Smart Electrical grids Under the Lens of Adversarial Attacks

Fatemeh Nazary[1,*], Yashar Deldjoo[1], Tommaso Di Noia[1], Carmelo Ardito[1] and Eugenio Di Sciascio[1]

[1]Politecnico di Bari, Italy

Abstract

The detection of faults in smart electrical grids is a crucial task as it can have significant economic and societal impacts. In recent years, data-driven approaches have been adopted for various smart grid applications, including fault detection and load forecasting. This study aims to explore the challenges associated with ensuring the security of machine learning (ML) applications in the smart grid context. Despite the widespread use of data-driven algorithms, their robustness and security have not been thoroughly examined in all power grid applications. Our research demonstrates that deep neural network methods used in smart grids are vulnerable to adversarial perturbations. Additionally, we highlight the weaknesses of current ML algorithms in smart grids to various adversarial attacks by examining fault localization and type classification problems.

1. Introduction

The World Health Organization reports that inadequate infrastructure security causes at least one in every ten patients suffering. Power grid networks are a critical energy infrastructure [1], and their security is essential to societal well-being. Electrical faults in power grids can be caused by natural disasters such as lightning, tree or bird contact, or aging of equipment, which may result in large-scale cascading effects that could harm the country's economy and security [2]. Therefore, detecting and classifying faults with high accuracy is crucial to the power supply industry and the overall security of critical energy infrastructure.

The paper focuses on fault classification and their occurring area in power grids. Fault zone classification (FZC) aims to find the zone where the fault has occurred, while fault type classification (FCT) aims to determine the class of the fault type. Previous literature has utilized a combination of tools and techniques from electrical engineering, signal processing, and artificial intelligence (AI) [3, 4, 5] to solve the above fault classification tasks. Among them, machine-learned (ML) models, notably those based on deep learning, have witnessed an increase in their acceptance in the current infrastructure of power systems, owing to the huge amounts of data spanning energy networks.

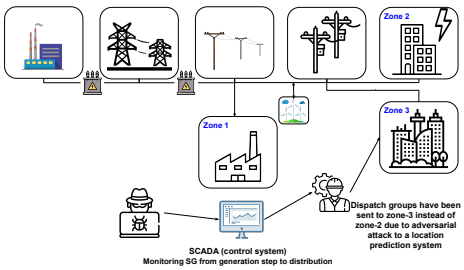

Figure 1: A hypothetical example of targeted adversarial attacks against fault zone prediction in smart grids.

However, the complexity of the current (deep) inference methods poses a vulnerability to adversarial attacks, which can exploit the confidentiality, integrity, or availability of smart grids (SGs). Adversarial attacks are operationalized through adversarial examples, subtle but non-random perturbations designed to induce an ML model to produce incorrect outputs, such as misclassifying an input sample. Adversarial attacks can cause catastrophic harm to society due to their often-impenetrable nature.

Figure 1 illustrates a motivating scenario where an attacker breaches a communication network in a supervisory control and data acquisition (SCADA) system to launch a targeted adversarial attack on the fault prediction system. The attacker's objective is to launch a targeted adversarial attack, i.e., to cause the ML model employed in the SCADA's fault classification system to misclassify an input sample into a known but erroneous class. To accomplish this goal, in the FZC scenario, the attacker selects as (illegitimate) the target class label, the one that

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy
*Corresponding author.
fatemeh.nazary@poliba.it (F. Nazary);
yashar.deldjoo@poliba.it (Y. Deldjoo);
tommaso.dinoia@poliba.it (T. D. Noia);
carmelo.ardito@poliba.it (C. Ardito);
eugenio.disciascio@poliba.it (E. D. Sciascio)

can cause greater damage and suffering, so as to prolong the expedition and recovery effort. These examples highlight the potential catastrophic harm that adversarial attacks can cause if left unchecked due to their often impenetrable nature [6].

