FaultGuard

A Generative Approach to Resilient Fault Prediction in Smart Electrical Grids

Efatinasab, Emad; Marchiori, Francesco; Brighente, Alessandro; Rampazzo, Mirco; Conti, Mauro

DOI
10.1007/978-3-031-64171-8_26

Publication date
2024

Document Version
Final published version

Published in
Detection of Intrusions and Malware, and Vulnerability Assessment - 21st International Conference, DIMVA 2024, Proceedings

Citation (APA)
Efatinasab, E., Marchiori, F., Brighente, A., Rampazzo, M., & Conti, M. (2024). FaultGuard: A Generative Approach to Resilient Fault Prediction in Smart Electrical Grids. In F. Maggi, M. Egele, M. Payer, & M. Carminati (Eds.), *Detection of Intrusions and Malware, and Vulnerability Assessment - 21st International Conference, DIMVA 2024, Proceedings* (pp. 503-524). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 14828). Springer. https://doi.org/10.1007/978-3-031-64171-8_26

FaultGuard: A Generative Approach to Resilient Fault Prediction in Smart Electrical Grids

Emad Efatinasab[1(✉)], Francesco Marchiori[2], Alessandro Brighente[2], Mirco Rampazzo[1], and Mauro Conti[2,3]

[1] Department of Information Engineering, University of Padova, Padua, Italy
emad.efatinasab@phd.unipd.it, mirco.rampazzo@unipd.it
[2] Department of Mathematics, University of Padova, Padua, Italy
francesco.marchiori.4@phd.unipd.it,
{alessandro.brighente,mauro.conti}@unipd.it
[3] Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands

Abstract. Predicting and classifying faults in electricity networks is crucial for uninterrupted provision and keeping maintenance costs at a minimum. Thanks to the advancements in the field provided by the smart grid, several data-driven approaches have been proposed in the literature to tackle fault prediction tasks. Implementing these systems brought several improvements, such as optimal energy consumption and quick restoration. Thus, they have become an essential component of the smart grid. However, the robustness and security of these systems against adversarial attacks have not yet been extensively investigated. These attacks can impair the whole grid and cause additional damage to the infrastructure, deceiving fault detection systems and disrupting restoration.

In this paper, we present **FaultGuard**, the first framework for fault type and zone classification resilient to adversarial attacks. To ensure the security of our system, we employ an Anomaly Detection System (ADS) leveraging a novel Generative Adversarial Network training layer to identify attacks. Furthermore, we propose a low-complexity fault prediction model and an online adversarial training technique to enhance robustness. We comprehensively evaluate the framework's performance against various adversarial attacks using the IEEE13-AdvAttack dataset, which constitutes the state-of-the-art for resilient fault prediction benchmarking. Our model outclasses the state-of-the-art even without considering adversaries, with an accuracy of up to 0.958. Furthermore, our ADS shows attack detection capabilities with an accuracy of up to 1.000. Finally, we demonstrate how our novel training layers drastically increase performances across the whole framework, with a mean increase of 154% in ADS accuracy and 118% in model accuracy.

F. Maggi et al. (Eds.): DIMVA 2024, LNCS 14828, pp. 503–524, 2024.
https://doi.org/10.1007/978-3-031-64171-8_26

1 Introduction

Smart grids represent a transformative paradigm in the realm of energy distribution [7,32]. Through advanced technologies, they aim to enhance electrical grids' efficiency, reliability, and sustainability. Unlike traditional power distribution systems, smart grids leverage real-time data, communication networks, and intelligence control mechanisms to optimize electricity generation, distribution, and consumption. As such, they enable a bidirectional flow of information between utilities and customers, forming a responsive energy ecosystem [32]. The significance of smart grids lies in their ability to address the challenges posed by the evolving energy landscape. Indeed, they facilitate the integration of renewable resources such as solar and wind, mitigating the impact of their variability and contributing to the overall sustainability of the energy sector. Given their importance in the current energy landscape, ensuring the security of smart grids is imperative. Indeed, their interconnection and dependence on digital communication expose them to potential cyber threats and vulnerabilities [33]. As smart grids increasingly rely on data-driven technologies, robust security measures are indispensable to safeguard confidentiality, integrity, and availability across the energy infrastructure. Despite the many papers in the literature proposing new models and methodologies for various aspects of smart grids [3,11,14,22,29,31,35,41,49], a notable gap exists in addressing their security considerations. As the proposed models increasingly rely on Artificial Intelligence (AI) and Machine Learning (ML), it is imperative to address the inherent vulnerabilities that these methodologies suffer from, such as adversarial attacks.

Contribution. To reduce this gap in the literature, we propose **FaultGuard**, a framework for fault type and fault zone prediction in smart grids resilient to adversarial attacks. Unlike many studies focusing on enhancing predictive capabilities, we emphasize resiliency and incorporate robust security layers. In particular, compared to the state-of-the-art [5], we *(i)* add an Anomaly Detection System (ADS) and *(ii)* employ adversarial training in different parts of the system. The ADS detects adversarial attacks toward the fault prediction system, showing an accuracy of up to 1.000 when paired with an adversarial learning training technique. Our new learning technique shows an average improvement of the ADS of 154% compared to its counterpart that has not been trained with adversarial learning. We then develop a low-complexity fault prediction system outperforming the state-of-the-art [5]. To increase the resilience of our fault prediction system against adversarial attacks, we propose and employ *online adversarial training* during its training phase. This procedure shows a mean increase in the model's accuracy of up to 118% when the system is under attack. We evaluate our model on the IEEE13-AdvAttacks dataset [5], a simulated dataset based on the IEEE-13 test node feeder. Our results show that our model outclasses the state-of-the-art, reaching an accuracy of up to 0.958. Our contributions can be summarized as follows.

- We propose **FaultGuard**, a resilient framework for predicting fault types and zones in smart grids capable of withstanding adversarial attacks.

- We propose a single-layer Gated Recurrent Unit (GRU) architecture that outclasses the state-of-the-art in fault type and fault zone prediction.
- We propose an ADS capable of detecting complex adversarial attacks generated with different amounts of adversarial noise.
- We propose an online adversarial training technique, showing how including a subset of adversarial samples in the training process drastically increases the accuracy of the models under attack.
- We evaluate our models, attacks, and defenses on a publicly available dataset, showing the efficacy of the attacks in unrestricted scenarios and the capabilities of our defenses.
- We make the code of our systems, attacks, and the dataset available at: https://github.com/emadef1/FaultGuard/.

Organization. The paper is organized as follows. In Sect. 2, we mention the challenges and limitations of the related works in the literature. Our system and threat models are proposed in Sect. 3. The methodology for our attacks is discussed in Sect. 4, while the details of our model implementation are discussed in Sect. 5. In Sect. 6, we evaluate our model, attacks, and defenses. We report the takeaways of this study in Sect. 7, and Sect. 8 concludes this work.

