Mitigating Over-Generalization in Anomalous Power Consumption Detection using Adversarial Training

SRINIDHI MADABHUSHI, University of Denver, USA
RINKU DEWRI, University of Denver, USA

Power consumption anomaly detection systems that use neural networks for prediction tasks are vulnerable to adversarial attacks, leading to unreliable performance and potential adverse effects on the power grid. Certain attack configurations can evade detection or trigger false alarms due to the neural networks' generalization tendencies, causing adaptation to attack values. This study investigates the effectiveness of adversarial training methods in improving detection performance against such attacks on power consumption data. Leveraging techniques like the Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), and Projected Gradient Descent (PGD), we assess the resilience of anomaly detection models. Through empirical experiments, we evaluate detection accuracy, adjustment capabilities, and prediction errors of these adversarially trained models across three datasets. Our results show significant improvements in detection performance, particularly in attack scenarios that normal prediction models would not detect. Additionally, we analyze the models' adaptability to anomalous data and quantify prediction errors, providing insights into their robustness and limitations. Integrating adversarial training techniques into anomaly detection models for power grids can reduce over-generalization to attack data, enhancing the detection of malicious demand manipulation attacks.

CCS Concepts: • **Computing methodologies** → **Neural networks**; • **Security and privacy** → **Intrusion detection systems**.

Additional Key Words and Phrases: Power consumption, Anomaly detection, Time series prediction, Neural networks, Adversarial attacks, Adversarial training

1 Introduction

In the era of smart grids and digitized power systems, timely detection of anomalous power consumption patterns is crucial for grid security, reliability, and efficiency. Anomalies can indicate events from equipment malfunctions to cyber attacks aimed at disrupting grid operations [26]. Traditional anomaly detection methods, based on statistical techniques or rule-based systems, often struggle with the dynamic nature of modern power grids. Machine learning techniques have emerged as promising tools for power consumption anomaly detection, leveraging historical patterns to identify deviations and enable proactive maintenance. However, these models are designed to generalize to unseen inputs—a behavior that, while usually beneficial, can be problematic. Without adversarial training, these prediction models often fail to detect certain obvious anomalies, such as a sudden 50 kW surge in power usage, and exhibit increased false alarm rates due to over-predicting consumption after such surges. This issue arises from the neural network's tendency to generalize to high-wattage inputs, referred to as over-generalization in this work, leading to unrealistic predictions exceeding 100 kW. This over-generalization compromises detection performance during attacks of varying intensities and becomes a vulnerability that adversaries can exploit to inject manipulated inputs that evade detection.

In this paper, we investigate the use of adversarial training to address the over-generalization behavior of prediction models in power consumption anomaly detection. We focus on adapting and integrating three adversarial attack generation methods—Fast Gradient Sign Method (FGSM) [15], Basic Iterative Method (BIM) [22], and Projected Gradient Descent (PGD) [27]—into the prediction model training process to enhance the model's ability to identify and mitigate anomalies in power consumption data. We employ three state-of-the-art neural network prediction models of increasing

Authors' Contact Information: Srinidhi Madabhushi, nidhi.madabhushi@du.edu, University of Denver, Denver, Colorado, USA; Rinku Dewri, rinku.dewri@du.edu, University of Denver, Denver, Colorado, USA.

complexity—Multi-Layer Perceptron (MLP), Long Short Term Memory (LSTM), and Convolutional Neural Network LSTM (CNN-LSTM)—to evaluate the impact of adversarial training on prediction responses. By customizing adversarial training, we aim to constrain the models' tendency to over-generalize to high-wattage inputs and improve detection performance under diverse attack scenarios. We summarize the contributions of this work as follows:

- We introduce a way to control the prediction model's generalization to new data in adversarial training methods. This significantly improves detection performance under various attack scenarios, enhancing the performance of a simple model like MLP by 70%, making it comparable to CNN-LSTM.
- We evaluate the performance of various adversarial training techniques with increasing complexity and find that while some prediction models benefit from improved performance, others may experience further deterioration.
- We observe that despite adversarial training, certain attack scenarios may still go undetected, indicating opportunities for stealthy attacks even with an adversarially robust model.

The subsequent sections of the paper are structured as follows: Sections 2, 3, 4 and 5 provides a comprehensive background on demand manipulation attacks, the utilization of anomaly detection models for their identification, related research, and the driving motivations behind our study. Section 6 elaborates on the methodology, describing the construction of attack profiles, and the changes we introduce in the traditional adversarial training process. In Section 7, we analyze the outcomes of adversarially trained models, offering valuable insights into their predictive responses. We highlight the key findings in Section 8 and present a discussion in Section 9, where we analyze performance trends across multiple datasets using additional metrics, evaluate ramp-based attacks, and explore the broader implications and limitations of this work. Lastly, Section 10 draws conclusions from our study and evaluates the efficacy of adversarially robust models for anomaly detection.