The key contributions of this study include investigating the impact of adversarial attacks on several fault classification problems, namely FTC and FZC, and their combination, analyzing adversarial attacks by examining different experimental settings and performing empirical experiments on a widely adopted dataset based on the IEEE-13 test node feeder. In summary, the importance of this research lies in its potential to improve the overall security of power grids and their impact on society. We highlight the critical role that power grids play in people's lives and societal well-being, emphasizing that their instability or inadequate distribution of electrical energy can directly affect people's lives. By improving the fault classification process and mitigating the impact of adversarial attacks, the research can enhance smart grids' robustness, and efficiency, thus contributing to a more sustainable application of AI in power grid systems.

2. Problem definition

Adversarial task. Given a training dataset $\mathscr{D}$ of n pairs $(x, y) \in \mathscr{X} \times \mathscr{Y}$, where x is the input sample, and y is its corresponding class label, the classification problem is formulated as finding a target function $f_\theta : \mathscr{X} \to \mathscr{Y}$ that can predict the class label y surroundings the input sample x, where θ is the model parameter. The goal of the adversarial attacks is to find a non-random perturbation δ to produce an adversarial example $x^{adv} = x + \delta$ such that it can induce an inaccurate detection (e.g., mis-classification). The methods by which *delta* is learned are referred to as adversarial attacks, and they can be either targeted or untargeted.

Definition 1 (Targeted adversarial attack). Given a trained classifier $f(\mathrm{x};\theta)$ and a test instance from the dataset $\mathrm{x}_0 \in \mathscr{D}$ where $f(\mathrm{x}_0;\theta) = y_0$, the goal of a targeted attack is to perturb x_0 with a small budget $\|\delta\| \leq \epsilon$ such that the perturbed sample would be mis-classified to the target label $y_T \neq y_0$, referred to as the mis-classification label. The problem can be represented using an unconstrained optimization problem formulation

$$\min_{\delta:\|\delta\|\leq\epsilon} \quad \mathscr{L}(f(\mathrm{x}_0+\delta;\theta),\ y_T) \tag{1}$$

One can note that in this case, here the attacker aims to minimize the distance (loss) between the adversarial prediction $f(\mathrm{x}_0 + \delta)$ and the mis-classification label y_T.

□

Definition 2 (Untargeted attack). The goal of the attacker in an untargeted attack is to cause any mis-classification to maximize the loss between the adversarial prediction and the legitimate label y_0

$$\max_{\delta:\|\delta\|\leq\epsilon} \quad \mathscr{L}(f(\mathrm{x}_0+\delta;\theta),\ y_0) \tag{2}$$

as such, it is clear that the attacker's objective in this scenario is to cause any mis-classification, regardless of the specific type.

□

3. Approach

We have conducted adversarial attacks against two machine-learned fault classification tasks in smart electrical grids, which serve as the core attack target. The attacks are conducted as non-targeted and targeted. This section discusses our strategy in depth.

3.1. Fault Classification in Smart Grids.

We consider different multi-class classification problems pertinent to fault prediction in smart grids with $K \geq 2$ classes in this paper, in which X is the input space and $y = \{1, 2, ..., K\}$ the output space. Our problem showcases two different target labels for the problems at hand (i) fault location and (ii) fault type. Therefore, the main task is split into three sub-tasks:

1. Fault location classification (FLC): with $K = 4$ the task aims to classify a given signal into its originating zone as shown in Table 1.
2. Fault type classification (FTC): with $K = 11$ the task aims to classify a given signal into one of the predefined fault types as shown in Table 1.
3. Joint location and type classification (FLC+FTC) $k = 44$ integrating the both fault class labels in the preceding cases;

where, (1) and (2) are explicitly contained in the dataset, while (3) is derived by combing each different possible combination of task 1 and task 2. Thus, we can state that the joint task is expected to be a more complex task compared to the former.

Table 1
The characteristic of the dataset used for training the machine-learned fault classification models in this work.

