Defense against Joint Poisoning and Evasion Attacks: A Case Study of DERMS

Zain ul Abdeen*[1], Ahmad Al-Tawaha*[1], Padmaksha Roy*[1], Rouxi Jia[1], Laura Freeman[2], Peter Beling[2], Chen-Ching Liu[1], Alberto Sangiovanni-Vincentelli[3], Ming Jin[1]

∗ equal contributions
[1]Department of Electrical and Computer Engineering, Virginia Tech, USA,
[2]National Security Institute, Virginia Tech, USA,
[3]Department of Electrical Engineering and Computer Sciences,UC Berkeley, USA
{zabdeen, atawaha, padmaksha, ruoxijia, laura.freeman, beling,ccliu, jinming}@vt.edu, {alberto}@berkeley.edu

Abstract

There is an upward trend of deploying distributed energy resource management systems (DERMS) to control modern power grids. However, DERMS controller communication lines are vulnerable to cyberattacks that could potentially impact operational reliability. While a data-driven intrusion detection system (IDS) can potentially thwart attacks during deployment, also known as the evasion attack, the training of the detection algorithm may be corrupted by adversarial data injected into the database, also known as the poisoning attack. In this paper, we propose the *first* framework of IDS that is robust against joint poisoning and evasion attacks. We formulate the defense mechanism as a bilevel optimization, where the inner and outer levels deal with attacks that occur during training time and testing time, respectively. We verify the robustness of our method on the IEEE-13 bus feeder model against a diverse set of poisoning and evasion attack scenarios. The results indicate that our proposed method outperforms the baseline technique in terms of accuracy, precision, and recall for intrusion detection.

Introduction

With the rapid digitization of societal-scale infrastructures, power systems are gradually being transformed into cyber-physical power systems (CPPSs), also known as smart grids. The use of distribution energy resources (DERs) such as rooftop photovoltaic and energy storage systems introduces variability in operations—uncontrolled variations in power injection can induce abrupt fluctuations in nodal voltages, jeopardizing system reliability (Liu and Stewart 2021; ul Abdeen et al. 2024). Thus, distributed energy resource management systems (DERMS) are increasingly deployed to manage the potential adverse impacts of DERs on distribution feeder voltages (Jain, Sahani, and Liu 2021). The centralized DERMS controller receives data streams from advanced metering infrastructure (AMI) and then decides upon optimal real and reactive power dispatch settings for inverter-based DERs (Dall'Anese, Dhople, and Giannakis 2014). However, the heavy reliance on communications exposes the system to cyberattacks (Case 2016). By targeting the DERMS communication channels, attackers can initiate falsified dispatch commands that cause severe voltage disturbances and damage substations or household equipment.

Cyber-vulnerability makes it imperative to study assessment and defense strategies (Ike et al. 2022). The denial-of-service (DoS) attack (Chen et al. 2022) and false data injection attack (FDIA) (Jafarigiv et al. 2021) are two commonly analyzed attacks on the DERMS controller. Methods to detect and mitigate these attacks in the cyber-layer have also been investigated (Huseinović et al. 2020; Raja et al. 2022), including recent works with machine learning (Guo et al. 2021; Hasnat and Rahnamay-Naeini 2021; Nguyen et al. 2021). Besides DoS and FDIA, a relatively low-probability but high-severity attack involves modifications to the DERMS controller algorithm after gaining unauthorized access (Jain, Sahani, and Liu 2021). As the software can be altered to disguise malicious command data packets, such attacks can be difficult to detect with a centralized method (Sun et al. 2021). To counteract, decentralized inverter-based IDSs are proposed in (Jain, Sahani, and Liu 2021; Urbina et al. 2016); specifically, a regression model for expected control commands is trained with historical data and the prediction error is subsequently leveraged for evasion attack detection. Nevertheless, a crucial vulnerability persists: the historical data may be adversarially manipulated by a data poisoning attack; as a consequence, the trained model may trigger false alarms or miss attack events when deployed in test time (Tian et al. 2022). This calls for an IDS that is robust to attacks that may occur at different stages.

In this paper, we focus on the challenging scenario where both poisoning (training phase) and evasion (testing phase) attacks can be staged. For the development of the defense mechanism, our key insight is that as the trained model will be used subsequently for evasion attack detection, such model should be trained robustly and in an end-to-end fashion. The contributions are summarized as follows:

- Development of a model-based IDS against joint poisoning and evasion attacks on DERMS;
- Formulation of a bilevel optimization problem, where the inner level robustly learns a model and the outer level finds an optimal threshold for the model-based prediction error;
- Evaluation of the proposed method in a range of attack scenarios and demonstration of improved robustness against the baseline method.

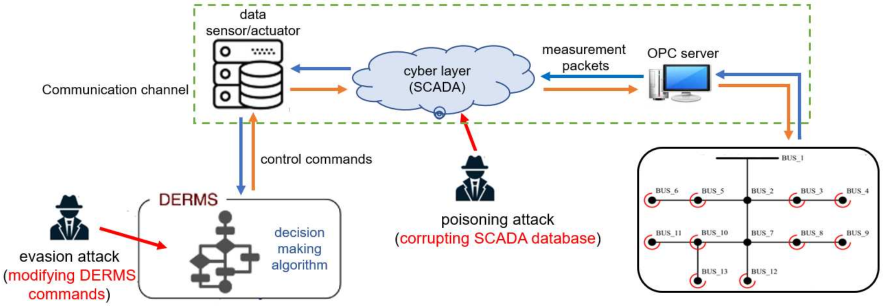

Figure 1: Cyber-physical power system and its vulnerability.