2 Threat Model

Demand manipulation attacks (MAD attacks) occur when an adversary manipulates the power grid's demand using consumer devices, typically in residential settings. These devices can be controlled directly by the attacker or by the consumer, often involving a botnet of devices to manipulate power demand rapidly [10]. Attackers can access high-wattage IoT devices in the same area and switch them on and off simultaneously, causing disruptive demand surges [37]. They can also send false messages to influence consumer behavior, like fake maintenance alerts during peak periods [32]. These attacks can cause frequency instability, line failures, and increased operating costs as ISOs are forced to purchase additional reserve power. We assume that an adversary has unauthorized access to IoT devices such as smart thermostats, water heaters, and electric vehicles, exploiting their vulnerabilities to simultaneously increase power usage. Historical attacks like the Mirai botnet, which compromised about 600,000 devices using weak passwords, illustrate this risk [3]. Controlling 600,000 smart thermostats, each with two 1 kW air conditioners, allows manipulation of up to 35 GW of power. Including other devices like electric water heaters (4.5 kW) and electric vehicles (7 kW), the controlled power could easily exceed 35 GW. Such coordinated surges could destabilize the grid, causing major operational, financial, and infrastructural impacts.

We assume a black-box threat model, where the attacker has no knowledge of the detection model's architecture, parameters, or internal behavior. Instead, the attacker crafts demand manipulation attacks by directly injecting high-wattage surges into the system through coordinated control of IoT devices. These injections, implemented as shift-based or ramp-based anomalies, are

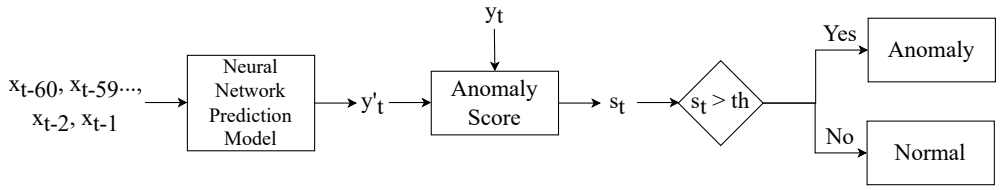

Fig. 1. Prediction-based anomaly detection process

generated without access to model gradients or training data and represent realistic attack patterns that exploit the model's generalization behavior indirectly.

3 Anomaly Detection Model

Anomaly detection in power consumption refers to identifying unusual patterns that deviate from expected energy usage behavior, which is particularly useful in detecting MAD attacks. It also enables timely interventions to enhance energy efficiency and reduce operational costs. Advanced methodologies, leveraging machine learning and deep learning, have been developed to automate and improve the accuracy of anomaly detection. This paper focuses on prediction-based anomaly detection systems for household power consumption data adopted from [46]. As depicted in Figure 1, the process involves three main stages: prediction, scoring, and thresholding. Historical power consumption values from the last 60 minutes ($x_{t-60}, x_{t-59}, ...x_{t-2}, x_{t-1}$) are input into a neural network prediction model, which forecasts future consumption (y'_t). The predicted (y'_t) and observed values (y_t) at time t are then compared, and an anomaly score (s_t) is computed to quantify the difference. This score (s_t) is evaluated against a predetermined threshold (th) to determine if a time instance is anomalous or not. Each of these components is described in detail in the following subsections.

3.1 Power Consumption Datasets

We use three datasets in this work, each offering different temporal resolutions and contextual settings to evaluate the robustness and generalizability of our approach. The first dataset, sourced from the UCI Machine Learning Repository [16], contains minute-level readings of household power consumption between 2007 and 2009. We focus on three attributes: date, time, and global active power. This dataset is primarily used for training, with data from 2007–2008 used for training and validation, and 2009 reserved for testing. Each input to the model comprises the previous 60 minutes of power consumption, forming a structured temporal sequence suitable for learning general consumption patterns. The second dataset [1] consists of one-minute interval power consumption data from a household in northeast Mexico, spanning the full year of 2023. Although it includes detailed weather information, we use only the active power readings (in kilowatts) for testing to assess performance across seasonal variations. The third dataset, obtained from the iFlex dynamic pricing experiment [17], features hourly electricity consumption data from Norwegian households exposed to time-varying electricity prices. We select a single household (participant_id = Exp_1) from Phase 1 of the study, covering the period from January 6 to March 20, 2020. Each test input uses the previous 60 hourly values to maintain consistency with the training setup. To handle missing values across all datasets, we apply linear interpolation, a method well-suited for time series data and consistent with our model's reliance on temporal continuity.

Table 1. Model architecture for MLP, LSTM and CNN-LSTM

Parameter	MLP	LSTM	CNN-LSTM
Convolution layer	n/a	n/a	(64, 2), (64, 2)
LSTM layer	n/a	32, 32, 64	128
Dense layer	100	n/a	32, 64
Dropout	n/a	0.07, 0.03	n/a
Number of epochs	30	20	30
Batch size	2048	2048	4096

3.2 Prediction Model Training

We selected three neural network models for time series prediction in an anomaly detection system, each increasing in complexity. First is the Multi-layer Perceptron (MLP) utilized by Chou et al., which belongs to the category of feedforward neural networks characterized by multiple layers with interconnected neurons [9]. The second model is the Long Short-Term Memory (LSTM) network, a type of recurrent neural network widely applied in time series prediction tasks. We employed a deep learning LSTM network proposed by Chahla et al. for power consumption prediction [8]. Lastly, we utilized a hybrid approach combining Convolutional Neural Network (CNN) and LSTM, known as CNN-LSTM, which has demonstrated improved performance compared to LSTM alone in power consumption prediction tasks [21].

For each model, we adopt the number of layers reported in prior literature and tune the remaining hyperparameters by varying baseline values—sourced from related studies—by powers of two. The hyperparameter search covers: number of nodes per layer (8–128), dropout rate (0 to 1), batch size (512–4096), and number of epochs (10–100 in steps of 10). Mean squared error on the validation set guides the selection, with the model yielding the lowest validation error chosen for further use.