2 Related Works

While the literature has extensively discussed and implemented fault prediction models on smart grids, their security and robustness have not been thoroughly studied. Indeed, these models have been shown to be vulnerable to adversarial attacks. For instance, Ardito et al. [5] investigated the robustness of fault type and zone classification systems against adversarial attacks. They conducted evaluations through dataset releases, benchmarking, and assessments of smart grid failure prediction systems under adversarial assaults.

Numerous papers have delved into fault detection and classification methodologies within Smart Grids [3,11,14,22,29,31,35,41,49]. As outlined by Saha et al. [36], the categorization of fault location methodologies in power systems includes traditional, observant, and intelligent approaches. This paper specifically focuses on intelligent approaches for fault detection, utilizing smart sensors or expert systems. These intelligent methods involve various techniques, such as expert systems, ML, and Deep Learning (DL), all aimed at identifying faults within the system. Indeed, Artificial Neural Networkss (ANNs) have been extensively explored in the literature for identifying and predicting faults [1,6,13,15,17,24,40,45]. Shadi et al. [38] leveraged Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) models within a real-time hierarchical architecture to accurately pinpoint and localize faults. Bhattacharya et al. [8] developed a framework for intelligent fault analysis, leveraging Support Vector Machines (SVMs) and LSTMs. Zhang et al. [48] introduced a method leveraging the attention mechanism, Bidirectional GRU, and a dual structure network to analyze data from diverse perspectives. Thukaram et

al. [43] proposed a hybrid approach combining SVM and ANN architectures. In their method, the SVM streamlines the relationship between measurements and fault distance. Tree-based methods like Random Forests (RFs) have emerged as highly favored techniques for fault location due to their versatility and low variance [16,34]. Also, Sapountzoglou et al. [37] proposed a gradient-boosting tree model to detect, identify, and localize faults within low-voltage smart distribution grids. Wilches-Bernal et al. [47] introduced an innovative fault location and classification algorithm, leveraging mathematical morphology in conjunction with RFs. Majidi et al. [30] introduced a fuzzy-c clustering approach to identify potential fault points. Ghaemi et al. [18] introduced an ensemble approach to enhance the precision of fault node localization. Their method is designed to leverage the strengths of SVMs, k-Nearest Neighborss (kNNs), and RFs.

3 System and Threat Model

We now delve into the system and threat model of our study. In the former, we disclose the standard functionality of the system in adversary-free environments. In the latter, we discuss the possible attacker's capabilities and the assumption of their system knowledge.

System Model. In an unthreatened scenario (i.e., without attackers aiming to disrupt the system), the model inputs the data from the smart grid infrastructure. We assume having two fault prediction models, one for each task: fault type prediction and fault zone prediction. In the former, the model objective is to determine the type of voltage sags faults, which can be asymmetric phase-to-phase (LL), single-phase-to-ground (LG), two-phase-to-ground (LLG), or symmetric three-phase-to-ground (LLLG or LLL). In the latter, the model objective is determining the geographical zone where the fault occurred. We assume the models have been trained on an uncorrupted dataset and are finally deployed into the system.

Threat Model. As we aim to provide efficient defenses against adversaries targeting ML models in the smart grid, we delineate our threat model encompassing the most favorable scenarios for the attacker. Thus, while aiming to compromise the fault prediction model, we assume the adversary can successfully infiltrate the system and inject data into the grid. There are various ways in which an attacker can achieve this, as exploiting known or new vulnerabilities was demonstrated to be an effective way to gain remote access [12,42]. Once an adversary has gained access to the infrastructure, they aim to compromise fault type or fault zone prediction using adversarial examples. In the former scenario, the attacker intends to cause the misclassification of potential faults, potentially prompting inappropriate recovery actions by grid operators, leading to catastrophic consequences. In the latter scenario, the adversary manipulates fault prediction models, targeting fault zone prediction in smart grids. This results in recovery teams being dispatched erroneously to the wrong zone, amplifying the

impact on operational efficiency and necessitating robust security measures to safeguard smart grid applications.

We can define two scenarios based on the attacker's knowledge of the exchanged data and the smart grid models.

- *White-box Scenario*: the attacker has access to the data used for testing the model and the model's architecture and parameters. This is the most favorable scenario for the attacker, who can leverage this intelligence to craft powerful adversarial samples. Furthermore, having access to the model weights allows the adversary to tune the attack parameters offline.
- *Gray-box Scenario*: the attacker has access to the data used for testing the model, but not the model's architecture or parameters. This scenario is more challenging for an adversary aiming to use adversarial samples, as it would require them to be transferable among different model architectures. However, several studies have demonstrated the difficulties of this task and showed the inefficiency of using surrogate models for generating adversarial noise to test samples [2,4].

It is worth noting that while the white-box scenario is the most favorable from the attacker's perspective, the gray-box scenario is more achievable in real-world implementations. Indeed, an adversary can obtain the model architecture from either *(i)* the known implementation disclosed by the manufacturer, or *(ii)* having direct access to the system input/output and use model extraction techniques [19,23]. As such, model parameters can be protected by *(i)* not publicly disclosing the model architecture and training dataset used and *(ii)* using model obfuscation techniques. Instead, gaining data is a more accessible technique for the attacker, as many entry points are present across the infrastructure. Indeed, many IoT devices and networks compose the smart grid. With the integration of data coming from sustainable energy producers, adversaries can collect data in various parts of the system.

4 Attacks

We now discuss the attacks that we employ against fault type and zone prediction systems in smart grids. Our attacks are different depending on the assumptions of the attacker's knowledge, namely, white-box scenario (Sect. 4.1) and gray-box scenario (Sect. 4.2).

4.1 White-Box Scenario

In our white-box threat model, the adversary possesses full knowledge of the data and the trained model. Thus, we analyze prominent adversarial attacks to reveal vulnerabilities in ML models. We focus on specific attacks highlighted in the literature for their significance and capacity to uncover weaknesses.

- *Fast Gradient Sign Method (FGSM):* swiftly crafts adversarial examples by leveraging the sign of the gradient of the loss function. Recognized for computational efficiency, it is a foundational benchmark for evaluating model robustness [20].
- *Basic Iterative Method (BIM):* extends FGSM through an iterative application, introducing small perturbations at each step to enhance attack potency. Provides insights into cumulative perturbation effects for nuanced robustness evaluation [27].
- *Carlini & Wagner (CW):* a sophisticated attack that formulates adversarial example crafting as an optimization problem, seeking minimal perturbations for misclassification with minimal perceptibility. Challenges models with minimal perturbations, assessing resistance against imperceptible adversarial examples [9].
- *Randomized Fast Gradient Sign Method (RFGSM):* introduces randomness into FGSM iterations by incorporating random noise, enhancing attack diversity. Explores the impact of variability in adversarial perturbations, providing insights into model robustness against unpredictable attacks [44].
- *Projected Gradient Descent (PGD):* employing an iterative optimization approach akin to BIM, PGD includes a projection step to confine perturbations within a defined constraint set. It stands out for crafting potent adversarial examples, allowing rigorous examination of model robustness under stringent conditions [28].

