AdaptEdge: Targeted Universal Adversarial Attacks on Time Series Data in Smart Grids

Sultan Uddin Khan, Mohammed Mynuddin, and Mahmoud Nabil, *Member, IEEE*

***Abstract*—Deep learning (DL) has emerged as a key technique in smart grid operations for task classification of power quality disturbances (PQDs). Even though these models have considerably improved the efficiency of power infrastructure, their susceptibility to adversarial attacks presents potential difficulties. For the first time, we introduce a novel algorithm called Adaptive Edge (AdaptEdge), which effectively employs targeted universal adversarial attack to deceive DL models working with time series data. The unique contribution of this algorithm is its ability to maintain a delicate balance between the fooling rate and the imperceptibility of perturbations to human observers. Our results demonstrate a fooling rate of up to 90.78% in the ResNet50 model—the highest achieved thus far—while maintaining an optimal signal-to-noise ratio (SNR) of 3dB and ensuring signal integrity. We implemented our algorithm across various advanced DL models and found considerable efficacy, demonstrating its adaptability and versatility across diverse architectures. The results of our study highlight the pressing need for developing more robust DL model implementations in the context of the smart grid. Additionally, our proposed approach demonstrates its effectiveness in addressing this need.**

***Index Terms*—Targeted attack, universal adversarial attack, time series data, smart grid, power quality disturbance, deep learning.**

NOMENCLATURE

AdaptEdge	Adaptive Edge
AMI	Advanced Metering Infrastructure
CNN	Convolutional Neural Network
DL	Deep Learning
FGSM	Fast Gradient Sign Method
IED	Intelligent Electronic Devices
LSTM	Long Short-Term Memory
PQDs	Power Quality Disturbances
SG	Smart Grid
SNR	Signal-to-Noise Ratio
TSD	Time Series Data
TUAA	Targeted Universal Adversarial Attack

Manuscript received 18 June 2023; revised 3 October 2023, 25 November 2023, 7 January 2024, and 12 February 2024; accepted 29 March 2024. Date of publication 2 April 2024; date of current version 23 August 2024. This work was supported in part by the National Science Foundation under Grant 2301553, and in part by Cisco under Grant CG 70615867. Paper no. TSG-00901-2023. *(Corresponding author: Sultan Uddin Khan.)*

Sultan Uddin Khan and Mohammed Mynuddin are with North Carolina A&T State University, Greensboro, NC 27411 USA (e-mail: skhan5@aggies.ncat.edu).

Mahmoud Nabil is with the ECE Department, North Carolina A&T State University, Greensboro, NC 27411 USA.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSG.2024.3384208.

Digital Object Identifier 10.1109/TSG.2024.3384208

I. INTRODUCTION

SMART Grid (SG), uses cutting-edge communication and information technology to facilitate a dependable and efficient energy supply . It allows for two-way communication between the utility and its customers, enabling them to manage their energy consumption more effectively and potentially even sell excess power back to the grid [1]. They use advanced technologies to monitor and control power distribution in real-time, including sensors, automation, and communication networks [2].

Advanced Metering Infrastructure (AMI) is used by smart meters installed at the load end to assess power usage and provide real-time data to the utility [3]. Effective communication allows for regulating energy production and consumption in real-time, enhancing grid stability, lowering energy waste, and using more renewable energy sources. While SG improves energy efficiency and consumer interaction, their interconnected nature makes them vulnerable to targeted universal adversarial attacks (TUAA). Such attacks could target the grid's sensors, automation systems, and communication networks, potentially leading to manipulated energy usage data or disruptions in grid operations.

The vulnerability of SG to TUAA is further complicated by the occurrence of power quality disturbances (PQDs). PQDs, which involve deviations in voltage, current, or frequency, can adversely affect electrical equipment's performance. These disturbances, while sometimes resulting from the inherent intermittency of renewable energy sources or system malfunctions, can also be exacerbated or mimicked by adversarial attacks. TUAA can exploit these PQDs as a cover, masking their manipulative activities within the grid. For instance, attackers might induce or simulate PQDs to disrupt the grid's frequency and voltage, thereby compromising grid stability and reliability. Additionally, cybersecurity threats can not only cause PQDs in SG by disrupting communication networks and control systems [4] but also open avenues for sophisticated adversarial attacks. These attacks can disrupt energy flow and data integrity, making it challenging to distinguish between genuine PQDs and those orchestrated as part of a TUAA. Thus, mitigating the negative impact of PQDs is crucial for maintaining the resilience and security of SG.

PQDs can have significant consequences that range from minor inconveniences to major economic and safety risks. Poor power quality can damage electronic equipment, cause production downtime, increase maintenance costs, reduce equipment lifespan, pose safety risks, and result in energy

waste. Additionally, non-compliance with regulations and standards can lead to legal or financial penalties. Machine learning and deep learning are increasingly used in SG to improve their efficiency, reliability, and security. By utilizing these algorithms, anomalies in the grid, like power quality disturbances or equipment failures, can be quickly detected, leading to reduced downtime and faster response times [5].

Previous research has explored the vulnerability of DL models to specific adversarial attacks in power systems [6], investigated the effectiveness of defense methods against untargeted attacks [7], and examined joint adversarial example and false data injection attacks in power system state estimation [8]. However, no study is currently on targeted universal adversarial perturbation of time series data (TSD). While Rathore et al. [9] proposed a targeted adversarial attack in TSD using the Fast Gradient Sign Method (FGSM), universal adversarial attacks present a greater threat as they can deceive the model across multiple inputs using a single perturbation. This makes them more effective and potentially more damaging. Therefore, this paper aims to address this gap by proposing a novel approach for generating targeted universal adversarial perturbations in TSD within the SG. Our approach can significantly impact the security of time series-based systems and applications. The key contributions of our manuscript are as follows:

- We propose the targeted universal adversarial attack (TUAA) on TSD. This attack methodology aims to deceive DL models operating on TSD by crafting adversarial examples that can fool the models across multiple inputs. Unlike targeted attacks that require specific perturbations for each input instance, the targeted universal attack utilizes a single perturbation to achieve its objective. This approach has the potential to be more effective and impactful, posing a greater threat to the security and reliability of time series-based systems. To the best of our knowledge, this is the first time such an attack has been implemented.
- A new algorithm, specifically designed to facilitate TUAA, primarily focusing on deceiving DL models operating on TSD, is introduced. Our approach allows for the generation of adversarial examples that can fool power system control centers in SG, demonstrating the vulnerability of this model to attack. This contribution highlights the potential security risks associated with TSD and provides a foundation for developing more robust models in the future.
- Our algorithm is the first to consider and successfully balance the crucial trade-off between imperceptibility (i.e., signal-to-noise ratio SNR) and fooling rate for launching TUAA on TSD. This careful balancing ensures effective attack success rates without compromising the stealthy nature of adversarial perturbations. This innovative perspective validates our algorithm's efficacy and paves the way for a new direction in adversarial machine learning research, especially about TSD.

The remainder of this paper is structured as follows. In Section II, we discuss the existing research on cyber attacks against the SG and highlight the gaps in the literature that our work aims to address. In Section III, we introduce the application of DL in smart grids and attack scenarios. We then present the threat model in Section IV to provide insight into the attack methodology. Section V explains the proposed TUAA on TSD in the SG. We describe the datasets used in our experiments and the model architecture in Section VI. We present the simulation results in Section VII, demonstrating the effectiveness of our proposed attacks against SG. We conclude the paper in Section VIII by summarizing our findings and discussing the implications of our work for the security of SG.

II. Related Work

A. Power Quality Disturbances Classification

Machine learning algorithms are revolutionizing the SG by offering a versatile set of applications, including demand forecasting, anomaly detection, grid optimization, energy theft detection, load balancing, power quality disturbances classification, and more. In [10], Multiple PQDs (MPQDs) are detected and categorized using a novel hybrid technique based on Stockwell transform (ST) and deep learning. In [11], the authors introduce a swift and accurate algorithm for monitoring PQDs in SG, amalgamating histogram and discrete wavelet transform techniques for feature extraction and employing machine learning for precise classification to enhance PQDs detection performance. Yigit, Yiğit et al. [12] employed a Convolutional Neural Network (CNN) structure with Gated Recurrent Unit for classifying PQDs signals. The authors demonstrated that, in their research, the performance of the VGG-16 and ResNet-50 models was very similar. In [13], authors suggest a method for classifying PQD using a DL-based CNN that incorporates an attention model. This model focuses on rescaling available data based on pixel count before pooling it to create an enhanced data set for deeper CNN analysis.

B. Cyber Attacks on the Smart Grid

Several studies have investigated the impact of cyber attacks on the SG, including untargeted and targeted attacks. In [14], Niazazari and Livani highlights how adversarial attacks can lead CNNs to misclassify events in SGs. Sayghe et al. [15] discuss adversarial attacks on Multilayer Perceptron for detecting false data injection. In [16], authors proposed an adversarial machine learning approach that utilizes black-box optimization techniques to generate dynamic load-altering attacks. In [17], researchers suggest Ensemble and Transfer Adversarial Attacks across diverse DL models. Tian et al. [18] proposed an adversarial attack crafting method based on a forward derivative that considers input element magnitude, attack impact on multiple regression output, and other controllable measurement meters. Cheng et al. [19] utilize different adversarial attack mechanisms to add noise signal to the input Phasor Measurement Units time series and show that current DL-based power system event classifiers are highly susceptible to such attacks, which could compromise the power transmission system's reliability. In [6], the authors demonstrate the vulnerability of current ML algorithms in power systems to adversarial examples and propose an efficient

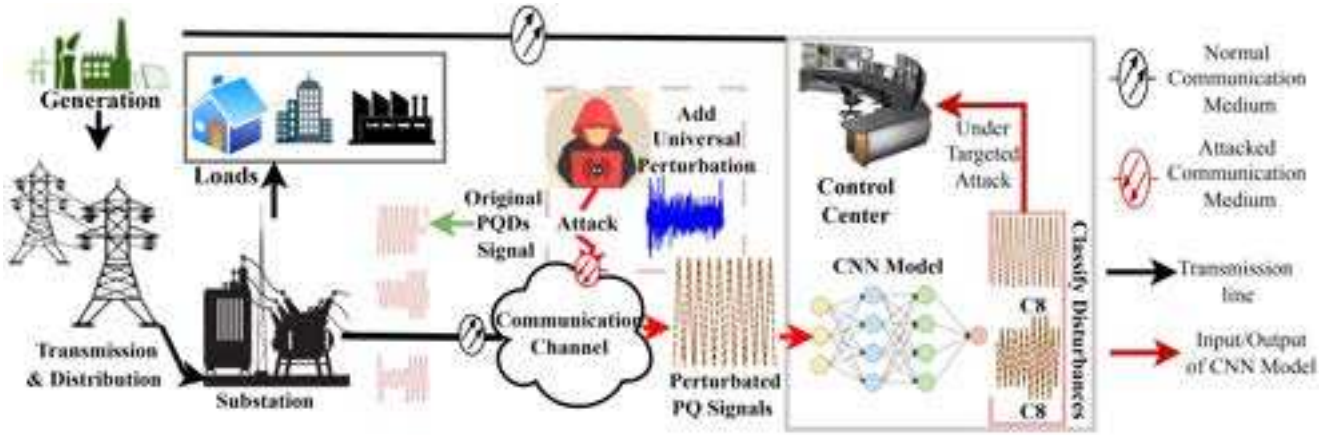


Fig. 1. Smart grid infrastructure illustrating the potential exploitation points for targeted adversarial attacks during signal transmission to the DL-based power control center.

