Letter

CoRE: Constrained Robustness Evaluation of Machine Learning-Based Stability Assessment for Power Systems

Zhenyong Zhang and David K. Y. Yau, *Senior Member, IEEE*

Dear Editor,

Machine learning (ML) approaches have been widely employed to enable real-time ML-based stability assessment (MLSA) of large-scale automated electricity grids. However, the vulnerability of MLSA to malicious cyber-attacks may lead to wrong decisions in operating the physical grid if its resilience properties are not well understood before deployment. Unlike adversarial ML in prior domains such as image processing, specific constraints of power systems that the attacker must obey in constructing adversarial samples require new research on MLSA vulnerability analysis for power systems. In this letter, we propose a novel evaluation framework to analyze the robustness of MLSA against adversarial samples with key considerations for damage (i.e., the ability of the adversarial data to cause ML misclassification), bad data detection, physical consistency, and limited attacker's capacity to corrupt data. Extensive experiments are conducted to evaluate the robustness of MLSA under different settings.

To achieve comprehensive context awareness of power systems integrated with renewable energy sources, household loads, and electric vehicles, internet-of-things (IoT) technologies have been widely integrated into power systems to enable autonomous monitoring, optimization, and control. However, the increasing complexity of the power system necessitates ever more complicated, if not outright infeasible, models for traditional analytical methods to achieve sufficient coverage and accuracy, which brings challenges to efficient and agile system operations. To address the problem, big data collected by IoT monitoring devices may allow data-driven ML approaches to control/optimize the power system's operation effectively. As a typical example, MLSA has been widely used to assess/predict whether the current operating condition (OC) of an electrical grid is system-wide stable or unstable subject to credible contingencies in real-time [1]. It enables the grid to operate close to its stability limits in meeting significantly fluctuating demand, frequent kicks-in and out of intermittent energy resources, extreme weather conditions, etc. On the other hand, a host of real-world incidents, such as BlackEnergy and Stuxnet among others [2], evidence that communication channels supporting the IoT are prone to attacks by malicious cyber actors [3] and [4].

Meanwhile, research has shown that ML models can be generally misled by adversarial samples to give wrong answers [5]. Interestingly, adversarial samples investigated for computer vision, due to their low energy, are found to escape notice completely by the human eye while fooling the ML. For example, a cat image, with only a single pixel strategically subverted, can be misclassified as a dog. As a result, the robustness of ML models and their verification [6] under adversarial scenarios has attracted a lot of attention. This robustness metric can play a major role in selecting the best ML models for specific applications. However, traditional robustness analysis is not suitable for power grids [7]. Venzke and Chatzirasileiadis [8] analyzed the robustness of a fully connected neural network trained for security assessment by formulating it as a mixed-integer linear programming problem. Ren and Xu [9] analyzed the robustness of MLSA under different norms and verified lower bounds of adversarial perturbations. Their results do not consider key power-domain properties, however.

To have real impacts in the power domain, adversarial samples must obey its constraints or they will fail to mislead the MLSA into wrong decisions. This letter addresses this key requirement that has not been well addressed in the state of the art.

Specifically, this letter considers the adversarial scenario shown in Fig. 1. Control centers usually adopt security measures such as air-gapped isolation or logical isolation by a virtual private network (VPN). We assume that the attacker uses pathways like compromised USB drives or existing VPN intrusions to subvert the MLSA by modifying data (e.g., sensor measurements) in the local supervisory control and data acquisition (SCADA) network. This data may be assessed for integrity before being used by the MLSA, e.g., a bad data detection (BDD) method in state estimation can be used to reject abnormal data. Hence, we require adversarial samples to bypass the BDD before they can do damage. Besides, the adversarial examples must be credible data that meets physical constraints like power balance and power limits. Lastly, it is assumed that the attacker will not be able to corrupt certain data due to limited capability or other available defenses in the system. Note that prior work [10] has advanced an approach to evaluate MLSA robustness considering physical constraints under a linearized dc model. However, it has not addressed the security of MLSA under a nonlinear ac model. To address the gap, we propose a constrained robustness evaluation (CoRE) framework that analyzes MLSA robustness under an ac model in the face of attacks that satisfy a set of practical constraints including effectiveness (to cause misclassification), BDD bypass, physical consistency, and attacker's limited data corruption capability. We also present extensive experiments to analyze the robustness of MLSA under different ML models and model parameters.