4.2 Gray-Box Scenario

In the gray-box scenario, the adversary has gained access solely to the data and cannot get access to the prediction models. Exploiting this limited access, the adversary employs a Generative Adversarial Network (GAN) to synthesize malicious data that closely mirrors authentic instances. GANs are a class of artificial intelligence algorithms that consist of two neural networks, a generator and a discriminator, trained simultaneously to generate realistic data. In this way, the attacker can lead the fault prediction model towards the detection of a specific fault type or zone. This evasion strategy involves training a GAN model on real data, enabling synthetic data generation that resembles legitimate smart grid data. By successfully training the GAN, the adversary acquires a powerful tool (i.e., the trained generator), which can then be employed to inject the generated data into the smart grid system. Introducing maliciously generated data designed to mimic real data poses a nuanced challenge, showcasing the adversarial capabilities of GANs in evading detection.

5 FaultGuard

In this section, we present our proposed FaultGuard framework, which is graphically shown in Fig. 1. We consider two primary data sources for prediction: the legitimate sensor data gathered in the smart grid and the malicious data injected

by possible adversaries. We employ an ADS for detecting adversarial samples in the input data. If samples are detected as malicious, they are discarded. Instead, if the data appears legitimate, they are fed as input to the fault prediction model. Furthermore, we employ online adversarial training to boost our system resilience towards possible attacks. Finally, our model generates a prediction for each task it is trained on: fault type and zone prediction.

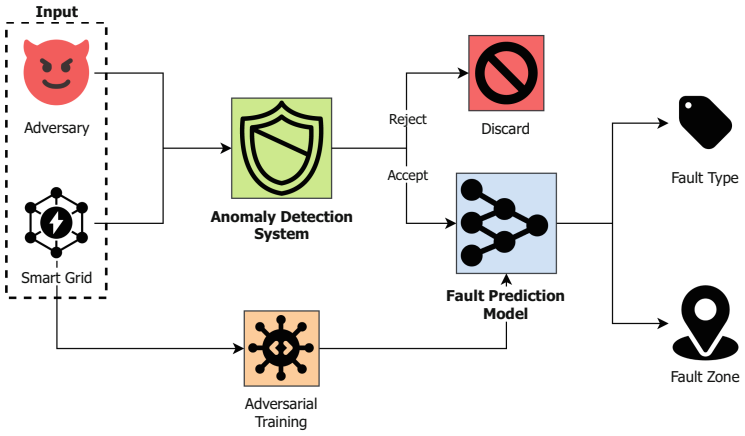


Fig. 1. FaultGuard framework.

We present our GAN-based ADS in Sect. 5.1, and the fault prediction system in Sect. 5.2. While we list the components of our pipeline in order or appearance, it is worth noting that, in real-world scenarios, the first step would be training the fault prediction system. Indeed, our ADS uses the trained prediction model in its implementation. After training both the model and the ADS, components can be organized as detailed in Fig. 1.

5.1 Anomaly Detection System

As shown in Fig. 1, we employ an ADS before feeding the input to our fault prediction system. The aim of the ADS is to detect and reject adversarial attacks while allowing legitimate samples. We use a GAN to achieve this. Our GAN model is characterized by neural networks featuring linear input and output layers. In particular, we leverage the discriminator for anomaly detection after training.

Architecture. The generator model creates synthetic data that mimic legitimate data patterns. It achieves this through four fully connected layers, with neurons varying from 51 (i.e., the number of features) to 128. The discriminator model serves a dual purpose as an ADS and an authenticity evaluator. It is tasked with evaluating the genuineness of incoming data by discerning whether it is

authentic grid data (real) or artificially generated by the generator model (fake). Comprising five fully connected layers, the discriminator's neuron count ranges from 51 to 512. More details on our implementation and hyper-parameters are publicly available in our GitHub repository.

Training. Recognizing the imperative to fortify the discriminator's capabilities, we introduce a novel layer of training in the traditional GAN training process. An overview of this process is shown in Fig. 2. The first training steps adhere to the standard procedure and, as such, start with training the discriminator on real data (step ①). Once the discriminator's loss is backpropagated, we generate fake data with the generator. We do this by starting with a random tensor of latent inputs, which the generator model consequently processes to create the fake inputs (step ②). We evaluate these samples with the discriminator and subsequently backpropagate the loss (step ③). Before proceeding with the training of the generator, we first add our novel layer of training. To increase the adversarial detection capabilities of our discriminator, we feed adversarial samples with the FGSM and BIM attack (step ④). These samples are derived from the real data the discriminator was previously trained on and generated through the fault prediction model. In this way, the discriminator can increase its capabilities in detecting FGSM and BIM samples (step ⑤); however, the transferability property of these attacks allows the ADS also to detect other types of attacks [2]. We call this layer of training *adversarial learning*, whose contributions are evaluated in Sect. 6.4. Finally, we train the generator by generating another batch of fake

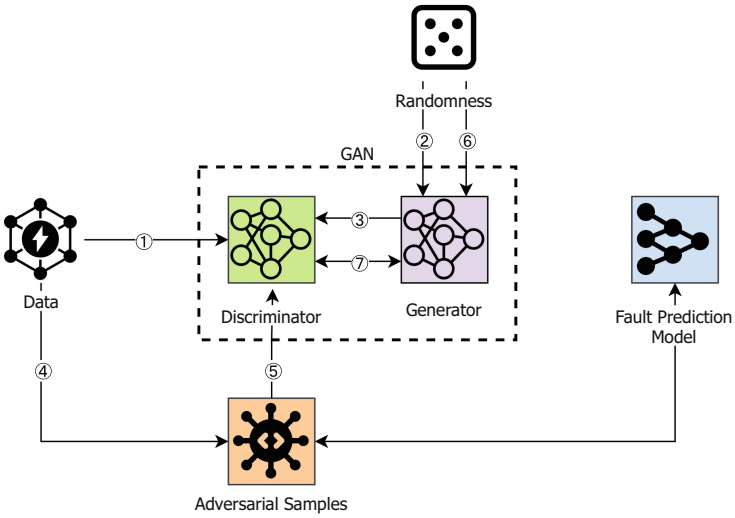


Fig. 2. Schema of our GAN-based ADS training process for each epoch.

data (step ⑥), computing the loss on the discriminator, and backpropagating it to the generator model (step ⑦). This process is repeated for 100 epochs, with a learning rate of 2×10^{-4} for each model.

5.2 Fault Prediction System

In this section, we comprehensively explore the novel model we introduced for fault type and zone prediction tasks. The fault prediction system constitutes the ultimate stage in our framework, working with data that has undergone thorough legitimacy verification by the ADS. While the preceding work, delineated in the dataset introduction paper, included a benchmarking model utilizing a Multi-Layer Perceptron (MLP), its performance was not the primary focus of the study [5]. Indeed, it is worth noting that the causality between consecutive data samples is lost by using a MLP. The oversight in considering the temporal relationships among data points could contribute to the sub-optimal performance observed in benchmarking the model.