TABLE I
COMPARISON WITH EXISTING TECHNIQUES

Previous works	TSD	Targeted	Universal	SNR/Fooling Rate Balance
[22]	✓	×	×	×
[9]	✓	✓	×	×
[23]	×	✓	✓	×
[24]	×	✓	✓	×
Our work	✓	✓	✓	✓

algorithm to generate such examples for categorical and sequential applications. Kosut, Oliver, et al. [20] perform the cyber attack on SG by malicious data injection technique. In [21], the researchers investigate how vulnerable a Long Short-Term Memory (LSTM) network and a CNN are to targeted, semi-targeted, and non-targeted adversarial attacks in predicting wind power outputs.

C. Comparison With Existing Technique

In [22], the algorithm focuses on generating adversarial attacks for individual power quality signals. While this provides a robust method for deceiving the model with different perturbations for each signal, it is not effective as a universal attack from the attacker's perspective because the attacker can misclassify the model with a single perturbation in a universal attack. Moreover, the authors in [22] adopt an untargeted universal attack approach on TSD; however, TUAA poses a greater danger than untargeted attacks. In their work, Rathore et al. [9] introduced a targeted adversarial attack on TSD using the FGSM. However, the universal adversarial attack poses a greater threat due to its ability to deceive the model across multiple inputs without crafting specific perturbations for each instance, unlike FGSM. This attack approach has the potential to be more impactful and cause widespread damage. In [23] and [24], authors have proposed algorithms for targeted universal adversarial attacks on image data, not in time series data. Moreover, their solution didn't perform optimized balancing to maintain a careful balance between the fooling rate and imperceptibility, a critical aspect for the stealthiness and success of attacks in time series data. We present the comparison of our work to the existing technique in Table I. Addressing this significant research gap, our paper aims to propose a novel method for generating TUAA to fool the DL model working with TSD in SG. Instead of generating adversarial signals tailored to specific samples, our AdaptEdge algorithm generates a universal perturbation that, when applied to any signal, causes the deep learning model to classify the signal as the attacker chooses. Through the AdaptEdge function, the perturbation intensity is fine-tuned to achieve a high fooling rate while maintaining an imperceptible perturbed signal. These distinctions highlight our proposed method's originality and expanded capabilities compared to the untargeted approach.

III. NETWORK AND THREAT MODEL

A. Network Model

The network model of the SG, depicted in Figure 1, encompasses various essential components. These include generators, transmission and distribution systems, substations, loads, communication channels, CNN models, and the control center. Generators are responsible for generating electrical power, utilizing diverse sources like fossil fuels, renewable energy, or nuclear power. The transmission and distribution systems facilitate the transfer and delivery of electricity from generators to end-users. Substations, located strategically throughout the grid system, regulate voltage levels for efficient electricity flow. Loads represent the devices and consumers that utilize electrical power, encompassing residential, commercial, and industrial appliances, lighting, and machinery. Communication channels, often employing optical fibers, enable seamless data exchange among grid components. At the core of the SG, the control center acts as a centralized hub for monitoring, managing, and responding to power generation, distribution, and load balancing.

The control center leverages state-of-the-art technologies and software applications to carry out these responsibilities effectively. Among these advanced technologies is a DL-based CNN. Integrated into the control center's software infrastructure, CNN performs vital tasks, including identifying and analyzing PQDs throughout the SG. By harnessing DL capabilities, CNN enables real-time analysis, enabling the control center to detect anomalies, diagnose faults, and promptly

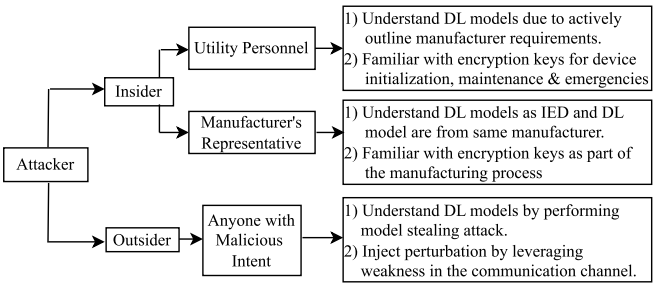


Fig. 2. Proposed Threat Model.

predict maintenance requirements. By comprehensively examining data collected from various components of the SG, the control center can make well-informed decisions efficiently. This utilization of DL techniques significantly enhances the operational efficiency and reliability of the entire SG, leading to more effective management of power resources.

B. Threat Model

In Figure 1, we consider that the SG functions as a Cyber-Physical System (CPS), where DL models process data obtained from the communication channel to enable intelligent decision-making. The system can optimize energy distribution, predict demand, detect anomalies, and automate maintenance tasks by analyzing these diverse data. Consequently, the SG can improve its operational efficiency, dependability, and responsiveness. Signal-based applications, such as PQDs identification or fault detection, heavily rely on the communication channel to capture transmitted signals for analysis by DL models. The attacker can manipulate the communication link to remain undetected and evade anomaly detection methods. In our threat model in figure 2, we consider both insider and outsider adversaries within a white-box system. The attacker possesses complete knowledge of the target model, including its architecture, parameters, and training data. An attacker with black box access and the capability to inject fake data is also powerful. A powerful attacker with black box access could perform several examples of direct attacks, such as the Gradient Estimation black-box attack [25] and the Word Substitution Ranking Attack [26]. These attacks demonstrate the potential for powerful attackers with black box access to perform direct and significant attacks on machine learning models. External attackers seeking to exploit the power grid system may attempt to gather white-box knowledge from various sources. These sources could include publicly available documents, specifications, or research publications related to the power grid infrastructure. According to Kerckhoff's principle, the security of a system should not rely on the secrecy of the algorithm or design but rather on the secrecy of the key. While some technical information may be accessible, obtaining a complete and up-to-date understanding of the system's intricacies could be challenging. In that case, external adversaries may rely on model stealing attacks [27]. This comprehensive knowledge enables attackers to craft tailored adversarial examples that exploit vulnerabilities within the model, simplifying the execution of successful attacks. In addition to outsider adversaries, we address the threat posed by insider adversaries with authorized access to the infrastructure. Insiders in our scenario can be employees of the company that manufactures Intelligent Electronic Devices (IED), and the DL models used at power control centers. Due to merging IED and DL model production, an insider within the manufacturing organization is already privy to intricate model details. Furthermore, utility personnel, who are also classified as insiders in this framework, comprehensively understand these DL models due to their active participation in the procurement process, in which they specify detailed requirements to the manufacturer and then implement the acquired DL model-based power control centers. The IED manufacturing and utility staff know encryption keys, given the threat model we described. This knowledge is inherent to the manufacturer's representative as part of the manufacturing process. Concurrently, utility personnel are familiar with encryption keys as a term incorporated in their contractual agreements with the manufacturer, ensuring device initialization, maintenance, and emergency interventions. Manufacturers may provide keys to ensure seamless integration and simulate real-world scenarios during IED deployment and testing. Moreover, given their positions, these insiders may have physical access to the devices, allowing for direct hardware tampering, implantation of malicious components, or firmware alterations, altering device configurations to compromise power control systems. Understanding these aspects is essential to comprehensively address and mitigate the risks associated with insider threats in critical power infrastructure environments.

IV. Universal Adversarial Attacks, Overview of Targeted Universal Adversarial Attacks in Smart Grid and Adaptive Edge Algorithm

A. Taxonomy of Universal Adversarial Attack

Universal Adversarial Perturbation (UAP): The Universal Adversarial Perturbation (UAP) refers to a vector that, when added to any signal within a specific dataset, results in misclassification by a deep learning model [28]. This perturbation is calculated by solving an optimization problem that aims to minimize the model's accuracy on a dataset with the perturbation applied. Given a dataset of n samples $X = x_1, x_2, \ldots, x_n$ and a target classifier f, the UAP vector r can be obtained by solving the following optimization problem:

$$r = \arg\min_{r} \sum_{i=1}^{n} \ell(f(x_i + r), y_i) \quad \text{subject to } |r|_p \leq \epsilon \tag{1}$$

where y_i is the true label of x_i, ℓ is a loss function, and ϵ is a hyperparameter that controls the magnitude of the perturbation. The objective of the optimization problem is to find the perturbation vector r that minimizes the loss of the target classifier f on the dataset.

On the other hand, in our problem formulation, the Targeted Universal Adversarial Perturbation (TUAP) seeks to generate a single perturbation that, when added to multiple input samples, causes a DL model to incorrectly classify into a specific

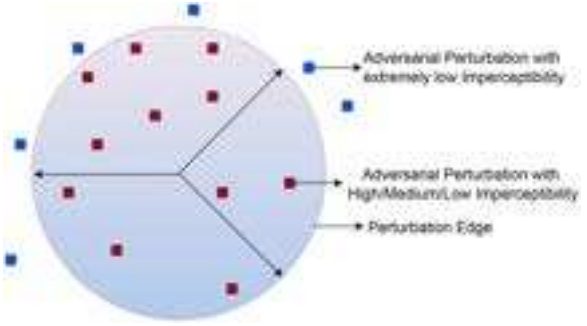


Fig. 3. Visualization of the Perturbation Edge.

predetermined class. The TUAP is formulated as the following optimization problem:

$$r_{\text{TUAP}} = \arg\max_{r} \left\{ \frac{1}{n} \sum_{i=1}^{n} \chi_{\{f(x_i+r) \ = y_{\text{target}}\}} \right\}$$
$$\text{subject to} \quad \max_{i} \{\text{SNR}(x_i, r) \ \} \geq \text{SNR}_{\text{min}} \text{ and } \|r\|_p \leq \epsilon \quad (2)$$

where y_{target} is the targeted class label for the adversarial attack, χ is the indicator function that equals 1 if the classifier f misclassifies the perturbed input $x_i + r$ as the target class y_{target}, and 0 otherwise, $\text{SNR}(x_i, r)$ is the signal-to-noise ratio between the original signal x_i and the perturbation r, SNR_{min} is the minimum acceptable SNR, and ϵ is the perturbation budget. The objective is to maximize the fooling rate, while also ensuring that the perturbation r remains imperceptible as measured by the SNR.