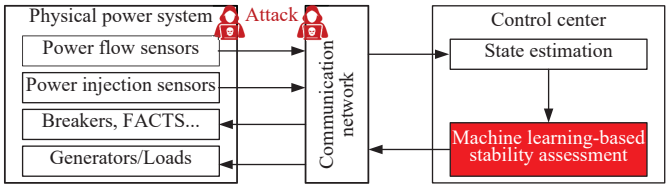

Fig. 1. Adversarial scenario of the MLSA. We assume that the field sensors and communication networks are vulnerable to cyber-attacks that compromise data integrity [2].

Problem formulation:

MLSA: Use of data-driven MLSA, supported by a control center with suitable computational and storage servers, is gaining interest among researchers and utility operators [11]. Such MLSA works as follows. It consists of two main stages: offline training and online prediction. First, detailed numerical stability analysis is conducted for a variety of OCs under different contingencies. The stability vs. instability of the OCs labels each corresponding data point. The training dataset is denoted by $\boldsymbol{S}=\{\boldsymbol{s}_1,\boldsymbol{s}_2,\ldots,\boldsymbol{s}_N\}$, where $\boldsymbol{s}_i=\{\boldsymbol{z}_i,y_i\}$ for $i\in\{1,2,\ldots,N\}$; $\boldsymbol{z}_i$ is an OC; and $y_i=0$ or 1, where "0" stands for "unstable" and "1" for "stable". In general, an OC is driven by power data such as power injections, power flows, and voltage phasor quantities. The offline training is carried out with the labeled training dataset to learn an optimized MLSA model f, whereas, in the online prediction stage, the current OC $\boldsymbol{z}_i$ is fed as real-time input to the MLSA model, which then determines if the OC is stable or not. In this letter, we mainly analyze the vulnerability of the prediction process of MLSA.

For ease of illustration, we use linear classifiers for the stability assessment in the following. We consider a binary classifier (a general multi-class classifier can be treated as an aggregation of binary classifiers), so that the MLSA model output has two possible values,

Corresponding author: Zhenyong Zhang.

Citation: Z. Y. Zhang and D. K. Y. Yau, "CoRE: Constrained robustness evaluation of machine learning-based stability assessment for power systems," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 2, pp. 557–559, Feb. 2023.

Z. Y. Zhang is with the State Key Laboratory of Public Big Data and the College of Computer Science and Technology, Guizhou University, Guiyang 550025, China (e-mail: zyzhangnew@gmail.com).

D. K. Y. Yau is with the Pillar of Information Systems Technology and Design, Singapore University of Technology and Design, Singapore 487372, Singapore (e-mail: david_yau@sutd.edu.sg).

Digital Object Identifier 10.1109/JAS.2023.123252

namely $\hat{y}_i = \mathcal{M}(\boldsymbol{z}_i) = \text{sign}(f(\boldsymbol{z}_i))$, where sign$(\cdot)$ is a sign function: $\mathcal{M}(\boldsymbol{z}_i) = 1$ if $f(\boldsymbol{z}_i) > 0$ and $\mathcal{M}(\boldsymbol{z}_i) = -1$ if $f(\boldsymbol{z}_i) \le 0$. Unlike adversarial analysis in other domains such as computer vision, we postulate detection and mitigation mechanisms for rejecting bad data in the training and testing phases of the MLSA. Therefore, adversarial samples must be able to bypass these defenses before they can compromise the MLSA results.

Adversarial attack: Assume a power transmission network having a set $\mathcal{N}$ of buses. Under an ac power flow model, the relationship between the measurements $(p_i, q_i, f_{ij}^p, f_{ij}^q)$ and state variables $(\theta_i, \theta_j, v_i, v_j)$ is nonlinear and can be formulated as follows:

$$p_i = v_i^2 \sum_{j\in\mathcal{K}_i} g_{ij} - v_i \sum_{j\in\mathcal{K}_i} v_j (b_{ij}\sin\theta_{ij} + g_{ij}\cos\theta_{ij}) \tag{1}$$

$$q_i = -v_i^2 \sum_{j\in\mathcal{K}_i} b_{ij} + v_i \sum_{j\in\mathcal{K}_i} v_j (b_{ij}\cos\theta_{ij} - g_{ij}\sin\theta_{ij}) \tag{2}$$

