2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA) | 979-8-3503-2385-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/TPS-ISA58951.2023.00021

Mitigating Targeted Universal Adversarial Attacks on Time Series Power Quality Disturbances Models

Sultan Uddin Khan[1], Mohammed Mynuddin[2], Isaac Adom[3], Mahmoud Nabil Mahmoud[4]
School of Electrical and Computer Engineering, North Carolina A & T State University, Greensboro, NC, USA
Email: skhan5@aggies.ncat.edu, mmynuddin@aggies.ncat.edu, iadom@aggies.ncat.edu, mnmahmoud@ncat.edu

***Abstract*—The utilization of deep learning models has been widely recognized for its significant contribution to the enhancement of smart grid operations, particularly in the domain of power quality disturbance (PQD) classification. Nevertheless, the emergence of vulnerabilities like targeted universal adversarial attacks can significantly undermine the reliability and security of deep learning models. These attacks can exploit the model's weaknesses, causing it to misclassify PQDs with potentially catastrophic consequences. In our previous research, we for the first time examined the vulnerability of deep learning models to targeted universal adversarial attacks on time series data in smart grids by introducing a novel algorithm that effectively attacks by maintaining a trade-off between fooling rate and imperceptibility. While this attack method demonstrated notable efficacy, it also emphasized the pressing need for robust defensive mechanisms to safeguard these critical systems. This paper provides a thorough examination and evaluation of different defense strategies, specifically adversarial training, defensive distillation, and feature squeezing, in order to identify the most effective method for mitigating targeted universal adversarial (TUA) attacks on time series data for three different types of imperceptibility (high, medium and low). Based on our analysis, adversarial training demonstrates a significant reduction in the success rate of attacks. Specifically, the technique reduced fooling rates by an average of 23.73% for high imperceptibility, 31.04% for medium imperceptibility, and a substantial 42.96% for low imperceptibility. These findings highlight the crucial role of adversarial training in enhancing the integrity of deep learning applications.**

***Index Terms*—targeted universal adversarial attack, time series data, adversarial training, smart grid, deep learning**

I. Introduction

Smart grids have revolutionized the electricity industry by incorporating advanced technologies and data-driven approaches to enhance energy efficiency, reliability, and sustainability. For instance, a study conducted by the U.S. Department of Energy found that smart grid technologies, including advanced metering infrastructure (AMI) and distribution automation, led to a 5% reduction in peak electricity demand and a 6% reduction in overall energy consumption [1]. This translates to significant cost savings and environmental benefits. However, power quality disturbances (PQDs) can have a detrimental impact on the performance of a smart grid system. PQDs in a power grid system refer to variations or deviations from the ideal or expected electrical waveform, which can lead to undesirable effects on the quality and reliability of electrical power. These disturbances can be caused by various factors, including voltage sags or dips, voltage swells or surges, voltage interruptions, harmonics, transients, and flicker [2]. Voltage sags occur when there is a brief reduction in voltage, which can affect sensitive equipment and cause malfunctions or downtime. Voltage swells, on the other hand, involve temporary increases in voltage, potentially damaging equipment. Voltage interruptions result in a complete loss of power for a short period, causing inconvenience and potential damage to electronic devices. Harmonics, which are unwanted frequencies, can distort waveforms and impact the efficient operation of equipment. Transients are short-duration voltage fluctuations caused by lightning, switching operations, or faults, which can cause equipment failures. Flicker refers to rapid changes in voltage that can lead to visual discomfort or affect the performance of sensitive equipment.

In a smart grid system, various methods are employed to detect PQDs, including the utilization of machine learning (ML) techniques [3]. Traditional methods involve using specialized monitoring devices, such as power quality analyzers and sensors, to measure voltage, current, and other electrical parameters [4]. These devices continuously monitor the grid and capture disturbances. ML algorithms enable the analysis of vast amounts of data from diverse sources, such as smart meters, sensors, and weather forecasts. This data-driven approach empowers the smart grid system to gain real-time insights, make intelligent decisions, identify patterns, predict electricity demand, detect anomalies, and optimize energy distribution, thus significantly improving grid efficiency and reliability. Moreover, ML algorithms facilitate load forecasting [5], energy storage management [6], and seamless integration of renewable energy sources, enabling the smart grid to dynamically adapt and respond to changing conditions. Continuously learning from data and refining its predictions, the ML-based smart grid system holds immense potential in revolutionizing the energy infrastructure.

Recent research has revealed a concerning vulnerability in various families of ML models known as adversarial examples [7]. Adversarial examples are carefully crafted inputs that are designed with the intention of misleading the target model into generating incorrect or unexpected outputs. This vulnerability extends across different types of ML models, including deep neural networks, support vector machines, and decision trees, among others [8]–[11]. These findings raise concerns regarding the reliability ML models, especially in safety-critical applications. The susceptibility to adversarial examples undermines the ability of ML models to make accurate and reliable predictions, potentially leading to severe

979-8-3503-2385-6/23/$31.00 ©2023 IEEE
DOI 10.1109/TPS-ISA58951.2023.00021

consequences. In the case of PQDs, the misclassification of such events could result in incorrect decision-making and inadequate response measures, jeopardizing the stability of the power grid system and the safety of connected devices [12], [13].

The potential danger lies in the striking similarity between adversarial signal attacks and benign signals, which can potentially deceive human operators and lead to a failure in detecting and recognizing ongoing attacks. Previous studies have examined the vulnerability of deep learning models to specific adversarial attacks in power systems [14] and evaluated defense mechanisms against non-targeted attacks [15]. However, there is a lack of research on targeted universal adversarial (TUA) perturbations in time series data within this context. In TUA, adversarial examples are crafted to be transferable across different instances of the same model architecture. In other words, these attacks are designed to be universally effective, meaning that the same adversarial example can mislead the target model, regardless of the data on which the model was trained. In this subsequent installment, we aim to further elaborate on our prior research [16] and introduce an investigation into diverse defense strategies. These strategies encompass adversarial training, defense distillation, and feature squeezing, intending to determine the most resilient countermeasure against TUA attacks specifically targeted at time series data. This work present a holistic defense strategy specifically designed to address the intricacies of time series data within the smart grid domain. The key contributions of our manuscript are as follows:

- Our research endeavors to lead the way in investigating strategies to counter targeted universal adversarial (TUA) attacks on time series data, thereby expanding the existing knowledge on adversarial attacks.
- Three well-known defense strategies, adversarial training, defensive distillation, and feature squeezing, are specifically adapted and evaluated by us. We use a moving average technique for feature squeezing to minimize background noise and maximize the smoothness of the waveform within each batch.
- After analyzing fooling rates, adversarial training is the best defense against targeted universal adversarial attacks on time series data. Adversarial training consistently lowered fooling rates to near zero. It decreased high imperceptibility by 23.73%, middle by 31.04%, and low by 42.96% on average. Defensive distillation showed promise, but adversarial training was more effective. Feature squeezing had inconsistent effects.