B. Adaptive Edge Algorithm

The Adaptive Edge (AdaptEdge) algorithm focuses on deceiving DL models analyzing TSD by creating adversarial perturbations that are both effective in misleading the model and imperceptible to human observers. The algorithm achieves this through a careful balance between the fooling rate and the SNR. The AdaptEdge algorithm introduces a novel approach to generating universal perturbations by dynamically adjusting the perturbation edge. This concept refers to the boundary within which perturbations can manipulate the model's predictions without becoming perceptible to humans. The conceptual representation of the perturbation edge, as illustrated in Figure 3, visually encapsulates the algorithm's essence. The semi-transparent sphere symbolizing the perturbation edge within the feature space delineates the limit of allowable perturbations, with the scattered red dots representing various adversarial perturbations confined within this boundary. This confinement ensures perturbations remain subtle, highlighting the algorithm's ability to conduct stealthy adversarial attacks without compromising the integrity of the signal. The dynamic adjustment mechanism of the AdaptEdge algorithm is mathematically encapsulated in the optimization problem formulated for TUAP, as shown in Equation (2). The essence of this problem is captured by an objective function aiming to maximize the fooling rate, $\arg\max_r\{\frac{1}{n}\sum_{i=1}^{n} \chi_(f(x_i + r) = y_{\text{target}})\}$, where r represents the adversarial perturbation vector designed to mislead the classifier f into incorrectly classifying

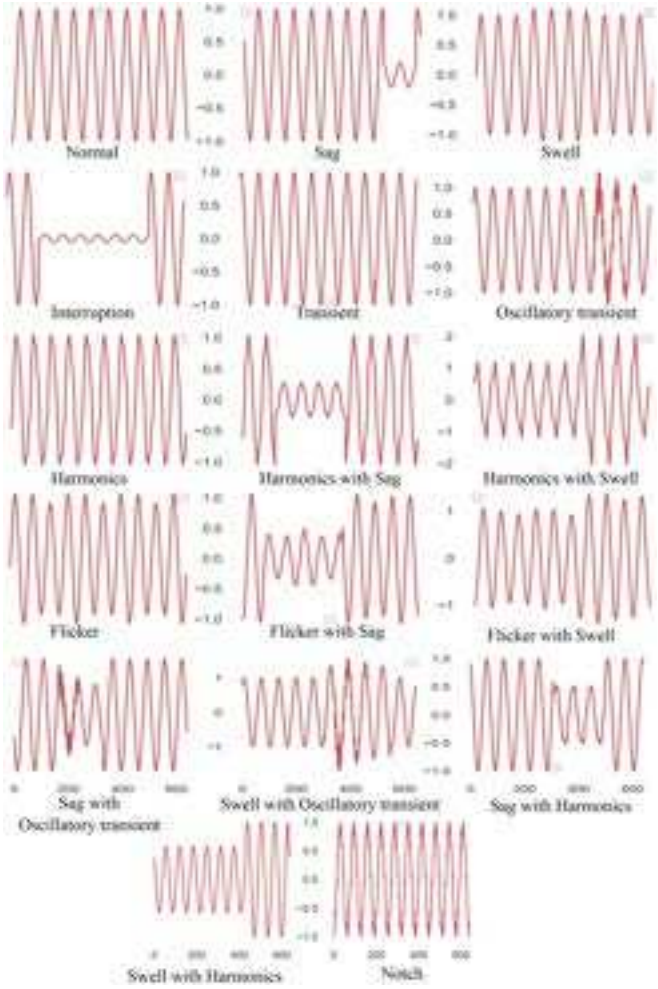


Fig. 4. Waveshape of Different PQDs.

the perturbed inputs $x_i + r$ as a specific target class y_{target}. The optimization problem is further constrained to ensure that the generated perturbations remain imperceptible. The first constraint, $\max_i\{\text{SNR}(x_i, r) \ \} \geq \text{SNR}_{\text{min}}$, ensures that the signal-to-noise ratio for any perturbed sample remains above a minimum threshold. Concurrently, the second constraint, $\|r\|_p \leq \epsilon$, controls the magnitude of the perturbation. Through these constraints, the AdaptEdge algorithm adeptly navigates the trade-off between maximizing the fooling rate and ensuring the perturbation's imperceptibility.

C. Overview of Targeted Universal Adversarial Attacks in Smart Grid

PQDs can be categorized into different classes: Normal, indicating a typical waveform free of anomalies; Sag, a momentary voltage drop; Swell, a transient voltage spike; Interruption, a momentary power outage; Transient, sudden surges typically caused by equipment failures or lightning; Oscillatory transient, a brief frequently decaying waveform deviation; Harmonics, integer multiples of the fundamental frequency causing distortion; Harmonics with Sag and Harmonics with Swell are respective combinations of harmonics with voltage declines and surges; Flicker, perceptible

voltage fluctuations causing illumination discomfort; Flicker with Sag and Flicker with Swell, voltage dips and surges combined with flicker, respectively; Sag with Oscillatory transient and Swell with Oscillatory transient, voltage dips and surges, and repetitive waveform deviations; Sag with Harmonics and Swell with Harmonics are harmonics-induced distortions accompanied by declines and surges,and Notch, a short disturbance in the waveform. When PQD occurs, it propagates through the transmission and distribution system to reach the substation. To effectively analyze these disturbances, they are processed through a DL model. The model is designed to account for the variability and complexity of PQD data, which can vary significantly across different substations due to factors such as local consumption patterns, the integration of renewable energy sources, etc.

The DL model processes this raw data, extracting key features and classifies disturbances with high precision. After classifying the PQD, the DL model communicates the results to the control center. This information empowers the autonomous power system controller to take decisive action, including voltage regulation, transitioning to backup power sources, activating or deactivating generating stations, and adjusting the load. The communication channel is crucial for transmitting raw signals between the substation and the control center. However, attackers can compromise this channel and introduce TUAA to the PQDs signals that can threaten the resilience of SG to potential cyber threats and manipulate the DL model's output by altering the signals received by the control center. Incorrect signals may lead to erroneous decisions, jeopardizing the power system's optimal and safe operation.

V. Methodology of Targeted Universal Adversarial Attack

The method for constructing a TUAA using the AdaptEdge algorithm is described in Algorithm 1. For greater clarity, we have illustrated the process of the Adaptive Edge Algorithm with a flowchart. Figure 5 depicts the algorithm's progression graphically. The algorithm begins by initializing parameters such as the initial fooling rate, universal perturbation, and the number of iterations. The signals belonging to the source class are stored in an array. The main loop of the algorithm continues until either the fooling rate surpasses a predefined threshold or the maximum number of iterations is reached. During each iteration, the algorithm calculates the fooling rate and SNR, and then calls the AdaptEdge function to dynamically adjust the perturbation edge value ε. The fooling rate, in the context of targeted attacks, is defined as the fraction of adversarial samples that were both misclassified by the model and classified specifically to the desired target class. Mathematically, let Y_{true} be the true labels of our test samples, Y_{adv} be the predicted labels of the adversarial samples, and Y_{target} be the desired target class for our adversarial attack. The fooling rate is given by:

$$\textit{fooling rate} = \frac{\sum_{i=1}^{N} \mathbb{I}\left(Y_{\text{true},i} \neq Y_{\text{adv},i} \text{ and } Y_{\text{adv},i} = Y_{\text{target}}\right)}{N}$$

Algorithm 1 Adaptive Edge (AdaptEdge) Algorithm

Require: `source_class`,`target_class`, `maximum_iterations`,initial perturbation edge ε,`fooling_rate_threshold`
Ensure: Universal perturbation for targeted adversarial attack

```
fooling_rate ← 0
universal_perturbation ← 0
i ← 0
arr ← source class signals from training samples
while fooling_rate<fooling_rate_threshold and i <maximum_iterations do
    Search for optimal ε based on AdaptEdge function
    Select one signal, x, at a time from arr
    if target_class ≠ source_class then
        Initialize pert_signal = x and a_p = 0
        Compute the gradients: ∇f_s(x), ∇f_t(x), and prediction
        while prediction[target] < c_t do
            Compute p_d, p_m, c_p
            Update a_p and pert_signal
            Re-compute gradients and prediction
        end while
        universal_perturbation = a_p
        Update universal_perturbation
        universal_perturbation= project(universal_perturbation, ε)
    end if
    i ← i + 1
    Apply universal_perturbation to all samples in the test dataset
    Calculate Signal_to_Noise_Ratio
    Calculate fooling_rate
end while
return universal_perturbation
```

Where $\mathbb{I}$ is the indicator function, which is 1 if the condition inside is true and 0 otherwise. N is the total number of test samples. In the context of the targeted attack, this fooling rate measures how often the adversarial perturbations caused a misclassification specifically towards the desired target class. A higher fooling rate indicates that the adversarial attack is more effective in guiding the misclassifications toward the target class.

Algorithm 2 introduces the AdaptEdge function, which dynamically adjusts the perturbation edge to enhance the effectiveness of deceiving DL models in the context of TSD. The hypothesis behind this approach is that modifying the perturbation edge dynamically can improve the attack's success without affecting the input signal's imperceptibility. Empirical studies demonstrated promising results, validating the effectiveness of our method in generating perturbations capable of deceiving DL models when applied to TSD. In the initial phase of our proposed algorithm, we establish a specific threshold for the SNR to classify degrees of imperceptibility as high, medium, or low. These levels are determined by modulating the perturbation edge value and visually inspecting

Algorithm 2 Adaptive Edge (AdaptEdge) Function

Require: `SNR_threshold`,`fooling_rate_threshold`, initial perturbation edge ε.