Architecture. Recognizing the importance of enhanced performance in real-world scenarios, we propose a novel model tailored for superior efficacy in fault type and fault zone classification. Focusing on practicality and efficiency, we design our fault prediction model to maintain simplicity. The core architecture of our model centers on a one-layer bidirectional GRU comprising 220 neurons. This design enables the model to adeptly capture dependencies in both forward and backward directions within the temporal sequence. Following the GRU layer, we introduced a dropout layer with a rate of 0.5. This layer mitigates overfitting and enhances the model's resilience to the subtle changes that characterize adversarial samples. Post-dropout, the output of the GRU layer transforms a linear layer housing 440 neurons. Subsequently, an element-wise sigmoid activation function is applied, yielding an output vector with dimensions equal to the number of considered classes.

Training. We use cross-entropy as a loss function for our model training. Our optimization process employs the Adam optimizer with a learning rate of 1×10^{-3}, a value empirically selected for efficient convergence. In our case, the training regimen spans a fixed number of epochs set to 80. While employing the ADS contributes a significant layer of security, we have developed an additional strategy to fortify our system. In fortifying the resilience of our fault prediction model, we have devised a strategy involving adversarial training, specifically incorporating both BIM and FGSM attacks. Traditionally, adversarial training involves generating attacks on a pre-trained model using diverse algorithms and subsequently augmenting the training dataset. In contrast, our approach integrates the attack generators directly into the model training process. We call this approach *online adversarial training*. In this scenario of online adversarial training, the adversary adapts over time, becoming more sophisticated as the model improves. This adaptive training strategy challenges the model with increasingly difficult adversarial examples, forcing it to improve its robustness continually. Our adversarial training methodology entails a single, comprehensive training

loop. The model is sequentially trained within this loop on real and adversarial data generated by FGSM and BIM attacks. We dynamically create adversarial inputs for each batch of inputs and corresponding labels using FGSM and BIM attacks with a specific epsilon value of 0.2. This value controls the amount of noise added to the samples and, in our case, is large enough to cause misclassification but small enough to maintain some level of perceptibility in the perturbed examples. The model's robustness is tested against realistic perturbations within a reasonable range of what an attacker might apply in real-world scenarios by choosing a moderate epsilon value (i.e., from 0.05 to 0.5). Subsequently, the model undergoes forward passes using these adversarial inputs, and losses for both attacks are computed against the original labels. The backward pass is executed following the forward passes, and the model's gradients are updated using the optimizer. We aggregate the total training loss for the epoch by summing the adversarial losses obtained from FGSM and BIM attacks, scaling the cumulative loss based on the batch size processed during each iteration. This step ensures an appropriate scaling of the total loss relative to the batch size. The accumulated loss is the foundation for computing the average training loss after each epoch. By integrating this new layer of training into the model, we can enhance its performance on real-world data, akin to a form of data augmentation.

6 Evaluation

We now delve into the evaluation of the attacks and FaultGuard system. Our evaluation comprehends all scenarios detailed in the previous sections. We first give details on the dataset used for our evaluation in Sect. 6.1. To provide a baseline evaluation to discuss the success of our attacks and defenses, we evaluate our models on different tasks in Sect. 6.2. We then evaluate our attacks in Sect. 6.3, and finally study the capabilities of FaultGuard in Sect. 6.4.

6.1 Dataset

In the electrical industry, a variety of simulation programs are being used to address fault prediction challenges, including PSCAD [10,25], MATLAB Simulink [46], RSCAD [39], and MATPOWER [21]. Despite the extensive use of these simulation tools in smart grid failure prediction systems, there is a lack of publicly available datasets generated from these tools. So we turn to the dataset introduced by Ardito et al. [5], the only publicly available dataset including substantial simulated fault data rooted in the IEEE-13 test node feeder. The IEEE-13 node test feeder includes a 4.16 kV voltage generator, 13 buses for fault simulation, and three-phase signal measurement. The distribution system is divided into four zones, which are used to identify the location of a fault that has occurred. This dataset comprises 51 features and two target classes: fault label and fault zone, incorporating both traditional and renewable energy sources. It has been carefully curated to serve as a benchmark for assessing the effectiveness of adversarial attacks against fault prediction systems in smart electrical grids. Moreover, we introduce a robust windowing technique to handle

our data effectively. This involves partitioning the dataset into segments, each of a predefined size. These windows are created by iteratively traversing the data with a step size equal to half of the window size. Specifically, we choose a window size of 16 s for our dataset. To enhance the dataset's quality and optimize it for our prediction models, we conduct essential preprocessing steps, chief among them being normalization. This critical process ensures consistent data quality and mitigates potential biases that may arise from variations in feature magnitudes. We divide our dataset into three subsets: training (85% of the dataset), validation (5% of the dataset), and test (10% of the dataset).

6.2 Baseline Evaluation

In this phase, we look at the evaluation of our fault prediction system. Initially, we gauge the baseline performance of our system without incorporating countermeasures or exposure to adversarial attacks. The training phase involves utilizing the training data, and subsequently, we evaluate the effectiveness of our GRU-based fault prediction system on the test set. The results are notable, with our model achieving a mean accuracy of 0.604 ± 0.01 for fault type prediction and an accuracy of 0.958 ± 0.01 for fault zone prediction. This performance marks a substantial improvement (a mean 33.11% increase) compared to the state-of-the-art [5]. When replicating this model from the literature, we implemented our preprocessing methods and adjusted the seed and computing environment to match those of our proposed model. These modifications may have influenced the discrepancies observed between the reported results in the paper and our findings. Also, we have chosen to integrate classical ML algorithms such as XGBoost, Random Forest, and Decision Tree into our analysis for two primary reasons. Firstly, they serve as a baseline for comparison against our proposed models, enabling us to gauge the performance and efficacy of our approaches. Secondly, their inclusion underscores the significance of considering causality between data points in this task, highlighting the importance of leveraging advanced techniques to capture temporal dependencies within the data. The detailed results are presented comprehensively in Table 1.

Table 1. Comparison of the model's accuracy.

Model	Accuracy	
	Fault Type	Fault Zone
MLP reproduced from [5]	0.407	0.800
MLP claimed by [5]	0.460	0.710
Decision Tree	0.522	0.818
Random Forest	0.543	0.831
XGBoost	0.560	0.841
GRU	**0.604**	**0.958**

Combinatorial Accuracy. While our fault prediction system demonstrates superior evaluation accuracy compared to existing literature, the practical implementation necessitates a nuanced approach to address misclassification events. Issuing notifications to grid operators for each detected fault could potentially result in multiple false alarms, leading to operational challenges. To address this concern, we adopt a strategy where we wait for the identification of multiple consecutive fault data batches before triggering a notification. This approach introduces the concept of *combinatorial accuracy*, which considers the number of consecutive faulty batches required to initiate an alert. This concept is crucial for balancing the trade-off between efficient fault detection and minimizing false alarms, ensuring system robustness in real-world scenarios. The formula for combinatorial accuracy reflects a geometric distribution and is expressed as follows:

