveal potential security issues in data-driven algorithms, this paper proposes an adversarial attack method based on conditional generative adversarial network (CGAN). Experimental results indicate that adding minute disturbances generated by CGAN to target samples can misleads the predictions of SE model. This method directs the attack based on preset label conditions during the model application.**

Keywords—data-driven algorithms, state estimation, conditional generative adversarial network, adversarial attack

I. INTRODUCTION

A. Background

With the gradual implementation of new power system architectures, the complexity of power system has significantly increased, manifesting in greater uncertainty, non-linearity, and expansion of state space dimensions. These development trends have greatly increased the difficulties of traditional model-driven methods in state estimation (SE) and control, making it challenging to meet the growing demands for real-time performance and accuracy in modern power grids. In contrast, data-driven SE methods construct empirical functions for SE by learning from a large amount of measurement results. By utilizing these constructed empirical functions, the complex physical modeling of power grid can be effectively omitted, thereby achieving fast and accurate predictions of bus states[1].

However, data-driven methods also introduce new challenges. As the training process of data-driven models relies on samples, it leads to a severe dependence on sample data. Consequently, the quality and representativeness of sample data have a crucial impact on model performance, leading to stringent requirements for sample data. Based on these factors, data-driven models face numerous challenges, including insufficient algorithm robustness, a tendency for overfitting, and limited generalization ability. These issues may threaten the safe operation of power system and could even become potential targets for cyber attacks.

This research is supported by the Project of State Grid Corporation of China SGHLDK00KJJS2400196 (Research on key technology of coordinated defense against cyber-physical attack in power monitoring system).

B. Related Work

Current data-driven SE algorithms include methods based on graph convolutional network (GCN), deep neural network (DNN), etc. In recent years, SE algorithms that combine model-driven and data-driven approaches have received significant attention from the academic community. A strategy that integrates the DC power flow method with GCN is proposed in Reference [2], enabling rapid extraction of data features while preserving the physical topological characteristics of the network. A method that combines physical models with machine learning for dynamic frequency prediction in power system is presented in Reference [3].

Current research on attacks targeting data-driven algorithms mainly focuses on poisoning and adversarial attacks. A flexible poisoning attack strategy is discussed in Reference [4], in which the model's learning outcome can be manipulated by injecting carefully crafted poisoned samples into the training data, thereby enhancing the effectiveness and stealthiness of the attack. The results of attacks on intelligent power grid voltage stability assessment using adversarial samples generated by six representative methods are presented in Reference [5]. Current methods for generating adversarial attack samples include the Fast Gradient Method and Projected Gradient Descent. However, research on using generative adversarial network (GAN) to create such samples is still in its early stages.

C. Contributions

This research focus on the application stage of power system SE model and utilizes a method based on conditional generative adversarial network (CGAN) to implement targeted attacks on the model's predictions. In the specific implementation process, this research first employs GAN to generate adversarial samples that mislead the predictions of the SE model. It then further utilizes CGAN to optimize the attack results on SE model. The method proposed in this research and its main contributions are as follows:

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

1) Power system SE based on DNN. This research trains a DNN with extensive simulation data to quickly and accurately estimate the state of various power grid nodes.

2) Adversarial attack method based on GAN. The adversarial attack samples generated by GAN can cause significant deviations in the estimation results after being input into the SE model.

3) Optimized adversarial attack method based on CGAN. Building upon the GAN, this research incorporates labels as constraint conditions during the network training process. This enables the adversarial samples to produce attack corresponding to the labels, resulting in directed deviations in SE results.

This experiment utilizes the IEEE New England 39-bus 10-generator system. The results demonstrate that the adversarial attack samples generated by CGAN can cause the power system SE predictions to deviate according to the labels, achieving the expected outcomes.

II. Proposed Methodology

A. Static SE Problem in Power Systems

Power system node states encompass voltage magnitudes and phase angles at each node, serving as key indicators of the power system. Currently, the Supervisory Control and Data Acquisition (SCADA) systems in power grids can collect active power, reactive power, and voltage magnitude at each node, but they are unable to directly measure phase angle information. Consequently, the phase angle states of the power grid can only be obtained through calculation. The objective of SE is to rapidly and accurately estimate the current state information of the system through extensive measurement data, thereby ensuring the system stability.