Ensure: Optimal ε for the required level of imperceptibility with the desired fooling rate

```
fooling_rate ← 0
ε ← small initial value
while fooling_rate≤fooling_rate_threshold or SNR≥SNR_threshold do
    Increase ε
    Calculate perturbation with new ε
    Update fooling_rate
    Update SNR
end while
while fooling_rate≥fooling_rate_threshold or SNR≥SNR_threshold do
    Decrease ε
    Calculate perturbation with new ε
    Update fooling_rate
    Update SNR
end while
return Optimal ε
```

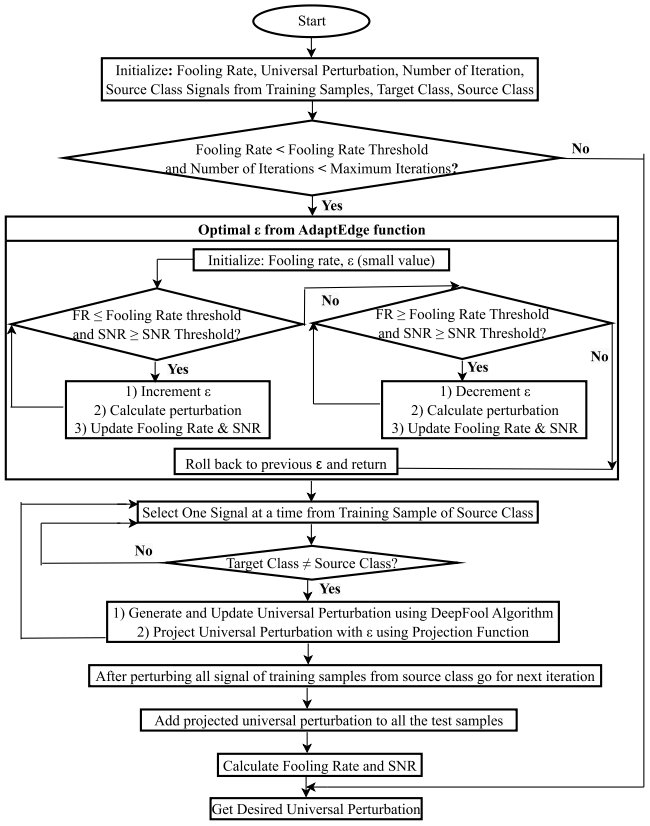


Fig. 5. Flow chart of targeted universal adversarial attack on time series data using AdaptEdge algorithm.

the resulting signal integrity. We infer that the attacker is aware of these established SNR thresholds and uses this information to navigate the levels of imperceptibility.

When we increase ε, we effectively expand the "edge" of the hypersphere, allowing perturbations to exist in a larger region. When ε is decreased, this boundary is shrunk, thereby confining the perturbations to a smaller region. In this context, the algorithm begins with a very small value for the ε and gradually increases it. The goal is to determine the maximum value of ε that preserves signal integrity while achieving high, medium, and low levels of imperceptibility.

After the SNR threshold and the fooling rate threshold have been set, our algorithm starts with an initial small value of ε. The experimentation begins with a high degree of imperceptibility for the given source-target class pair. If the fooling rate meets or exceeds the threshold, the algorithm further minimizes ε to determine if a desired fooling rate is achieved using a value smaller than the initial ε. If the desired rate of deception is achieved, ε is further decreased to ensure an extremely high level of imperceptibility (i.e., high SNR). This reduction continues until a fooling rate equal to or greater than the criterion is reached. This optimal ε represents the optimal radius and thus defines the optimal perturbation edge. In contrast, if a fooling rate below the threshold is obtained using the initial ε, the algorithm increases ε to determine if the rate of deception approaches or surpasses the threshold and ε increases until the threshold is reached or exceeded. The algorithm then starts decreasing ε to a value between the incremented ε and the initial ε to check whether the imperceptibility can be increased with the desired fooling rate. If the fooling rate again equals or exceeds the threshold, ε will continue to decrease until it falls below the threshold. When the fooling rate falls below the threshold, the algorithm ceases reducing the ε and resets it to its previous value. This dynamic modification of ε permits optimal adversarial perturbation while preserving signal integrity. The algorithm demonstrates a sophisticated strategy that consistently and adaptively seeks the optimal perturbation edge value. By intelligently balancing imperceptibility and deception rate, it precisely navigates the search space, always converging on a minimal point. Our empirical evaluations show that the proposed linear search method is highly effective despite its apparent simplicity. It yielded positive results in nearly 93.75% of the cases we tested (Please refer to Table II).

Algorithm 1 only attempts to generate a perturbation if the current source signal is not classified as the desired target class, as it prevents the algorithm from wasting computational resources on unnecessary perturbations. However, a mechanism is implemented if it is impossible to modify the source class to the target class. If there is no significant progress toward the desired misclassification after the maximum number of iterations, the algorithm will identify this situation and proceed to the next signal. If not, the algorithm employs the DeepFool algorithm [29] to generate an adversarial perturbation for the current signal and update the universal perturbation. The gradient determines how to perturb the input to maximize classifier output change. The algorithm iteratively computes the gradient of the model's output with respect to the input data and modifies the input data to reduce the model's confidence in the true class, and it repeats until the classifier output changes. If the target class

TABLE II
SIMULATION RESULTS FOR THE TARGETED UNIVERSAL ADVERSARIAL ATTACK USING ADAPTEDGE ALGORITHM

Source Class	Target Class	FR (HI)	FR (MI)	FR (LI)	Source Class	Target Class	FR (HI)	FR (MI)	FR (LI)
C1	C2	5.87	10.34	38.18	C6	C14	14.18	37.68	61.20
C1	C7	8.323	54.11	73.04	C7	C9	6.28	6.21	5.81
C1	C9	14.37	63.22	89.49	C7	C14	9.81	28.95	67.12
C1	C12	35.16	40.92	61.62	C7	C8	37.92	44.60	45.59
C1	C14	15.42	55.82	84.74	C7	C13	27.763	53.01	65.74
C1	C17	54.62	66.51	70.06	C8	C13	3.82	5.62	5.99
C1	C6	4.22	33.98	57.24	C8	C9	7.34	5.19	3.68
C1	C8	66.293	82.03	90.00	C8	C14	5.88	6.49	6.18
C1	C11	19.373	32.23	41.72	C9	C8	13.1	15.70	15.597
C1	C13	51	75.25	83.88	C9	C14	14.97	16.24	17.33
C1	C16	8.01	28.20	34.26	C9	C13	49.1	49.13	42.78
C2	C6	0.26	12.22	55.54	C10	C6	0.22	6.02	50.79
C2	C8	63.82	81.88	85.78	C10	C8	34.54	55.69	79.35
C2	C11	9.773	31.36	31.36	C10	C13	46.65	65.36	77.42
C2	C13	55.06	73.78	83.24	C10	C9	0.34	5.73	7.82
C2	C15	27.84	28.57	35.60	C10	C4	2.39	4.08	23.61
C2	C17	54.29	68.33	68.97	C10	C7	1.04	36.19	64.46
C2	C3	29.16	28.86	38.08	C10	C9	10.371	9.99	47.61
C2	C7	26.84	50.78	75.12	C10	C14	16.803	71.04	90.78
C2	C9	14.29	37.21	63.40	C10	C17	47.55	65.72	73.58
C2	C12	31.13	41.19	51.14	C11	C6	4.88	39.50	61.80
C2	C14	12.243	54.99	87.70	C11	C8	60.64	81.54	89.90
C2	C16	1.96	7.37	30.84	C11	C12	0.133	0.52	12.51
C3	C7	4.54	39.91	63.92	C11	C14	28.28	76.55	89.85
C3	C9	15.22	28.66	65.88	C11	C16	0.22	5.15	25.71
C3	C12	35.05	57.44	70.26	C11	C7	4.93	44.57	68.57
C3	C14	12.78	57.71	85.17	C11	C9	11.69	19.05	61.50
C3	C6	3.74	37.21	61.10	C11	C13	50.693	70.19	80.17
C3	C8	61.873	82.12	88.11	C11	C15	32.14	37.82	39.98
C3	C11	33.71	34.71	34.46	C11	C17	59.7	65.84	65.84
C3	C13	56.44	71.12	81.11	C12	C8	75.823	88.92	90.54
C3	C17	61.763	67.43	68.60	C12	C13	21.75	25.13	23.99
C4	C7	28.88	58.96	73.38	C12	C17	62.583	69.11	69.11
C4	C9	2.08	0.58	0.08	C12	C9	3.823	6.06	23.52
C4	C14	10.813	12.86	12.23	C12	C14	12.43	12.56	11.34
C4	C8	30.56	41.28	48.09	C13	C9	1.403	2.76	3.26
C4	C13	41.15	43.22	29.89	C13	C8	32.2	54.21	60.82
C5	C6	4.6	39.03	64.77	C13	C14	1.4	2.76	3.26
C5	C8	79.533	86.83	86.83	C14	C6	0.38	0.84	4.21
C5	C11	30.64	31.31	31.31	C14	C9	1.65	0.84	0.20
C5	C13	55.12	76.36	84.00	C14	C8	17.08	20.12	20.91
C5	C16	0.54	1.40	19.25	C14	C13	42.11	44.50	46.40
C5	C3	23.183	23.56	44.65	C15	C8	19.71	15.00	12.73
C5	C7	4.49	45.20	69.70	C15	C12	1.52	13.83	31.40
C5	C9	17.76	54.82	86.22	C15	C14	1.52	13.83	31.40
C5	C12	34.79	44.50	54.29	C15	C9	37.88	65.58	85.90
C5	C14	11.96	17.19	50.98	C15	C13	26.14	44.69	40.26
C5	C17	53.01	64.82	68.98	C16	C7	77.3	26.44	53.94
C6	C7	0.83	34.67	62.98	C16	C9	6.26	20.77	63.18
C6	C9	3.76	4.90	8.16	C16	C14	24.96	66.97	89.65
C6	C13	59.88	72.09	80.33	C16	C8	67.833	87.05	89.48
C6	C17	61.78	66.74	66.74	C16	C13	48.12	80.35	85.80
C6	C8	74.24	86.32	86.32	C17	C13	19.54	35.82	52.38
C6	C11	23.8	30.53	36.11	C17	C8	44.75	52.00	58.71
C6	C14	14.18	37.68	61.20	C17	C14	9.703	57.98	87.38

* FR-Fooling Rate, HI- High Imperceptibility, MI- Medium Imperceptibility, LI- Low Imperceptibility.

differs from the source class, the algorithm initializes the perturbation signal from a given input signal x and begins with zero accumulated perturbation a_{p}. It computes gradients with respect to the intended source and target classes for the signal. The loop iterates until the confidence predicted for the target class exceeds a specified threshold, c_t. The algorithm determines the perturbation direction p_d throughout each iteration, and the perturbation magnitude p_m, as well

as the accumulated perturbation a_{p}, are then updated using current iteration perturbation c_p. This iterative process refines the perturbation until it is sufficient to fool the classifier. After iterations, the algorithm returns the accumulated perturbation to incorrectly classify the input signal as the target class. The updated perturbation is then projected onto the ε-radius ball using a projection function to restrict the perturbation's maximum strength. In this context, the projection function is a mathematical operation that confines the updated adversarial perturbation within a defined boundary, specifically a hypersphere of radius ε. The projection function scales the given values to fit within a hypersphere of radius ε in L2-norm space, ensuring that the magnitude of the adversarial perturbation does not exceed the specified ε limit. The essence of an adversarial perturbation is its magnitude and direction, which indicates how an input sample is modified to cross a classifier's decision boundary. The projection operation primarily modifies the magnitude while preserving the direction, thereby not significantly altering the adversarial characteristics of the perturbation. While the direction of the perturbation is crucial, limiting its magnitude ensures that the perturbations remain discreet, thereby enhancing the imperceptibility of the attack. The projection ensures that the magnitude does not exceed a predetermined threshold, making it difficult to detect and effective. The algorithm applies the current universal perturbation to the entire test dataset and predicts the class labels of the perturbed dataset after each iteration. The SNR is then computed to evaluate the imperceptibility of the disturbance, and the fooling rate is recalculated to determine if it has reached the desired threshold. Ultimately, the algorithm produces a universal adversarial perturbation that can cause effective misclassification towards the target class while preserving the perceptual quality of the signals. This comprehensive and adaptive strategy ensures a delicate balance between high imperceptibility and the desired fooling rate, making the algorithm significantly contribute to adversarial machine learning.