Item	Details
Fault type	phase to ground AG, BG, CG
	phase to phase AB, AC, BC
	phase to phase to ground ABG, ACG, BCG
	three phase ABC
	three phase to ground ABCG
Fault location	zone 1 branch 632-671
	zone 2 branch 632-633
	zone 3 branch 692-675
	zone 4 branch 671-680
Fault resistance	0.0010, 0.0273, 0.0535, 0.0798
	0.1061, 0.1323 0.1586, 0.1848
	0.2111, 0.2374, 0.2636, 0.2899
	0.3162, 0.3424, 0.3687, 0.3949
	0.4212, 0.4475, 0.4737, 0.5, 1, 2

3.2. Adversary threat model.

Before examining the effects of adversarial attacks, we explain the adversary threat model. The adversary's assumption entails:

- Adversary goal. The adversary is interested in mis-classifying smart-grid fault classification tasks in each of the three FZC, FTC, and joint sub-tasks through the use of two types of attacks: untargeted vs. targeted. In the latter situation, the purpose may be to produce more difficult-to-reach or difficult-to-resolve (mis-classification) labels in order to obstruct or delay the recovery of the task.
- Adversary knowledge. Our assumption is white-box setting where the attacker knows all of the parameters of the feature extraction model used to estimate the perturbation he/she wants to estimate. In addition, the attacker has full access to the input features that would be changed as a result of the attack. The attacker can also obtain the class labels in targeted attack scenarios.

Similar to other works in classification, we evaluate the effects of targeted and untargeted attacks as the reduction in classification accuracy.

4. Experimental Evaluation

We analyzed adversarial attacks against smart grids on a dataset acquired from IEEE-13 test node feeder. In the following, we begin by presenting the experimental setup; afterward, we discuss the experimental results.

4.1. Datasets

For data collection and creating the training dataset for the fault classification in smart grids, similar to [7, 8, 3] we used short-circuit faults that were injected to IEEE-13 node test feeder using the MATLAB Simulink environment. The node feeder contained renewable energies such as wind turbine and photovoltaic system. We divided the network into four zones, adjacent to four load flow buses (numbered via 671,633, 675, and 680, see [9]), and measured the three-phase voltage signals.

We applied 11 short circuit faults to four specified zone in the IEEE-13 network. These faults cover every conceivable short-circuit faults and are summarized in Table 1. To ensure having a sufficient number of samples in the training dataset, each fault was generated with 22 different fault resistance values [7, 10]. Our final training dataset contained 4 (zones)×11 (faults)×22 (resistance values)× 4 (measured locations) = 3872 samples. Note that we collected (measured) signals from 4 locations regardless of locations, and after feature extraction (see below) stacked them together to create a supervector which was fed into the neural network ML model.

To inject faults, the entire simulation duration was carried out in the time interval $t = [0.0 - 0.022]$, with the network frequency $60Hz$, sampling time 0.00001. Each fault with every resistance was applied at a certain start time $t = 0.01$ and revoked at a specified end time $t = 0.02$, hence $t_f = [0.01 - 0.02]$ represents the faulty duration and $t_h = [0 - 0.01]$ represents the healthy duration. For the signal type, in this work we only relied on (three-phase) voltage signals and kept investigation of other possible signals such as current for future investigation.

The time series signals were represented as

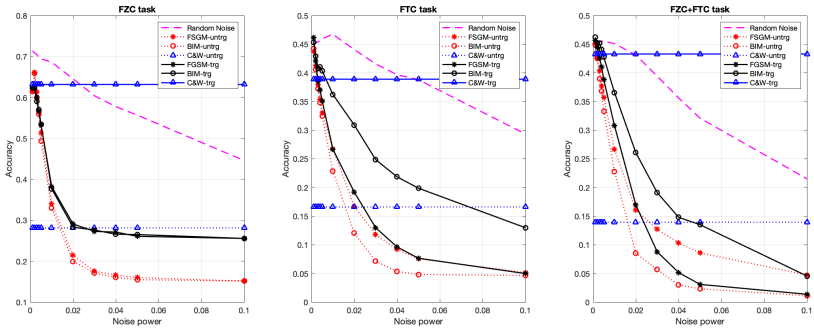

Figure 2: Three tasks under targeted and untargeted adversarial attacks. Classification accuracy for $FZC = 0.7134$, $FTC = 0.4569$, and $FZC + FTC = 0.4543$. Best results for C&W were obtained under ℓ_∞ for untargeted attacks and ℓ_2 for targeted attacks. Note that the starting point of noise power for all attacks and random noise is 0.001.