The rest of the paper is organized as follows. Section presents related works on cyber-defense mechanisms for power grids. Section discusses backgrounds on CPPS and the attack model. The defense strategy is presented in Section . Section conducts numerical evaluation of the proposed method and discusses the results. Finally, Section concludes the work.

Related work

Cyberattacks on power grids include DoS, delay attacks, man-in-the-middle attacks, replay attacks, FDIAs, just to name a few (Peng et al. 2019; Tuyen et al. 2022). Various defense mechanisms have been proposed (Sun, Hahn, and Liu 2018; Peng et al. 2019), among which machine learning (ML) techniques are promising (Berghout, Benbouzid, and Muyeen 2022). Existing data-driven IDSs can be categorized into supervised learning (Wang et al. 2019), semi-supervised learning (Farajzadeh-Zanjani et al. 2021), unsupervised learning (Karimipour et al. 2019), self-supervised learning (Zhang et al. 2021), and reinforcement learning (Kurt et al. 2018). To improve data efficiency, model-based defense mechanisms can leverage physics and have been developed to detect evasion attacks (Karimipour and Dinavahi 2017; Jain, Sahani, and Liu 2021).

Nevertheless, the effectiveness of ML-based IDS can be significantly reduced by poisoning attacks, which have become an emerging threat (Tian et al. 2022). Different from an evasion attack that occurs during deployment (test time), a poisoning attack can misguide the model training by manipulating the training data, thus yielding a falsified model for deployment. Defending against poisoning attacks is challenging and less investigated for power system cybersecurity (Zografopoulos, Konstantinou, and Hatziargyriou 2022). Furthermore, very few works have addressed the scenario of joint poisoning and evasion attacks, leaving a substantial gap in the literature. The challenge is to reason about the propagation of error induced by poisoning attacks to test time performance and then design a learning framework that is robust to such error propagation.

The present study initiates the first study in this important direction. Our technique hinges on bilevel optimization (Liu et al. 2021). The methodology design is inspired by the recent line of research on end-to-end optimization (Kotary et al. 2021), which provides a principled way to design the training of a model in view of its consequent usage during test time.

Power system and attack model

Cyber-physical power system

The physical layer of a CPPS consists of the feeders and DERMS controller and the cyber layer represents the information and communication technology (ICT) or the supervisory control and data acquisition (SCADA) system and the communication paths for data exchange (see Fig. 1 for an illustration with the IEEE-13 bus feeder model) (Jain, Sahani, and Liu 2021). The open platform communication (OPC) server receives measurement packets from the feeder and prepares the data for SCADA to access in the cyber layer. Measurement data are sent to the DERMS controller through a firewall to check for discrepancies. The control actions, such as the optimal real and reactive power setpoints, are computed by the DERMS controller and sent back to the feeder for actuation. An IDS is deployed within the cyber layer to constantly check for the integrity of data and control commands.

Attack model

Due to the heavy reliance on ICT, the attack surface is wide in practice and may include data integrity injection at substations and over communication links, distributed attacks by manipulating endpoint devices such as smart meters and smart appliances, or even a more powerful attack such as spear phishing attacks that gain access to communication paths or modify the DERMS controller software (Adepu, Kandasamy, and Mathur 2018). The goal of the attacker can be to falsify the dispatch control commands to cause voltage violations and exact damage to the physical systems. Nevertheless, among all the possibilities, some of the attacks can be more severe (e.g., taking over control centers) than others (e.g., attacking smart appliances). As a consequence, the attack may range in severity due to the ability of the attacker.

In this study, we consider two types of threats: evasion attacks and poisoning attacks. While these threats can be implemented with one or a combination of the aforementioned attacks, the key difference is the time when the attack is staged: an evasion attack occurs during deployment to evade IDS, whereas a poisoning attack may be conducted in an earlier stage during model training to corrupt the IDS.

Our assumption of the attacker is comparable to the existing works on data integrity attacks (Liang et al. 2016); in particular, we assume that the data used for training or during actual operations can be maliciously manipulated. We remark that the attacker considered in our study is stronger than some existing works on model-based defense (Jain, Sahani, and Liu 2021; Ghaeini et al. 2018) in the sense that prior works focus on evasion attacks while we consider the additional mode of poisoning attacks. This stronger attack model seems more practical due to the increasing use of ML in modern IDS and the various security loopholes in database systems (Tian et al. 2022; Ike et al. 2022).