VI. Simulation Result

A. Description of Power Quality Disturbances

In the context of SG, noise can originate from various sources. Internal components, such as transformers and power electronics, can contribute to noise, while external factors, such as electromagnetic interference and environmental disturbances, can contribute to additional fluctuations. This noise is typically stochastic and can follow a variety of statistical distributions, including Gaussian and Poisson. Gaussian noise might model uncertainties from electronic devices and their inherent thermal noise, whereas Poisson distributions are more appropriate for representing event-driven noise, such as that from random fault occurrences. Notably, these distributions are approximations of the original noise distribution in SG. Depending on the specific system and environment, the actual noise distribution may be more complex and variable. Its fluctuating nature makes it challenging to distinguish from adversarial perturbations. Adversarial perturbations can be designed to have similar statistical properties, subtly altering the signal without changing its fundamental characteristics. This similarity to legitimate system noise poses difficulties in detection, as grid operators and security systems are accustomed to continuous and pervasive noise. Well-designed adversarial perturbations can thus camouflage themselves within the existing "noise floor," evading detection by mimicking legitimate noise patterns.

PQD manifests in many intriguing forms, ranging from voltage sags and interruptions to more elusive phenomena like flickers, swells, and spikes. This diverse array extends to oscillatory transients, harmonics, notches, and complex combinations. The provided equations offer simplified representations of these signals, which may be subject to variations based on the specific characteristics of each disturbance. The mentioned parameters represent the key variables related to each signal class, although additional parameters may exist for more extensive modeling and analysis. It is important to note that these PQDs can have distinct effects on the power system and connected equipment, giving rise to various operational issues and potential damage. Understanding and addressing these PQDs is crucial for maintaining a reliable and efficient power system.

B. Datasets and Deep Learning Model

To assess the performance of the proposed TUAA for PQDs, we apply ResNet50 as a DL model in our case. The mathematical model and parameters of PQDs proposed in [30] are employed, where the PQ models set the sampling frequency of signals to 3200 Hz, the fundamental frequency to 50 Hz, the number of total cycles to 10, and the amplitude to 1. Consequently, the input signal vectors have a fixed length of 640, although the actual signal is continuous and uninterrupted. For all 17 classes of signals, Table A displays signal types, mathematical equations, and parameters. A publicly available [22] labeled dataset focuses on processing and analyzing relatively clean, class-balanced data. A class-balanced, publicly accessible labeled dataset concentrates on processing and analyzing relatively pure data. Using 15000 signals from each class, the dataset contains 255,000 signals with an SNR of 30dB. All of the samples are separated into 17 PQD. We take 207000 training samples, 23000 validation samples, and 25000 testing samples. To assure the randomness and robustness of our model, we randomize the order of data samples. The labels are transformed to encode them using one-hot encoding. Signals are reshaped to ensure that each time point is regarded as a separate feature, preserving the temporal dependencies. After training for ten epochs, our model demonstrates excellent performance, achieving a test accuracy of 99.22%.

C. Experimental Setup

1) Hardware Requirements: The experiments were conducted on a state-of-the-art computational system. The central processing unit (CPU) is an Intel Core i9-9920X, which operates on a 64-bit x86 architecture. The CPU boasts 12 cores per socket, facilitating multi-threaded operations with two threads per core, resulting in a total of 24 logical CPUs. Cache memory is distributed across different levels: 32K

for L1d and L1i, 1024K for L2, and a substantial 19712K for L3. Complementing the system's computation prowess is a robust memory setup of 125GB RAM. Graphics and computationally intensive tasks are delegated to four NVIDIA Quadro RTX 6000 GPUs, enhancing the system's parallel processing capabilities.

2) Software Requirements: The experiments conducted in this study utilized a specific set of software tools and libraries. Ubuntu 18.04.5 LTS (Bionic Beaver), well-regarded for its stability and wide compatibility, was the operating system, and Python (version 3.6) was the primary programming language to construct our computational environment. The system employs NVIDIA's CUDA toolkit for GPU-accelerated tasks, specifically version 10.2.89, complemented by the NVIDIA driver version 470.94. The GPUs efficiently handle a variety of processes, from system operations such as Xorg to computational tasks written in Python. We used Keras 2.2.4 with TensorFlow 1.13.1 as the backend for DL tasks and model implementations and Pycharm 2023.2 (Community Edition) as the integrated development environment. Keras preprocessing (version 1.1.2) facilitated data augmentation and preprocessing stages. Using Matplotlib (version 3.3.3), data visualization was achieved. We relied heavily on Numpy (version 1.19.5) for numerical computations and Pandas (version 1.1.4) for data manipulation and management. Lastly, we utilized Scikit-learn (version 0.23.0) for specific machine-learning tasks and data preprocessing processes. Researchers attempting to replicate or extend our findings must use these precise versions to maintain consistency with our experimental design.

3) Hyperparameter Settings: In our experiments, the DL model utilizes the categorical cross-entropy loss function and the Nadam optimizer with the following parameters: $lr = 0.002$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and stability term $\epsilon = 1e-08$. The learning rate's decay schedule was set to 0.004. Based on visual inspection, the imperceptibility levels of adversarial perturbation were classified as follows: low imperceptibility at SNR values of 3 dB, medium at 5 dB, and high at 7 dB. The corresponding thresholds for the fooling rate were 30%, 50%, and 70%, respectively. The initial value for ε before starting the experiment for low, medium, and high imperceptibility is 0.5, 2.5, and 4.5, respectively. In addition, our algorithm was limited to a maximum of 150 iterations to ensure convergence.

4) Result and Discussion: Our study explores TUAA on TSD in the SG. Our algorithm proves to be effective across various applications due to the inherent similarities in TSD. Through empirical analysis, we demonstrate the successful execution of universally targeted adversarial attacks on TSD using our proposed algorithm. In Table II, the simulation evaluates the model's behavior under various combinations of source and target classes, where the source class refers to the original classification and the target class represents the intended misclassification. Our proposed algorithm can misclassify these signals into 14 target classes- sag, swell, interruption, oscillatory transient, harmonics, harmonics with sag, harmonics with swell, flicker with sag, flicker with swell, sag with oscillatory transient, swell with oscillatory transient, sag with harmonics, swell with harmonics, and notch- that could significantly compromise the power system's stability, dependability, and overall performance. In Figure 7, we present a confusion matrix that demonstrates the effectiveness of our proposed algorithm in a specific adversarial scenario where 'Normal' is the source class and 'Harmonic with Sag' is the target class. The matrix provides a clear visual representation of how well our model performs in this particular circumstance, and this is clearly demonstrated by the high fooling rate of 90%. Our method generates a universal adversarial perturbation for each pair of source and target classes in the training sample set. When added to test samples, these perturbations induce the model to misclassify them into the intended target class. In table II, we present all combinations of source and target classes that resulted in a fooling rate greater than zero. This data has been divided into three distinct categories of human imperceptibility: high, medium, and low. When examining the waveshape of high imperceptibility in Figure 6(a), we observe that the adversarial perturbations are so minute that they are nearly undetectable, allowing the clean sample to be the most prominent component of the image. Therefore, we classify these instances as having a high degree of imperceptibility. The adversarial perturbations become marginally more apparent for the waveshape of medium imperceptibility in Figure 6(b). However, the original, clean waveforms are still readily apparent. This is because the perturbations closely match the shape of the clean sample, merging with the background noise. Therefore, these instances are classified as medium imperceptibility. If we examine the waveshape of low imperceptibility in the figure in Figure 6(c), the adversarial perturbations are noticeably more pronounced while retaining a degree of subtlety. Analytically, these disturbances are still adversarial; the waveform characteristics do not deviate significantly from the normal pattern. There are spikes in peak values, but the waveforms resemble those of conventional waveforms with added noise. Due to the inherent dynamics of the system, the waveform in power systems can display a variety of complexities. Several peaks, dips, or distortions may spontaneously manifest due to load characteristics, unexpected load increase, or abrupt load reduction. Considering the inherent variation of power system waveforms, This imperceptibility will also make it exceedingly difficult for a human observer to distinguish an attack from typical system noise. Table II displays the range of fooling rates obtained by our proposed algorithm, which ranges from a minuscule 0.08% to an impressive 90.78% for varied degrees of imperceptibility. When analyzing the fooling rates across the three levels of imperceptibility, certain patterns and distinctions emerge. For instance, the fooling rate tends to increase as the level of imperceptibility decreases from HI to LI for various class combinations. This may suggest that universal adversarial perturbations are more effective at misleading the model with low imperceptibility. However, In certain instances, such as the transition from C7 to C9, the fooling rate is nearly constant across all three levels of imperceptibility. This may indicate the inherent robustness of particular class transitions, regardless of their imperceptibility. In a few cases, such as the transition from C1 to C6, the

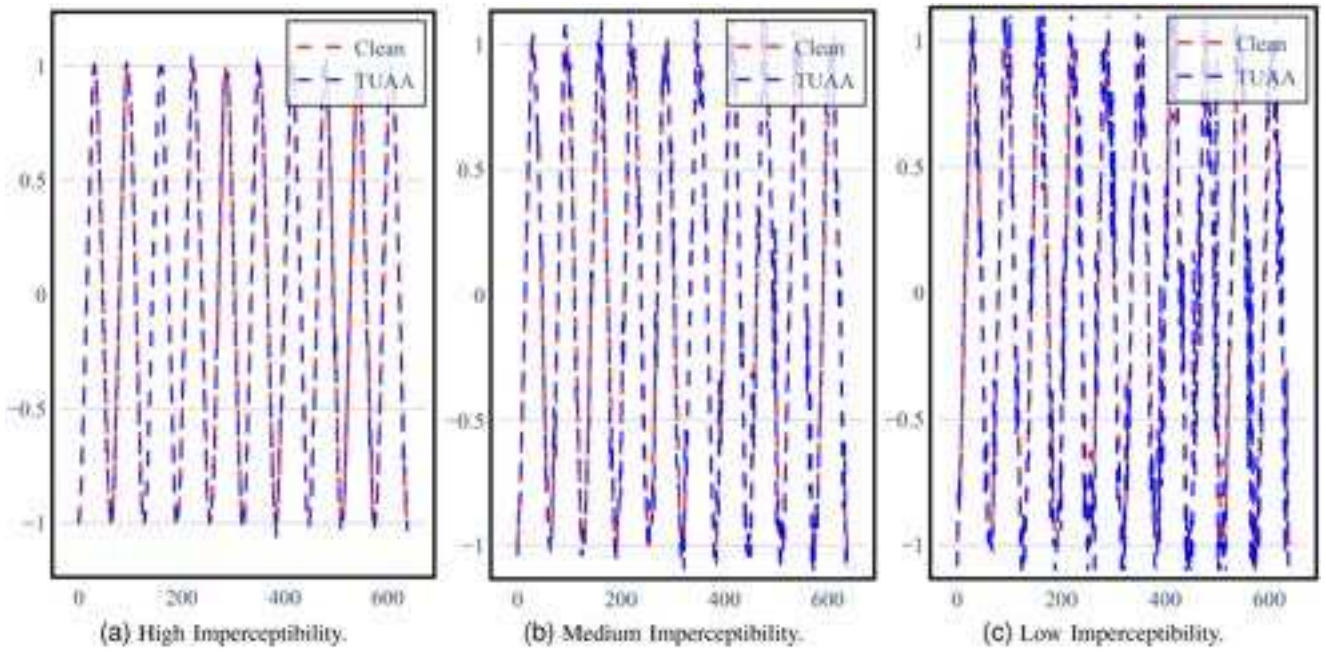


Fig. 6. Waveforms of power quality disturbances after targeted universal adversarial attack with high, medium, and low imperceptibility.