In power systems, the mathematical model for SE is a measurement equation that reflects the interrelationships among network structure, line parameters, state variables, and real-time measurements [6]. For a network with N nodes, the system's operational state is described by a set $\{x = V_{1:N}, \theta_{1:N}\}$, where V_i and θ_i are the voltage magnitude and voltage phase angle of each node, respectively. Within the real-time data collection capabilities of existing SCADA systems, the main types of data that can be effectively collected include active and reactive power injected into buses and voltage magnitudes. These measurement results can be described by a set $\{z = P_{1:N}, Q_{1:N}, V_{1:N}\}$, where P_i and Q_i are the measured net injected active and reactive power at the nodes, respectively. Therefore, the relationship between measurement results and node states can be expressed as:

$$z = h(x) + v \tag{1}$$

In equation (1), z is the measurement vector from the SCADA system, x is the node state vector, v is the measurement error vector, and $h(\bullet)$ is the nonlinear function mapping node states to measurement data.

In power systems, $h(\bullet)$ is a nonlinear function of x [7], expressed as shown in equations (2) and (3):

$$P_i = \sum_{j \in i} V_i V_j \left(G_{ij} cos\theta_{ij} + B_{ij} sin\theta_{ij} \right) \tag{2}$$

$$Q_i = \sum_{j \in i} V_i V_j \left(G_{ij} sin\theta_{ij} + B_{ij} cos\theta_{ij} \right) \tag{3}$$

To solve this nonlinear function, the most classic approach is the weighted least squares (WLS) iterative method [8], which describes the problem as an optimization formula shown in equation (4):

$$\min \left\| h(x) - z \right\|_2 \tag{4}$$

According to article [9], this optimization formula can be expressed as the objective function in equation (5):

$$x^* = arg \ min_x \left[z - h(x) \right]^T R^{-1} \left[z - h(x) \right] \tag{5}$$

In equation (5), R is a diagonal covariance matrix representing error weights, and x^* is the estimated value of the state vector.

B. Data-driven SE Model

Disregarding measurement noise, the relationship between state estimates and measurement results can be described as $z = h(x)$. Consequently, there must exist an inverse function $h(\bullet)^{-1}$ that maps measurement results to state estimates. Since voltage magnitudes can be accurately measured in the system, the purpose of the data-driven algorithm is to construct an empirical function $f_\theta : z \to x$.

The data-driven algorithm utilizes n sets of historical data $\{N = X_{1:n}, Z_{1:n}\}$ obtained from power grid measurements or software simulations. These data are fed into the model for repeated iterative training. Without constructing physical topology, the algorithm mines data features and constructs an empirical function by minimizing the loss function (6) [10].

$$\theta^* = \arg \min_\theta \frac{1}{n} \sum_{i=1}^{n} \left(f_\theta(z_i) - x_i \right)^2 \tag{6}$$

C. Adversarial Sample Generation Method Based on CGAN

CGAN is a variant of GAN. GAN can transform random noise into fake samples that share similar data characteristics with original samples. The training of GAN is an adversarial process, primarily consisting of training a generator and a discriminator. The generator aims to produce data that is as realistic as possible, while the discriminator works to distinguish between real and fake data. The training of the generator and the discriminator occurs simultaneously. As training progresses, the generator becomes increasingly adept at generating realistic samples, while the discriminator must improve its discrimination ability. Eventually, if the generator is sufficiently trained, the discriminator will be unable to distinguish between real and generated samples, at which point the generator and discriminator reach a dynamic equilibrium.

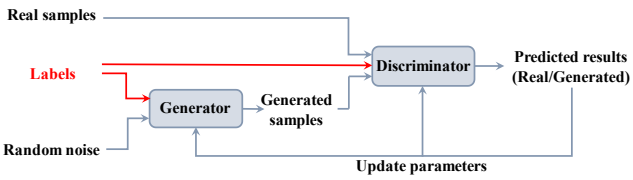


Fig. 1. Comparison of working principles between GAN and CGAN

In this research, CGAN is used to solve the optimization problem of adversarial attack. Its working principle compared to GAN is illustrated in the red part of Figure 1. CGAN adds label constraints to the basic GAN structure, with both the generator and discriminator inputs including label values. This enables the generator to produce disturbances that include features corresponding to the label data, while the discriminator determines the authenticity of real and generated samples under the same label conditions. With the addition of labels, the network training process changes from unsupervised learning to supervised learning.