Final architecture details including layer configurations, dropout rates, and training parameters—are summarized in Table 1. Layer order matches the model sequence, and CNN-LSTM includes two convolutional layers with specified filters and kernel sizes. All models take 60 input features (previous 60 minutes of consumption) and predict a single output value. We use mean squared error as the loss function, Adam as the optimizer, and TensorFlow as the backend.

These initial architectures are trained on clean power consumption data without attacks. Once baseline models are established, attack data is introduced for anomaly detection evaluation. These models are then improved using our customized adversarial training process, detailed in Section 6.2.

3.3 Anomaly Score

An anomaly score gives the extent to which the data point should be considered as an anomaly. We use the score (s_t) as adopted by Wang et al. [46] and is computed as shown in Equation 1, where $predicted_t$ and $observed_t$ are the prediction and observed values at time t and T represents all times until t, i.e. $\{1, 2, 3, ..., t-2, t-1\} \in T$.

$$s_t = \frac{|predicted_t - observed_t|}{avg_{i \in T}(|predicted_i - observed_i|)}, \tag{1}$$

3.4 Thresholding Mechanism

The threshold represents a value applied to an anomaly score, above which a data point is identified as an anomaly. To determine the most suitable threshold for our data, we employ a percentile-based approach, generating ten thresholds that vary according to percentile values. This method allows us

to explore thresholds that best align with the characteristics of our dataset. The chosen percentiles include 60%, 70%, 80%, 90%, 95%, 98%, 99%, 99.5%, 99.9%, and 99.999%, ranging from the strictest threshold to the most lenient. These percentiles are calculated based on the anomaly scores derived from the validation data. After evaluating the performance of different thresholds in terms of detection and false alarm rates, we select the 80th percentile threshold for MLP and CNN-LSTM (0.92 for both models), and the 70th percentile threshold for LSTM (0.74).

4 Related Work

The integration of machine learning in cyber-physical systems (CPS) and power grids has significantly advanced anomaly detection and forecasting capabilities, but it has also introduced new challenges, particularly regarding adversarial robustness. This section explores three key areas of related work: the application of machine learning techniques in anomaly detection within power grids, the vulnerabilities of these techniques to adversarial attacks, and the development of robust time-series models to mitigate such threats. Additionally, we provide a comparison with existing literature to position this work within the broader context of adversarial robustness research in CPS.

4.1 Anomaly Detection in Power Grids

Machine learning plays a crucial role in enhancing anomaly detection in power grids by employing a variety of techniques, including supervised, semi-supervised, and unsupervised methods, tailored to address the unique operational and security challenges faced by these critical infrastructures. These approaches are widely applied across different processes of the grid, such as monitoring voltage stability in transmission lines, identifying abnormal load patterns in distribution networks, and ensuring the integrity of smart meters against cyber attacks. Deep learning models like Long Short-Term Memory networks (LSTMs) and convolutional neural networks (CNNs) are particularly effective in handling time-series data, enabling the detection of anomalies such as false data injection attacks and demand manipulation with high precision [26]. Moreover, hybrid methods combining techniques like regression and neural networks enhance the granularity of anomaly detection by adapting to complex and non-linear behaviors in grid operations [9].

These machine learning models are not only pivotal in identifying anomalies but also in facilitating predictive maintenance, allowing utilities to anticipate and rectify potential failures before they escalate into severe outages. In communication networks, machine learning aids in identifying denial-of-service attacks, ensuring the stability of real-time grid control systems like SCADA and WAMS [44]. At a centralized level, anomaly detection integrates data from physical components and network layers, enabling the detection of coordinated and multifaceted attacks, which are increasingly prevalent in modern smart grids [26]. The importance of machine learning in power grids cannot be overstated as it ensures the resilience of the grid against both internal operational disruptions and external cyber-physical threats [47]. With increasing complexity and the proliferation of IoT devices, machine learning offers robust, adaptive, and scalable solutions to safeguard the reliability, efficiency, and security of power grids in an evolving threat landscape.

4.2 Vulnerabilities of Machine Learning in Cyber-physical Systems

While machine learning has significantly enhanced the detection of anomalies in cyber-physical systems (CPS), it also introduces new vulnerabilities. These arise from the susceptibility of ML models to adversarial attacks [18, 45], which can compromise their effectiveness in ensuring the security and reliability of CPS. Adversarial attacks exploit neural networks by introducing subtle, carefully crafted perturbations to input data, which remain undetectable to human observers or automated systems [23]. These perturbations manipulate the network's decision-making process,

resulting in incorrect or misleading outputs while maintaining the appearance of normality. This creates significant security risks, as adversarial examples can disrupt critical CPS operations or evade detection mechanisms [31]. To address this challenge, researchers emphasize the importance of retraining ML models with adversarial samples [35], highlighting the need for adaptive defense strategies to safeguard CPS against evolving cyber risks. Rashid et al. investigated the use of adversarial training methods to enhance the resilience of deep learning models deployed for cyberattack detection in IoT-driven smart cities, demonstrating improved detection accuracy and reduced vulnerabilities [11, 33]. Additionally, Mode and Hoque explored the susceptibility of Prognostics and Health Management (PHM) systems to adversarial attacks, revealing vulnerabilities in models based on LSTM, GRU, and CNN architectures [28]. A case study on Decentral Smart Grid Control (DSGC) systems demonstrated how adversarial examples, such as label-flipping attacks, can significantly degrade the accuracy of deep learning models classifying stable and unstable grid states, highlighting the profound impact of adversarial ML on energy CPS operations [30]. Similarly, studies have demonstrated the susceptibility of machine learning models in CPS to adversarial examples, including complex time-series anomaly detection and decision-making tasks, where such attacks exploit system vulnerabilities to disrupt operations, underscoring the urgency of robust adversarial defense mechanisms in these critical applications [24, 41, 50].