fooling rate is highest at the MI level and reduces at the LI level. Such occurrences necessitate a deeper investigation into the potential causes and particular characteristics of the adversarial perturbations applied. Figure 8 illustrates this dynamic by depicting a trade-off curve between imperceptibility and fooling rate for the swell-harmonics (source-target). The curve reveals that increasing the perturbation magnitude increases the fooling rate while decreasing the imperceptibility.Beginning with a low ε, the algorithm gradually increases it to increase the fooling rate. However, it constantly monitors the SNR to ensure it does not degrade significantly. As the perturbation edge increases, the fooling rate increases, which indicates that the perturbation is more effective at deceiving the model. Simultaneously, the graph indicates a decrease in SNR as ε increases. If the SNR exceeds a predetermined threshold, the algorithm will recalculate to maintain acceptable noise levels. The performance sweet spot attained by the AdaptEdge algorithm is denoted by the trajectory in Figure 8, where the deceiving rate exhibits consistent growth without the SNR becoming critically low. The demonstrated trade-offs between the fooling rate and SNR, as it evolves, shed light on the algorithm's efforts to optimize results. In the SG, the control center primarily bases its decision-making on precisely interpreting signals from substations. This includes decisions regarding load balancing, error detection, and other crucial operational duties. As the fooling rate increases, the control center's DL model misclassifies a greater proportion of signals. These misclassifications can result in erroneous interpretations, such as incorrectly identifying a fault or estimating the load incorrectly. This directly impacts the decisions made by the control center, which may result in suboptimal or even detrimental operational commands. An optimal SNR ensures that adversarial perturbations are nuanced enough to remain undetected but influential enough to cause the desired

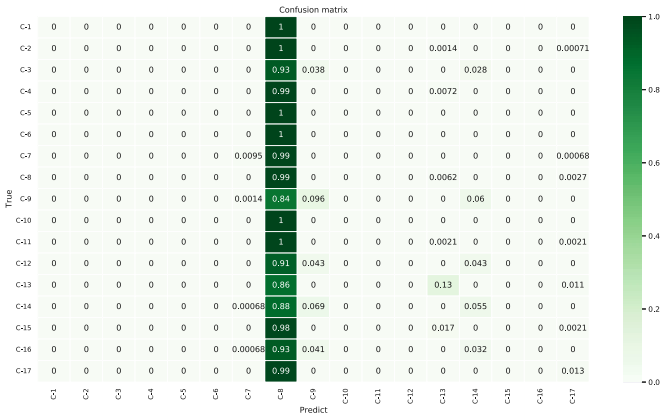


Fig. 7. Confusion matrix for the targeted universal adversarial attack.

misclassification. As the AdaptEdge algorithm modifies, the perturbation edge determines the adversarial attack's potency. A more pronounced perturbation can result in a higher fooling rate but may also reduce the SNR. This balance is crucial as it indicates the attack's ability to mislead the control center without triggering alarms due to observable signal corruption. Our proposed algorithm can misclassify models into 14 distinct target classes, requiring 42 universal perturbations tailored to one of the three degrees of imperceptibility. In our threat model, we have hypothesized that the attacker could be a member of the intelligent IED manufacturing industry or a knowledgeable entity from the utility company, both of which have extensive knowledge of power systems. Under these conditions, the adversary can choose the degree of imperceptibility that best serves their objectives. If the attacker desires a significant disruption, it may select a targeted class, considering the power flow characteristics during the perturbation period and a lower

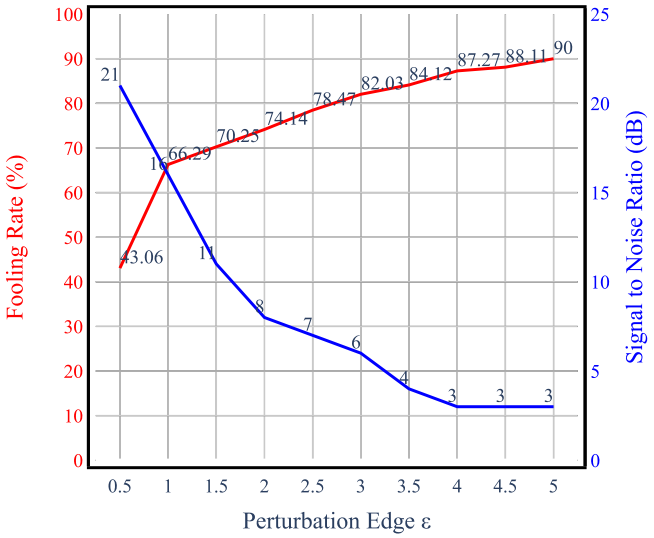


Fig. 8. Trade-off between fooling rate and SNR with an increase of perturbation edge for the combination of normal (source class) and harmonics with sag (target class) using the AdaptEdge algorithm.

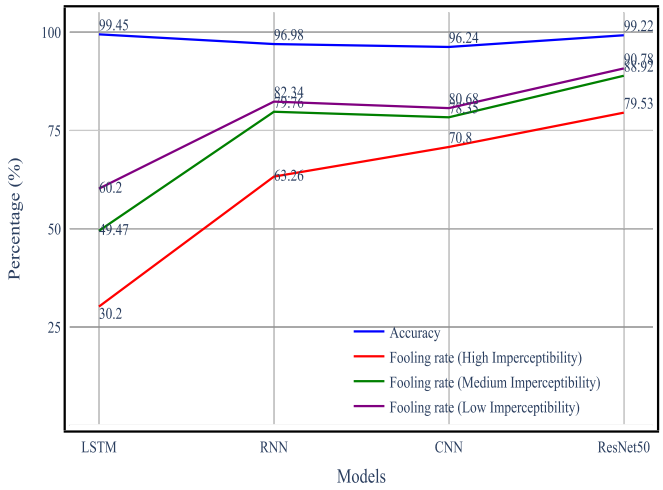


Fig. 9. Comparative Evaluation of AdaptEdge Algorithm Across Different Models.

level of imperceptibility. Alternatively, if the attacker aims to ensure long-term stealthiness, they may sacrifice the fooling rate to maximize imperceptibility. The attacker may opt for medium imperceptibility, which has a relatively higher fooling rate but a noise level comparable to high imperceptibility. This versatility offers the attacker many options for launching their attack. Even contemplating the lower fooling rate of some source and target class pairs in Table II, it should be noted that even a single erroneous decision could cause significant disruptions in critical systems such as the SG. Because of the interconnected nature of power grids, a cyberattack, even on a critical substation, can trigger a chain reaction that shuts down the entire system. When the compromised substation connects a large power plant to the grid or serves as a major hub in the distribution network, the effects are magnified. In applications of this magnitude, ensuring that the DL model resists even the smallest adversarial attacks is crucial to ensure seamless operation. In addition, it is noteworthy that certain source-target pairings deviate from the norm of increasing fooling rate with decreasing imperceptibility, such as sag-swell and swell-flicker with sag, etc. This demonstrates the adaptability and flexibility of the proposed algorithm. Instead of following the trade-off curve shown in Figure 8, the algorithm is designed to iteratively explore the possibility of a higher fooling rate while simultaneously increasing imperceptibility. This characteristic enhances the robustness and efficacy of our algorithm, making it a formidable instrument in the domain of adversarial perturbations.

5) Comparative Evaluation of AdaptEdge Algorithm Across Different Deep Learning Models: To validate the efficacy of our proposed algorithm, we have expanded our experiments to include advanced DL models. As there are no comparable techniques designed specifically for TUAA on TSD, evaluating our algorithm on multiple models serves as comparative validation. In Figure 9, despite test accuracies ranging from 96.236% for DNN to 99.45% for LSTM, all models are susceptible to TUAA. This demonstrates that test accuracy alone cannot indicate the resistance of a model to TUAA. The LSTM model is distinguished by its high test accuracy of 99.452% and substantially lower fooling rates at all levels of imperceptibility compared to other models. This indicates that LSTM may have some inherent resistance to TUAA. However, even this model exhibits a 60.2% fooling rate at low levels of imperceptibility, indicating that additional work is required to make it resilient. ResNet50 has a high test accuracy of 99.22%, but it is the most susceptible to adversarial attacks. At low imperceptibility, its fooling rate skyrockets to 90.78%. This makes it the least robust model among those tested, and deployment of it in security-sensitive applications raises significant concerns. Despite having comparable test accuracies (96.236% for DNN and 96.98% for RNN), their fooling rates differ at high levels of imperceptibility. RNN has a 7% reduced fooling rate at higher imperceptibility than DNN. However, the susceptibility of both models increases as imperceptibility decreases, emphasizing that neither model is genuinely resistant to TUAA. There is a consistent trend across all models that the fooling rate increases as the level of imperceptibility decreases. This demonstrates the tradeoff between the imperceptibility and efficacy of an attack. High imperceptible attacks are typically less effective at deceiving the model. One important observation is the universal susceptibility of all four models to TUAA, irrespective of their architecture and test accuracy. This may indicate a fundamental vulnerability in how neural networks interpret the feature space, making them susceptible to carefully crafted perturbations. Even for the LSTM model, which has the lowest fooling rate, the rate rises from 30.20% at high imperceptibility to 60.20% at low imperceptibility; this indicates that it is still difficult to achieve both high imperceptibility and high fooling rates. According to these observations, even though neural networks may perform exceptionally well under benign conditions, their performance can deteriorate substantially in

the presence of adversarial perturbations. When deploying such models in real-world applications, it is essential to account for these weaknesses.