$$f_{ij}^p = v_i^2 g_{ij} - v_i v_j (b_{ij}\sin\theta_{ij} + g_{ij}\cos\theta_{ij}) \tag{3}$$

$$f_{ij}^q = -v_i^2 b_{ij} + v_i v_j (b_{ij}\cos\theta_{ij} - g_{ij}\sin\theta_{ij}) \tag{4}$$

where p_i and q_i $(i, j \in \mathcal{N})$ denote respectively the active and reactive power injection of bus i; $\theta_{ij} = \theta_i - \theta_j$ is the phase angle difference between bus i and j; v_i and v_j are respectively the voltage magnitude of bus i and j; f_{ij}^p and f_{ij}^q are respectively the active and reactive power flows of branch $\{i, j\}$; g_{ij} and b_{ij} are respectively the conductance and susceptance of branch $\{i, j\}$; and $\mathcal{K}_i$ is the set of neighbor buses connected to bus i. Normally, the concise form of the ac model can be written as $\boldsymbol{m} = \boldsymbol{h}(\boldsymbol{v}, \boldsymbol{\theta}) + \boldsymbol{\eta}$, where $\boldsymbol{v} = \{v_1, v_2, \ldots, v_n\}$ and $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_n\}$ are the state variables; $\boldsymbol{m}$ denotes the sensor measurements of p_i, q_i; $\boldsymbol{h}(\cdot)$ represents the nonlinear relationship between the state variables and sensor measurements; and $\boldsymbol{\eta}$ is Gaussian distributed measurement noise. Generally, the state variables are estimated using the measurements $\boldsymbol{m}$ by minimizing the mean square of the residual $\hat{\boldsymbol{v}}^*, \hat{\boldsymbol{\theta}}^* = \underset{\hat{\boldsymbol{v}},\hat{\boldsymbol{\theta}}}{\text{argmin}} (\boldsymbol{m} - \boldsymbol{h}(\hat{\boldsymbol{v}}, \hat{\boldsymbol{\theta}}))^T (\boldsymbol{m} - \boldsymbol{h}(\hat{\boldsymbol{v}}, \hat{\boldsymbol{\theta}}))$, where $\hat{\boldsymbol{v}}^*$ and $\hat{\boldsymbol{\theta}}^*$ are optimal estimates of state variables $\boldsymbol{v}$ and $\boldsymbol{\theta}$. BDD is usually conducted for the state estimates to detect abnormal data, given by $\|\boldsymbol{m} - \hat{\boldsymbol{m}}\|_2 < \beta$, where $\hat{\boldsymbol{m}} = \boldsymbol{h}(\hat{\boldsymbol{v}}^*; \hat{\boldsymbol{\theta}}^*) = [\hat{\boldsymbol{p}}^T, \hat{\boldsymbol{q}}^T, \hat{\boldsymbol{f}}_p^T, \hat{\boldsymbol{f}}_q^T]^T$; $\|\cdot\|_2$ is the ℓ_2 norm of a vector; and β is a predefined threshold. If the inequality holds, the measurements bypass the BDD.

Based on the diagram in Fig. 1, suppose that the input to the MLSA model is $\boldsymbol{z} = [\hat{\boldsymbol{v}}^*; \hat{\boldsymbol{\theta}}^*; \hat{\boldsymbol{p}}; \hat{\boldsymbol{q}}; \hat{\boldsymbol{f}}_p; \hat{\boldsymbol{f}}_q]$, Correspondingly, the adversarial perturbation of $\boldsymbol{z}$ is formulated as $\boldsymbol{r} = [\Delta\hat{\boldsymbol{v}}^*; \Delta\hat{\boldsymbol{\theta}}^*; \Delta\hat{\boldsymbol{p}}; \Delta\hat{\boldsymbol{q}}; \Delta\hat{\boldsymbol{f}}_p; \Delta\hat{\boldsymbol{f}}_q]$. The adversarial perturbation on the measurement, say $\boldsymbol{m}$, is $\Delta\boldsymbol{m} = [\Delta\hat{\boldsymbol{p}}; \Delta\hat{\boldsymbol{q}}; \Delta\hat{\boldsymbol{f}}_p; \Delta\hat{\boldsymbol{f}}_q]$. Note that the attacker cannot arbitrarily change the measurement $\boldsymbol{m}$. For the adversarial perturbation to be impactful, $\Delta\boldsymbol{m}$ must correspond to feasible power data and be able to bypass bad data detection and mitigation in the power system in question.