The challenge in measuring detection performance for power grid applications is the lack of anomalous data to train detection systems [6, 26]. Modern machine learning and deep learning approaches can generalize better with new data, but this adjustment may lead to the inability to detect attack instances, categorizing them as the new norm. Previous studies have shown that deep learning models may generalize too well, resulting in failure to identify anomalies [7, 14, 34]. Additionally, there is a lack of evaluation of a model's tolerance to adversarial inputs, leading to false negatives or undetected anomalies, and their impact on the target system. This gap allows certain undetected demand manipulation attacks to remain feasible in power consumption systems that utilize state-of-the-art neural networks as forecasting models, highlighting the importance of further research into adversarial robustness and its implications [25].

4.3 Adversarial Robustness of Time Series Models

Adversarially robust models for time-series prediction in the domain of power consumption data have become a focal point of recent research due to their critical role in anomaly detection systems. These forecasting models, including neural networks, have undergone rigorous adversarial training to enhance their resilience against attacks. Adversarial training incorporates adversarially perturbed data during the training process, improving the model's ability to withstand malicious inputs designed to induce errors in prediction output. Techniques such as the Fast Gradient Sign Method (FGSM)[15], Basic Iterative Method (BIM)[22], Projected Gradient Descent (PGD)[27], and DeepFool[29] exploit the model's gradient information to generate adversarial examples with minimal but effective input perturbations, challenging the model's accuracy without noticeable changes to human analysts.

Studies have consistently highlighted the vulnerability of time-series anomaly detection models to adversarial perturbations [4, 25, 49], demonstrating that even minimal perturbations can significantly impact state-of-the-art models. For instance, adversarial transformation networks and perturbation-based methods have shown how subtle yet targeted modifications can degrade the performance of classification and prediction models across various architectures [20, 48]. Furthermore, data-specific considerations, such as feature selection, dimensionality reduction, and statistical constraints, have been found to inadvertently increase model susceptibility to adversarial examples, emphasizing the need for robust data handling [2, 5]. Efforts to address these challenges through frameworks like TSA-STAT and the benchmarking of defense mechanisms,

such as regularization-based approaches, have underscored the potential of adversarial training to enhance resilience against both white-box and black-box attacks [5, 36]. This growing body of work underscores the critical importance of integrating adversarial robustness into the design of anomaly detection systems for power consumption, ensuring power grid reliability and security in the face of evolving adversarial threats.

4.4 Comparison with Existing Literature

While several studies have explored adversarial training for anomaly detection in cyber-physical systems, only a few have directly targeted power grid data, particularly in the context of demand manipulation attacks (MAD attacks). Among these, most approaches focus on perturbing the entire input sequence or rely on standard adversarial objectives without accounting for the model's tendency to over-generalize. In contrast, our work introduces a novel adversarial training strategy that specifically addresses the issue of over-generalization—where models adapt too readily to anomalous inputs, thereby increasing false negatives. We propose a targeted perturbation method that modifies only the last observed consumption value, aligning with real-world energy usage anomalies caused by abrupt IoT device activation. Additionally, our approach is evaluated across multiple neural architectures (MLP, LSTM, CNN-LSTM) and integrates multiple adversarial training methods (FGSM, BIM, PGD), demonstrating its robustness and adaptability. Table 2 provides a detailed comparison between our work and existing studies across several criteria, including model architecture, attack methods, and evaluation metrics.

To summarize, this work differs from existing literature in the following key ways:

(1) Focus on MAD attacks in household-level power consumption data, a threat model underexplored in adversarial training research.
(2) Emphasis on controlling model generalization to prevent adaptation to adversarial patterns, thereby improving anomaly detection and reducing false negatives.
(3) Introduction of a localized perturbation strategy that realistically mimics IoT-driven consumption anomalies.
(4) Use of realistic, domain-informed perturbation magnitudes (4–13 kW), instead of fixed epsilon values.
(5) Broad evaluation across multiple models, attack methods and power consumption datasets, highlighting the efficiency of the proposed approach.

5 Motivation

Before delving into adversarial training methods, we first evaluated the baseline performance of anomaly detection using three forecasting models: MLP, LSTM, and CNN-LSTM, all trained on unperturbed, original data. The average detection rate for MLP and LSTM was 55% and 51% respectively, with corresponding false alarm rates of 48% and 44%, highlighting limitations in their ability to reliably detect anomalies across varied attack scenarios. These findings revealed how each model responded to anomalous consumption patterns, laying the groundwork for interpreting the results of our subsequent experiments with adversarial training methods. These baseline challenges—particularly the relatively low detection rates and high false alarms—motivated our investigation into adversarial training as a means to enhance detection performance.