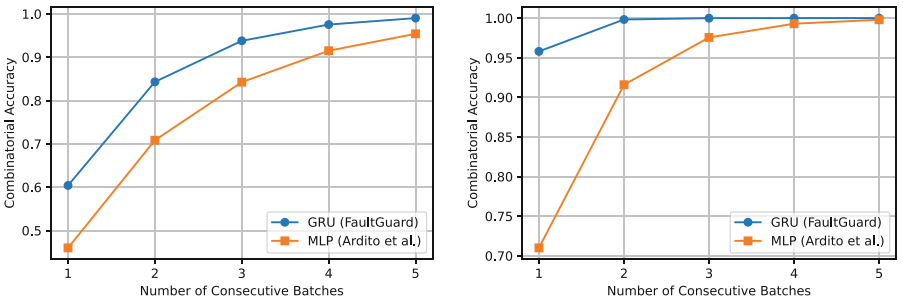
$$combinatorial\_accuracy = \left(1 - (1 - accuracy)^{batches}\right). \tag{1}$$

This formulation ensures that a higher accuracy value is associated with each notification, providing a more reliable indication of actual fault occurrences. The relationship between this accuracy value and the number of consecutive faulty batches is thoroughly analyzed and illustrated in Fig. 3, offering valuable insights into the system's performance under this combinatorial accuracy framework. As evident, the simple strategy of awaiting confirmation from another faulty batch of data significantly enhances the model's accuracy for fault type prediction, improving from 0.604 to a score of 0.843. Likewise, for fault zone classification, accuracy rises from 0.958 to 0.998. This approach drastically reduces the probability of issuing a false alarm notification to grid operators. Specifically, the probability is minimized to 13.7% for fault type prediction and 0.2% for fault zone prediction. This underscores our methodology's effectiveness in elevating accuracy and mitigating the risk of generating false alarms, contributing to a more reliable fault prediction system. Therefore, we opt for a notification delay parameter of two batches of unauthorized data, as it strikes a balanced trade-off between minimizing the false alarm rate and the timely data collection.

6.3 Attacks Evaluation

We now evaluate our attacks against the fault prediction models. We divide the evaluation into the scenarios discussed in the threat model in Sect. 3, i.e., white-box attacks and gray-box attacks.

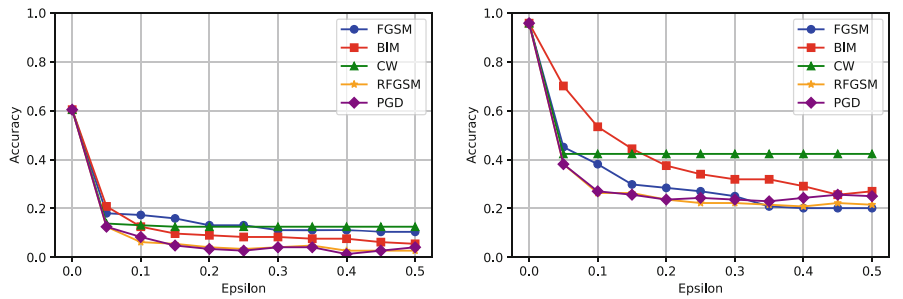
White-box Evaluation. In this evaluation, we comprehensively assess the efficacy of the white-box attacks, as discussed in Sect. 4. To implement these attacks, we leverage the TorchAttacks library [26], probing the baseline system to evaluate the susceptibility of our model without incorporating any countermeasures or defenses. The attacks are executed with varying epsilon values, signifying the strength of each attack and the degree of perturbation introduced. Specifically, we explore epsilon values ranging from 0.05 to 0.50. The outcomes of these



(a) Fault type prediction. (b) Fault zone prediction.

Fig. 3. Combinatorial accuracy of fault prediction tasks when delaying notification by the number of consecutive faulty batches.

attacks across different tasks are visually presented in Fig. 4. Notably, even with a minimal epsilon of 0.05, a significant decline in the performance of all models is evident. In this case, the accuracy of the prediction model drops to (an average of) 0.155 for fault type and 0.467 for fault zone. For the reproduced MLP model from [5], the accuracy drops to (an average of) 0.070 for fault type and 0.178 for fault zone under the FGSM attack. This evaluation underscores the baseline system's vulnerability to white-box attacks, shedding light on the need for robust defenses and countermeasures to fortify our fault prediction model against adversarial threats.



(a) Fault type prediction accuracy. (b) Fault zone prediction accuracy.

Fig. 4. Model's accuracy at varying epsilon values on the white-box attacks.

Gray-box Evaluation. In this section, we thoroughly examine the effectiveness of our gray-box attack, as detailed in Sect. 4.2. This attack is executed using the generator component within our GAN model, originally designed for anomaly

detection. The key aspect of our approach is the use of binary cross-entropy loss, a fundamental element in the GAN framework. The successful training of our GAN relies on specific hyper-parameters, including 150 training epochs and a learning rate of 2×10^{-4}, which governs the optimization process. After completing the training of our GAN, we generate a considerable volume of synthetic malicious data, totaling around 1500 batches. These synthetic data samples, crafted from random noise by our trained generator, are fed into our fully trained fault prediction system for classification. The results show that the generator produces data that can be classified as belonging to 9 out of 11 classes in the fault type prediction task and 3 out of 4 classes in the fault zone prediction task. Effectively, this translates to an Attack Success Rate (ASR) of 0.818 for fault type prediction and 0.750 for fault zone prediction. This underscores that even without prior knowledge of the fault prediction models, an adversary with access to data can create deceptive, malicious data that can elude classic ADS employed in the smart grid.

6.4 FaultGuard Evaluation

In this section, we evaluate the performance of FaultGuard under the attacks described in the previous section. First, we evaluate the effectiveness of the ADS module against white-box attacks. Then, we evaluate our ADS against white-box and gray-box attacks.

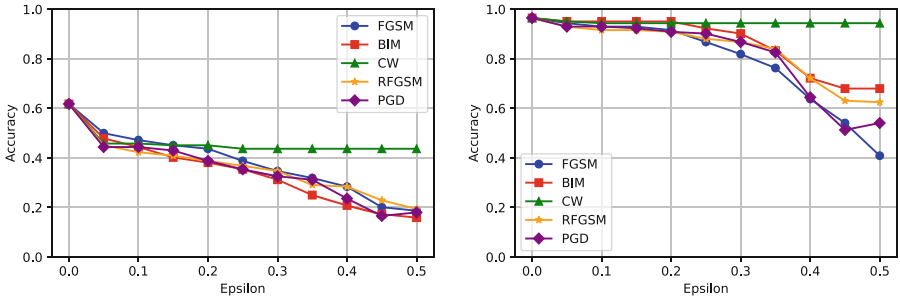
ADS Evaluation. In this section, we conduct an experimental evaluation of our ADS, a critical component of our defense strategy against GAN-based attacks and various white-box attacks, as detailed in Sect. 5. We initiate the process by utilizing the training dataset to train our GAN model. We retain the trained discriminator, a crucial element for our subsequent evaluation. The evaluation involves merging generated malicious data with authentic test data from our dataset, simulating malicious attempts alongside real data. These merged datasets are then subjected to the discriminator for analysis, utilizing a predefined threshold (0.5) for discerning determinations between legitimate and anomalous data. The results underscore the discriminator's effectiveness, achieving a mean accuracy rate of 0.991 ± 0.005 standard deviation when classifying real data from malicious data generated by our GAN in the fault type prediction task and a mean accuracy rate of 0.972 ± 0.005 standard deviation in the fault zone prediction task. We subject our model to different white-box adversarial attacks in the subsequent phase. We generate adversarial attacks targeting the fault type and fault zone prediction models, individually applying these attacks to the models. The results are detailed in Table 2, highlighting the ADS's resilience and the contributions of our adversarial learning training layer. The ADS achieves a mean accuracy rate of 0.979 ± 0.050 standard deviation when classifying real data from malicious data generated by our white-box attacks in the fault type prediction task and a mean accuracy rate of 0.821 ± 0.050 standard deviation in the fault zone prediction task.