The remaining sections of this paper are organized as follows. In Section II, we describe the related work. Section III is a brief presentation of TUA attacks on neural networks. Section IV focuses on the overview of defense methods against adversarial targeted attacks. The experiments, results, and discussions are presented comprehensively in Section V. Finally; Section VI summarizes the conclusions drawn from the study.

II. Related Work

Numerous scholarly investigations have been conducted to examine the effects of cyber attacks on the smart grid, encompassing both untargeted and targeted attacks. In [17], the authors investigate the effects of adversarial attacks on convolutional neural network-based frameworks for event cause analysis. In their study, Kosut, Oliver, et al. [18] conducted a cyber attack on a smart grid system using a technique known as malicious data injection. In their study, Cheng et al. [19] employed various adversarial attack techniques to introduce noise signals into the input time series of Phasor Measurement Units (PMUs). Their findings demonstrate that existing deep learning-powered event classifiers for power systems are highly vulnerable to these attacks, posing a potential threat to the reliability of power transmission systems.

Concurrently, efforts are being made to develop robust defenses, including adversarial training [20], input transformations, and [21], which aim to enhance the model's ability to resist adversarial attacks. In [22], the authors examine the security challenges associated with neural network-based state estimation in the smart grid. It specifically addresses the problem of adversarial attacks targeting neural network-based state estimation and presents a highly efficient adversarial attack method. Li, Jiangnan, et al. [23] investigate the security threat of false data injection attacks (FDIAs) in power system state estimation, highlighting the limitations of existing machine learning techniques, and propose an adversarial-resilient DNN-based approach that incorporates random input padding to effectively mitigate adversarial attacks while maintaining detection performance. Reference [24] explores the security implications of FDI attacks in distributed demand response systems within smart grids, revealing the vulnerability of deep learning-based FDI attack detection methods to adversarial machine learning attacks. Before our efforts, there was, to the best of our knowledge, a conspicuous lack of research addressing targeted universal adversarial attacks on time series data. Our earlier work [16] paved the way in this direction by introducing a novel algorithm tailored to this specific challenge. The present study explores defenses against such attacks on time series data, venturing deeper into uncharted territory. Specifically, we adapt and evaluate prominent defense methods to establish their efficacy within the context of time series data, thereby bridging a significant research gap in the field.

III. Targeted universal adversarial attack

In our prior research [16], we showcased a TUA attack on time series data related to PQDs. We explored the threat model of TUA attack, considering factors such as the attacker's knowledge, capabilities, and goals.

A. Threat Model

- **Attacker's Knowledge.** The attacker's level of knowledge can vary between a white-box setting, where they possess comprehensive information about the trained model, including its learning algorithm, structure, parameters, hyper-parameters, and training data, and a

black-box setting, where the attacker only has limited knowledge and is aware only of the model's inputs and outputs. This suggests that the attacker could potentially be an insider with access to certain information about the model's architecture and data or an outsider with limited information about the model's inner workings. We are focusing on white box settings in this research.

- **Attacker's Goals.** In a TUA attack, the attacker's objective is to find a single perturbation value that, when added to most clean samples, leads the model to predict the targeted class intentionally. The attacker aims to achieve this by exploiting the model's vulnerabilities and generating subtle perturbations that can deceive the model consistently across different instances.

B. Adaptive Edge algorithms

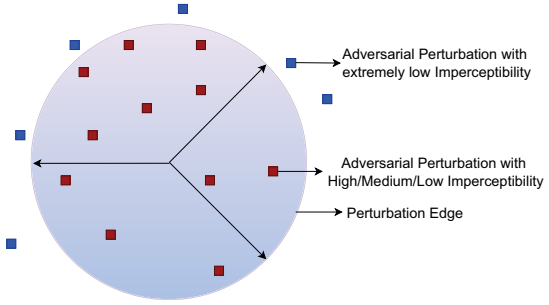

Fig. 1: Visualization of the Perturbation Edge

Our previous research introduces the methodology for constructing a TUA attack utilizing the Adaptive Edge (AdaptEdge) algorithm. The algorithm is designed to effectively deceive deep learning models that operate on time series data by striking a balance between the fooling rate and the imperceptibility of the attack to human observers. To achieve this, the algorithm relies on two essential metrics. Firstly, the fooling rate measures the attack's success in causing the deep learning model to misclassify the majority of clean time series data into the targeted class. Secondly, the signal-to-noise ratio (SNR) provides insights into the imperceptibility of the adversarial perturbation. This approach is based on the hypothesis that dynamically modifying the *perturbation edge* can improve the attack's effectiveness while maintaining the imperceptibility of the input signal. Figure 1 illustrates a conceptual depiction of the *perturbation edge* within a feature space with multiple dimensions. The semi-transparent sphere symbolizes the *perturbation edge*, which, in the present context, corresponds to the limit of permissible perturbations. Adversarial perturbations confined within the boundaries of this hypersphere exhibit subtlety and effectively impact the predictions of the model while simultaneously maintaining their imperceptibility to human observers. The scattered red dots contained within the sphere serve as representations of diverse adversarial perturbations. The fact that they are confined within the hypersphere suggests that they adhere to the boundary established by the *perturbation edge*. Any location beyond the boundaries of this sphere would indicate a conspicuous or excessive disturbance, which is likely to be perceptible by humans and thus unsuitable for stealthy adversarial attacks. The arrows originating from the central point symbolize possible directions in which data may be altered. The magnitude of their size, upon reaching the boundary of the sphere, signifies the utmost degree to which data can be altered in that specific direction while still adhering to the confines of the perturbation boundary. The visualization depicts blue dots situated beyond the perturbation boundary, indicating a significantly low imperceptibility. Therefore, a human observer would be able to perceive these perturbations. In the feature space, the *perturbation edge* is defined as the radius of the smallest hypersphere that encloses the data. This hypersphere's confinement of perturbations creates a boundary that plays a vital role in preserving the integrity of the signal. Analogous to how the "edge" of a circle in two-dimensional space represents its boundary, adversarial perturbations in multidimensional feature space are confined within a hypersphere. This restriction ensures that the perturbations remain subtle and imperceptible to human observers while effectively influencing the model's predictions.