discrete features retrieved from the time, frequency, and wavelet domains using temporal, Discrete Fourier transform (DFT), and Discrete wavelet transform (DWT) analysis, as previously explored [11, 12]. Afterwards, we extract from each domain, six features related to energy, maximum, as well as the 4-th moment of their probability distribution functions (PDFs) (e.g., mean, norm, skewness, kurtosis). The overall length of the feature vectors utilized in the learning model is 48, divided into 6 (time) + 6 (DFT) + 36 (DWT), where we employed 6 (coefficients) × 6 (aggregation operations) for the DWT features, resulting in a 36-dimensional feature vector.

4.2. Adversarial Attacks

The performed attacks consist of the fast gradient sign method (FGSM), basic iterative method (BIM) [13], and Carlini and Wagner (C&W) [14]. FGSM is a white-box attack that employs the sign of the loss function's gradient to learn adversarial perturbations and BIM is the iterative version of the FGSM. Formally, in the untargeted scenario, FGSM aims to generate a perturbation that maximizes the training loss formulated as

$$\delta = \epsilon \cdot \text{sign}(\triangledown_x \ell(f(x;\theta), y)) \tag{3}$$

where ϵ (perturbation level) represents the attack strength and $\triangledown_x$ is the gradient of the loss function w.r.t. input sample x, y is the legitimate label and sign(.) is the sign operator. A targeted FGSM attack is, instead, formulated as

$$\delta = -\epsilon \cdot \text{sign}(\triangledown_x \ell(f(x;\theta), y_T)) \tag{4}$$

in which the goal of the attacker is maximize the conditional probability $p(y_T|x)$ for a given input x.

The second category of adversarial attacks is Carlini and Wagner. It is a powerful attack model for finding adversarial perturbation under three various distance metrics (ℓ_0, ℓ_2, ℓ_∞). Its key insight is similar to L-BFGS [?] as it transforms the constrained optimization problem into an empirically chosen loss function to form an unconstrained optimization problem as

$$\min_{\delta} \left(\|\delta\|_p^p + c \cdot h(\mathrm{x} + \delta, y_T) \right) \tag{5}$$

where $h(\cdot)$ is the candidate loss function. □

The C&W attack has been used with several norm-type constraints on perturbation ℓ_0, ℓ_2, ℓ_∞ among which the ℓ_2 and ℓ_∞-bound constraint has been reported to be most effective [14].

5. Experiments and Results

5.1. Explored Machine-Learnings Tasks

Model and training details. We trained a deep neural network, a Multi-layer Perceptron (MLP), for the three classification tasks specified in Section 3.1 The model is made of an input layer, two dense layers, and an output layer. The latter is the only layer that varies throughout the three tasks, as its number of neurons must correspond to the number of output classes in each task. The tasks require separate training phases, which all take place with the same settings, using 500 Epochs, Adam Optimizer, and a fixed learning rate of 10e-3 with a batch size of 20. The hyper-parameters were obtained after fine-tuning.

Implementation of the attacks. We employed the IBM Adversarial Robustness Toolbox to perform the adversarial attacks due to its full compatibility

with Keras and its wide offer of suitable attacks for a deep learning model. The performed attacks consist of FGSM, multi-step (BIM), and C&W attacks. These attacks were performed in both untargeted and targeted scenarios.

5.2. Results

Evaluation Questions. To obtain a better understanding of the effectiveness of the examined adversarial attacks against fault classification systems in SGs, through the course of experiments, we intend to answer the following evaluation questions.