D. Adversarial Attack on data-driven SE Models

TABLE I. CGAN-BASED DATA-DRIVEN SES ATTACK PROCESS

CGAN-Based Data-Driven SE Attack Process
Input training database P_{data} and noise P_z;
Classify samples in database P_{data} and assign label T_i to each category;
Initialize state estimation training database P_{data};
Pretrain the SE model f_θ using equation (6);
for Epoch = 1, 2,⋯ **do**
for iteration = 1, 2,⋯ **do**
Extract mini-batch noise $z^{(1)},\cdots,z^{(m)}$ from the generator's prior random noise $P_z(z)$;
Embed respective label $T^{(i)}$ into each noise $z^{(i)}$;
Extract mini-batch samples $x^{(1)}$, ⋯ ,$x^{(m)}$ from the real data distribution $P_{data}(x)$;
Generate adversarial samples using equations (7) and (11);
Update generator parameters using stochastic gradient descent: $\nabla_\theta \frac{1}{m}\sum_{i=1}^{m} log(1-D((G(z^{(i)}\|y)+x^{(i)})\|y))$
Determine the probability of real samples using equation (12);
Update discriminator parameters using stochastic gradient descent: $\nabla_{\theta,d} \frac{1}{m}\sum_{i=1}^{m}[log\ D\ (x^{(i)}+log(1-D(G(z^{(i)}\|y)+x^{(i)})\|y)]$
end for
end for
Input the obtained adversarial samples into the SE model using equation (10), and determine whether there is a deviation in the predictions;

For a power grid system with N nodes, the data collected by the SCADA system from each node can be described as $Z=[Z_1,Z_2,...,Z_N]^T$, where Z_i contains the P, Q, and V measurement results of the i-th node. The generator network transforms m input random noise vectors $r=[r_1,r_2,...,r_m]^T$ into disturbances $K=[K_1,K_2,...K_N]^T$. Let $F(\bullet)$ represent the mapping from the real sample Z_i and disturbance K_i to the adversarial attack sample $\tilde{Z}_i$. $G(\bullet)$ and $D(\bullet)$ describe the output functions of the generator and discriminator, respectively. The principle of adversarial attack on nodes with I-dimensional vectors can be described as:

$$Z^i(I,:)+=K^i \Rightarrow \tilde{Z}^i=F(Z^i,K^i) \tag{7}$$

$$K^i=G\left(r^i\right) \tag{8}$$

$$s=D\left(\tilde{Z}^i,Z^i\right) \tag{9}$$

$$L=f_\theta\left(\tilde{Z}^i\right) \tag{10}$$

In equation (9), s represents the discriminator's score for determining real and fake samples. This score reflects the probability that the input sample comes from the real sample set. When the network is sufficiently trained, this score will stabilize at 0.5. In equation (10), L represents the prediction result $\{V_{1:N},\theta_{1:N}\}$ of the SE model. When adversarial samples $\tilde{Z}$ are input, L's predictions will exhibit significant deviations, thus achieving the attack.

CGAN adds label constraints to the foundation of GAN. Given an input label y, the working principles of the generator and discriminator are as follows:

$$K^i=G\left(r^i\,|\,y\right) \tag{11}$$

$$s=D\left(\left(\tilde{Z}^i,Z^i\right)|\,y\right) \tag{12}$$

Equation (11) represents the disturbance generated by the generator under the label y, while equation (12) represents the discriminator's judgment of real and generated sample under the condition of label y. The implementation process of the SE attack algorithm based on the CGAN is shown in Table I.

III. CASE STUDY

A. Experiment Setup

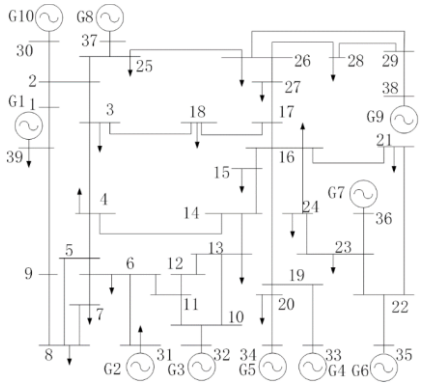


Fig. 2. Topology of the New England 39-bus system

This experiment focuses on the New England 39-10 bus system, utilizing MATPOWER for simulation. The simulation yields power, voltage magnitude, and phase angle information for 10,000 samples across 39 buses. These simulated samples are then used to train both the SE model and the CGAN. The New England 39-bus system network comprises 19 load buses and 10 generator buses, with its topological structure illustrated in Figure 2.