Table 2. ADS accuracy for each attack and considering the Adversarial Learning (AL) layer. Results are averaged for each ϵ value.

Task	AL	Accuracy				
		FGSM	BIM	CW	RFGSM	PGD
Fault Type	✗	0.674 ± 0.032	0.261 ± 0.067	0.386 ± 0.006	0.703 ± 0.009	0.700 ± 0.007
	✓	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{0.897 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
Fault Zone	✗	0.274 ± 0.032	0.241 ± 0.011	0.232 ± 0.000	0.300 ± 0.004	0.293 ± 0.008
	✓	$\mathbf{1.000 \pm 0.000}$	$\mathbf{0.999 \pm 0.001}$	0.108 ± 0.002	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$

Despite the ADS's robust performance against various white-box adversarial attacks, it exhibits vulnerability to the CW attack, a sophisticated adversarial technique known for its intricacy. While the ADS may occasionally miss some batches of adversarial attacks, particularly those crafted using the intricate CW technique, the subsequent layer of defense comes into play. The adversarial learning layer ensures the fault prediction models' resilience against these advanced attacks. In only one case (i.e., fault zone + AL against CW attacks), we notice a drop in performance with the inclusion of our novel layer. This standalone case is caused by the missing inclusion of the CW attack in the layer, and more details are given in Sect. 7. As elucidated in upcoming sections, our fault prediction models showcase remarkable resilience even when confronted with the CW attack, successfully averting a significant decline in system performance. The multi-layered defense strategy underscores a comprehensive approach aimed at enhancing the overall robustness and effectiveness of the fault prediction system in smart electrical grids.

Fault Prediction Model Evaluation. This section comprehensively evaluates our fault prediction systems, augmented with online adversarial training as a defense mechanism. As observed in Sect. 6.3, our model, while surpassing the state-of-the-art, remains vulnerable to white-box adversarial attacks. We introduce a novel training layer to fortify our models against such attacks, extensively discussed in Sect. 5.2. After integrating this new training layer, we assess our models' performance in fault type and zone prediction critical tasks. The outcomes of this evaluation are presented in Fig. 5. Notably, the resilience of our models against adversarial attacks, particularly in comparison to models without the defense mechanism, exhibits a substantial improvement. This enhancement is particularly pronounced when facing sophisticated attacks like CW, which have demonstrated the ability to bypass our ADS system. The results underscore the efficacy of the introduced adversarial training layer in bolstering the model's robustness, significantly mitigating the impact of adversarial attacks on fault prediction performance. This defense mechanism is a crucial safeguard, ensuring our fault prediction systems' continued reliability and effectiveness, even in the face of sophisticated and intricate adversarial challenges. A summary of the results of our models and their improvement when paired with the defense mechanism is shown in Table 3.



(a) Fault type prediction accuracy. (b) Fault zone prediction accuracy.

Fig. 5. Model's accuracy at varying epsilon values on the white-box attacks when equipped with the defense mechanism.

Table 3. Comparison of model performance under adversarial attacks ($\epsilon = 0.05$) before and after employing our Online Adversarial Training (OAT).

Task	OAT	Accuracy					
		Baseline	FGSM	BIM	CW	RFGSM	PGD
Fault Type	✗	0.604	0.180	0.208	0.138	0.125	0.125
	✓	**0.618**	**0.500**	**0.479**	**0.458**	**0.451**	**0.444**
Fault Zone	✗	0.958	0.451	0.701	0.423	0.381	0.381
	✓	**0.965**	**0.944**	**0.951**	**0.951**	**0.930**	**0.930**

Computational Cost. All experiments in this paper have been conducted on Kaggle as a cloud resource with the following configurations: Intel Xeon (2.20 GHz), NVIDIA Tesla P100 (3584 Cuda cores, 16 GB), 30 GB of RAM, Linux Debian with Python 3.10.12[1]. The defense strategies have shown commendable performance through lightweight models and two white-box adversarial attacks. However, a notable trade-off is the introduction of additional computational load. The ADS and fault prediction systems have been augmented with two adversarial data generators and incorporated two new loss calculations for the prediction models to enhance robustness. As a result, there is a significant increase in computational demands, leading to longer training times for the models. The changes significantly affect training times, as depicted in Table 4, comparing standard models to those fortified with defense mechanisms. Despite the increased computational cost, it is crucial to consider the enhanced robustness and resilience brought by these defenses. Furthermore, this training procedure is needed only when initially deploying the model in the system and does not require maintenance. This computational trade-off underscores the ongoing challenge of bal-

[1] Additional details on packages' versions are available at: https://github.com/emadef1/FaultGuard/blob/main/requirements.txt.

Table 4. Model training times with and without Online Adversarial Training (OAT).

Task	OAT	Training Time
Fault Type	✗	1.18 min
	✓	78.90 min
Fault Zone	✗	1.29 min
	✓	95.00 min

ancing model efficiency with the imperative to defend against adversarial threats in the smart grid domain.

7 Takeaways

Fault prediction systems are a well-researched topic in the literature. However, researchers often neglect the security aspect of these systems, making their implementation problematic due to the high chances of errors when dealing with adversaries. Therefore, we present a summary of the key takeaway messages, making it easier for practical implementation in real-world scenarios. By doing so, we assist practitioners in effectively applying these systems and offer researchers recommended best practices for their studies.

Takeaway 1 – *Fault prediction systems are vulnerable to adversarial attacks, regardless of the scenario's assumptions on the attacker's knowledge.*

As discussed in Sect. 6.3, adversarial attacks are particularly effective against the models tested in this study. Indeed, model accuracy dropped by up to 74.34% with an epsilon value of just 0.05. While these values are valid only for white-box scenarios, it is worth noting that even in gray-box scenarios (i.e., having access to the data), the attacker has great leverage on the system, reaching ASRs up to 0.818. As such, adversarial attacks require particular consideration when designing a ML-based fault prediction system.

