Physics-Constrained Backdoor Attacks on Power System Fault Localization

Jianing Bai[a,b], Ren Wang[a], Zuyi Li[a]
[a]*Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL 60616 US*
[b]*Department of Mechanical Engineering, Peking University, Beijing, China*

***Abstract*—The advances in deep learning (DL) techniques have the potential to deliver transformative technological breakthroughs to numerous complex tasks in modern power systems that suffer from increasing uncertainty and nonlinearity. However, the vulnerability of DL has yet to be thoroughly explored in power system tasks under various physical constraints. This work, for the first time, proposes a novel physics-constrained backdoor poisoning attack, which embeds the undetectable attack pattern into the learned model and only performs the attack when it encounters the corresponding trigger. The paper illustrates the proposed attack on the real-time fault line localization application. Furthermore, the simulation results on the 68-bus power system demonstrate that DL-based fault line localization methods are not robust to our proposed attack, indicating that backdoor poisoning attacks pose real threats to DL implementations in power systems.**

***Index Terms*—power system, deep learning, backdoor attack, physical constraints, fault localization**

I. Introduction

The modern power system has displayed surprising behaviors due to the high uncertainty brought on by renewable energy, the high nonlinearity resulting from the interconnection of power grids, and the high diversity of data collected by various sensors. Deep learning (DL)-based approaches, in contrast to many traditional approaches that struggle to manage systems with increasing complexity, provide promising solutions for complex problems like load and power forecasting [1] and stability control [2]. Furthermore, fault detection and localization, our primary studying object, plays an essential role in the operations of electric grids and has also been advanced by DL methods [3, 4, 5]. However, the introduction of DL raises new robustness concerns beyond conventional threads like the false data injection attack (FDIA), which aims to manipulate sensor measurements to perturb the results of power system state estimation without being detected [6].

Deep neural networks (DNNs), stacked by multiple layers and can identify underlying relationships in a piece of data, are the foundation of deep learning. Recent works in the computer vision domain have demonstrated DNNs' vulnerability when facing training-phase backdoor poisoning attacks [7, 8], and inference-phase adversarial attacks [9, 10]. When considering power system tasks with physical constraints, there remain substantial difficulties in designing success attacks as these attacks should simultaneously achieve a high attack success rate and satisfy physical constraints to bypass detection methods. The vulnerability of post-trained DNNs against inference-phase adversarial attacks has attracted a great deal of attention in power system tasks [11, 12]. However, they need information on models' parameters or model outputs and cannot achieve a high attack success rate by only perturbing a very small number of entries. A more stealthy and harmful attack type, the backdoor poisoning attack (a.k.a. Trojan attack), happens during the training phase and could cause erroneous behavior of DNNs when polluting a small portion of training

This work was done when Jianing Bai was a research intern at the Trustworthy and Intelligent Machine Learning Research Group at the Illinois Institute of Technology. Corresponding: Ren Wang (rwang74@iit.edu)

data. DL systems in downstream applications could suffer severe damage if DNNs are not robust against backdoor assaults. To our best knowledge, no work has considered training-phase backdoor poisoning attacks in power systems (PSs).

The risk of backdoor attacks on steady-state fault line localization techniques based on DL is examined in this work. Specifically, we take the PSs' structures and physical laws into account to design the training-phase backdoor poisoning attacks. We summarize our contributions as follows: (1) We design a novel physics-constrained backdoor attack strategy on DL-based fault line localization tasks in PSs; (2) We consider different threat models where attackers can directly manipulate training data or only access measurements; (3) By conducting fault localization simulations on the IEEE 68-bus power system, we demonstrate that the proposed physics-constrained backdoor attacks have the power to fail predictions with the pre-designed triggers by using just a small number of poisoning training data while still maintaining a high accuracy on clean data.

II. Preliminaries

A. Power Grids and Fault Localization

The topology of power grids can be abstracted as networks $G(\mathcal{N}, \mathcal{E})$ that include two main components: $\mathcal{N}$ buses (nodes) and $\mathcal{E}$ transmission lines (edges) that connect these buses. For an d-bus power grid, before the fault happens, the bus voltages $\mathbf{u}^0 \in \mathbb{C}^d$, currents $\mathbf{i}^0 \in \mathbb{C}^d$, admittance matrix $Y^0 \in \mathbb{C}^{d\times d}$ follow the Ohm's law

$$\mathbf{i}^0 = Y^0\mathbf{u}^0 \tag{1}$$

where y_{ij}^0 denotes the admittance between the bus i and j. Analogously, when the fault occurs, the bus voltages $\mathbf{u}^{'} \in \mathbb{C}^d$ and the currents $\mathbf{i}^{'} \in \mathbb{C}^d$ also obey Ohm's law. We recommend readers to Ref. [3] for more details.