To categorize degrees of imperceptibility as high, medium, or low, we set a specific threshold for the SNR. These levels are established by methodically varying the perturbation edge value, followed by a visual examination of the integrity of the resulting signal. We assume that the attacker is aware of these predetermined SNR thresholds and employs this knowledge to maneuver through the imperceptibility levels. The symbol "ε" is often used to indicate the radius of this hypersphere. By increasing the "edge" of the hypersphere, perturbations can exist over a wider volume. The experiment commences with a high imperceptibility for the specific source-target class pair. If the rate of fooling meets or surpasses the predetermined threshold, the algorithm proceeds to reduce the value of in order to ascertain whether a desired rate of fooling can be achieved using a smaller value than the initial . Once the desired level of deception is attained, the value of ε is subsequently reduced to guarantee a significantly elevated level of imperceptibility, specifically regarding a high signal-to-noise ratio (SNR). The reduction process persists until a fooling rate equal to or surpasses the predetermined criterion is achieved. The optimal value of ε signifies the ideal radius, establishing the optimal boundary for perturbation. On the other hand, in the event that fooling rate lower than the specified threshold is achieved utilizing the initial ε value, the algorithm proceeds to increment ε to assess whether the rate of deception approaches or surpasses the threshold. This process continues until the threshold is reached or surpassed. Subsequently, the algorithm proceeds to decrease the value of ε within the range bounded by the incremented ε and the initial ε in order to assess the potential for enhancing imperceptibility while maintaining the desired fooling rate. In the event that the rate of deception once again reaches or surpasses the predetermined threshold, the variable ε will persistently decrease until it reaches a value that is lower than the threshold. Once the rate of deception drops

below a predetermined threshold, the algorithm discontinues its process of diminishing ε and instead restores it to its previous value. The dynamic modification of the ε parameter allows for the optimal creation of adversarial perturbations while maintaining the signal's integrity. We refer the reader to [16] for further details on the Adaptive Edge (AdaptEdge) algorithm and its underlying hypothesis.

IV. Overview of Defense Method

Defending against adversarial examples involves an ongoing arms race between defenses and attacks due to the significant consequences that successful attacks may have on critical infrastructure. Resilient defense mechanisms are imperative in this context. In this work, we investigate and compare three defense techniques against TUA attacks. The first notable defense technique is adversarial training [7], introduced by Goodfellow et al. It involves augmenting the training data with adversarial samples generated using the Fast Gradient Sign Method (FGSM), enabling the model to better handle adversarial inputs. The second defense technique we explore is defense distillation [25]. Defensive distillation involves training a secondary model to approximate the output probabilities of the original model. This process forces the model to learn a more smoothed and generalized decision boundary, making it harder for adversarial perturbations to cause significant misclassifications. Lastly, we investigate the defense technique of feature squeezing [26]. Feature squeezing reduces the input data's dimensionality and quantizes it to a lower precision. By doing so, it effectively removes some of the fine-grained details that adversarial attacks typically exploit. This regularization process enhances the model's ability to resist adversarial perturbations and improves its overall robustness. This technique involves applying a moving filter across an original image and adjusting the value of the central pixel to the median value of the pixels within the filter. Suppose the discrepancy between the predicted outcome of the unaltered image and the predicted outcome of a compressed image using either of the two techniques surpasses a predetermined threshold. In that case, it can be inferred that the provided input is likely to be an adversarial instance. To mitigate the impacts of adversarial perturbations on time series data, we employ a waveform smoothing mechanism as our strategy. This mechanism aims to smooth out the fluctuations caused by adversarial perturbations in the time series data, making it more resistant to the influence of these perturbations and enhancing the model's robustness. We apply a moving average technique for the purpose of effectively reducing noise and enhancing the smoothness of individual waveforms within a given batch. Fundamentally, the process involves recalibrating each value in the waveform data by considering its adjacent values within a specified window or size. The resultant waveforms exhibit a "squeezed" characteristic, which enhances their resilience against adversarial perturbations.

The schematic representation of the defense strategy employed in our investigation is depicted in figure 2. Our research aims to thoroughly analyze and compare the three aforementioned approaches to identify the most effective strategy for mitigating TUA attacks on time series data. In order to gain insight into the alignment of our defense methodology with the overall process, it is necessary to examine the procedure for classifying PQDs. Power quality disturbances (PQDs) can arise from a wide range of sources, including household appliances, industrial equipment, and commercial facilities, and can manifest in various expressions. When these disruptions arise, they propagate throughout the transmission and distribution infrastructure until they reach the substation. When these disruptions arise, they are detected through the utilization of measurement devices such as Phasor Measurement Units (PMUs) or Intelligent Electronic Devices (IEDs) situated within the substation. These intelligent devices establish communication with the control center using the communication network utilizing the communication panel. Therefore, signals will be transmitted from a substation to the control center through direct or optical fiber communication, which may involve transmitting the signals through other substations. Wireless communication is occasionally utilized; nevertheless, the signal loss in such instances tends to be slightly higher. The communication panels of the control center will be responsible for receiving signals from various substations. Subsequently, the disruptions are inputted into a deep learning algorithm, extracting fundamental characteristics such as frequency, amplitude, and waveform configuration. The model facilitates the classification of disturbances, which utilizes extracted features. This enables the control center to implement mitigation measures effectively. Nevertheless, when confronted with an adversarial attack, the attacker employs TUA perturbations to the PQD signals in order to manipulate the outputs of the deep learning model. In the event of success, this outcome has the potential to result in inaccurate grid management decisions, thereby introducing the possibility of instability and disruptions. In light of such attacks, our defensive strategy becomes active. As a result, despite the presence of TUA attacks, the model consistently maintains accurate classification. This ensures the reliability of communication between the transmission and distribution system and the control center, thereby upholding the stability of the grid.

V. Simulation Results and Discussion

A. Dataset Description

To evaluate the efficacy of the proposed TUA attack for PQDs, we utilize ResNet50 as the deep learning model in our study. In this model, the PQDs are characterized by a sampling frequency of 3200 Hz, a fundamental frequency of 50 Hz, a total cycle count of 10, and an amplitude of 1. As a result, the input signal vectors are constrained to a constant length of 640, even though the underlying signal is continuous and uninterrupted. Table I presents the various signal types for all 17 classes of signals. In this study, a labeled dataset consisting of 255,000 power-quality signals is utilized. The dataset is publicly available [12] and has a signal-to-noise ratio (SNR) of 30 dB. It is worth noting that the dataset is balanced regarding class distribution. 25% of the data, which has been