Takeaway 2 – *Complex attacks, such as CW, are more difficult to be detected by the ADS.*

As shown in Table 2, our ADS can detect most attacks with perfect accuracy, except for the CW attack. While in the fault type prediction task we obtained an accuracy of 0.897 with the addition of the adversarial learning layer, in the fault zone task, including this layer effectively worsened its performance (from 0.232 to 0.108). It is noteworthy that adversarial learning is performed only with the FGSM and BIM attacks. However, even when including CW in the training process, there were no improvements w.r.t. the standard ADS model against the same attack. For these reasons, complex attacks require special attention, as improperly implementing the ADS might not be enough to detect them.

Takeaway 3 – *Complex attacks, such as CW, are more effective against good-performing models.*

Another correlation regarding the CW attack that can be extracted from our evaluation is that attacks are more effective against good-performing models. Indeed, the fault type prediction model (baseline accuracy of 0.604) was less affected by the attacks w.r.t. the fault zone prediction model (baseline accuracy of 0.958). As such, while researchers and practitioners strive to reach the highest possible accuracy score, the threat of adversarial attacks becomes more pronounced. This highlights the paradoxical relationship between model performance and susceptibility to attacks. Consequently, the pursuit of high accuracy should be accompanied by a heightened awareness of potential vulnerabilities and the implementation of robust defense mechanisms.

Takeaway 4 – *Different forms of adversarial training greatly improve the models' resistance against adversaries.*

One of the significant contributions of this paper is the proposal of novel adversarial training techniques. Indeed, including the adversarial learning layer on the ADS significantly improved its performance, and the online adversarial training performed on the fault prediction model made it more resilient against attacks. These findings underscore the effectiveness and versatility of diverse adversarial training methodologies in improving model robustness and defense capabilities, contributing valuable insights to advancing secure and resilient machine learning models.

8 Conclusions

In the smart grid, fault prediction systems are promising tools that can ensure energy delivery and reliability. However, despite the growing interest in the literature, the security aspect of these systems is often neglected. This oversight can lead to safety issues and delays, making the implementation of those systems counterproductive.

Contribution. In this paper, we introduced **FaultGuard**, a resilient framework designed for fault type and zone classification tasks. To ensure the security of our system, we incorporated an ADS with a unique GAN training layer to detect attacks. Additionally, our approach involved a low-complexity fault prediction model and employed an online adversarial training technique to bolster robustness. We thoroughly evaluated the framework's performance, assessing its resilience against diverse adversarial attacks using the publicly available IEEE13-AdvAttack dataset, a simulated dataset derived from the IEEE-13 test node feeder. FaultGuard outclassed the state-of-the-art, reaching accuracy values up to 0.958 and being natively resilient against adversarial attacks. Furthermore, our ADS detected attack attempts with accuracies up to 1.000.

Future Work. While still outclassing the state-of-the-art and other ML models in the same task, improving the accuracy for fault type prediction is needed to further strengthen its contribution. Indeed, the accuracy value achieved in this study is highly dependent on the dataset employed. As such, creating a richer dataset is the most significant way to improve these results. Also, since this paper focused on evasion attacks towards fault prediction systems, studying poisoning attacks might provide powerful insights into the resilience of these systems. By delineating a new system and threat model accounting for this threat, it might be possible to study the vulnerabilities of the models and thus improve their security. Finally, by combining these results with those obtained in this paper, we would be able to provide a complete overview of the resilience of ML-based fault prediction models, aiding practitioners in safely deploying these systems.

References

1. Al-Shaher, M.A., Sabry, M.M., Saleh, A.S.: Fault location in multi-ring distribution network using artificial neural network. Electric Power Syst. Res. **64**(2), 87–92 (2003)
2. Alecci, M., Conti, M., Marchiori, F., Martinelli, L., Pajola, L.: Your attack is too dumb: formalizing attacker scenarios for adversarial transferability. In: Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses, pp. 315–329 (2023)
3. Andresen, C., Torsæter, B.N., Haugdal, H., Uhlen, K.: Fault detection and prediction in smart grids, pp. 1–6 (2018). https://doi.org/10.1109/AMPS.2018.8494849
4. Apruzzese, G., Anderson, H.S., Dambra, S., Freeman, D., Pierazzi, F., Roundy, K.: Real attackers don't compute gradients: bridging the gap between adversarial ml research and practice. In: 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), pp. 339–364 (2023)
5. Ardito, C., Deldjoo, Y., Di Noia, T., Di Sciascio, E., Nazary, F.: Ieee13-advattack a novel dataset for benchmarking the power of adversarial attacks against fault prediction systems in smart electrical grid. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM 2022, pp. 3817-3821. ACM (2022). https://doi.org/10.1145/3511808.3557612
6. Aslan, Y.: An alternative approach to fault location on power distribution feeders with embedded remote-end power generation using artificial neural networks. Electr. Eng. **94**, 125–134 (2012)
7. Bayindir, R., Colak, I., Fulli, G., Demirtas, K.: Smart grid technologies and applications. Renew. Sustain. Energy Rev. **66**, 499–516 (2016)
8. Bhattacharya, B., Sinha, A.: Intelligent fault analysis in electrical power grids. In: 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 985–990 (2017). https://doi.org/10.1109/ICTAI.2017.00151
9. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 39–57. IEEE Computer Society, May 2017. https://doi.org/10.1109/SP.2017.49
10. Chakraborty, S., Das, S.: Application of smart meters in high impedance fault detection on distribution systems. IEEE Trans. Smart Grid **10**(3), 3465–3473 (2019). https://doi.org/10.1109/TSG.2018.2828414