Defense strategy

Decentralized detection

In model-based IDSs, the expected command is compared against the actual command received, and an anomaly is detected when the difference between these values is larger than a threshold (Jafarigiv et al. 2021; Jain, Sahani, and Liu 2021). As the cyberattack may directly target the DERMS software to make malicious data packets appear legitimate, a centralized IDS may be evaded, while a decentralized approach that uses locally available measurements may be more difficult to deceive. While our framework can incorporate more complex models such as neural networks (Jafarigiv et al. 2021; ul Abdeen et al. 2022), due to limited computational power at the inverter level, a simple model such as linear regression is preferred. Following (Jain, Sahani, and Liu 2021; Zeng et al. 2021), a linear regression model is used to predict the expected control commands based on local load and maximum charging and discharging rate values. For instance, the expected control command for real power dispatch set point is given by

$$P_D^{pred} = \alpha_D^{1,p} + \alpha_D^{2,p} * p_L + \alpha_D^{3,p} * q_L + \alpha_D^{4,p} * p_{Dmax}, \quad (1)$$

where p_L and q_L represent the active and reactive power demands, respectively, and p_{Dmax} is the maximum generation limit of the inverter. Here, $\{\alpha_D^{j,p}\}_{j=1,\ldots,4}$ are coefficients of the regression model. Note that we can write (1) as $P_D^{pred} = \alpha \cdot x$, where $x = [1, p_L, q_L, p_{Dmax}]^\top$ and $\alpha = [\alpha_D^{1,p}, \alpha_D^{2,p}, \alpha_D^{3,p}, \alpha_D^{4,p}]^\top$.

In the following, we denote x_i as the feature vector for data point i (so $\alpha \cdot x_i$ is the expected command) and p_i as the actual command. For each data point (x_i, p_i), if the absolute difference $|\alpha \cdot x_i - p_i|$ between the expected and actual commands is larger than a threshold τ, then we consider that an anomaly has occurred; otherwise, the data point is considered normal. Similar models can be instantiated for other control commands, such as reactive power dispatch set points for PV inverters and charging or discharging rates for energy storage inverters. To streamline the presentation, we will focus on the real power dispatch set point for illustration.

Bilevel formulation of defense

Problem setup. Let $\mathcal{D}_1 = \{(x_i, p_i)\}_{i=1}^{n_1}$ be an unlabeled dataset, where $x_i \in \mathbb{R}^4$ is the feature vector and p_i is the actual command. Suppose we also have access to a dataset that contains labels regarding whether an attack has occurred, i.e., $\mathcal{D}_2 = \{(x_i, p_i, y_i)\}_{i=1}^{n_2}$, where $y_i \in \{-1, +1\}$ is the label with $+1$ indicating the anomaly and -1 indicating the normal condition. In practice, as cyberattack data are rare and difficult to obtain, we expect that the amount of unlabeled data to be much larger than the amount of cyberattack data, namely $n_1 \gg n_2$. Based on our attack model, a certain (but unknown) percentage of the dataset $\mathcal{D}_1$ may be poisoned; thus the actual measurements p_i cannot be trusted. We assume that the labels y_i in $\mathcal{D}_2$ are authentic, since they are often carefully cross-checked by experts, although it is possible to extend our framework to consider corrupted labels as well.

Due to the presence of poisoned data during training, the conventional pipeline that first estimates the model parameter α with $\mathcal{D}_1 \cup \mathcal{D}_2$ and then uses the learned model to detect evasion attacks may no longer be effective (Jain, Sahani, and Liu 2021; Jafarigiv et al. 2021). Our strategy is differentiated from prior works in two-fold: *1)* the training algorithm to obtain α^* should be robust to poisoning attacks on $\mathcal{D}_1$, and *2)* as α^* is used in the downstream decision task—evasion attack detection—so the search of the prediction model should be aware of this task. We address these two aspects as follows.

Robust training against poisoning attacks. To design a robust training algorithm, we formulate the following optimization problem:

$$\arg\min_{\alpha,\delta} \frac{1}{2} \sum_{(x_i,p_i)\in\mathcal{D}_1} (\alpha \cdot x_i - p_i + \delta_i)^2 + \lambda \|\delta\|_1, \quad (2)$$

where $\|\cdot\|_1$ is the ℓ_1 norm, λ is a hyperparameter, and $\delta = [\delta_1, \ldots, \delta_{n_1}]^\top$ is hypothetical bad data vector, which is introduced to counterbalance the potential attacks on p_i. The training loss consists of two terms: the squared loss of reconstruction error and a penalty on the sparsity of δ. The overall problem is convex; in fact, it is strongly convex due to the presence of the squared loss, thus the optimal solution is unique. Under certain conditions, it has been shown that we can exactly recover the poisoned data by solving (2) (Jin et al. 2020). However, it is difficult to determine the best way to set λ—a larger value may induce a sparser δ but also a higher loss on the reconstruction error, and vice versa. While prior works set this number by hand, we propose to set this number so that it supports the ultimate task assigned for the model: evasion attack detection.