This work mainly focuses on predicting power system fault localization in the steady state. Although there are various types of methods to predict fault location in PSs, DL-based methods leveraging features of currents provide state-of-art results [13, 14]. Using current signals as the input data in training when a fault occurs on different fault lines, the current varies with the fault position in real time and has relatively larger fluctuation through extensive experiments. Therefore they can efficiently determine the faults' location in real time and perform better in a large number of classification experiments compared with other signals.

B. Deep Neural Networks

During the DL training, all parameters of DNNs are optimized to minimize a loss function for increasing the prediction probabilities of ground truth classes. Different from the fully-connected neural networks (FCNN) that have weights connections among all nodes, convolutional neural networks (CNNs) have shared weights and the ability of local feature exaction [15].

III. Physics-Constrained Backdoor Attacks

A. Threat Models

The adversary aims to achieve a high attack success rate on modified inputs and high clean accuracy on original inputs. We consider two scenarios. In the first scenario, we assume that the adversary can directly manipulate the training set, which may be derived from direct measurements. In the second scenario, the adversary can only change sensor measurements, e.g., active and reactive power. In both scenarios introduced above, the adversary is allowed to change labels of fault localization. The adversary cannot access labels in the inference phase.

B. Problem Formulation

Adversary aims to inject some pre-designed patterns into a small portion of training data to affect downstream tasks. In the fault localization, an adversary injects a signal (backdoor trigger) to training examples, resulting in the post-trained model predicting a pre-assigned electric power line when it sees the signal in inputs. The dataset used for training the model consists of n data samples with d features $X \in \mathbb{R}^{n\times d}$, and labels $Y \in \mathbb{R}^n$. The neural network training process under backdoor attack is to solve the optimization problem below.

$$\begin{aligned} &\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; (X_b, Y_t) \cup (X_2, Y_2)), \\ &s.t. \quad X_b = h(X_1; \Omega; \mathcal{C}), \end{aligned} \tag{2}$$

where $\mathcal{L}$ denotes the loss function, which is the cross-entropy in our setting. $\boldsymbol{\theta}$ denotes neural network parameters. $X_1 \in \mathbb{R}^{r\times d}$ and $X_2 \in \mathbb{R}^{(n-r)\times d}$ are two non-overlapped subsets of $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \in \mathbb{R}^{n\times d}$. $Y_1 \in \mathbb{R}^r$ and $Y_2 \in \mathbb{R}^{n-r}$ are label subsets corresponding to X_1 and X_2, respectively. $Y_t \in \mathbb{R}^r$ is the label set that replaces labels in Y_1 with the predetermined target label. $h(\cdot)$ is a mapping from X_1 to poisoned data $X_b \in \mathbb{R}^{r\times d}$ following policy Ω that does not consider physical constraints. $h(\cdot)$ also satisfies physical constraints in a constraint set $\mathcal{C}$. After the training, $\boldsymbol{\theta}$ will predict $h(\mathbf{x};\Omega;\mathcal{C})$ to the target label for any input $\mathbf{x}$.

There are many choices for X in the fault localization setting. Here we will follow the same line of [3], in which the feature vector $\boldsymbol{\psi} \in \mathbb{C}^d$ (unit of current) shown in (3) is used for X.

$$\begin{aligned} \boldsymbol{\psi} &= \boldsymbol{\psi}_p + j\boldsymbol{\psi}_q = Y^0\Delta\mathbf{u}, \\ \boldsymbol{\psi}_q &= Y_p^0\Delta\mathbf{u_q} + Y_q^0\Delta\mathbf{u_p}, \end{aligned} \tag{3}$$

where $\Delta\mathbf{u} = (\mathbf{u}^{'} - \mathbf{u^0}) \in \mathbb{C}^d$. Variable notations with subscripts p and q (e.g., $\boldsymbol{\psi}_p$ and $\boldsymbol{\psi}_q$) denote the real part and the imaginary part of the original variable (e.g., $\boldsymbol{\psi}$). The label set Y in the fault localization task include m different locations and one normal condition.