There are instances where the anomaly detection system fails to identify obvious anomalies in the data. For example, even when a household experiences a significant surge of 100 kW, indicating an anomalous event, the detection system may overlook it. This failure stems from the prediction model's tendency to adapt to anomalous input data, thereby generating predictions that align closely with these anomalies. Consequently, the prediction model begins to perceive such anomalies

Table 2. Comprehensive comparison of this work with related studies on adversarial robustness in cyber-physical systems

Criteria	This work (2025)	[43] (2024)	[40] (2023)	[42] (2022)	[19] (2021)
CPS application	Power grid utility	Distributed energy resources (DER)	Smart grid	Smart grid	Water treatment, distribution
Model Type	MLP, LSTM, CNN-LSTM	Gaussian-NB, Decision trees, and DNN	Autoencoder, CNN-RNN, Feedforward networks	CNN classifier	Rule-based RNN
Anomaly Detection Task	Power consumption demand manipulation	Network communication traffic	Electricity theft	Power quality disturbances	Industrial CPS sensor and actuator data
Motivation	Reduce over-generalization & improve detection accuracy	Address data imbalance for better detection	Enhance theft detection robustness & generalization	Improve PQ classification & transfer attacks	Detect stealthy attacks & improve resilience
Adversarial Attack	FGSM, BIM, PGD	FGSM	L-BFGS, BIM	DeepFool	FGSM
Mitigation	Adversarial training	Adversarial training	Ensemble defense	Adversarial training	Adversarial training

as the new normal, making them undetectable to the detection system. To address this issue, we aim to mitigate the model's tendency to excessively adjust to new data through adversarial training. By doing so, we seek to restrain the model's adaptation to anomalous input, thereby enhancing its ability to identify and flag such anomalies accurately.

When the prediction model's input contains a higher frequency of anomalous instances, the predicted values often mirror the behavior of these anomalies. Consequently, false alarms may be triggered even in the absence of a genuine attack, as the model tends to align with the prevailing anomalous patterns. This phenomenon results in significant discrepancies between the observed and predicted values. The proliferation of false alarms can lead to alarm fatigue among users, diminishing the effectiveness of the detection system. By subjecting the model to adversarial training, we aim to minimize its reliance on anomalous patterns, thereby reducing the occurrence of false alarms and enhancing the system's overall reliability. The baseline performance assessment served as a vital benchmark for evaluating the impact of adversarial training on model robustness and anomaly detection capabilities, which we discuss in detail in Section 7.

5.1 Over-Generalization in Neural Networks for Power Grid Applications

Over-generalization is a critical vulnerability in neural networks, where models learn overly broad representations that fail to distinguish between legitimate and malicious input patterns. In power grid forecasting, where the input is time-series data dominated by normal household

consumption, over-generalization leads models to incorrectly treat attack-induced anomalies (e.g., sudden consumption surges) as normal behavior [25]. This misinterpretation undermines anomaly detection systems and poses a serious threat in the context of demand manipulation attacks (MAD attacks). This issue arises because models such as autoencoders and sequence predictors are designed to learn generalized temporal patterns, but when trained exclusively on normal data, they often extend these patterns too broadly. Gao et al. show that even structurally abnormal sequences can be reconstructed with low error due to the model's strong generalization capacity [13]. Since these abnormal sequences often superficially resemble normal data, the model treats them as valid, masking potential anomalies. Song et al. observe similar behavior in predictive models, where both the encoder and decoder contribute to this failure by accurately processing unseen inputs that fall outside the training distribution [38]. To address this, Sun et al. introduce memory modules that anchor reconstructions to prototypical normal patterns, thereby amplifying reconstruction error for truly novel or attack-induced inputs [39]. Without such constraints, models tend to overfit to dominant trends and fail to flag deviations—especially when the anomalies are rare or artificially injected.

To counteract this form of over-generalization, our work introduces a customized adversarial training procedure with the following key properties:

(1) **Targeted Perturbation**: We perturb only the last observed input value in the time-series, which has the highest influence on the model's forecast. This contrasts with traditional adversarial training, which perturbs the entire input sequence.
(2) **Realistic Perturbation Magnitudes**: Perturbations are sampled between 4–13 kW to mimic real-world activation of high-wattage IoT devices, unlike the static or uniform ϵ used in typical adversarial examples.
(3) **Minimizing, Not Maximizing, Loss**: Rather than amplifying the prediction error (as in conventional adversarial training), we invert the gradient sign to minimize deviation from the expected (unperturbed) target, encouraging the model to resist anomalous inputs.
(4) **Enhanced Learning Signal via SSE**: We use Sum of Squared Errors (SSE) instead of MSE to give greater weight to each adversarial sample during training.
(5) **Low-Ratio, High-Impact Training**: With only 5% adversarial samples in the training set, the model gains resilience to high-frequency or high-magnitude anomalies, showing that carefully designed perturbations can guide the model away from overgeneralization even under realistic data scarcity.

6 Methodology

In this section, we present an overview of our methodology, covering the creation of attack profiles, the modified adversarial training process and the justification for perturbing only the last value.

6.1 Attack Profiles for Testing

An attack profile represents a dataset that consists of attack instances and is created by adding perturbations to the test dataset. We generate 2000 attack profiles using a perturbation model that randomly chooses the time instances to be injected with extra wattage representing an attack instance. The following are the steps for generating the attack profiles.

(1) For proportion p, randomly choose the list of indexes I (or timestamps) in the dataframe that will be treated as attack instances from the entire test dataset D.

$$I = random(n = length(D), r = length(D) * p) \tag{2}$$

where n is the total number of indexes in the dataset D, r is the number of attack instances to select from D and $random$ chooses r values from 0 to n.