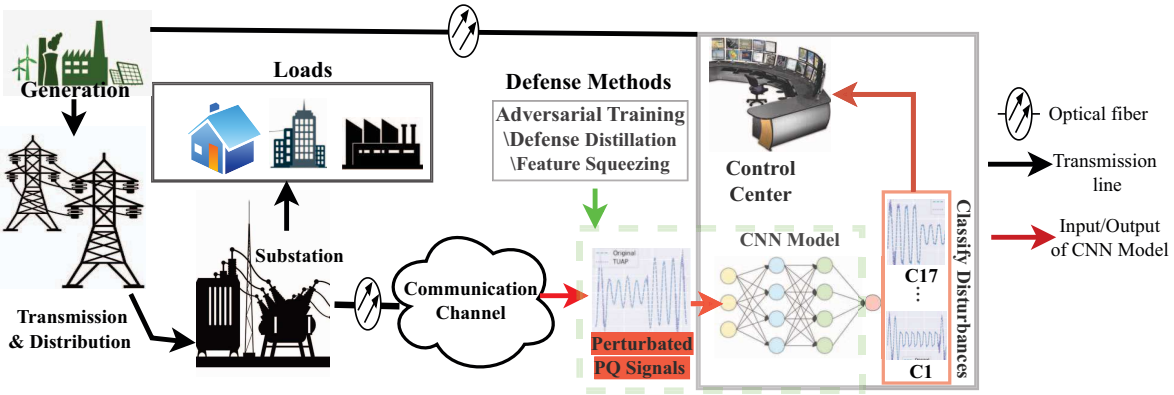

Fig. 2: Overview of Defense Applying Different Methods

TABLE I: Power quality disturbances signals name and corresponding classes

Class	Signal Name	Class	Signal Name	Class	Signal Name
C-1	Normal	C-7	Harmonics	C-13	Sag with Oscillatory transient
C-2	Sag	C-8	Harmonics with Sag	C-14	Swell with Oscillatory transient
C-3	Swell	C-9	Harmonics with Swell	C-15	Sag with Harmonics
C-4	Interruption	C-10	Flicker	C-16	Swell with Harmonics
C-5	Transient/Impulse/Spike	C-11	Flicker with Sag	C-17	Notch
C-6	Oscillatory transient	C-12	Flicker with Swell		

randomly shuffled, is allocated for testing purposes, while the remaining 75% is designated for training. The deep learning model mentioned above has been trained for the purpose of the PQD assessment task, which involves the classification of signals into 17 distinct classes. After undergoing ten epochs of training, our model exhibits high performance, attaining a test accuracy of 99.26%.

B. Result and Analysis

This section will comprehensively examine the effectiveness of three defense mechanisms- adversarial training, defensive distillation, and feature squeezing- for power quality disturbance classification models vulnerable to TUA attacks. The preliminary phase of the discussion focuses on a comparative depiction of clean and adversarial examples employed in the initial stage of the TUA attack. These adversarial examples aim to possess qualities undetectable by the human visual system while still possessing sufficient strength to mislead the deep learning model, resulting in inaccurate classification and preserving the integrity of the signal. Upon analysis of Figure 3, it is evident that the adversarial perturbations are of such small magnitude that they are practically imperceptible, thereby enabling the clean sample to dominate the image. Hence, these instances are categorized as possessing a high level of imperceptibility. Figure 4 exhibits a slight increase in the visibility of adversarial perturbations. Nevertheless, the unaltered, original waveforms remain clearly visible. This phenomenon occurs because the perturbations closely resemble the shape of the clean sample, thereby blending with the background noise. Hence, these occurrences are categorized as having a medium level of imperceptibility. In Figure 5, the adversarial perturbations exhibit a higher level of visibility while still maintaining a certain level of subtlety. From an analytical perspective, it can be observed that these disturbances continue to exhibit adversarial behavior, as the waveform characteristics do not display substantial deviations from the standard pattern. The peak values exhibit random increases, while the waveforms bear a resemblance to conventional waveforms with the inclusion of noise. The aforementioned similarity holds significant importance, as it greatly complicates the task of differentiating an attack from the regular noise generated by a system for a human observer. In our previous research, we conducted a comprehensive series of experiments, thoroughly investigating all possible combinations of source and target classes. By employing our novel algorithm, we effectively manipulated our model to misclassify 14 out of the 17 intended classes. In the context of defense evaluations, we selected the 14 pairs of source and target classes that exhibited the most significant attack success rates. To illustrate the effectiveness of the various defense mechanisms against adversarial attacks, we provide both tabulated results and visual representations in our analysis. Before and after utilizing adversarial training, defensive distillation, and feature squeezing, the table provides a comprehensive breakdown of the fooling rates. The accompanying bar chart depicts the decline in fooling rates after the application of each method, which can help in understanding the trend and relative efficacy of these defenses.

Table II displays a comparative examination of the fooling rate prior to and following the implementation of adversarial training (AT) for three distinct levels of imperceptibility (High,

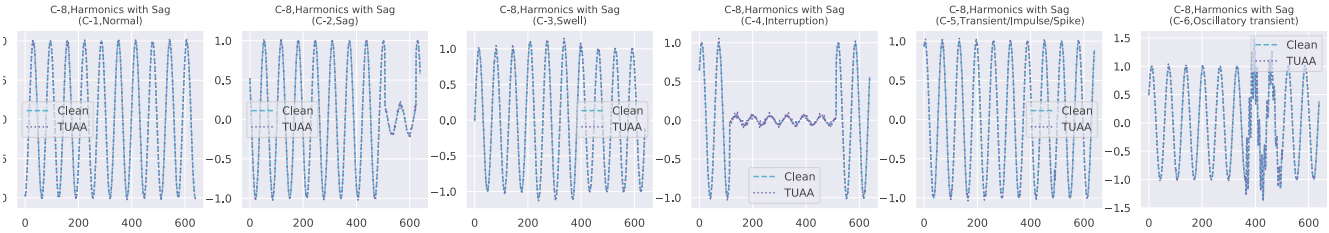

Fig. 3: Waveforms of power quality disturbances after TUA attack for high imperceptibility

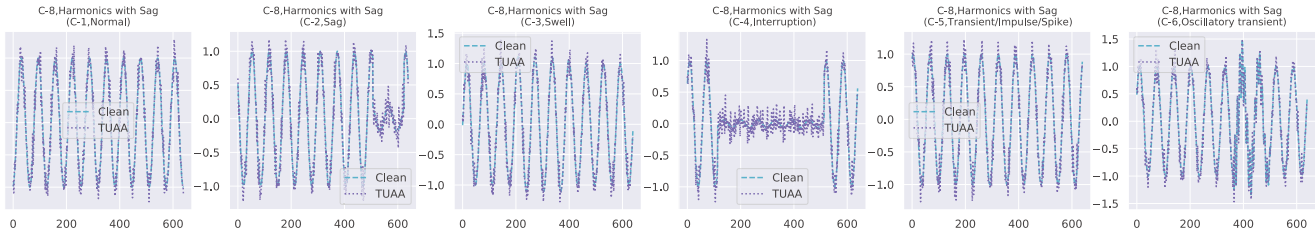

Fig. 4: Waveforms of power quality disturbances after TUA attack for medium imperceptibility