B. Analysis of SE Results Based on DNN

This research employs DNN to achieve SE. During the model training phase, a sample set of 10,000 instances containing P, Q, V, θ is randomly allocated to construct the test set and training set. Using supervised learning, the inputs $\{P_{1:N},Q_{1:N},V_{1:N}\}$ and outputs $\{V_{1:N},\theta_{1:N}\}$ are mapped to establish a corresponding relationship, and a regression prediction model is trained. The trained model predicts the test set states, and its performance is evaluated by calculating the Mean Absolute Error (MAE) between the predictions and actual values to ensure accuracy. During the data preprocessing stage, the sample data is normalized to improve the training efficiency and accuracy of the model.

The predictions of the SE model are shown in Figure 3. The degree of overlap between the curves indicates that the predictions are generally consistent with the actual results. The MAE of the predicted voltage magnitude value compared to the true value is 0.0177, which accounts for 0.59% of the measurement range, indicating it is less than 1%. The MAE of the predicted phase angle compared to the true value is 0.0188, which accounts for 0.62% of the measurement range, indicating it is less than 1%. These results meet the accuracy requirements, demonstrating that the DNN can accurately achieve SE for the power system.

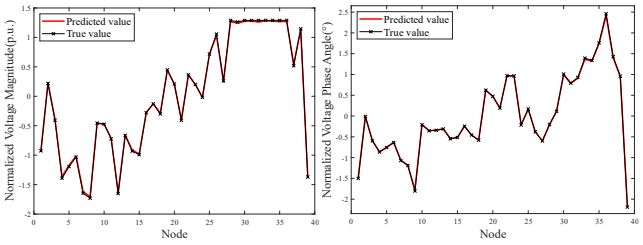


Fig. 3. Predictions of the SE model based on DNN

C. Adversarial Attack on SE Model Based on GAN

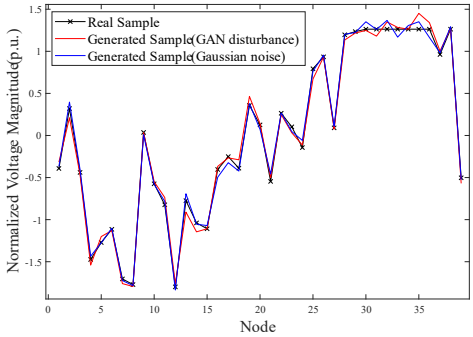
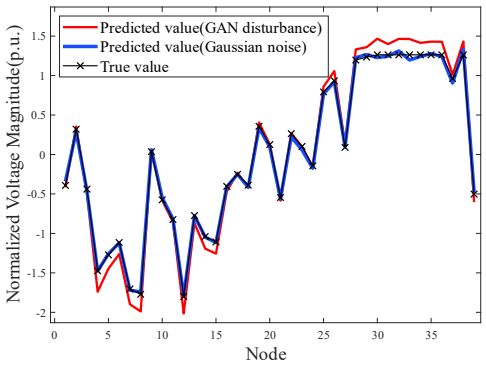


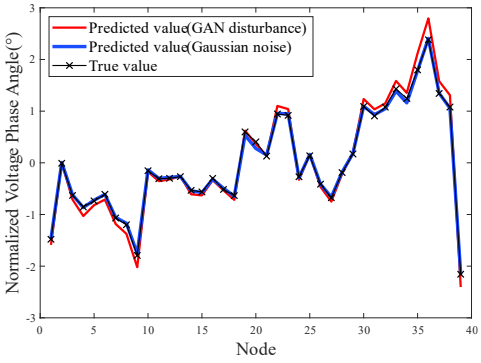
Fig. 4. The comparative analysis of the two types of attack sample

To comprehensively analyze the attack effects of disturbances generated by GAN, this experiment also applies Gaussian noise, which is commonly generated in industry, to attack the SE model and compares the results of the two attack methods. Gaussian noise consists of random measurement errors following a normal distribution. Adding Gaussian noise with a mean of 0 and a standard deviation of 0.05 to the real samples produces attack samples. The comparison of the two types of attack samples obtained from the experiment is illustrated in Figure 4.