(2) Add the perturbation ϵ to the chosen attack instances to create the attack profile D'.

$$D' = D[I] + \epsilon \tag{3}$$

(3) Repeat the above steps for each proportion $p \in \{0.05, 0.1, 0.15, \ldots, 0.9, 0.95, 1.0\}$. The value of ϵ varies depending on the type of attack being generated. We consider two types of injection attacks: shift and ramp. In the shift attack, ϵ remains constant across all selected attack instances and is chosen from the set $\{1, 2, 3, \ldots, 99, 100\}$. In contrast, for the ramp attack, a single ϵ value is selected from the set of 100 values, but the injection increases incrementally within each continuous attack window until it reaches the target injection of ϵ.

The rationale for constructing such attack profiles is to cover different types of attacks, including point and shift attacks occurring with different frequencies represented by the proportions. Lower proportions mimic point attacks, assuming that randomly chosen anomalous instances are spread across the one year of test data. As the proportions increase, the occurrences of these anomalies are closer to each other, resulting in more continuous and time window-based attacks. By covering a range of these proportions, we can address different types of attacks along with different intensities represented by the adders. For ramp attacks, this construction enables the simulation of gradual and progressive injections, where magnitude increases over time instead of appearing suddenly. Such profiles reflect more subtle adversarial behavior intended to evade detection by resembling typical consumption patterns. While our primary focus is on presenting results using shift attacks, we include a separate section on ramp attacks to demonstrate that detection performance remains comparable, while also highlighting key distinctions between the two attack types.

6.1.1 Attack Effect on Power Grid. The attack profiles are generated for a single household to enable edge detection of MAD attacks. When scaled to multiple households in the same geographical area, synchronized device switching can disrupt grid operations, causing line failures, increased operating costs, or even blackouts [37]. We perform three attack variations on the WSCC 9-bus system, mimicking different proportions of anomalous data: point anomalies, frequent occurrences, and profiles where all values are anomalous (i.e. $p = 1.0$). Using the PowerWorld simulator, we assess the frequency instability caused by each attack under a worst-case scenario of low inertia (inertia constants (H) for generators 2 and 3 is 5s and 10s respectively) and a 30% load increase as performed in [37]. Although the household dataset uses minute-level data and the simulator operates in seconds, we align the time units for consistency. Each attack increases the load by 30 MW, representing the load increase of a power grid (multiple households). Figure 2 shows the impact of three attacks, each initiated during stable grid operation. The point attack causes frequency disturbances but remains within the safe frequency range ($58.2Hz$ to $61.2Hz$ [12]). Increasing the frequency of the attack, as seen in the second variation, also allows the frequency to stabilize after disturbances but leads to higher operating costs. The third attack, with a sustained increase, causes a generator failure due to frequency dropping below $58.2Hz$. If high-frequency attacks go undetected, they can increase operating costs and therefore, detecting such attacks is crucial.

6.2 Adversarial Training of Prediction Models

To incorporate adversarial training into the selected prediction models, we utilize three prominent adversarial attack generation techniques: Fast Gradient Sign Method (FGSM) [15], Basic Iterative Method (BIM) [22], and Projected Gradient Descent (PGD) [27]. These techniques generate adversarial inputs, which are then used to train the prediction models. Through this process, the models become more adept at recognizing and mitigating the effects of adversarial inputs,

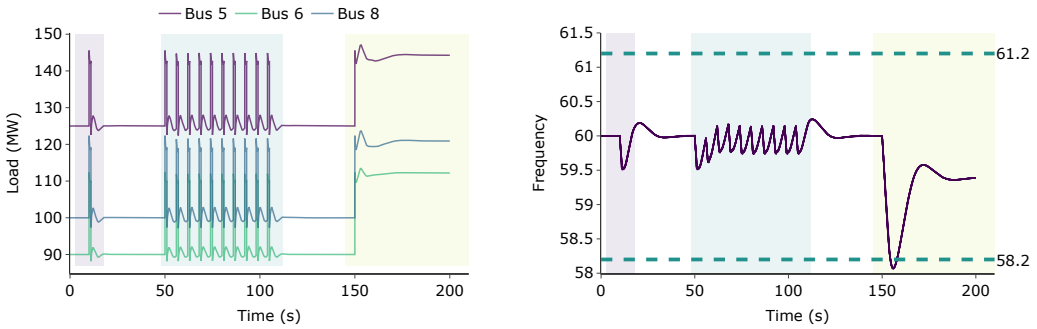

Fig. 2. Load increase performed using three types of attacks (left) and frequency response to each attack (right). Each attack is highlighted with a different color.

thereby improving their overall robustness and reliability. The perturbation added to the input data is randomly selected from a range between 4 kW and 13 kW, simulating the activation of a high-wattage household device [37]. We introduce three key adaptations to traditional adversarial training methods, tailored to the time-series power consumption prediction task:

(1) **Targeted Perturbation of the Input Sequence:** Instead of perturbing all time steps as in standard approaches, we modify only the final observed power consumption value in the input sequence, keeping the preceding 59 values unchanged. The input becomes $[x_{t-60}, x_{t-59}, \ldots, x_{t-2}, x'_{t-1}]$, where x' is the perturbed value. This localized strategy sharpens the model's response to sudden anomalies, limits overgeneralization, and enables simpler models like MLP to reach robustness comparable to CNN-LSTM.