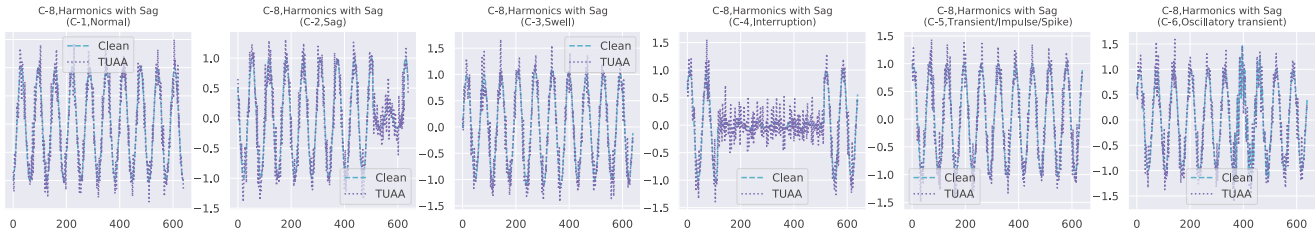

Fig. 5: Waveforms of power quality disturbances after TUA attack for low imperceptibility

TABLE II: Comparison of the fooling rate before and after adversarial training

Source	Target	FR (HI-BAT)	FR (HI-AAT)	FR (MI-BAT)	FR (MI-AAT)	FR (LI-BAT)	FR (LI-AAT)
C-1	C-2	5.87%	1.13%	10.34%	0.01%	38.18%	0.00%
C-5	C-3	23.183%	0.02%	23.56%	0.02%	44.65%	0.00%
C-10	C-4	2.39%	0.00%	4.08%	0.00%	23.61%	0.00%
C-5	C-6	4.6%	0.11%	39.03%	0.04%	64.77%	0.00%
C-4	C-7	28.88%	1.90%	58.96%	2.85%	73.38%	10.92%
C-1	C-8	66.293%	34.03%	82.03%	48.80%	90.00%	44.51%
C-1	C-9	14.37%	8.54%	63.22%	18.80%	89.49%	26.34%
C-1	C-11	19.373%	1.88%	32.23%	0.32%	41.72%	0.01%
C-3	C-12	35.05%	0.75%	57.44%	1.09%	70.26%	0.27%
C-16	C-13	48.12%	12.95%	80.35%	25.90%	85.80%	12.35%
C-16	C-14	24.96%	3.06%	66.97%	10.29%	89.65%	5.57%
C-11	C-15	32.14%	4.63%	37.82%	8.64%	39.98%	0.00%
C-1	C-16	8.01%	4.71%	28.20%	5.35%	34.26%	5.65%
C-10	C-17	47.55%	6.66%	65.72%	3.20%	73.58%	6.91%

* BAT- Before Adversarial Training, AAT- After Adversarial Training, HI- High Imperceptibility, MI- Medium Imperceptibility, LI- Low Imperceptibility.

Medium, Low) across a range of source-target combinations. The color-coding scheme has been employed in the provided table to enhance the comprehensibility of the findings. The third and fourth columns utilize various shades of green, with darker shades indicating a higher fooling rate for high imperceptibility and the darkest shade representing 100%. As the percentage decreases, there is a corresponding increase in the lightness of the green shade. In a similar manner, the representation of columns 5 and 6 is achieved through the utilization of shades of yellow, thereby adhering to the aforementioned shading principle for medium imperceptibility. Columns 7 and 8 utilize red shades, wherein a deeper hue signifies a higher fooling rate for low imperceptibility, while the intensity of the color diminishes as the percentages decrease. The attack success rates exhibit a noticeable decrease after undergoing adversarial training, thus demonstrating the efficacy of employing adversarial training as a resilient defensive approach against adversarial attacks. If we focus on figure 6,7 and 8,

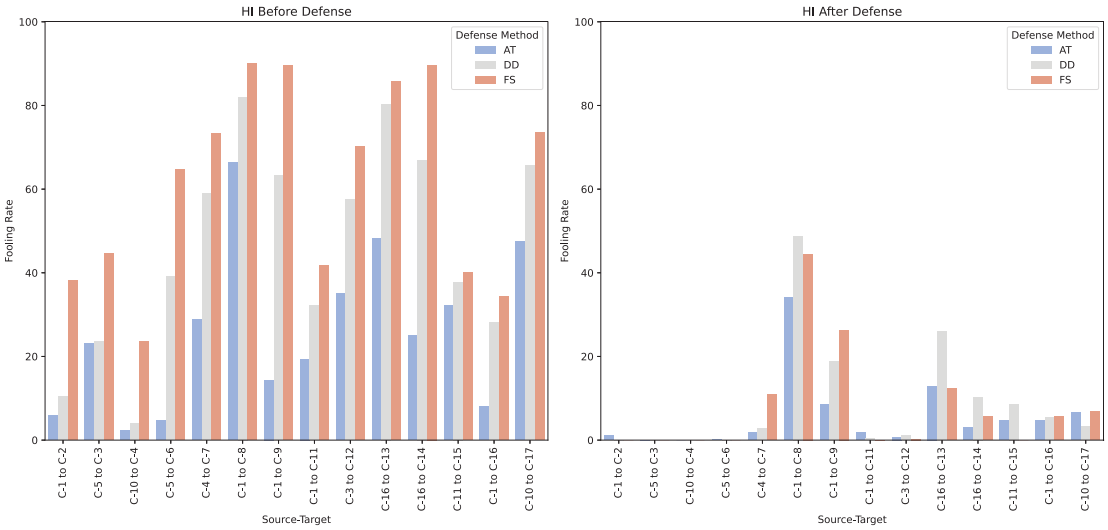

Fig. 6: Graphical representation of fooling rate before and after adversarial training for high imperceptibility

the fooling rates decreased significantly after the adversarial training in cases of a combination of source and target classes such as C1-C2, C5-C3, C10-C4, and C5-C6. For example, for the source-target combination C1-C2, the fooling rate under high imperceptibility went down from 5.87% to 1.13% after adversarial training. Under medium imperceptibility, it went down from 10.34% to 0.01%, and for low imperceptibility, the reduction was even more striking, from 38.18% to 0%. Such reductions demonstrate the efficacy of the adversarial training method in teaching the model to classify adversarially perturbed examples correctly. Nevertheless, the reduction in fooling rate was not as significant for specific source-target combinations, namely C1-C8, C1-C9, C16-C13, and C16-C14. In the case of C1-C8, under the influence of medium imperceptibility, the rate of deception decreased from 82.03% to 48.80% after the application of adversarial training. Similarly, under the influence of low imperceptibility, the rate of deception decreased from 90% to 44.51%. Although there was a notable decrease, the rates did not exhibit as substantial a decline as in other instances.