Task-aware learning for evasion attack detection. The model parameter α is used to detect evasion attacks by checking the prediction error, which provides a proper goal to guide the search of hyperparameter λ. Furthermore, the detection threshold τ needs to be tuned to support this task. To this end, a bilevel optimization problem is formulated:

$$\begin{aligned} \min_{\lambda,\tau} \quad & \sum_{(x_i,y_i,p_i)\in\mathcal{D}_2} \ell(|\bar{\alpha} \cdot x_i - p_i| - \tau, y_i) \\ \text{s. t.} \quad & (\bar{\alpha}, \delta^*) \text{ is the optimal solution to (2)} \end{aligned} \quad (3)$$

where $\ell(t, y) = \log(1 + \exp(-ty))$ is the logistic loss. In the above formulation, the inner level determines the model parameters $\bar{\alpha}$ and hypothetical bad data vector δ^*, while the outer level determines the hyperparameter λ used within the inner level and the detection threshold τ. The optimal τ depends on the learned model $\bar{\alpha}$, which in turn depends on the hyperparameter λ. Since the inner-level problem variable is included in the upper-level problem, in the case of poisoning attacks, the attack error may propagate into the evasion attack performance. Thus, the outer level also plays a role in rectifying the learned model to ensure that τ properly accounts for the potential corrupted model. Above all, the outer-level has fewer decision parameters than the inner-level, which agrees with the imbalanced data sizes $(n_1 \gg n_2)$.

Algorithm

To solve (3), we can use gradient descent on the outer-level variables, while solving the inner-level problem exactly in each iteration (see Algorithm 1). Let $L(\lambda, \tau) = \sum_{(x_i,y_i,p_i)\in\mathcal{D}_2} \ell(|\bar{\alpha}\cdot x_i - p_i| - \tau, y_i)$ denote the outer-level objective. The gradients of L with respect to λ and τ are given by:

$$\frac{\partial L}{\partial \tau} = \sum_{(x_i,y_i,p_i)\in\mathcal{D}_2} \frac{y_i \exp(-y_i(|r_i| - \tau))}{1+\exp(-y_i(|r_i|-\tau))}, \tag{4}$$

and

$$\frac{\partial L}{\partial \lambda} = \sum_{(x_i,y_i,p_i)\in\mathcal{D}_2} \frac{-y_i \mathrm{sign}(r_i)\exp(-y_i(|r_i|-\tau))}{1+\exp(-y_i(|r_i|-\tau))} x_i^\top \frac{\partial \bar{\alpha}(\lambda)}{\partial \lambda}, \tag{5}$$

where $r_i = \bar{\alpha}\cdot x_i - p_i$ and $\mathrm{sign}(r_i) = 1$ if $r_i \geq 0$ and 0 otherwise. As $\bar{\alpha}$ is a function of λ, the key is to obtain the gradient $\frac{\partial \bar{\alpha}(\lambda)}{\partial \lambda}$.

Implicit gradient. The difficulty of obtaining the implicit gradient from the usual Karush–Kuhn–Tucker (KKT) conditions is due to the presence of $\|\cdot\|_1$, which is not differentiable. In the following, we provide a closed-form solution to the implicit gradient $\frac{\partial \bar{\alpha}(\lambda)}{\partial \lambda}$. We start with a proposition that reformulate the inner-level problem as an optimization over the Huber loss:

$$f_{\text{Huber}}(z;\lambda) = \begin{cases} \frac{1}{2}z^2 & |z| \leq \lambda \\ \lambda(|z| - \frac{1}{2}\lambda) & |z| > \lambda \end{cases}.$$

Proposition 1. *Suppose that* $(\bar{\alpha}, \delta^*)$ *is the solution to* (2) *and let* $\bar{\alpha}'$ *be the solution to* $\min_\alpha \sum_{(x_i,p_i)\in\mathcal{D}_1} f_{\mathit{Huber}}(\alpha \cdot x_i - p_i; \lambda)$*. Then, we have* $\bar{\alpha} = \bar{\alpha}'$*, and the i-th component of* δ^* *is given by:*

$$\delta_i^* = \mathrm{sign}(p_i - \bar{\alpha}\cdot x_i)\max(0, |\bar{\alpha}\cdot x_i - p_i| - \lambda).$$

The implication of the above result is that we can eliminate the inner-level variable δ and exclusively focus on α by changing the loss function. The following result provides the closed-form solution to the implicit gradient.

Theorem 1. *Suppose that* $(\bar{\alpha}, \delta^*)$ *is the solution to* (2)*. Let* $\mathcal{I}_1 = \{i : |\bar{\alpha}\cdot x_i - p_i| < \lambda\}$*,* $\mathcal{I}_2 = \{i : \bar{\alpha}\cdot x_i - p_i \leq -\lambda\}$*, and* $\mathcal{I}_3 = \{i : \bar{\alpha}\cdot x_i - p_i \geq \lambda\}$ *be partition of dataset* $\mathcal{D}_1$*. Additionally, let* $A = \sum_{i\in\mathcal{I}_1} x_i x_i^\top \in \mathbb{R}^{d\times d}$ *and suppose that A is invertible. Then, we have that*

$$\frac{\partial \bar{\alpha}}{\partial \lambda} = \lambda A^{-1}\left(\sum_{i\in\mathcal{I}_2} x_i - \sum_{i\in\mathcal{I}_3} x_i\right). \tag{6}$$