(2) **Realistic and Interpretable Perturbation Magnitudes:** Rather than using a fixed epsilon, we inject domain-relevant perturbations (4–13 kW) that mimic high-wattage IoT device activation. Only 5% of the training data is perturbed to balance robustness and generalization, preventing overfitting. These adversarial examples are constructed by minimizing prediction error (unlike typical loss-maximizing strategies), with targets set to the true, unperturbed values—enhancing interpretability and real-world relevance.

(3) **Loss Function Adjustment and Sample Emphasis:** To emphasize learning from rare but impactful adversarial instances, we replace Mean Squared Error (MSE) with Sum of Squared Errors (SSE), which assigns greater weight to high-error samples. Even with only 5% adversarial examples, this helps the model generalize to both simulated attacks and real anomalies, without overfitting to adversarial data alone.

All three prediction models (MLP, LSTM, CNN-LSTM) are trained with FGSM, BIM, and PGD adversarial examples using these adaptations and are evaluated against their non-adversarial counterparts and each other. We have made the trained models, along with a sample usage script, available online[1].

6.2.1 Fast Gradient Sign Method (FGSM). The Fast Gradient Sign Method (FGSM) [15] is a straightforward yet potent adversarial attack technique commonly employed to generate adversarial examples for neural networks. It perturbs the input data by taking a small step in the direction of

[1]https://github.com/crisp-du/pgad/

the gradient of the loss function with respect to the input, while ensuring that the magnitude of the perturbation is bounded. Mathematically, the FGSM attack can be expressed as follows:

$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla_x J(x, y_{\text{true}})), \tag{4}$$

Here, x_{adv} represents the adversarial example, x denotes the original input, ϵ signifies the perturbation magnitude, $(\nabla_x J(x, y_{\text{true}}))$ denotes the gradient of the loss function J with respect to the input x, y_{true} denotes the true label associated with the input x, and $\text{sign}(\cdot)$ denotes the sign function, which extracts the sign of its argument.

In our context, we aim to construct an adversarial input that minimizes the loss compared to the observed consumption at time t, thus inverting the sign before ϵ. Therefore, we modify the FGSM equation such that only the most recent consumption (60^{th} input observation as shown in Algorithm 1) is perturbed:

$$x'_{t-1} = x_{t-1} - \epsilon \cdot \text{sign}(\nabla_X J(X, y_t)), \tag{5}$$

Here, X represents the entire input series consisting of consumption values in the last 60 minutes (from time $t-1$ to $t-60$). The steps carried out to apply FGSM to obtain the adversarial data for training is described in Algorithm 1.

6.2.2 Basic Iterative Method (BIM). The Basic Iterative Method (BIM) [22], also known as the iterative FGSM, is an iterative variant of the FGSM. It applies the FGSM multiple times with smaller step sizes, aiming to generate stronger adversarial examples. We modify the equation for BIM with the goal of decreasing the loss compared to the observed consumption.

$$x'_{t-1,i+1} = \text{clip}_\epsilon(x'_{t-1,i} - \alpha \cdot \text{sign}(\nabla_X J(X'_i, y_t))), \tag{6}$$

Here, $x'_{t-1,i}$ represents the adversarial example at iteration i for the most recent time stamp $t-1$, α denotes the step size of each iteration, and $\text{clip}_\epsilon(.)$ denotes the function that clips the values of the argument to be within a range of ϵ around the original input.

6.2.3 Projected Gradient Descent (PGD). The Projected Gradient Descent (PGD) is an iterative optimization method used to generate adversarial examples, and it is an extension of the BIM technique. PGD iteratively applies small perturbations to the input data, and the resulting adversarial example is projected back into the ball of the original input using a projection function. Similar to FGSM and BIM, the goal of PGD is to minimize the loss compared to the observed consumption and therefore, we modify it as follows.

$$x'_{t-1,i+1} = \text{proj}_\epsilon(x'_{t-1,i} - \alpha \cdot \text{sign}(\nabla_X J(X'_i, y_t))), \tag{7}$$

In this equation, $x'_{t-1,i}$ represents the adversarial example at iteration i for the most recent time stamp $t-1$, α denotes the step size of each iteration, and $\text{proj}_\epsilon(.)$ represents the projection function that constrains the adversarial example within a specified ϵ-neighborhood of the original input.

6.3 Justification for Perturbing the Last Observed Value

Building on the three key adaptations introduced in Section 6.2, this section details and justifies our approach of applying localized perturbations—specifically to the last input value—during adversarial training for time-series prediction models.

Traditional adversarial training often perturbs the entire input sequence, which can lead the model to internalize adversarial patterns as normal, reducing detection performance. In contrast, our method perturbs only the final input value in the sequence, leveraging its stronger influence on the model's output due to the temporal nature of forecasting. This localized perturbation helps

Algorithm 1 Fast Gradient Sign Method (FGSM)

function FGSM($model, x, y, chosen\_inds, \epsilon$)
  using GradientTape context manager to record operations:
    $predictions \leftarrow model(x, training = True)$
    $loss \leftarrow$ sum_of_squared_error$(y, predictions)$
  $gradients \leftarrow$ compute gradients using $(loss, x)$
  $signs \leftarrow$ sign$(gradients[$chosen_inds$, 60])$
  $x\_adv \leftarrow x$
  $x_adv[$chosen_inds$, 60] \leftarrow x_adv[$chosen_inds$, 60] - \epsilon * signs$
  set any negative values in x_adv to 0
  return x_adv
end function

the model interpret such deviations as anomalies rather than adapting to them, thereby preserving its ability to detect unexpected consumption patterns. This strategy also reflects the nature of real-world anomalies, which typically manifest as abrupt spikes or drops in power consumption—such as from switching high-wattage IoT devices (e.g., water heaters, air conditioners, EV chargers) on or off—rather than gradual shifts. By modeling perturbations ranging from 4 kW to 13 kW, we enhance both the realism and interpretability of the training process. Additionally, preserving earlier time steps ensures that the adversarial signal is temporally localized, maintaining sequence integrity and reducing the risk of false negatives.