6) Possible Countermeasures Against Targeted Universal Adversarial Attacks: In [31], we evaluate in depth three widely used defense mechanisms: adversarial training, defensive distillation, and feature squeezing. Our experimental results shed light on their strengths and limitations in TSD for SG against TUAA. Adversarial training entails augmenting the training data with adversarial samples. Defensive distillation is training a secondary model to approximate the output probabilities of the primary model. This process forces the model to learn a more uniform and seamless decision boundary, making it more difficult for adversarial perturbations to lead to significant misclassifications. Feature squeezing reduces the dimensionality of the input data and quantifies it with a reduced precision. By doing so, some of the fine-grained details that adversarial attacks typically exploit are effectively removed. This regularization process enhances the model's ability to resist adversarial perturbations and improves its robustness. In our experiment, we found that adversarial training reduced the fooling rate by an average of 23.73% for high imperceptibility, 31.04% for medium imperceptibility, and an impressive 42.96% for low imperceptibility, establishing itself as a better countermeasure. The effectiveness of defensive distillation is notable, but it does not consistently outperform adversarial training. The effectiveness of feature squeezing has been demonstrated, particularly in high and medium imperceptibility levels, but its performance is less consistent. Both adversarial training and defensive distillation consistently defend against adversarial attacks, with adversarial training showing a minor edge. Feature compression results in a wider variety of outcomes, especially at low imperceptibility. In terms of versatility, it is typically observed that adversarial training outperforms other methods. This is a significant step towards ensuring the safety and dependability of such vital systems, but the search for a fail-safe system is far from complete. Future research efforts must enhance these defense mechanisms or develop new techniques for constructing resilient smart grid systems.

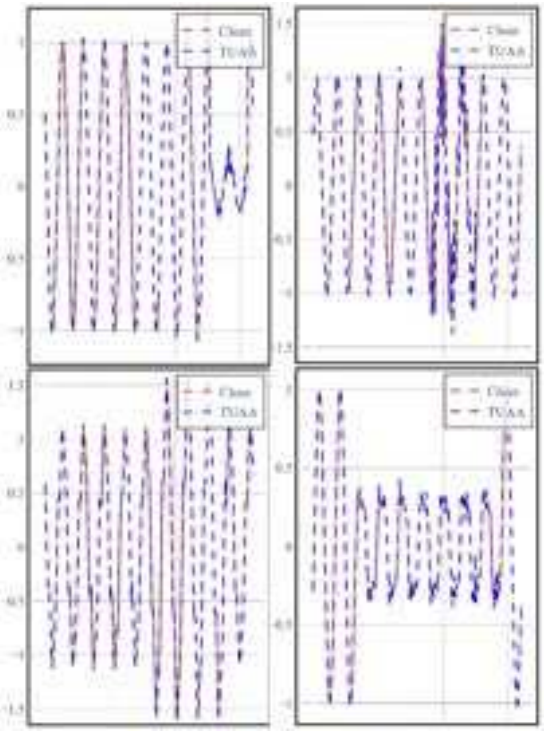
VII. Conclusion

This research presents a complete analysis of the TUAA on DL models employed for classifying PQD in SG. The outcome of the attack yields a maximum fooling rate of 90.78% for the ResNet50 model. Our approach has been expanded to encompass more advanced DL models. Our proposed algorithm's performance yielded a fooling rate of 82.34% for the RNN, 60.2% for the LSTM, and 80.68% for the CNN. These findings showcase the effectiveness and adaptability of our algorithm across different DL models. The observed fooling rate underscores the substantial threat posed by TUAA in SG where even a relatively minor fooling rate can lead to severe and far-reaching repercussions. The authors conducted a thorough examination and subsequently incorporated three distinct imperceptibility criteria to validate the efficacy of their adversarial perturbations. The present study conducted a thorough investigation, revealing that attacks across all imperceptibility criteria, render them challenging for human observers to detect. Based on the findings, our research emphasizes the significance of creating robust DL models for accurately categorizing PQD in SG. Several promising avenues emerge as we contemplate the future of this research, each with the potential to increase the significance of our work. Future researchers can design a robust defense mechanism to detect TUAA on TSD. The optimization of our current algorithm to increase the fooling rate and imperceptibility of attacks can be a secondary objective. By addressing these obstacles, we hope to ensure the continued growth and development of SG and increase the security and dependability of critical electrical infrastructure.

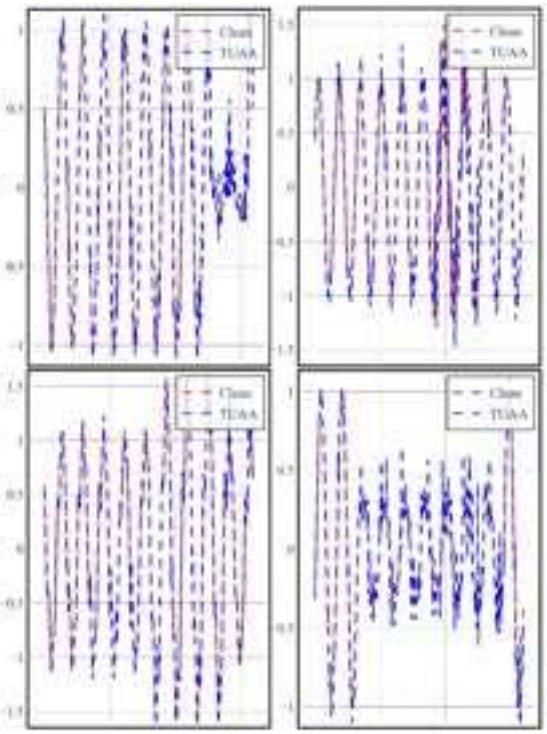
Appendix A
Mathematical Model Of PQD

PQ disturbance	Mathematical Equations	Parameters
C1: Normal	$v(t) = A\sin(wt-\phi)$	Frequency = 50 Hz A = 1
C2: Sag	$v(t) = A(1-\alpha(u(t-t1)-u(t-t2)))\sin(wt-\phi)$	$0.1 \leqslant \alpha \leqslant 0.9$ $T \leqslant t2-t1 \leqslant 9\ T$
C3: Swell	$v(t) = A(1+\beta(u(t-t1)-u(t-t2)))\sin(wt-\phi)$	$0.1 \leqslant \beta \leqslant 0.8$ $T \leqslant t2-t1 \leqslant 9\ T$
C4: Interruption	$v(t) = \mathrm{A}(1-\rho(u(t-t1)-u(t-t2)))\sin(wt-\phi)$	$0.1 \leqslant \rho \leqslant 0.9$ $T \leqslant t2-t1 \leqslant 9\ T$
C5: Transient/Impulse/Spike	$v(t) = A(\sin(wt)+\mathrm{sign}(\sin(wt))) \times\left\{\sum_{n=0}^{9} k[u(t-(t1-0.02n)) -u(t-(t2-0.02n))]\right\}$	$0 \leqslant t1t2 \leqslant 0.5\ T$ $0.01 \leqslant t2-t1 \leqslant 0.05\ T$ $0.1 \leqslant\ K \leqslant 0.4$
C6: Oscillatory transient	$v(t) = A[\sin(wt-\varphi) +\beta_e^{-(t-t_l)/\tau}\sin(w_n(t-t_l) -\vartheta)((u(t-t_l)-u(t-t_l)))]$	$0.5T \leq t_{ll}-t_l \leq \frac{N}{3.33}T$ $8ms \leq \tau \leq 40$ ms $-\pi \leq \vartheta \leq \pi$ $w_n = 2\pi f_n$
C7: Harmonics	$v(t) = A[\sin(wt-\phi)+\sum\alpha_n\sin(nwt-\vartheta_n)]$	$\sum \alpha n^2 = 1$
C8: Harmonics with Sag	$v(t) = A(1-\alpha(u(t-t1)-u(t-t2)))[\sin(wt$	$0.05 \leq \alpha_{n'} \leq 0.15$ $n' = \{3,5\}$
C9: Harmonics with Swell	$v(t) = A(1-\beta(u(t-t1)-u(t-t2)))[\sin(wt-\phi)$	$0.05 \leq \alpha_{n'} \leq 0.15$ $n' = \{3,5\}$
C10: Flicker	$v(t) = A\left[1+\lambda\sin\left(w_f t\right)\right]\sin(wt-\phi)$	$0.05 \leqslant \lambda \leqslant 0.1$ $8 \leqslant f_f \leqslant 25$ Hz $w_f = 2\pi f_f$
C11: Flicker with Sag	$v(t) = A\left[1+\lambda\sin\left(w_f t\right)-\alpha(\mathrm{u(t-t1)} -\mathrm{u(t-t2)})\right]\sin(wt-\phi)$	$0.05 \leqslant \lambda \leqslant 0.1$ $8 \leqslant f_f \leqslant 25$ Hz $w_f = 2\pi f_f$
C12: Flicker with Swell	$v(t) = A\left[1+\lambda\sin\left(w_f t\right)-\beta(\mathrm{u(t-t1)} -\mathrm{u(t-t2)})\right]\sin(wt-\phi)$	$0.05 \leqslant \lambda \leqslant 0.1$ $8 \leqslant f_f \leqslant 25$ Hz $w_f = 2\pi f_f$
C3: Sag with Oscillatory transient	$v(t) = A[\sin(wt-\varphi)(1-\alpha(u(t-t1)-u(t-t2))) +\beta_e^{-(t-t_l)/\tau}\sin(w_n(t-t_l) -\vartheta)((u(t-t_u)-u(t-t_l)))]$	same as sag & oscillatory transient
C14: Swell with Oscillatory transient	$v(t) = A[\sin(wt-\varphi)(1-\beta(u(t-t1)-u(t-t2))) +\beta_e^{-(t-t_l)/\tau}\sin(w_n(t-t_l) -\vartheta)((u(t-t_{ll})-u(t-t_l)))]$	same as swell & oscillatory transient
C15: Sag with Harmonics	$v(t) = A[1-\alpha(u(t-t1)-u(t-t2)))[\sin(wt-\varphi) +\sum_{n'=3}^{5}\alpha_{n'}\sin(n'wt-\vartheta_{n'})]$	same as sag & Harmonics
C16: Swell with Harmonics	$v(t) = A[1-\alpha(u(t-t1)-u(t-t2)))[\sin(wt-\varphi) +\sum^{5}\alpha_{n'}\sin(n'wt-\vartheta_{n'})]$	same as swell & Harmonics
C17: Notch	$v(t) = A[\sin(wt-\varphi)-\mathrm{sign}(\sin(wt-\varphi)) \times\left\{\sum_{n=0}^{N_c-1} k[u(t-(t_c+s.n)) -u(t-(t_d+s.n))]\right\}$	$0 \leqslant t1t2 \leqslant 0.5\ T$ $0.01 \leqslant t2-t1 \leqslant 0.05\ T$ $0.1 \leqslant\ K \leqslant 0.4$