Algorithm 1: Bilevel optimization algorithm for (3)

```
Input: stepsize $\beta_\tau$ and $\beta_\lambda$, iterations $K$, initial values of $\tau_1$
    and $\lambda_1$
for $k = 1, \ldots, K$ do
    Solve the lower level problem (2) to obtain $\alpha_k$
    Obtain $\frac{\partial L}{\partial \tau}$ and $\frac{\partial L}{\partial \lambda}$ at $\tau_k$ and $\lambda_k$ using 4 and 5, respectively
    Update the value of $\tau$ and $\lambda$ using gradient descent
        $\tau_{k+1} = \tau_k - \beta_\tau \frac{\partial L}{\partial \tau},$
        $\lambda_{k+1} = \lambda_k - \beta_\lambda \frac{\partial L}{\partial \lambda}$
end for
Output: $\tau_K$ and $\alpha_K$
```

Numerical evaluation

Experimental setup. We follow the same procedure as (Jain, Sahani, and Liu 2021) to obtain the datasets, with $n_1 = 1000$ for the unlabeled dataset $\mathcal{D}_1$ and $n_2 = 200$ for the labelled dataset D_2. In $\mathcal{D}_1$, we consider the cases where 10% and 30% of the data are poisoning attacked. Dataset $\mathcal{D}_2$ consists of 20% of data with label $+1$, i.e., anomaly. We also vary the levels of corruption by changing the true measurement of p_i by the percentages of 40%,70%, and 100%, with random noises of small magnitudes added upon the obtained values.

During the training stage, we solve (3) with the datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ to obtain the solution (α^*, τ^*). During testing, we use the decentralized detection method outlined in Sec. to detect evasion attacks. We evaluate the performance of our method in terms of metrics including the accuracy, precision and recall. Specifically, let TP, TN, FP, and FN denote the true positives, true negatives, false positives, and false negatives, respectively. Then, we have that

$$\text{accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}},$$
$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}}, \quad \text{recall} = \frac{\text{TP}}{\text{TP+FN}}.$$

Baseline method. As a comparison, we also implement the approach from (Jain, Sahani, and Liu 2021). To briefly recap their method, a model parameter $\bar{\alpha}$ is first learned by solving a standard least-square regression problem on $\mathcal{D}_1$. Then, the threshold τ is manually designed based on the obtained $\bar{\alpha}$. To make a fair comparison, we also fine-tune the threshold based on $\mathcal{D}_2$.

Results and discussions. Tables 1 and 2 show the performance metrics of our proposed approach and the baseline method corresponding to 10% and 30% of poisoning attacks on $\mathcal{D}_1$, respectively. We report the mean and the standard deviation over 10 independent runs. In general, it can be observed that the proposed method has improved accuracy, precision, and recall compared to the baseline. The improvement is more substantial in the case where 30% of $\mathcal{D}_1$ are poisoning attacked (Table 2). This is expected as the baseline method uses linear regression to learn the model, which is well-known to be vulnerable to outliers or adversarial data.

As the *evasion attack magnitudes* increase from 40% to 100%, an interesting trend can be observed that the performance of each method (ours and baseline) increases. This is because for attacks with larger magnitudes, the differences between the expected and actual commands may easily surpass the detection threshold, even if the prediction model is not reliable.

As the *poisoning attack magnitudes* increase from 40% to 100%, there is a clear trend that the performance of the baseline method drops. In contrast, in many cases, we can actually observe a slight increase in performance as the poisoning attack magnitudes increase from 40% to 70%. This benefits from the robust training procedure in the inner-level problem, which can more easily detect adversarial data with a large deviation from normal. However, as the attack magnitude further increases from 70% to 100%, even a few undetected outliers may significantly bias the training outcome, thus we can see a slight decrease in performance in some cases.

For both methods, we observe a higher precision than the recall. This can be attributed to the fact that the datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ have imbalanced labels—the amount of anomaly data is fewer that the amount of normal data. This is generally to be expected, as anomalies often occur rarely. It may be an interesting direction for future work to test methods such as cost-sensitive learning or X-risk optimization to optimize for compositional measures (Yang 2022).

Last, we visualize the performance of the IDSs in a case with 70% and 100% attack magnitudes by evasion and poisoning attacks, respectively, as shown in Fig 2. As shown in the top figure that plots the difference between actual and expected commands, there are many instances of disagreements between our method and the baseline. Further examinations in the bottom two subplots indicate that in most cases, our method is able to accurately detect anomalies while avoiding false positives. In contrast, the baseline methods create multiple instances of false positives and false negatives, due to the corrupted prediction model affected by the poisoning attacks.

Conclusion

Cybersecurity is a tug of war—as the attacker's capability grows, so should the defender's, and it has strategic value in assuming a strong attacker so the defense can be assessed and designed commensurately. In this paper, we envisioned joint poisoning and evasion attacks on both the cyber and physical layers of a power system, a scenario that has not been systematically studied in the literature. As a countermeasure, we design a defense mechanism by formulating a bilevel optimization problem, where the inner level and outer level work jointly but with different goals. In particular, the inner-level problem accounts for robust training against poisoning attacks, whereas the outer-level problem addresses evasion attacks by guiding the inner-level training through implicit gradients. The robustness of the method is evaluated under different attack scenarios and compared with a baseline model. In the future, we plan to evaluate our model on a real-time digital simulator for power systems. Another interesting direction is to learn a nonlinear model through convexification to account for AC power flows (Jin, Lavaei, and Johansson 2018).