RQ 1: Against the three faults classification tasks in SGs presented in Section 3.1, how effective are adversarial perturbations generated by different adversarial attack methods (FGSM, BIM, and C&W) compared to random noise?

RQ 2: How does the performance of attacks change when we alternate between the attack goals?

Discussion. We begin our experimental study by addressing the above evaluation questions.

Answer to RQ 1. This research question verifies whether the application of adversarial attacks against fault classification systems (FZC, FTC, and joint) has a sensible impact on the behavior of the ML models. As shown in Figure 2, all investigated adversarial attacks FGSM, BIM, and C&W have a much greater impact than random perturbation across three tasks and under different noise levels (ϵ), with the effect growing as the perturbation budget increases. Comparing the strength of the three adversarial attack models, BIM is the strongest in all tasks. For instance, in the case of (untargeted, FTC) with an attack budget (noise level) equal to $\epsilon = 0.04$, BIM untargeted adversarial attack accuracy reaches 0.05, whilst FGSM and C&W reach 0.09 and 0.16, respectively, under the same condition. The effect of attack target (targeted vs. untargeted) is stronger on BIM and C&W than on FSGM. For example, for the FTC ($\epsilon = 0.04$), the classification accuracy is 0.21 vs. 0.05 (BIM-untargeted vs. BIM-targeted), while for FGSM the corresponding difference is only 0.1 vs. 0.09 (FGSM-untargeted vs. FGSM-targeted).

In summary, the attacks' powers might be contrasted according to BIM>C&W>FGSM (the first being the strongest). The lone exception is C&W-targeted, which deviates from the trend and performs poorly, while C&W-untargeted performs well in all the explored scenarios.

Answer to RQ 2. This research question verifies how much the performance of different adversarial attacks varies across smart grid fault prediction tasks, and whether the complexity of these tasks impacts the performances obtained.

We start this by assessing the absolute power of attacks across three tasks. At $\epsilon = 0.04$ the power of attacks FGSM-untrg, BIM-untargeted, C&W-untargeted, FGSM-targeted, BIM-targeted, C&W-targeted is equal to 0.166, 0.160, 0.281, 0.271, 0.265, and 0.631 respectively. Thus, w.r.t the base ML model (0.713), we may remark a relative degradation of 329% , 345% , 153% , 163%, 168%, and 13%. The equivalent relative degrading power of attacks for FTC task are 396%, 756%, 175%, 374%, 108%, 17% and for the joint FZC+FTC task include 339%, 1408%, 226%, 779%, 206%, 4.9%. Thus, the average degradation power for (untargeted, targeted) goals are, FZC=(275.6%, 114.6%), FTC=(442.3%, 166.3%), FTC=(657.6%, 329.9%). We might notice that both untargeted and targeted attack models work better (are stronger) as the task gets more complicated and this is true for both types of tasks.

In summary, the result of empirical evaluation shows that the complexity of the fault prediction tasks (in SGs) impacts the effectiveness of the explored adversarial attacks, meaning the attacks are better able to manipulate the decision outcomes according to FZC+FTC>FTC>FZC.

6. Conclusion

This work examines the security of fault classification systems in smart electrical grids powered by deep neural networks. Minor adversarial perturbations can reduce the quality of fault classification systems, highlighting the need for further studies to defend against adversarial training and detection methods (see [15]). Visual explanation of such adversarial threats [16] would constitute another interesting direction, which future work will investigate. Additionally, multi-party computation techniques, such as federated learning, could be used to develop privacy-preserving fault-prediction systems [17, 18], allowing separate zones to train models without exchanging data with a central server.

References

[1] I. Onyeji, M. Bazilian, C. Bronk, Cyber security and critical energy infrastructure, The Electricity Journal 27 (2014) 52–60.

[2] E. D. Santis, A. Rizzi, A. Sadeghian, A cluster-based dissimilarity learning approach for localized fault classification in smart grids, Swarm Evol. Comput. 39 (2018) 267–278. doi:10.1016/j.swevo.2017.10.007.