When the attack samples are input into the SE model for prediction, a representative attack result is shown in Figure 5. As evident from Figure 5, when adversarial samples are fed into the model, the predictions exhibit significant deviations. The MAE between the voltage magnitude attack results and the true values is 0.1038, whereas the MAE for normal predictions is 0.0177. Similarly, the MAE between the phase angle attack results and the true values is 0.1058, while the MAE for normal predictions is 0.0188. The results of the adversarial attack show a prediction error that is six times greater than the normal rate. After introducing Gaussian noise, the MAE between the predicted voltage magnitude and the true values is only 0.0229, while the MAE between the predicted phase angle and the true values is 0.0237. The adversarial attack results are 4.6 times greater than the Gaussian noise attack. This indicates that the adversarial samples generated by the GAN exhibit superior attack effectiveness. The comparison of attack results is presented in Table II.



(a) Voltage magnitude



(b) Phase angle

Fig. 5. Comparison of attack results from two sample types

TABLE II. COMPARISON OF ATTACK RESULTS

MAE of the predicted value	V	θ
Original Data	0.0177	0.0188
Adversarial Sample with GAN	0.1038	0.1058
Gaussian White Noise(0.05p.u.)	0.0229	0.0237

D. Adversarial Attacks on SE Model Based on CGAN

After demonstrating the vulnerability of SE model in the previous section, this research further employs CGAN to optimize adversarial samples. During the training process of the CGAN, 4000 samples with voltage magnitudes over 1.05 per-unit and 4000 samples below 0.95 per-unit were incorporated, each with assigned labels. After CGAN training, the adversarial samples that cause the SE model to predict values beyond the upper and lower bounds are shown in Figure 6. In Figure 6, the disturbances amplitudes generated by the CGAN are all less than 0.01, keeping the real samples within the normal voltage fluctuation range.

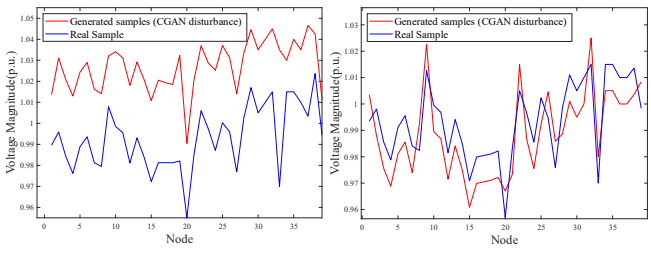
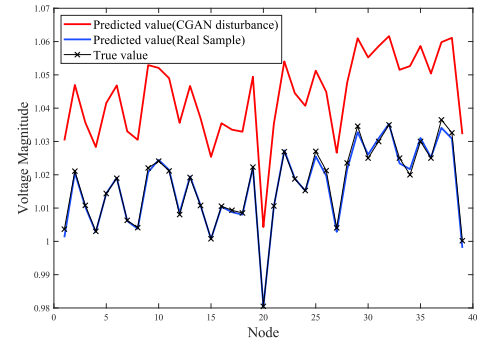
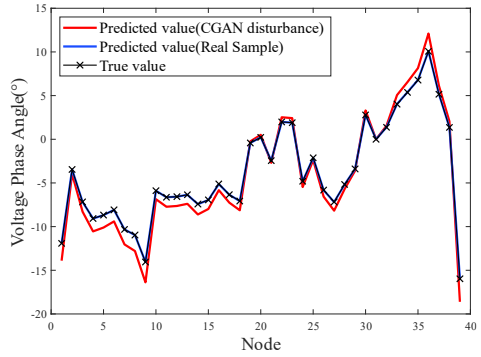