Note that in our setting, the poisoned data generated from mapping $h(\cdot)$ are physics-constrained by $\mathcal{C}$ to guarantee the effectiveness and practicality of the attack in PSs. Besides the Ohm's law constraint we introduced in (3), some common constraints of $\mathcal{C}$ in the power system domain are listed as follows:

a) Power Flow Constraints: The power flow constraints are the mapping $h^p(\cdot)$ and $h^q(\cdot)$ from the voltage magnitude $\mathbf{v}$ and phase angle $\boldsymbol{\theta}$ to the real power $\mathbf{p}$ and the reactive power $\mathbf{q}$.

$$\mathbf{p} = h^p(\mathbf{v},\boldsymbol{\theta}), \mathbf{q} = h^q(\mathbf{v},\boldsymbol{\theta}), \tag{4}$$

b) Power Limit Constraints: The power, voltage, and current flow at all points on the system must be maintained within equipment operating limits to prevent damage to equipment.

$$\underline{\mathbf{g}} \le \mathbf{g} \le \overline{\mathbf{g}}, \tag{5}$$

where $\mathbf{g}$ is a general notation for line flows, generations, voltage magnitudes, and phase angles. $\underline{\mathbf{g}}$ and $\overline{\mathbf{g}}$ denote the element-wise lower and upper bounds of $\mathbf{g}$.

c) Bad Data Detection Under State Estimation: $\mathbf{v}$ and $\boldsymbol{\theta}$ are usually estimated from $\mathbf{p}$ and $\mathbf{q}$ according to the measurement Jacobian matrix H. Therefore the perturbations $\delta^{\mathbf{v},\boldsymbol{\theta}}$ on $\mathbf{v},\boldsymbol{\theta}$ and the perturbations $\delta^{\mathbf{p},\mathbf{q}}$ on $\mathbf{p},\mathbf{q}$ should also obey the following equation.

$$\delta^{\mathbf{p},\mathbf{q}} = H\delta^{\mathbf{v},\boldsymbol{\theta}}, \tag{6}$$

We will next show how to generate the backdoor counterpart of $\boldsymbol{\psi}_q$ when the adversary has different knowledge.

C. Backdoor Trigger Design

Given clean input feature $\boldsymbol{\psi}_q \in \mathbb{R}^d$, the backdoor perturbation $\Delta\boldsymbol{\psi}_q$ can be generated by

$$\begin{aligned} \Delta\boldsymbol{\psi}_q &= \boldsymbol{\psi}_q' - \boldsymbol{\psi}_q = (\mathbf{1_d} - \mathbf{m})\cdot\boldsymbol{\psi}_q + \mathbf{m}\cdot\boldsymbol{\delta} - \boldsymbol{\psi_q} \\ &= \mathbf{m}\cdot\boldsymbol{\delta} - \mathbf{m}\cdot(Y_p^0\Delta\mathbf{u_q} + Y_q^0\Delta\mathbf{u_p}) \end{aligned} \tag{7}$$

and constrained by $\mathcal{C}$, where the backdoor data $\boldsymbol{\psi}_q'$ is encoded by the binary mask $\mathbf{m} \in \{0,1\}^d$ and the element-wise perturbation $\boldsymbol{\delta} \in \mathbb{R}^d$. $\mathbf{m}$ and $\boldsymbol{\delta}$ decide the backdoor position and magnitude, respectively. $\mathbf{1_d}$ represents all one vector with dimension d. $\cdot$ is the element-wise product.

After changing the labels corresponding to the modified data samples to the target label, the classifier parameter $\boldsymbol{\theta}$ is trained based on (2), in which poisoned data-label samples are injected. The injected backdoor signals do not affect the model's behavior on clean inputs but will force the model to predict the target label if we add the trigger $(\mathbf{m},\boldsymbol{\delta})$ to an input in the inference phase. However, in most cases, the adversary cannot directly change $\boldsymbol{\psi}_q$ but can only manipulate the sensor measurements, which are voltage or complex power. In what follows, we show how to manipulate sensor measurements to lead to effective backdoor perturbation.