CoRE framework: We propose a CoRE framework for MLSA models. An adversarial sample of a specific $\boldsymbol{z}_i$ must satisfy the following constraints:

1) Misclassification constraint ($\mathbf{C}_1$): The adversarial sample needs to cause a wrong ML prediction with respect to the ground truth, i.e.,

$$\mathcal{M}(\boldsymbol{z}_i + \boldsymbol{r}_i) \ne \mathcal{M}(\boldsymbol{z}_i), \quad \text{s.t. } \mathcal{M}(\boldsymbol{z}_i) = y_i \tag{5}$$

where $\boldsymbol{r}_i = \left[\Delta\hat{\boldsymbol{v}}_i^*; \Delta\hat{\boldsymbol{\theta}}_i^*; \Delta\boldsymbol{m}_i\right]$. $\Delta\hat{\boldsymbol{v}}_i^*$, $\Delta\hat{\boldsymbol{\theta}}_i^*$, and $\Delta\boldsymbol{m}_i$ denote adversarial perturbations on the state variables and measurements $\boldsymbol{m}_i$ with respect to the OC $\boldsymbol{z}_i$. Considering the linearized ML model $f(\boldsymbol{z}) = \boldsymbol{w}^T\boldsymbol{z} + b$, the misclassification constraint can be rewritten as

$$\begin{aligned} &\boldsymbol{w}^T(\boldsymbol{z}_i + \boldsymbol{r}_i) + b > 0 \text{ if } f(\boldsymbol{z}_i) \le 0, \quad \text{s.t. } \mathcal{M}(\boldsymbol{z}_i) = y_i \\ &\boldsymbol{w}^T(\boldsymbol{z}_i + \boldsymbol{r}_i) + b \le 0 \text{ if } f(\boldsymbol{z}_i) > 0, \quad \text{s.t. } \mathcal{M}(\boldsymbol{z}_i) = y_i. \end{aligned} \tag{6}$$

2) Consistency constraint ($\mathbf{C}_2$): As discussed, effective adversarial samples cannot violate physical consistency such as power balance and limits for generation, loads, power flows, etc. Power balance will be maintained if the BDD constraint (i.e., $\mathbf{C}_3$) is satisfied. To meet the power limit constraint, the adversarial perturbation should satisfy

$$\Delta\boldsymbol{m}^{\min} \le \Delta\boldsymbol{m}_i \le \Delta\boldsymbol{m}^{\max} \tag{7}$$

where $\Delta\boldsymbol{m}^{\min}$ and $\Delta\boldsymbol{m}^{\max}$ are respectively the lower and upper limits of the adversarial perturbation on the measurement.

3) BDD constraint ($\mathbf{C}_3$): Here, the spatial mutual dependency of power data is considered. The adversarial perturbations need to bypass the BDD that filters out bad data. According to [12], the BDD is circumvented through

$$\Delta\boldsymbol{m}_i = \boldsymbol{h}(\hat{\boldsymbol{v}}_i^* + \Delta\hat{\boldsymbol{v}}_i^*, \hat{\boldsymbol{\theta}}_i^* + \Delta\hat{\boldsymbol{\theta}}^*) - \boldsymbol{h}(\hat{\boldsymbol{v}}_i^*, \hat{\boldsymbol{\theta}}_i^*) \tag{8}$$

where $\hat{\boldsymbol{v}}_i^*$ and $\hat{\boldsymbol{\theta}}_i^*$ are state estimates under the measurement $\boldsymbol{m}_i$.

4) Corruption constraint ($\mathbf{C}_4$): Unlike other domains such as computer vision, inputs to the MLSA model may not be easily observed or corrupted if the power system in question is not fully open. The traditional adversarial assumption that the attacker has good knowledge of inputs to the ML model may not hold in the context of power systems, e.g., certain measurements are not known because there are no available physical and cyber channels for the attacker to access the data. Suppose the set of such data is denoted by $\mathcal{P}$. Then, the corruption constraint is defined by

$$\Delta\boldsymbol{m}_i = 0, \quad \forall i \in \mathcal{P}. \tag{9}$$

Overall, the robustness of MLSA for a specific OC $\boldsymbol{z}_i$ is evaluated by solving the following problem:

$$\begin{aligned} \delta(\boldsymbol{z}_i; f) = &\min_{\Delta\hat{\boldsymbol{v}}_i^*, \Delta\hat{\boldsymbol{\theta}}_i^*} \|\Delta\boldsymbol{m}_i\| \\ &\text{s.t. } (5)/(6) - (9) \end{aligned} \tag{10}$$

where the operation $\|\cdot\|$ can be the ℓ_0, ℓ_1, ℓ_2, and ℓ_∞ norm. We can see that the minimization problem is nonlinear and nonconvex. The *fmincon* package, a nonlinear optimizer in MATLAB, is adopted to compute the optimal solution. Overall, the RoCE of an MLSA model f is defined as

$$\xi_{\text{adv}}(f) = \sum_{i=1}^{T} \frac{\delta(\boldsymbol{z}_i; f)}{\|\boldsymbol{z}_i\|} \tag{11}$$

where T is the number of test OCs. The RoCE framework for MLSA models is given in Fig. 2.

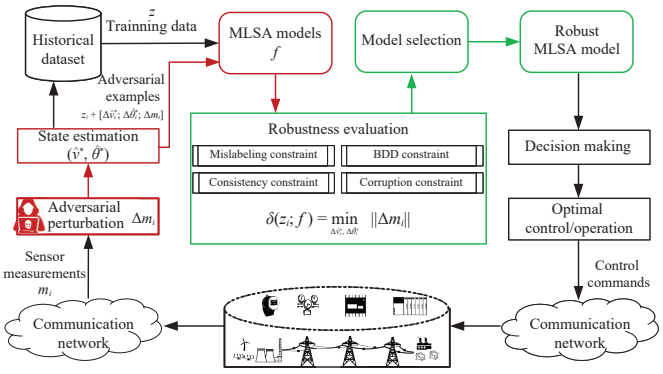

Fig. 2. The constrained robustness evaluation framework for MLSA models.

Experimental results: We conduct experiments based on an IEEE 68-bus power system [13], which consists of 68 buses, 83 transmission lines, and 16 generators. For diversity, the active loads are obtained by sampling from a multivariate Gaussian distribution with a Monte Carlo method. The nonlinear ac model is considered to calculate active and reactive power injections and power flows.

The transient stability of generators is analyzed to generate the dataset. The stability is violated if the difference between any phase angles of generators is larger than 180° at any point in time during the simulation period. We create four different three-phase line outages, corresponding to the branches {63, 62}, {59, 58}, {25, 54}, {31, 30}, respectively, to act as contingencies, denoted by **CT1**, **CT2**, **CT3**, and **CT4**. Each contingency lasts for 10 s. For each contingency, a total of 12 000 OCs are collected, in which 10 000 samples are used for training and 2000 samples for testing.

The MLSA is implemented with the support vector machine

(SVM). The overall robustness (11) is computed with 10000 OCs. The robustness value $\xi_{\text{adv}}(f)$ is calculated as an ℓ_2 norm. According to our analysis, it is possible for the RoCE problem (10) to not have any feasible solution, which means that the adversarial perturbation violates one or more of the constraints ($\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$, or $\mathbf{C}_4$). Therefore, we define an additional metric to quantify the robustness of MLSA models $\xi_{\text{roc}}(f)$ = # of robust OCs/# of test OCs, where a robust OC is one that does not have feasible adversarial perturbations according to Problem (10). Note that the value of $\xi_{\text{roc}}(f)$ is the ratio of robustness OCs among all test OCs.

First of all, the robustness of RoCE is evaluated for MLSA under different contingencies. Here the penalty factor for the SVM model is fixed to 10. We randomly select 8 measurements that cannot be corrupted for each adversarial example. Datasets of the four contingencies **CT1**, **CT2**, **CT3**, and **CT4** are used. Table 1 shows the evaluation results. We find that the robustness of MLSA increases when more constraints are added to the optimization Problem (10) under the contingencies **CT1**, **CT2**, and **CT4**. Although the robustness metric $\xi_{\text{adv}}(f)$ with the constraints $\mathbf{C}_1 \wedge \mathbf{C}_2 \wedge \mathbf{C}_3 \wedge \mathbf{C}_4$ is smaller than that with the constraints $\mathbf{C}_1 \wedge \mathbf{C}_2 \wedge \mathbf{C}_3$ under the contingency **CT3**, the robustness metric $\xi_{\text{roc}}(f)$ with the former constraints is larger than that with the latter constraints. We believe that this result is obtained because the measurements that cannot be corrupted are far away from the classification boundary.