Fig. 6. Adversarial examples generated by CGNA

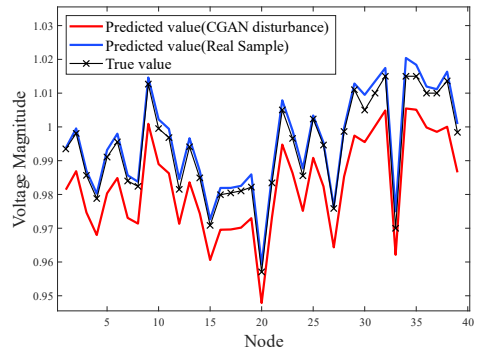


(a) Voltage magnitude

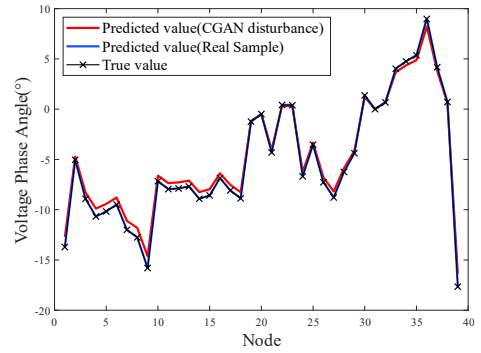


(b) Phase angle

Fig. 7. Exceeding upper bound adversarial sample attack result



(a) Voltage magnitude



(b) Phase angle

Fig. 8. Exceeding lower bound adversarial sample attack result

Representative attack results after inputting these adversarial samples into the SE model are shown in Figures 7-8. As can be seen from Figures 7-8, when disturbances exceeding the upper and lower bounds are added to the original samples, the predictions of the SE model exhibit upper and lower bound violations. The maximum value of the upper bound attack result is 1.0614. The MAE between the upper bound attack results and the true values is 0.508, while the average prediction error is 0.0051. The attack results are 100 times the prediction error. The minimum value of the lower bound attack result is 0.9479. The MAE between the lower bound attack results and the true values is 0.2531, while the average prediction error is 0.0055. The attack results are 46 times the prediction error. The results demonstrate that adversarial examples can achieve the expected attack outcomes based on preset label values.

IV. Conclusion

This research demonstrates the effectiveness of adversarial samples in attacking data-driven SE model, revealing the security risks inherent in data-driven algorithms. Experimental results show that adversarial samples generated by GAN can cause significant deviations in the predictions of data-driven SE model. Small disturbances with specific labels produced by CGAN can cause SE results to exceed 1.05 or fall below 0.95, achieving targeted bias. These findings provide a theoretical basis for developing effective defense strategies in the future.

References

[1] Y. Zhang, H. Zhang, C. Li, et al., “Review on deep learning applications in power system frequency analysis and control,” Proc. CSEE, vol. 41, no. 10, pp. 3392–3406, 2021.

[2] Z. Wu, Q. Wang, J. Hu, and Y. Tang, “Integrating model-driven and data-driven methods for fast state estimation,” Int. J. Electr. Power Energy Syst., vol. 139, p. 107982, 2022.

[3] Q. Wang, F. Li, Y. Tang, and Y. Xu, “Integrating model-driven and data-driven methods for power system frequency stability assessment and control,” IEEE Trans. Power Syst., vol. 34, no. 6, pp. 4557–4568, 2019.

[4] W. Jiang, H. Li, S. Liu, Y. Ren, and M. He, “A Flexible Poisoning Attack Against Machine Learning,” in Proc. 2019 IEEE International Conference on Communications (ICC), Shanghai, China, 2019, pp. 1-6.

[5] Q. Song, C. Ren, R. Tan, and Y. Xu, “Understanding credibility of adversarial examples against smart grid: a case study for voltage stability assessment,” in Proc. 2021 12th ACM Int. Conf. Future Energy Syst., Virtual Event, 2021, pp. 95–106.

[6] F. C. Schweppe, “Power system static-state estimation, part III: implementation,” IEEE Trans. Power Appar. Syst., vol. PAS-89, no. 1, pp. 130–135, 1970.

[7] Z. Wu, Q. Wang, J. Hu, and Y. Tang, “Integrating model-driven and data-driven methods for fast state estimation,” Int. J. Electr. Power Energy Syst., vol. 139, p. 107982, 2022.

[8] S. Shanmugapriya and D. Maharajan, “A fast Broyden's approximation-based weighted least square state estimation for power systems,” Int. J. Numer. Model. Electron. Netw. Devices Fields, vol. 34, no. 2, pp. 1–7, 2020.

[9] J. Hu, Q. Wang, Y. Ye, Z. Wu, and Y. Tang, “A high temporal-spatial resolution power system state estimation method for online DSA,” IEEE Trans. Power Syst., vol. 39, no. 1, pp. 877–889, 2024.

[10] J. Zhang, J. Hu, C. Miao, and Q. Wang, “A GAN-based adversarial attack method for data-driven state estimation,” in Proc. 2023 IEEE 6th Int. Electr. Energy Conf. (CIEEC), Hefei, China, 2023, pp. 3655–3659.