Table III presents a comprehensive analysis of the fooling rates before and after implementing a technique known as defensive distillation. Comprehensively, it can be observed that the deception rates exhibit a consistent decline after the implementation of defensive distillation techniques. In the context of high imperceptibility from figure 6, it is observed that the fooling rate for the source-target pair C-1 to C-2 decreases from 5.87% to 2.28%. Considerable decreases are also evident for pairs such as C-5 to C-3, wherein the rate decreases from 23.183% to a negligible 0.02%. In the context of medium imperceptibility from figure 7, it has been observed that there is a significant decrease in the rate at which the system is deceived. The rate for the C-5 to C-6 pair experiences a significant reduction, dropping from 39.03% to a mere 0.01%. Certain reductions exhibit a more moderate nature, such as the C-1 to C-8 pair, wherein the rate experiences a decline from 82.03% to 52.77%. Defensive distillation demonstrates its effectiveness in achieving low imperceptibility. From figure 8, in the case of the C-1 to C-2 pair, the observed rate experiences a decline from 38.18% to 0.00%.Nevertheless, certain pairs, such as C-1 to C-16, exhibit a comparatively less significant reduction, declining from 34.26% to 2.92%. The application of defensive distillation typically reduces the fooling rate; however, the extent of this reduction is contingent upon the specific source-target pair. Certain pairs undergo a significant decrease, approaching nearly 0.00%, whereas others observe a comparatively more moderate decline. Defensive distillation has demonstrated notable efficacy in mitigating imperceptibility levels that are classified as high or medium. The attack success rates within these categories frequently decrease significantly, approaching negligible values after implementing the aforementioned technique. In the context of low imperceptibility, it is observed that although there is a noticeable decrease, certain pairs exhibit fooling rates that do not decline to the same extent.

Table IV presents a comparative analysis of the fooling rate before and after applying feature squeezing for three different levels of imperceptibility (High, Medium, and Low) across various source-target combinations. In instances of high imperceptibility from figure 6, a noticeable decrease in the rate of successful deception is observed for numerous Source-Target pairs following the implementation of feature squeezing. For example, in pair C-5 to C-3, the fooling rate decreases from 23.183% to 0.00%. Nevertheless, in certain instances, such as the C-1 to C-8 pair, the reduction observed is minimal, decreasing from 66.293% to 60.72%. Medium imperceptibility is characterized by a notable decrease observed in various instances. From figure 7, it can be observed that the rate experiences a significant decrease from 63.22% to 17.77% after the application of compression for the C-1 to

TABLE III: Comparison of the fooling rate before and after defensive distillation

Source	Target	FR (HI-BDD)	FR (HI-ADD)	FR (MI-BDD)	FR (MI-ADD)	FR (LI-BDD)	FR (LI-ADD)
C-1	C-2	5.87%	2.28%	10.34%	0.02%	38.18%	0.00%
C-5	C-3	23.183%	0.02%	23.56%	0.05%	44.65%	0.02%
C-10	C-4	2.39%	0.00%	4.08%	0.00%	23.61%	0.00%
C-5	C-6	4.6%	0.09%	39.03%	0.01%	64.77%	0.00%
C-4	C-7	28.88%	7.58%	58.96%	5.97%	73.38%	7.44%
C-1	C-8	66.293%	44.77%	82.03%	52.77%	90.00%	56.48%
C-1	C-9	14.37%	5.22%	63.22%	19.59%	89.49%	21.05%
C-1	C-11	19.373%	0.38%	32.23%	0.00%	41.72%	0.00%
C-3	C-12	35.05%	3.72%	57.44%	1.26%	70.26%	0.02%
C-16	C-13	48.12%	12.41%	80.35%	2.77%	85.80%	3.56%
C-16	C-14	24.96%	4.88%	66.97%	8.68%	89.65%	11.67%
C-11	C-15	32.14%	4.83%	37.82%	1.26%	39.98%	3.60%
C-1	C-16	8.01%	5.17%	28.20%	3.01%	34.26%	2.92%
C-10	C-17	47.55%	6.13%	65.72%	7.87%	73.58%	5.13%

* BDD- Before Defensive Distillation, ADD- After Defensive Distillation, HI- High Imperceptibility, MI- Medium Imperceptibility, LI- Low Imperceptibility.

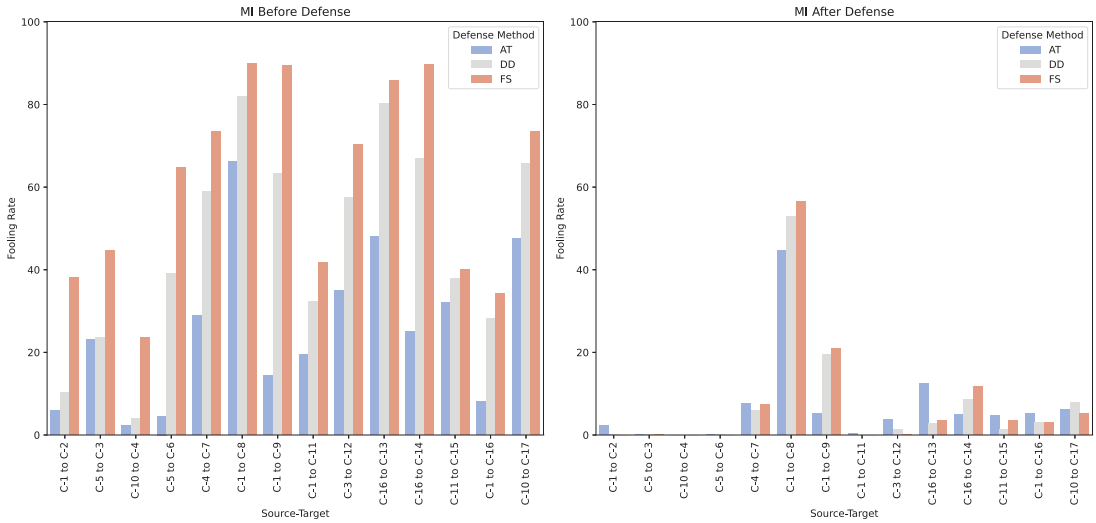

Fig. 7: Graphical representation of fooling rate before and after adversarial training for medium imperceptibility