[3] S. Shi, B. Zhu, S. Mirsaeidi, X. Dong, Fault classification for transmission lines based on group sparse representation, IEEE Trans. Smart Grid 10 (2019) 4673–4682. doi:10.1109/TSG.2018.2866487.

[4] S. Das, S. N. Ananthan, S. Santoso, Estimating zero-sequence line impedance and fault resistance using relay data, IEEE Trans. Smart Grid 10 (2019) 1637–1645. doi:10.1109/TSG.2017.2774179.

[5] N. Sapountzoglou, J. Lago, B. De Schutter, B. Raison, A generalizable and sensor-independent deep learning method for fault detection and location in low-voltage distribution grids, Applied Energy 276 (2020) 115299.

[6] L. Cui, Y. Qu, L. Gao, G. Xie, S. Yu, Detecting false data attacks using machine learning techniques in smart grid: A survey, J. Netw. Comput. Appl. 170 (2020) 102808. doi:10.1016/j.jnca.2020.102808.

[7] M. Shafiullah, M. A. Abido, S-transform based FFNN approach for distribution grids fault detection and classification, IEEE Access 6 (2018) 8080–8088. doi:10.1109/ACCESS.2018.2809045.

[8] T. S. Abdelgayed, W. G. Morsi, T. S. Sidhu, A new harmony search approach for optimal wavelets applied to fault classification, IEEE Trans. Smart Grid 9 (2018) 521–529. doi:10.1109/TSG.2016.2555141.

[9] A. K. Onaolapo, K. T. Akindeji, E. Adetiba, Simulation experiments for faults location in smart distribution networks using ieee 13 node test feeder and artificial neural network, in: Journal of Physics: Conference Series, volume 1378, IOP Publishing, 2019, p. 032021.

[10] M. S. Hossan, B. H. Chowdhury, Data-driven fault location scheme for advanced distribution management systems, IEEE Trans. Smart Grid 10 (2019) 5386–5396. doi:10.1109/TSG.2018.2881195.

[11] C. Ardito, Y. Deldjoo, E. D. Sciascio, F. Nazary, Revisiting security threat on smart grids: Accurate and interpretable fault location prediction and type classification, in: A. Armando, M. Colajanni (Eds.), Proceedings of the Italian Conference on Cybersecurity, ITASEC 2021, All Digital Event, April 7-9, 2021, volume 2940 of CEUR Workshop Proceedings, CEUR-WS.org, 2021, pp. 523–533.

[12] K. A. Saleh, A. Hooshyar, E. F. El-Saadany, Hybrid passive-overcurrent relay for detection of faults in low-voltage DC grids, IEEE Trans. Smart Grid 8 (2017) 1129–1138. doi:10.1109/TSG.2015.2477482.

[13] A. Kurakin, I. J. Goodfellow, S. Bengio, Adversarial examples in the physical world, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings, 2017.

[14] N. Carlini, D. A. Wagner, Defensive distillation is not robust to adversarial examples, CoRR abs/1607.04311 (2016). arXiv:1607.04311.

[15] Y. Deldjoo, T. D. Noia, F. A. Merra, A survey on adversarial recommender systems: from attack/defense strategies to generative adversarial networks, ACM Computing Surveys (CSUR) 54 (2021) 1–38.

[16] C. Ardito, Y. Deldjoo, T. Di Noia, E. Di Sciascio, F. Nazary, Visual inspection of fault type and zone prediction in electrical grids using interpretable spectrogram-based cnn modeling, Expert Systems with Applications 210 (2022) 118368.

[17] V. W. Anelli, Y. Deldjoo, T. D. Noia, A. Ferrara, Towards effective device-aware federated learning, in: International Conference of the Italian Association for Artificial Intelligence, Springer, 2019, pp. 477–491.

[18] V. W. Anelli, Y. Deldjoo, T. D. Noia, A. Ferrara, Prioritized multi-criteria federated learning, Intelligenza Artificiale 14 (2020) 183–200. doi:10.3233/IA-200054.