Table 1. Robustness of Mlsa Across 4 Contingencies Under Different Constraints. "$\wedge$" Means "and"

		CT1	CT2	CT3	CT4
$\xi_{\text{adv}}(f)$	**Accuracy**	98.80%	98.34%	96.20 %	99.60%
	$\mathbf{C}_1$	0.005 46	0.003 52	0.002 28	0.017 45
	$\mathbf{C}_1 \wedge \mathbf{C}_2$	0.005 46	0.003 52	0.002 28	0.017 45
	$\mathbf{C}_1 \wedge \mathbf{C}_2 \wedge \mathbf{C}_3$	0.042 26	0.011 67	0.051 68	0.101 61
	$\mathbf{C}_1 \wedge \mathbf{C}_2 \wedge \mathbf{C}_3 \wedge \mathbf{C}_4$	0.042 27	0.012 56	0.0122	0.101 61
$\xi_{\text{roc}}(f)$	$\mathbf{C}_1$	0	0	0	0
	$\mathbf{C}_1 \wedge \mathbf{C}_2$	0	0	0	0
	$\mathbf{C}_1 \wedge \mathbf{C}_2 \wedge \mathbf{C}_3$	78.12%	85.23%	92.98%	84.52%
	$\mathbf{C}_1 \wedge \mathbf{C}_2 \wedge \mathbf{C}_3 \wedge \mathbf{C}_4$	79.20%	87.00%	96.34%	85.00%

Besides, from Table 1, it seems that the BDD constraint increases the value of $\xi_{\text{adv}}(f)$ by several times (e.g., 8, 3, 20, 9 times of that with constraints $\mathbf{C}_1 \wedge \mathbf{C}_2$ under **CT1**, **CT2**, **CT3**, and **CT4**, respectively). Since BDD is widely used in power systems, attackers face increased difficulty in constructing effective adversarial samples in practice. Table 1 also presents the number of robust OCs as the value $\xi_{\text{roc}}(f)$. It appears that the BDD and corruption constraints can reduce the attacker's ability to construct successful adversarial samples by 80%, which implies that the defenses can reject most of the adversarial samples.

Different from the prior work [10], we also analyze the impact of the corruption constraint (i.e., $\mathbf{C}_4$) on the values of $\xi_{\text{adv}}(f)$ and $\xi_{\text{roc}}(f)$. The penalty factor of SVM is fixed to 10. The dataset for the contingency **CT2** is used. All constraints $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$, and $\mathbf{C}_4$ are considered. A set of experiments are conducted by varying the set of measurements that cannot be corrupted. In Table 2, **PM1**, **PM2**, **PM3**, and **PM4** mean that there are respectively 10, 8, 5, and 3 randomly selected measurements that cannot be corrupted for each adversarial sample. We find that the values of $\xi_{\text{adv}}(f)$ and $\xi_{\text{roc}}(f)$ increase when there are more protected measurements, which indicates that the robustness of MLSA can be enhanced by the conventional wisdom of protecting critical measurements.

Table 2. Robustness of Mlsa Emphasizing on the Corruption Constraint (i.e., $\mathbf{C}_4$)

	PM1	PM2	PM3	PM4
$\xi_{\text{adv}}(f)$	0.020 00	0.012 56	0.092 56	0.092 34
$\xi_{\text{roc}}(f)$	89.76%	87.00%	77.22%	74.00%

We further evaluate the robustness of MLSA with different SVM parameters. The contingency **CT2** is used as an example and there are 8 measurements randomly selected that cannot be corrupted for each adversarial sample. From Table 3, the robustness $\xi_{\text{adv}}(f)$ of MLSA with **PF** (penalty factor) = 0.1 is the largest, whereas the robustness is the highest with **PF** = 1 according to the metric $\xi_{\text{roc}}(f)$. The robustness $\xi_{\text{adv}}(f)$ seems to decrease when the accuracy of the SVM model increases. It seems that the robustness of MLSA is affected by the model parameter, which should be carefully set considering the security issue.