11. Chen, K., Huang, C., He, J.: Fault detection, classification and location for transmission lines and distribution systems: a review on the methods. High Voltage **1**(1), 25–33 (2016). https://doi.org/10.1049/hve.2016.0005
12. Chen, T.M., Abu-Nimeh, S.: Lessons from stuxnet. Computer **44**(4), 91–93 (2011)
13. Coser, J., do Vale, D.T., Rolim, J.G.: Design and training of artificial neural networks for locating low current faults in distribution systems. In: 2007 International Conference on Intelligent Systems Applications to Power Systems, pp. 1–6 (2007). https://doi.org/10.1109/ISAP.2007.4441599
14. De La Cruz, J., Gómez-Luna, E., Ali, M., Vasquez, J.C., Guerrero, J.M.: Fault location for distribution smart grids: literature overview, challenges, solutions, and future trends. Energies **16**(5) (2023). https://doi.org/10.3390/en16052280
15. Dehghani, F.: CIRED -open access. Proc. J. **2017**(3), 1134–1137 (2017)
16. El Mrabet, Z., Sugunaraj, N., Ranganathan, P., Abhyankar, S.: Random forest regressor-based approach for detecting fault location and duration in power systems. Sensors **22** (2022). https://doi.org/10.3390/s22020458
17. Farias, P.E., de Morais, A.P., Rossini, J.P., Cardoso, G.: Non-linear high impedance fault distance estimation in power distribution systems: a continually online-trained neural network approach. Electric Power Syst. Res. **157**, 20–28 (2018). https://doi.org/10.1016/j.epsr.2017.11.018
18. Ghaemi, A., Safari, A., Afsharirad, H., Shayeghi, H.: Accuracy enhance of fault classification and location in a smart distribution network based on stacked ensemble learning. Electric Power Syst. Res. **205**, 107766 (2022). https://doi.org/10.1016/j.epsr.2021.107766
19. Gong, X., Wang, Q., Chen, Y., Yang, W., Jiang, X.: Model extraction attacks and defenses on cloud-based machine learning models. IEEE Commun. Mag. **58**(12), 83–89 (2020)
20. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples (2015)
21. He, M., Zhang, J.: A dependency graph approach for fault detection and localization towards secure smart grid. IEEE Trans. Smart Grid **2**(2), 342–351 (2011). https://doi.org/10.1109/TSG.2011.2129544
22. Hussain, N., Nasir, M., Vasquez, J.C., Guerrero, J.M.: Recent developments and challenges on ac microgrids fault detection and protection systems-a review. Energies **13** (2020). https://doi.org/10.3390/en13092149
23. Jagielski, M., Carlini, N., Berthelot, D., Kurakin, A., Papernot, N.: High accuracy and high fidelity extraction of neural networks. In: 29th USENIX security symposium (USENIX Security 20), pp. 1345–1362 (2020)
24. Javadian, S., Nasrabadi, A., Haghifam, M.R., Rezvantalab, J.: Determining fault's type and accurate location in distribution systems with dg using MLP neural networks. In: 2009 International Conference on Clean Electrical Power, pp. 284–289 (2009). https://doi.org/10.1109/ICCEP.2009.5212044
25. Jiang, H., Zhang, J.J., Gao, W., Wu, Z.: Fault detection, identification, and location in smart grid based on data-driven computational methods. IEEE Trans. Smart Grid **5**(6), 2947–2956 (2014). https://doi.org/10.1109/TSG.2014.2330624
26. Kim, H.: Torchattacks: a pytorch repository for adversarial attacks. arXiv preprint arXiv:2010.01950 (2020)
27. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world (2017)
28. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks (2019)

29. Mahfouz, M.M., El-Sayed, M.A.: Smart grid fault detection and classification with multi-distributed generation based on current signals approach. IET Gener. Transm. Distrib. **10**(16), 4040–4047 (2016). https://doi.org/10.1049/iet-gtd.2016.0364
30. Majidi, M., Arabali, A., Etezadi-Amoli, M.: Fault location in distribution networks by compressive sensing. IEEE Trans. Power Delivery **30**(4), 1761–1769 (2015). https://doi.org/10.1109/TPWRD.2014.2357780
31. Mousa, M., Abdelwahed, S., Klüss, J.: Review of diverse types of fault, their impacts, and their solutions in smart grid, pp. 1–7, April 2019. https://doi.org/10.1109/SoutheastCon42311.2019.9020355
32. Muqeet, H.A., Liaqat, R., Jamil, M., Khan, A.A.: A state-of-the-art review of smart energy systems and their management in a smart grid environment. Energies **16**(1), 472 (2023)
33. Nafees, M.N., Saxena, N., Cardenas, A., Grijalva, S., Burnap, P.: Smart grid cyber-physical situational awareness of complex operational technology attacks: a review. ACM Comput. Surv. **55**(10), 1–36 (2023)
34. Okumus, H., Nuroglu, F.M.: A random forest-based approach for fault location detection in distribution systems. Electr. Eng. **103**(1), 257–264 (2021)
35. Rahman Fahim, S., K. Sarker, S., Muyeen, S.M., Sheikh, M.R.I., Das, S.K.: Microgrid fault detection and classification: machine learning based approach, comparison, and reviews. Energies **13** (2020). https://doi.org/10.3390/en13133460
36. Saha, M.M., Izykowski, J.J., Rosolowski, E.: Fault location on power networks. Springer Science & Business Media (2009)
37. Sapountzoglou, N., Lago, J., Raison, B.: Fault diagnosis in low voltage smart distribution grids using gradient boosting trees. Electric Power Syst. Res. **182**, 106254 (2020). https://doi.org/10.1016/j.epsr.2020.106254
38. Shadi, M.R., Ameli, M.T., Azad, S.: A real-time hierarchical framework for fault detection, classification, and location in power systems using PMUS data and deep learning. Int. J. Electr. Power Energy Syst. **134**, 107399 (2022). https://doi.org/10.1016/j.ijepes.2021.107399
39. Shafiullah, M., Abido, M.A.: S-transform based FFNN approach for distribution grids fault detection and classification. IEEE Access **6**, 8080–8088 (2018). https://doi.org/10.1109/ACCESS.2018.2809045
40. Souza, J., Rodrigues, M., Schilling, M., Do Coutto Filho, M.: Fault location in electrical power systems using intelligent systems techniques. IEEE Trans. Power Delivery **16**, 59–67 (2001). https://doi.org/10.1109/61.905590
41. Stefanidou-Voziki, P., Sapountzoglou, N., Raison, B., Dominguez-Garcia, J.: A review of fault location and classification methods in distribution grids. Electric Power Syst. Res. **209**, 108031 (2022). https://doi.org/10.1016/j.epsr.2022.108031
42. Sullivan, J.E., Kamensky, D.: How cyber-attacks in ukraine show the vulnerability of the us power grid. Electr. J. **30**(3), 30–35 (2017)
43. Thukaram, D., Khincha, H., Vijaynarasimha, H.: Artificial neural network and support vector machine approach for locating faults in radial distribution systems. IEEE Trans. Power Delivery **20**(2), 710–721 (2005). https://doi.org/10.1109/TPWRD.2005.844307
44. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P.: Ensemble adversarial training: attacks and defenses (2020)
45. Usman, M.U., Ospina, J., Faruque, M.O.: Fault classification and location identification in a smart distribution network using ann. In: 2018 IEEE Power & Energy Society General Meeting (PESGM), pp. 1–6 (2018). https://doi.org/10.1109/PESGM.2018.8586471

46. Wang, X., et al.: High impedance fault detection method based on variational mode decomposition and teager-kaiser energy operators for distribution network. IEEE Trans. Smart Grid **10**(6), 6041–6054 (2019). https://doi.org/10.1109/TSG.2019.2895634
47. Wilches-Bernal, F., Jiménez-Aparicio, M., Reno, M.J.: An algorithm for fast fault location and classification based on mathematical morphology and machine learning. In: 2022 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), pp. 1–5 (2022). https://doi.org/10.1109/ISGT50606.2022.9817473
48. Zhang, F., Liu, Q., Liu, Y., Tong, N., Chen, S., Zhang, C.: Novel fault location method for power systems based on attention mechanism and double structure GRU neural network. IEEE Access **8**, 75237–75248 (2020). https://doi.org/10.1109/ACCESS.2020.2988909
49. Zidan, A., Khairalla, M., Abdrabou, A.M., Khalifa, T., Shaban, K., Abdrabou, A., El Shatshat, R., Gaouda, A.M.: Fault detection, isolation, and service restoration in distribution systems: State-of-the-art and future trends. IEEE Trans. Smart Grid **8**(5), 2170–2185 (2017). https://doi.org/10.1109/TSG.2016.2517620