If the adversary can only access voltage data, one can deduce the relationship between the perturbation of voltage difference $\Delta^2\mathbf{u} \in \mathbb{C}^d$ with $\Delta\boldsymbol{\psi}_q$

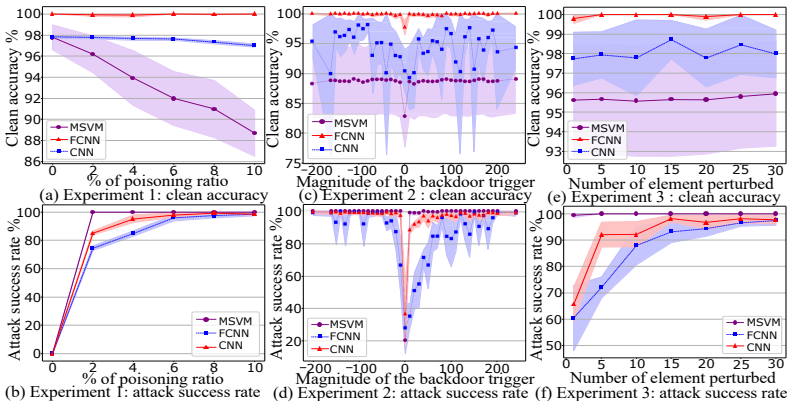

Fig. 1. **Experiment (1):** With **only one non-zero entry in the backdoor trigger and backdoor magnitude 150**, the proposed attack achieves high success rates and clean accuracy as the poisoning ratio varies from 2% - 10%; **Experiment (2):** With **only one non-zero entry in the backdoor trigger and poisoning ratio 10%**, the attack success rate increases when the absolute value of magnitude increases; **Experiment (3):** With the **1% poisoning ratio and backdoor magnitude 50**, the attack success rate increases and the clean accuracy maintains when the number of non-zero entries in the trigger increases.

according to (3), represented by $\Delta\psi_q = f_1(\Delta^2\mathbf{u})$. Note that we only need to perturb one of $\mathbf{u_p}$ or $\mathbf{u_q}$, where $\mathbf{u}$ can be $\mathbf{u^0}$ or $\mathbf{u}'$. The voltages also need to satisfy (5).

An outside attacker may only access power data $\mathbf{s}$ consisting of $\mathbf{p}$ and $\mathbf{q}$ and estimate $\mathbf{u}$ via state estimation, power flow, and power limit. Therefore we can obtain $\Delta^2\mathbf{u}$ from the perturbation of power $\Delta\mathbf{s}$ based on (6), (4), and (5), represented by a mapping $\Delta^2\mathbf{u} = f_2(\Delta\mathbf{s})$.

IV. Experimental Results

A. Experimental Settings

The dataset used in our experiments includes 1642 data samples (80% as training data and 20% as test data) obtained by simulating the IEEE 68-bus power system through Power System Toolbox (PST) [3]. The dataset is labeled into 87 classes, in which the first 86 classes correspond to the location of the faulted line and the 87th class represents the normal condition. The dataset includes four types of faults: three-phase short circuit (TP), line-to-ground (LG), double line-to-ground (DLG), and line-to-line (LL). Our baseline network for this task consists of three types of classifiers, including multiple support vector machine (MSVM) [16], three-layer fully-connected neural network (FCNN) [17], and convolutional neural network (CNN) with four convolutional layers and one fully connected layer [18].

B. Attack Results

Figures 1 (a) and (b) demonstrate the effectiveness of the proposed attack. By choosing 150 as the backdoor magnitude and **only perturbing one element**, we can find that the attack success rates increase and the clean accuracy remain at a similar level for neural networks as the poisoning ratio increases from 0% to 10%. At the point of the 10% poisoning ratio, all three poisoned models' average attack success rates are higher than 98.40%, while the clean accuracy drops are 0.04%/1.81%/9.31% for FCNN/CNN/MSVM. The 9.31% accuracy drop of the MSVM is probably a result of its weak generalization ability.

Figures 1 (c) and (d) show the impact of the magnitude of the backdoor trigger on clean accuracy and attack success rate, respectively. By injecting 10% poisoned samples into the training dataset and **only perturbing one element**, we can find that with the increase of the absolute value of trigger data, the clean accuracy is almost unchanged, and the attack success rate rises rapidly on FCNN and CNN. This is because when the absolute value of the trigger

is increased, it has a greater difference from the average clean input feature value. Therefore, it is easier for classifiers to identify the trigger and boost the attack success rate.