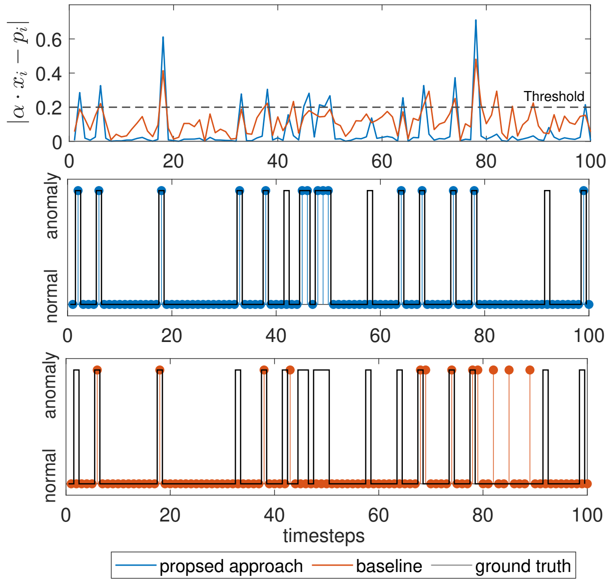

Figure 2: Visualization of the detection performance. Top plot: time series of the difference between predicted and actual commands (blue: ours, red: baseline). Middle/bottom plots: detection of anomaly based on the proposed method-/baseline technique (stem plots). The ground truth is marked with the square wave.

Acknowledgment

This project is funded by This research was supported in part by the C3.ai Ditial Transformation Institute, and by the Director of Cybersecurity, Energy Security, and Emergency Response, specifically through the Cybersecurity for Energy Delivery Systems program of the U.S. Department of Energy under contract DE-CR0000021.

References

Adepu, S.; Kandasamy, N. K.; and Mathur, A. 2018. Epic: An electric power testbed for research and training in cyber physical systems security. In *Computer Security*, 37–52. Springer.

Berghout, T.; Benbouzid, M.; and Muyeen, S. 2022. Machine learning for cybersecurity in smart grids: A comprehensive review-based study on methods, solutions, and prospects. *International Journal of Critical Infrastructure Protection*, 100547.

Case, D. U. 2016. Analysis of the cyber attack on the Ukrainian power grid. *Electricity Information Sharing and Analysis Center (E-ISAC)*, 388: 1–29.

Chen, X.; Zhou, J.; Shi, M.; Chen, Y.; and Wen, J. 2022. Distributed resilient control against denial of service attacks in DC microgrids with constant power load. *Renewable and Sustainable Energy Reviews*, 153: 111792.

Dall'Anese, E.; Dhople, S. V.; and Giannakis, G. B. 2014. Optimal dispatch of photovoltaic inverters in residential dis-

		40% evasion attack magnitude			70% evasion attack magnitude			100% evasion attack magnitude		
method	poisoning	accuracy	precision	recall	accuracy	precision	recall	accuracy	precision	recall
proposed	40%	**86.4(2.4)**	**93.9(1.25)**	**59.3(3.5)**	**93.8(3.8)**	**96.3(2.2)**	**85.5(6.0)**	**93.1(2.5)**	**96.0(1.5)**	**82.1(5.7)**
approach	70%	**84.4(3.7)**	**91.8(1.8)**	**60.7(4.3)**	**94.5(2.5)**	**96.8(1.4)**	**85.0(7.1)**	**95.4(1.7)**	**97.3(1.0)**	**88.0(3.7)**
	100%	**85.1(3.1)**	**92.2(1.7)**	**61.5(3.1)**	**92.6(1.8)**	**95.8(1.0)**	**79.5(5.1)**	**94.6(2.0)**	**96.8(1.1)**	**86.9(5.7)**
baseline	40%	85.3(2.5)	86.2(16.7)	55.8(4.2)	93.6(3.9)	96.3(2.2)	85.1(5.5)	92.2(3.3)	95.6(1.9)	80.0(6.5)
approach	70%	82.9(3.5)	78.7(22.0)	56.7(5.3)	93.8(2.1)	96.5(1.4)	83.6(6.0)	95.0(1.4)	97.1(1.0)	87.0(2.8)
	100%	82.5(3.2)	78.6(21.6)	54.4(3.2)	91.1(1.3)	95.1(1.0)	79.5(5.2)	93.4(2.1)	95.7(2.0)	84.4(4.9)

Table 1: Performance of the proposed method and baseline method with 10% of dataset $\mathcal{D}_1$ under poisoning attacks. Each row indicate the performance of the corresponding method when the training data is poison attacked with magnitude 40%, 70%, or 100%. The mean and the standard deviation (in paranthesis) are reported over 10 independent runs. We mark better performance measures by boldface.