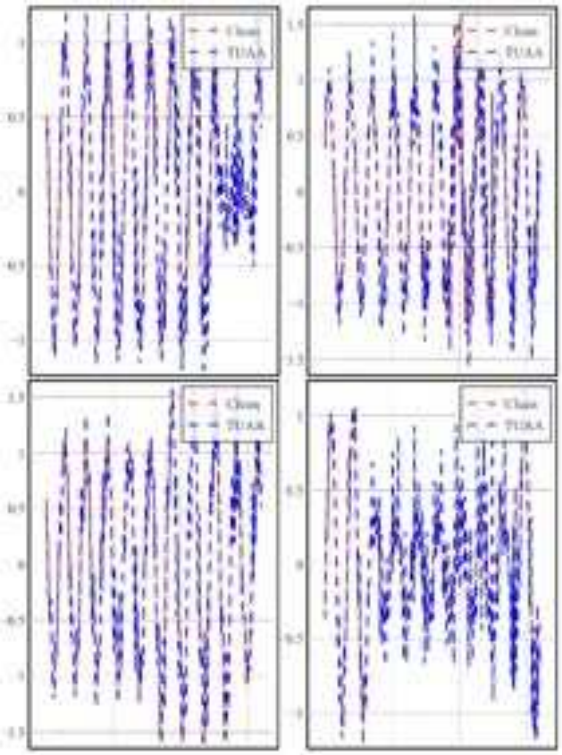
APPENDIX B
WAVEFORMS OF PQD AFTER TUAA WITH HIGH IMPERCEPTIBILITY



APPENDIX C
WAVEFORMS OF PQD AFTER TUAA WITH MEDIUM IMPERCEPTIBILITY



APPENDIX D
WAVEFORMS OF PQD AFTER TUAA WITH LOW IMPERCEPTIBILITY



REFERENCES

[1] A. Faizan, "Difference between traditional power grid and smart grid," *Electr. Academia*, 2017.

[2] G. Dileep, "A survey on smart grid technologies and applications," *Renew. Energy*, vol. 146, pp. 2589–2625, Feb. 2020.

[3] D. Bian, M. Kuzlu, M. Pipattanasomporn, and S. Rahman, "Analysis of communication schemes for advanced metering infrastructure (AMI)," in *Proc. IEEE PES Gen. Meeting Conf. Expo.*, 2014, pp. 1–5.

[4] C.-C. Sun, A. Hahn, and C.-C. Liu, "Cyber security of a power grid: State-of-the-art," *Int. J. Electr. Power Energy Syst.*, vol. 99, pp. 45–56, Jul. 2018.

[5] R. A. D. Oliveira and M. H. Bollen, "Deep learning for power quality," *Electr. Power Syst. Res.*, vol. 214, Jan. 2023, Art. no. 108887.

[6] Y. Chen, Y. Tan, and D. Deka, "Is machine learning in power systems vulnerable?" in *Proc. IEEE Int. Conf. Commun., Control, Comput. Technol. Smart Grids (SmartGridComm)*, 2018, pp. 1–6.

[7] R. Huang and Y. Li, "Adversarial attack mitigation strategy for machine learning-based network attack detection model in power system," *IEEE Trans. Smart Grid*, vol. 14, no. 3, pp. 2367–2376, May 2023.

[8] J. Tian, B. Wang, Z. Wang, K. Cao, J. Li, and M. Ozay, "Joint adversarial example and false data injection attacks for state estimation in power systems," *IEEE Trans. Cybern.*, vol. 52, no. 12, pp. 13699–13713, Dec. 2021.

[9] P. Rathore, A. Basak, S. Nistala, and V. Runkana, "Untargeted, targeted and universal adversarial attacks and defenses on time series," 2021, *arXiv:2101.05639*.

[10] C. Cui, Y. Duan, H. Hu, L. Wang, and Q. Liu, "Detection and classification of multiple power quality disturbances using stockwell transform and deep learning," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, Oct. 2022.

[11] F. Ucar, O. F. Alcin, B. Dandil, and F. Ata, "Power quality event detection using a fast extreme learning machine," *Energies*, vol. 11, no. 1, p. 145, 2018.

[12] E. Yiğit, U. Özkaya, Ş. Öztürk, D. Singh, and H. Gritli, "Automatic detection of power quality disturbance using convolutional neural network structure with gated recurrent unit," *Mobile Inf. Syst.*, vol. 2021, pp. 1–11, Jul. 2021.

[13] I. Topaloglu, "Deep learning based a new approach for power quality disturbances classification in power transmission system," *J. Electr. Eng. Technol.*, vol. 18, pp. 77–88, Jan. 2023.
[14] I. Niazazari and H. Livani, "Attack on grid event cause analysis: An adversarial machine learning approach," in *Proc. IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf. (ISGT)*, 2020, pp. 1–5.
[15] A. Sayghe, J. Zhao, and C. Konstantinou, "Evasion attacks with adversarial deep learning against power system state estimation," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, 2020, pp. 1–5.
[16] E.-N. S. Youssef, F. Labeau, and M. Kassouf, "Adversarial dynamic load-altering cyberattacks against peak shaving using residential electric water heaters," *IEEE Trans. Smart Grid*, vol. 15, no. 2, pp. 2073–2088, Mar. 2024.
[17] G. Zhang and B. Sikdar, "Ensemble and transfer adversarial attack on smart grid demand-response mechanisms," in *Proc. IEEE Int. Conf. Commun., Control, Comput. Technol. Smart Grids (SmartGridComm)*, 2022, pp. 53–58.
[18] J. Tian, B. Wang, J. Li, and C. Konstantinou, "Adversarial attack and defense methods for neural network based state estimation in smart grid," *IET Renew. Power Gener.*, vol. 16, no. 16, pp. 3507–3518, 2022.
[19] Y. Cheng, K. Yamashita, and N. Yu, "Adversarial attacks on deep neural network-based power system event classification models," in *Proc. IEEE PES Innov. Smart Grid Technol.-Asia (ISGT Asia)*, 2022, pp. 66–70.
[20] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, "Malicious data attacks on smart grid state estimation: Attack strategies and countermeasures," in *Proc. 1st IEEE Int. Conf. Smart Grid Commun.*, 2010, pp. 220–225.
[21] R. Heinrich, C. Scholz, S. Vogt, and M. Lehna, "Targeted adversarial attacks on wind power forecasts," 2023, *arXiv:2303.16633*.
[22] J. Tian, B. Wang, J. Li, and Z. Wang, "Adversarial attacks and defense for CNN based power quality recognition in smart grid," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 2, pp. 807–819, Mar./Apr. 2022.
[23] Q. Zhang, Y. Ding, Y. Tian, J. Guo, M. Yuan, and Y. Jiang, "AdvDoor: Adversarial backdoor attack of deep learning system," in *Proc. 30th ACM SIGSOFT Int. Symp. Softw. Test. Anal.*, 2021, pp. 127–138.
[24] H. Hirano, A. Minagi, and K. Takemoto, "Universal adversarial attacks on deep neural networks for medical image classification," *BMC Med. Imag.*, vol. 21, p. 9, Jan. 2021.
[25] A. N. Bhagoji et al., "Practical black-box attacks on deep neural networks using efficient query mechanisms," in *Proc. 15th Eur. Conf. Comput. Vis.*, 2018, pp. 8–14.
[26] C. Wu, R. Zhang, J. Guo, M. De Rijke, Y. Fan, and X. Cheng, "PRADA: Practical black-box adversarial attacks against neural ranking models," *ACM Trans. Inf. Syst.*, vol. 41, pp. 1–27, Apr. 2023.
[27] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction APIs," in *Proc. 25th USENIX Secur. Symp.*, 2016, pp. 601–618.
[28] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1765–1773.
[29] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2574–2582.
[30] R. Igual, C. Medrano, F. J. Arcega, and G. Mantescu, "Integral mathematical model of power quality disturbances," in *Proc. 18th Int. Conf. Harmonics Qual. Power (ICHQP)*, 2018, pp. 1–6.
[31] S. U. Khan, M. Mynuddin, I. Adom, and M. N. Mahmoud, "Mitigating targeted universal adversarial attacks on time series power quality disturbances models," in *Proc. 5th IEEE Int. Conf. Trust, Privacy Security. Intell. Syst., Appl.*, 2023, pp. 91–100.

Sultan Uddin Khan received the B.Sc. degree in electrical and electronic engineering from the Chittagong University of Engineering and Technology, Chittagong, Bangladesh, in 2011. He is currently pursuing the M.Sc. degree with the Department of Electrical and Computer Engineering, North Carolina A&T State University, USA. Before pursuing the Master of Science degree, he developed his skills in the field of power systems by working with Dhaka Power Distribution Company Ltd., one of the largest power distribution companies in Bangladesh, as a Power System Protection and Automation Engineer. He became profoundly involved in large-scale projects involving the modernization, protection, and smart grid technologies of power systems. His career spans more than seven years and is distinguished by extensive experience and practical knowledge that bridges the divide between power systems engineering and contemporary automation processes. His research interests include cyber security, smart grid, the application of deep learning in power systems, unmanned aerial vehicles, machine learning, and deep learning for cyber security.

Mohammed Mynuddin received the B.Sc. degree in electrical and electronic engineering from the Chittagong University of Engineering and Technology, Chittagong, Bangladesh, in 2011, and the M.Sc. degree in electrical engineering from Georgia Southern University, Statesboro, GA, USA, in 2020. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, North Carolina A&T State University, USA. His research interests include cyber security, connected and autonomous vehicles, unmanned aerial vehicles, smart grid, machine learning, and deep learning for cyber security.

Mahmoud Nabil (Member, IEEE) received the Bachelor of Science and Master of Science degrees (with Hons.) in computer engineering from Cairo University, Egypt, in 2012 and 2016, respectively, and the Ph.D. degree in electrical and computer engineering from Tennessee Tech University, Cookeville, Tennessee, in August 2019. He currently holds the position of an Assistant Professor with the Department of Electrical and Computer Engineering, North Carolina A&T State University. He is an accomplished researcher and has authored and coauthored numerous publications in prestigious venues. His research work has been published in renowned journals, such as IEEE INTERNET OF THINGS JOURNAL, IEEE TRANSACTIONS OF DEPENDABLE AND SECURE COMPUTING, IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, and IEEE TRANSACTIONS OF MOBILE COMPUTING. He has received significant funding for his research projects from esteemed national agencies and organizations, including the National Science Foundation, the Department of Transportation, Air Force Research Laboratory, NASA, Intel, Cisco, and Lockheed Martin. With diverse research interests, his areas of expertise include security and privacy in unmanned aerial systems, smart grids, machine learning applications, vehicular ad hoc networks, and blockchain applications. He has also contributed to leading conferences including the International Conference on Communication, International Conference on Pattern Recognition, and International Conference on Wireless Communication.