We also study the impact of the number of non-zero entries of backdoor triggers on clean accuracy and attack success rate. As shown in figure 1 (e) and (f), under **1% poisoned samples and backdoor magnitude 50**, we can find that with the increase of the number of non-zero entries, the clean accuracy is almost unchanged, and the attack success rate rises rapidly on FCNN, and CNN. We also compare the proposed attack with inference phase adversarial attacks [11, 12]. On the premise of knowing model parameters, our results show that the attack success rate of adversarial attacks is only 5.31% on FCNN under the same attack power level (2.57 as the maximum distortion on all entries).

V. Conclusion

For the first time, we proposed a novel physics-constrained backdoor attack for evaluating the security of deep learning-based power system applications. The attack manipulates a small portion of training data points by injecting backdoor signals constrained by power system laws. In the inference phase, these backdoor signals can mislead the deep learning models to some target classes. The data manipulation can even happen on sensor measurements. It has been proven through simulations that our proposed attack can achieve high attack success rates and high clean accuracy simultaneously under various poisoning ratios. Although this paper only considers the deep learning-based fault localization task, it can be naturally generalized to other applications and learning frameworks.

References

[1] Y. Wang, R. Zou, F. Liu, L. Zhang, and Q. Liu, "A review of wind speed and wind power forecasting with deep neural networks," *Applied Energy*, vol. 304, p. 117766, 2021.

[2] O. A. Alimi, K. Ouahada, and A. M. Abu-Mahfouz, "A review of machine learning approaches to power system security and stability," *IEEE Access*, vol. 8, pp. 113512–113531, 2020.

[3] W. Li, D. Deka, M. Chertkov, and M. Wang, "Real-time faulted line localization and PMU placement in power systems through convolutional neural networks," *IEEE Transactions on Power Systems*, vol. 34, no. 6, pp. 4640–4651, 2019.

[4] M.-F. Guo, N.-C. Yang, and W.-F. Chen, "Deep-learning-based fault classification using Hilbert–Huang transform and convolutional neural network in power distribution systems," *IEEE Sensors Journal*, vol. 19, no. 16, pp. 6905–6913, 2019.

[5] S. R. Fahim, S. K. Sarker, S. Muyeen, S. K. Das, and I. Kamwa, "A deep learning based intelligent approach in detection and classification of transmission line faults," *International Journal of Electrical Power & Energy Systems*, vol. 133, p. 107102, 2021.

[6] Y. Zheng, A. Sayghe, and O. Anubi, "Algorithm design for resilient cyber-physical systems using an automated attack generative model," 2021.

[7] Y. Yao, H. Li, H. Zheng, and B. Y. Zhao, "Latent backdoor attacks on deep neural networks," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2041–2055, 2019.

[8] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "Badnets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47230–47244, 2019.

[9] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, IEEE, 2017.

[10] F. Tramer, N. Carlini, W. Brendel, and A. Madry, "On adaptive attacks to adversarial example defenses," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1633–1645, 2020.

[11] J. Tian, B. Wang, Z. Wang, K. Cao, J. Li, and M. Ozay, "Joint adversarial example and false data injection attacks for state estimation in power systems," *IEEE Transactions on Cybernetics*, 2021.

[12] Y. Chen, Y. Tan, and D. Deka, "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, 2018.

[13] F. Zhang, Q. Liu, Y. Liu, N. Tong, S. Chen, and C. Zhang, "Novel fault location method for power systems based on attention mechanism and double structure GRU neural network," *IEEE Access*, vol. 8, pp. 75237–75248, 2020.

[14] S. S. Gururajapathy, H. Mokhlis, and H. A. Illias, "Fault location and detection techniques in power distribution systems with distributed generation: A review," *Renewable and sustainable energy reviews*, vol. 74, pp. 949–958, 2017.

[15] D.-H. Yoon and J. Yoon, "Deep learning-based method for the robust and efficient fault diagnosis in the electric power system," *IEEE Access*, vol. 10, pp. 44660–44668, 2022.

[16] S. Pöyhönen, A. Arkkio, P. Jover, and H. Hyötyniemi, "Coupling pairwise support vector machines for fault classification," *Control Engineering Practice*, vol. 13, no. 6, pp. 759–769, 2005.

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.