TABLE IV: Comparison of the fooling rate before and after feature squeezing

Source	Target	FR (HI-BFS)	FR (HI-AFS)	FR (MI-BFS)	FR (MI-AFS)	FR (LI-BFS)	FR (LI-AFS)
C-1	C-2	5.87%	4.99%	10.34%	0.00%	38.18%	38.18%
C-5	C-3	23.183%	0.00%	23.56%	0.00%	44.65%	0.00%
C-10	C-4	2.39%	0.00%	4.08%	0.00%	23.61%	3.60%
C-5	C-6	4.6%	2.14%	38.65%	38.65%	64.77%	64.77%
C-4	C-7	28.88%	19.34%	58.96%	57.22%	73.38%	73.38%
C-1	C-8	66.293%	60.72%	82.03%	82.00%	90.00%	90.00%
C-1	C-9	14.37%	12.68%	63.22%	17.77%	89.49%	63.22%
C-1	C-11	19.373%	0.00%	32.23%	23.66%	41.72%	41.7%
C-3	C-12	35.05%	0.60%	57.44%	5.96%	70.26%	70.26%
C-16	C-13	48.12%	48.12%	80.35%	80.35%	85.80%	85.80%
C-16	C-14	24.96%	18.96%	66.97%	66.97%	89.65%	89.65%
C-11	C-15	32.14%	4.48%	37.82%	26.14%	39.98%	38.83%
C-1	C-16	8.01%	0.04%	28.20%	0.04%	34.26%	3.69%
C-10	C-17	47.55%	47.55%	65.72%	65.72%	73.58%	73.58%

* BFS- Before Feature Squeezing, AFS- After Feature Squeezing, HI- High Imperceptibility, MI- Medium Imperceptibility, LI- Low Imperceptibility.

C-9 pairing. However, certain pairs exhibit no alteration, such as the transition from C-16 to C-13, which remains constant at 80.35%. The variability of the impact of feature squeezing is observed in the context of low imperceptibility. From figure 8, the pair consisting of C-1 to C-16 exhibits a decrease from 34.26% to 3.69%, whereas certain pairs, such as C-16 to C-14, remain unaltered at 89.65%. Feature squeezing has demonstrated significant efficacy in various scenarios, particularly when applied to imperceptibility levels categorized as high or medium. This phenomenon is clearly demonstrated

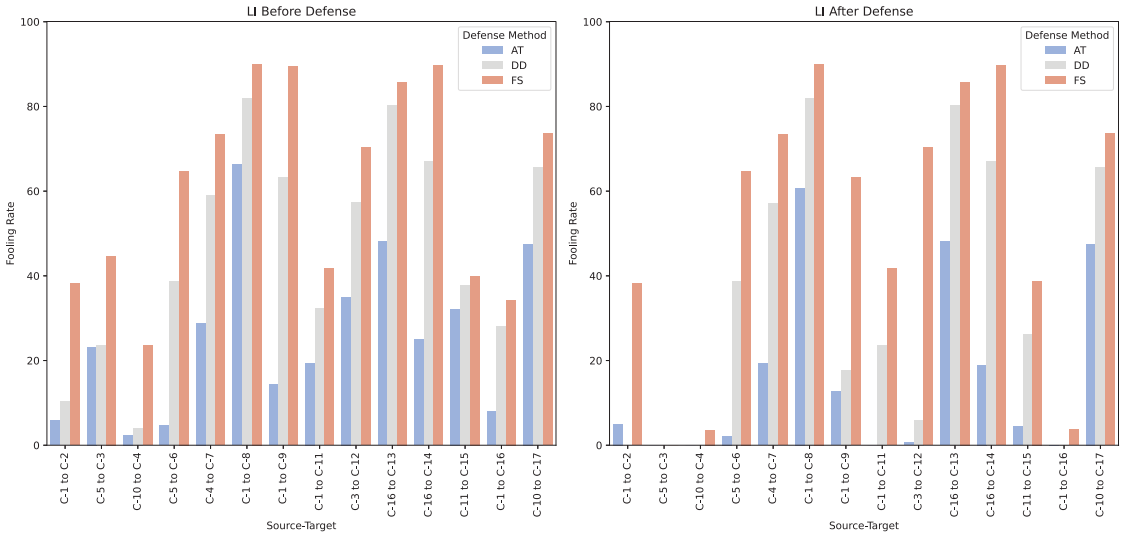

Fig. 8: Graphical representation of fooling rate before and after adversarial training for low imperceptibility

by the significant decrease in deception rates following the implementation of feature squeezing. The findings regarding low imperceptibility are somewhat inconclusive. There exist instances wherein the application of feature squeezing leads to a notable reduction in the rate of successful fooling attempts. However, it is worth noting that several other cases have been observed where no discernible improvement in the fooling rate is observed. The efficacy of feature squeezing appears to rely on the specific combination of source and target and the chosen level of imperceptibility. The efficacy of this approach varies depending on specific combinations, with some combinations yield greater benefits than others.

In each of the three discussed defense methods, the variations in the decrease of the fooling rate can be attributed to multiple potentially interrelated causes. The observed variations can be attributed to the distinctive characteristics in the pairings of source and target classes. Every source-target pair represents distinct categories of power quality disturbances, each exhibiting its own distinctive waveform characteristics and features. A model's susceptibility to misclassifying a particular class of disturbance as another can be significantly influenced by the degree of similarity or dissimilarity between the distinctive features of these classes. The model may encounter increased difficulty in distinguishing between the source and target when their characteristics exhibit greater dissimilarity, even applying a defense technique. As a result, this may lead to a comparatively smaller decrease in the rate of deception.

Based on the preceding discourse, we will present a comparative analysis of the efficacy exhibited by three distinct defense techniques. In our analysis, we have taken into account several factors, including the mean reduction in fooling rates, the consistency across different levels of imperceptibility, and the lowest achievement in fooling rates. Adversarial training generally provides the most notable decrease in fooling rates across all levels of imperceptibility. The performance of defensive distillation is noteworthy, although it does not consistently surpass that of adversarial training. Feature squeezing has demonstrated effectiveness, particularly in the context of high imperceptibility and medium imperceptibility levels, but its performance is less consistent. Both adversarial training and defensive distillation demonstrate consistency in their defense against adversarial attacks, with adversarial training exhibiting a slight advantage. Feature squeezing exhibits a greater range of outcomes, particularly in low imperceptibility. In terms of versatility, it is generally observed that adversarial training outperforms other methods. However, in situations where computational resources or time constraints are limiting factors, defensive distillation, and feature squeezing may be more appealing due to their less computationally intensive nature. Hence, it can be inferred that adversarial training demonstrates superior efficacy as a defense mechanism compared to the other two methods, specifically in mitigating the fooling rate. It consistently decreases the attack success rate across various source-target pairs and at all levels of imperceptibility. Defensive distillation has demonstrated efficacy, although it may not consistently attain the same level of low fooling rate as adversarial training. Feature squeezing exhibits less consistency, particularly at lower levels of imperceptibility, thereby diminishing its reliability as a defensive technique. Nevertheless, the selection of the methodology may vary based on specific use cases, available computational resources, and the significance of safeguarding against a particular level of imperceptibility. For example, if the objective is to mitigate high imperceptibility adversarial attacks effectively, both Adversarial training and feature squeezing demonstrate considerable promise.