Table 3. Robustness of Mlsa With Different Penalty Factors (**PF**)

		PF = 0.1	PF = 1	PF = 10
$\xi_{\text{adv}}(f)$	**Accuracy**	96.63%	96.79%	97.66%
	$\mathbf{C}_1$	0.015 23	0.009 34	0.010 60
	$\mathbf{C}_1 \wedge \mathbf{C}_2$	0.015 23	0.009 34	0.010 60
	$\mathbf{C}_1 \wedge \mathbf{C}_2 \wedge \mathbf{C}_3$	0.064 33	0.031 38	0.013 40
	$\mathbf{C}_1 \wedge \mathbf{C}_2 \wedge \mathbf{C}_3 \wedge \mathbf{C}_4$	0.075 58	0.051 68	0.012 56
$\xi_{\text{roc}}(f)$	$\mathbf{C}_1$	0	0	0
	$\mathbf{C}_1 \wedge \mathbf{C}_2$	0	0	0
	$\mathbf{C}_1 \wedge \mathbf{C}_2 \wedge \mathbf{C}_3$	87.54%	92.00%	86.12%
	$\mathbf{C}_1 \wedge \mathbf{C}_2 \wedge \mathbf{C}_3 \wedge \mathbf{C}_4$	88.45%	93.43%	87.00%

Acknowledgments: This work was supported in part by the Guizhou Provincial Science and Technology Projects (ZK[2022] 149), the Special Foundation of Guizhou University (GZU) ([2021] 47), the Guizhou Provincial Research Project for Universities ([2022]104), the GZU cultivation project of the National Natural Science Foundation of China ([2020]80), and Shanghai Engineering Research Center of Big Data Management.

References

[1] F. Thams, A. Venzke, R. Eriksson, and S. Chatzivasileiadis, "Efficient database generation for data-driven security assessment of power systems," *IEEE Trans. Power Syst.*, vol. 35, no. 1, pp. 30–41, Jan. 2019.

[2] G. Liang, S. R. Weller, J. Zhao, F. Luo, and Z. Y. Dong, "The 2015 Ukraine blackout: Implications for false data injection attacks," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 3317–3318, Apr. 2017.

[3] Z. Zhang, *et al*., "Analysis of moving target defense against false data injection attacks on power grid," *IEEE Trans. Inform. Forens. Secur.*, vol. 15, no. 1, pp. 2320–2335, Feb. 2020.

[4] Z. Zhang, *et al*., "On feasibility of coordinated time-delay and false data injection attacks on cybercphysical systems," *IEEE Internet Things J.*, vol. 9, no. 11, pp. 8720–8736, Jun. 2022.

[5] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Security Privacy*, Jun. 2017, pp. 39–57.

[6] W. Xiang, *et al*. "Verification for machine learning, autonomy, and neural networks survey," arXiv preprint arXiv: 1810.01989, 2018.

[7] Y. Chen, Y. Tan, and D. Deka, "Is machine learning in power systems vulnerable?" in *Proc. IEEE Smart Grid Comm.*, Oct. 2018, pp. 1–6.

[8] A. Venzke and S. Chatzivasileiadis, "Verification of neural network behaviour: Formal guarantees for power system applications," *IEEE Trans. Smart Grid*, vol. 12, no. 1, pp. 383–397, Jan. 2017.

[9] C. Ren and Y. Xu, "Robustness verification for machine learning-based power system dynamic security assessment models under adversarial examples," *IEEE Trans. Control Network Syst.*, 2022, DOI: 10.1109/TCNS.2022.3145285.

[10] Z. Zhang, M. Sun, R. Deng, C. Kang, and M.-Y. Chow, "Physics-constrained robustness verification of intelligent security assessment for power systems," *IEEE Trans. Power Syst.*, 2022, DOI: 10.1109/TPWRS.2022.3169139.

[11] I. Konstantelos, *et al*., "Implementation of a massively parallel dynamic security assessment platform for large-scale grids," *IEEE Trans. Smart Grid*, vol. 8, no. 3, pp. 1417–1426, Mar. 2016.

[12] J. Zhao, L. Mili, and M. Wang, "A generalized false data injection attacks against power system nonlinear state estimator and countermeasures," *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 4868–4877, May 2018.

[13] B. Pal and B. Chaudhuri, *Robust Control in Power Systems*. New York, USA: Spinger, 2006.