		40% evasion attack magnitude			70% evasion attack magnitude			100% evasion attack magnitude		
method	poisoning	accuracy	precision	recall	accuracy	precision	recall	accuracy	precision	recall
proposed	40%	**82.8(3.4)**	**91.06(1.8)**	**58.8(4.3)**	**94.1(2.2)**	**96.6(1.3)**	**83.7(4.1)**	**94.0(2.8)**	**96.5(1.7)**	**84.7(5.3)**
approach	70%	**82.6(4.1)**	**90.8(2.1)**	**61.0(3.4)**	**95(1.31)**	**97.1(0.7)**	**85.6(3.6)**	**92.0(3.5)**	**95.3(2.1)**	**82.1(4.8)**
	100%	**83.6(2.4)**	**91.3(1.2)**	**63.2(3.2)**	**93.4(2.3)**	**96.2(1.2)**	**83.2(6.3)**	**92.9(2.5)**	**95.9(1.4)**	**81.6(6.5)**
baseline	40%	81.5(3.6)	84.3(16.5)	55.6(3.4)	91.8(3.6)	95.4(2.0)	78.0(5.6)	93.7(3.0)	96.4(1.8)	84.1(5.2)
approach	70%	75.3(5.4)	48.1(18.5)	48.8(2.2)	87.5(4.2)	79.8(11.2)	73.4(8.5)	89.2(4.0)	91.8(4.7)	76.9(5.4)
	100%	71.1(3.6)	44.5(2.4)	47.3(1.6)	76.9(9.9)	65.3(11.7)	59.0(10.1)	83.5(4.9)	75.3(9.2)	69.4(8.2)

Table 2: Performance of the proposed method and baseline method with 30% of dataset $\mathcal{D}_1$ under poisoning attacks. See Table 1 for other descriptions.

tribution systems. *IEEE Transactions on Sustainable Energy*, 5(2): 487–497.

Farajzadeh-Zanjani, M.; Hallaji, E.; Razavi-Far, R.; Saif, M.; and Parvania, M. 2021. Adversarial semi-supervised learning for diagnosing faults and attacks in power grids. *IEEE Transactions on Smart Grid*, 12(4): 3468–3478.

Ghaeini, H. R.; Antonioli, D.; Brasser, F.; Sadeghi, A.-R.; and Tippenhauer, N. O. 2018. State-aware anomaly detection for industrial control systems. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, 1620–1628.

Guo, L.; Zhang, J.; Ye, J.; Coshatt, S. J.; and Song, W. 2021. Data-Driven Cyber-Attack Detection for PV Farms via Time-Frequency Domain Features. *IEEE Transactions on Smart Grid*, 13(2): 1582–1597.

Hasnat, M. A.; and Rahnamay-Naeini, M. 2021. Detecting and locating cyber and physical stresses in smart grids using the k-nearest neighbour analysis of instantaneous correlation of states. *IET Smart Grid*, 4(3): 307–320.

Huseinović, A.; Mrdović, S.; Bicakci, K.; and Uludag, S. 2020. A survey of denial-of-service attacks and solutions in the smart grid. *IEEE Access*, 8: 177447–177470.

Ike, M.; Phan, K.; Sadoski, K.; Valme, R.; and Lee, W. 2022. SCAPHY: Detecting Modern ICS Attacks by Correlating Behaviors in SCADA and PHYsical. In *2023 IEEE Symposium on Security and Privacy (SP)*, 362–379. IEEE Computer Society.

Jafarigiv, D.; Sheshyekani, K.; Kassouf, M.; Seyedi, Y.; Karimi, H.; and Mahseredjian, J. 2021. Countering FDI Attacks on DERs Coordinated Control System Using FMI-Compatible Cosimulation. *IEEE Transactions on Smart Grid*, 12(2): 1640–1650.

Jain, A. K.; Sahani, N.; and Liu, C.-C. 2021. Detection of Falsified Commands on a DER Management System. *IEEE Transactions on Smart Grid*, 13(2): 1322–1334.

Jin, M.; Lavaei, J.; and Johansson, K. H. 2018. Power grid AC-based state estimation: Vulnerability analysis against cyber attacks. *IEEE Transactions on Automatic Control*, 64(5): 1784–1799.

Jin, M.; Lavaei, J.; Sojoudi, S.; and Baldick, R. 2020. Boundary defense against cyber threat for power system state estimation. *IEEE Transactions on Information Forensics and Security*, 16: 1752–1767.

Karimipour, H.; Dehghantanha, A.; Parizi, R. M.; Choo, K.-K. R.; and Leung, H. 2019. A deep and scalable unsupervised machine learning system for cyber-attack detection in large-scale smart grids. *IEEE Access*, 7: 80778–80788.

Karimipour, H.; and Dinavahi, V. 2017. Robust massively parallel dynamic state estimation of power systems against cyber-attack. *IEEE Access*, 6: 2984–2995.

Kotary, J.; Fioretto, F.; Van Hentenryck, P.; and Wilder, B. 2021. End-to-end constrained optimization learning: A survey. *arXiv preprint arXiv:2103.16378*.

Kurt, M. N.; Ogundijo, O.; Li, C.; and Wang, X. 2018. Online cyber-attack detection in smart grid: A reinforcement learning approach. *IEEE Transactions on Smart Grid*, 10(5): 5174–5185.

Liang, G.; Zhao, J.; Luo, F.; Weller, S. R.; and Dong, Z. Y. 2016. A review of false data injection attacks against modern power systems. *IEEE Transactions on Smart Grid*, 8(4): 1630–1638.