VI. Conclusion

This study offers a comprehensive analysis of TUA attacks on time series data within the dynamic field of smart grid systems. In this study, we aimed to analyze and compare the effectiveness of three defensive techniques - adversarial training, defensive distillation, and feature squeezing - in order to identify the most optimal defense strategy against adversarial intrusions. The results of our study provide a detailed analysis of the benefits and constraints associated with each approach. Adversarial training is particularly distinguished by its consistent ability to reduce the effectiveness of adversarial attacks across a wide range of source-target pairs and different levels of imperceptibility. However, it is important to recognize that it is not capable of completely eliminating the rate of deception in all possible combinations of source and target. Although defensive distillation is viable, it may not always exhibit the same level of defensive strength as adversarial training. Feature squeezing may encounter occasional challenges, especially when confronted with subtle adversarial perturbations that are difficult to detect. Based on the diverse array of difficulties posed by adversarial perturbations, our research indicates that the defense approach requires further development with the enhancement of defense mechanisms achieved through the augmentation of a more diverse sample, thereby potentially strengthening their ability to withstand a wider range of adversarial techniques. This study emphasizes the necessity of implementing such improvements by drawing attention to the enduring presence of vulnerabilities, even in the face of our most robust existing defenses. The future research directions have a multitude of promising prospects. Given that adversarial training does not currently offer an all-inclusive solution, it is imperative for future research efforts to focus on investigating the underlying factors that contribute to these persistent vulnerabilities. There is potential in integrating current defensive techniques or implementing new, innovative strategies to develop a comprehensive defense. In the face of the complex challenges we encounter, our research shows our dedication to strengthening smart grid systems against hostile attacks, envisioning a future in which power systems embody reliability, security, and steadfast resilience. This study serves as a foundational basis for future academic pursuits, initiating the pursuit of enhanced and cohesive defense strategies against the continuous development of adversarial attacks in power systems and associated industries.

Acknowledgment

The work in this project is supported by the National Science Foundation Grant number 2301553 and Cisco Grant CG#70615867.

References

[1] Smartgrid. Advanced metering infrastructure and customer systems. US Department of Energy Office of Electricity and Energy Reliability. 2016; .

[2] Bollen MH, Gu IY. Signal processing of power quality disturbances. John Wiley & Sons; 2006.

[3] Topaloglu I. Deep learning based a new approach for power quality disturbances classification in power transmission system. Journal of Electrical Engineering & Technology. 2023;18(1):77–88.

[4] Broshi A. Monitoring power quality beyond en 50160 and iec 61000-4-30. In: 2007 9th International Conference on Electrical Power Quality and Utilisation; IEEE; 2007. p. 1–6.

[5] Bacanin N, Stoean C, Zivkovic M, et al. On the benefits of using metaheuristics in the hyperparameter tuning of deep learning models for energy load forecasting. Energies. 2023;16(3):1434.

[6] Abedi S, Kwon S. Rolling-horizon optimization integrated with recurrent neural network-driven forecasting for residential battery energy storage operations. International Journal of Electrical Power & Energy Systems. 2023;145:108589.

[7] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:14126572. 2014;.

[8] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:170606083. 2017;.

[9] Xiao H, Xiao H, Eckert C. Adversarial label flips attack on support vector machines. In: Ecai 2012. IOS Press; 2012. p. 870–875.

[10] Xiao H, Biggio B, Nelson B, et al. Support vector machines under adversarial label contamination. Neurocomputing. 2015;160:53–62.

[11] Calzavara S, Lucchese C, Tolomei G. Adversarial training of gradient-boosted decision trees. In: Proceedings of the 28th ACM international conference on information and knowledge management; 2019. p. 2429–2432.

[12] Tian J, Wang B, Li J, et al. Adversarial attacks and defense for cnn based power quality recognition in smart grid. IEEE Transactions on Network Science and Engineering. 2021;9(2):807–819.

[13] Zhao T, Yue M, Wang J. Robust power system stability assessment against adversarial machine learning-based cyberattacks via online purification. IEEE Transactions on Power Systems. 2023;.

[14] Chen Y, Tan Y, Deka D. Is machine learning in power systems vulnerable? In: 2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm); IEEE; 2018. p. 1–6.

[15] Huang R, Li Y. Adversarial attack mitigation strategy for machine learning-based network attack detection model in power system. IEEE Transactions on Smart Grid. 2023;.

[16] Khan SU, Mynuddin M, Nabil M. Adaptedge: Targeted universal adversarial attacks on time series data in smart grids. 2023;.

[17] Niazazari I, Livani H. Attack on grid event cause analysis: An adversarial machine learning approach. In: 2020 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT); IEEE; 2020. p. 1–5.

[18] Kosut O, Jia L, Thomas RJ, et al. Malicious data attacks on smart grid state estimation: Attack strategies and countermeasures. In: 2010 first IEEE international conference on smart grid communications; IEEE; 2010. p. 220–225.

[19] Cheng Y, Yamashita K, Yu N. Adversarial attacks on deep neural network-based power system event classification models. In: 2022 IEEE PES Innovative Smart Grid Technologies-Asia (ISGT Asia); IEEE; 2022. p. 66–70.

[20] Kurakin A, Goodfellow I, Bengio S. Adversarial machine learning at scale. arXiv preprint arXiv:161101236. 2016;.

[21] Guo C, Rana M, Cisse M, et al. Countering adversarial images using input transformations. arXiv preprint arXiv:171100117. 2017;.

[22] Tian J, Wang B, Li J, et al. Adversarial attack and defense methods for neural network based state estimation in smart grid. IET Renewable Power Generation. 2022;16(16):3507–3518.

[23] Li J, Yang Y, Sun JS, et al. Towards adversarial-resilient deep neural networks for false data injection attack detection in power grids. arXiv preprint arXiv:210209057. 2021;.

[24] Guihai Z, Sikdar B. Adversarial machine learning against false data injection attack detection for smart grid demand response. In: 2021 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm); IEEE; 2021. p. 352–357.

[25] Papernot N, McDaniel P, Wu X, et al. Distillation as a defense to adversarial perturbations against deep neural networks. In: 2016 IEEE Symposium on Security and Privacy (SP); 2016. p. 582–597.

[26] Xu W, Evans D, Qi Y. Feature squeezing: Detecting adversarial examples in deep neural networks. CoRR. 2017;abs/1704.01155. Available from: http://arxiv.org/abs/1704.01155.