Liu, C.-C.; and Stewart, E. M. 2021. Electricity Transmission System Research and Development: Distribution Integrated with Transmission Operations.

Liu, R.; Gao, J.; Zhang, J.; Meng, D.; and Lin, Z. 2021. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Nguyen, B. L.; Vu, T. V.; Guerrero, J. M.; Steurer, M.; Schoder, K.; and Ngo, T. 2021. Distributed dynamic state-input estimation for power networks of Microgrids and active distribution systems with unknown inputs. *Electric Power Systems Research*, 201: 107510.

Peng, C.; Sun, H.; Yang, M.; and Wang, Y.-L. 2019. A survey on security communication and control for smart grids under malicious cyber attacks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 49(8): 1554–1569.

Raja, D. J. S.; Sriranjani, R.; Parvathy, A.; and Hemavathi, N. 2022. A Review on Distributed Denial of Service Attack in Smart Grid. In *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, 812–819. IEEE.

Sun, C.-C.; Hahn, A.; and Liu, C.-C. 2018. Cyber security of a power grid: State-of-the-art. *International Journal of Electrical Power & Energy Systems*, 99: 45–56.

Sun, R.; Mera, A.; Lu, L.; and Choffnes, D. 2021. SoK: Attacks on industrial control logic and formal verification-based defenses. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, 385–402. IEEE.

Tian, Z.; Cui, L.; Liang, J.; and Yu, S. 2022. A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning. *ACM Computing Surveys (CSUR)*.

Tuyen, N. D.; Quan, N. S.; Linh, V. B.; Vu, T. V.; and Fujita, G. 2022. A Comprehensive Review of Cybersecurity in Inverter-based Smart Power System amid the Boom of Renewable Energy. *IEEE Access*.

ul Abdeen, Z.; Yin, H.; Kekatos, V.; and Jin, M. 2022. Learning neural networks under input-output specifications. In *2022 American Control Conference (ACC)*, 1515–1520. IEEE.

ul Abdeen, Z.; Zhang, X.; Gill, W.; and Jin, M. 2024. Enhancing Distribution System Resilience: A First-Order Meta-RL algorithm for Critical Load Restoration. In *2024 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, 129–134. IEEE.

Urbina, D. I.; Giraldo, J. A.; Cardenas, A. A.; Tippenhauer, N. O.; Valente, J.; Faisal, M.; Ruths, J.; Candell, R.; and Sandberg, H. 2016. Limiting the impact of stealthy attacks on industrial control systems. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 1092–1105.

Wang, D.; Wang, X.; Zhang, Y.; and Jin, L. 2019. Detection of power grid disturbances and cyber-attacks based on machine learning. *Journal of information security and applications*, 46: 42–52.

Yang, T. 2022. Algorithmic Foundation of Deep X-Risk Optimization. *arXiv preprint arXiv:2206.00439*.

Zeng, Y.; Chen, S.; Park, W.; Mao, Z. M.; Jin, M.; and Jia, R. 2021. Adversarial unlearning of backdoors via implicit hypergradient. *arXiv preprint arXiv:2110.03735*.

Zhang, J.; Pan, L.; Han, Q.-L.; Chen, C.; Wen, S.; and Xiang, Y. 2021. Deep learning based attack detection for cyber-physical system cybersecurity: A survey. *IEEE/CAA Journal of Automatica Sinica*, 9(3): 377–391.

Zografopoulos, I.; Konstantinou, C.; and Hatziargyriou, N. D. 2022. Distributed Energy Resources Cybersecurity Outlook: Vulnerabilities, Attacks, Impacts, and Mitigations. *arXiv preprint arXiv:2205.11171*.

Proofs

Proof of Proposition 1

Given α, the optimization with respect to δ can be decomposed into a series of smaller optimization problems:

$$\min_{\delta_i} \quad \frac{1}{2}(\alpha \cdot x_i - p_i + \delta_i)^2 + \lambda|\delta_i|, \tag{7}$$

for each $i = 1, .., n_1$, which has a closed-form solution

$$\delta_i^* = \operatorname{sign}(p_i - \alpha \cdot x_i) \max(0, |\alpha \cdot x_i - p_i| - \lambda). \tag{8}$$

Plugging in the above into the objective (7), we can see that the objective is equal to:

$$\frac{1}{2}(\alpha \cdot x_i - p_i + \delta_i)^2 + \lambda|\delta_i| = f_{\text{Huber}}(\alpha \cdot x_i - p_i; \lambda), \tag{9}$$

Proof of Theorem 1

Note that the subgradient of Huber loss is given by:

$$\frac{\partial}{\partial z} f_{\text{Huber}}(z; \lambda) = \begin{cases} z & |z| < \lambda \\ -\lambda & z \leq -\lambda \\ \lambda & z \geq \lambda \end{cases} \tag{10}$$

Then, by the KKT conditions:

$$\sum_{i \in \mathcal{I}_1} (\bar{\alpha} \cdot x_i - p_i) x_i - \lambda \left(\sum_{i \in \mathcal{I}_2} x_i - \sum_{i \in \mathcal{I}_3} x_i \right) = 0. \tag{11}$$

Hence, by taking the differentials of the above condition, we can obtain the closed-form solution (6).