From a computational standpoint, limiting perturbations to a single time step reduces training overhead and enables even simple models like MLP to achieve robustness comparable to more complex architectures like CNN-LSTM. Training with just 5% adversarial samples using this method proves sufficient for resilience against stronger attacks. Finally, we replace the Mean Squared Error (MSE) loss with the Sum of Squared Errors (SSE) to further emphasize adversarial samples during training, making the optimization process more focused. As demonstrated in Section 7, this approach improves detection of both high-frequency and subtle anomalies without increasing false alarm rates—crucial for robust anomaly detection in smart grid systems.

7 Results

In this section, we delve into the results of adversarially trained models, examining their performance, adjustment behavior, and explore adversarial training results when predetermined perturbations are used.

7.1 Detection and False Alarm Rates

We use two metrics, detection rate (DR) and false alarm rate (FAR), to evaluate the performance of the anomaly detection system before and after the adversarial training. Detection rate measures the proportion of anomalous instances that are correctly identified and it is also known as the true positive rate or recall. False alarm rate measures the proportion of normal instances that are incorrectly classified as anomalies and it is also known as the false positive rate. To evaluate the performance of these metrics, we use the attack profiles generated using the perturbation model described in Section 6.1

When the prediction model is trained on normal data without adversarial training, certain attack profiles evade detection by the anomaly detection system. Specifically for MLP, as the frequency of attacks increases, the detection and false alarm rates deteriorate. The results for the original MLP model are depicted in the top two plots in Figure 3. Similarly, there are problematic areas for

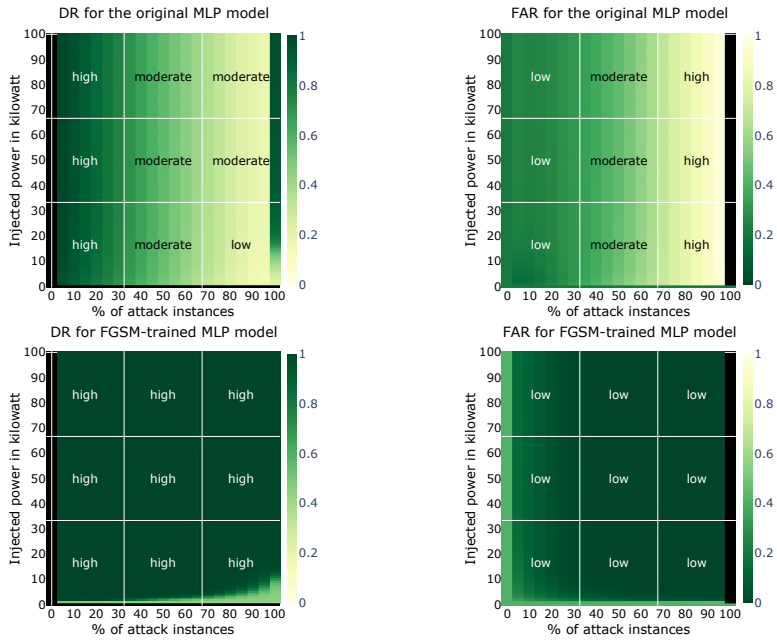

Fig. 3. Comparison of detection and false alarm rates between original and FGSM trained MLP models

LSTM and CNN-LSTM as well. Both exhibit poor performance in detection and false alarm rates for high-frequency and low-intensity attacks, i.e., the bottom-right corner of the heat map plots, which represent high-frequency attacks ranging from 70% to 100% and low-wattage injections ranging from 0 kW to 20 kW.

However, when all three models undergo adversarial training, the detection and false alarm rates across different attack profiles improve significantly. As shown in Figure 3, MLP exhibits a substantial increase in detection rates as well as a significant drop in false alarm rates across various attack profiles, indicating better performance with adversarial training. This detailed performance improvement for MLP across attack profiles is reflected in the overall performance, as seen in Figure 4, particularly when BIM and PGD are used. LSTM also shows improvement when FGSM is used, as depicted in Figure 4. However, when LSTM is adversarially trained using BIM or PGD, it results in a drop in the performance of the detection rates. The types of attack profiles that experience performance degradation when using BIM are those with higher wattage injections (top-right corner of the heat map plots). Conversely, for PGD, if there is a mixture of normal and anomalous instances, the detection model struggles to correctly recognize the attack instances. There is a notable improvement in false alarm rates for BIM, while false alarm rates for PGD also improve, except in cases of high-frequency attacks with high wattage injections. As for CNN-LSTM, adversarial training has not significantly improved its detection performance, as the original CNN-LSTM itself achieves a high detection rate, leaving less room for improvement.

A prevalent issue with the performance of the original models, lacking adversarial training, arises in the case of high-frequency and low-intensity attacks (bottom-right corner of the heat map). Even the best-performing model, CNN-LSTM, struggles to detect this category due to the challenge of distinguishing small injections from normal increases in consumption. Following the adversarial training process, as observed in Table 3, improvements have been observed in this area

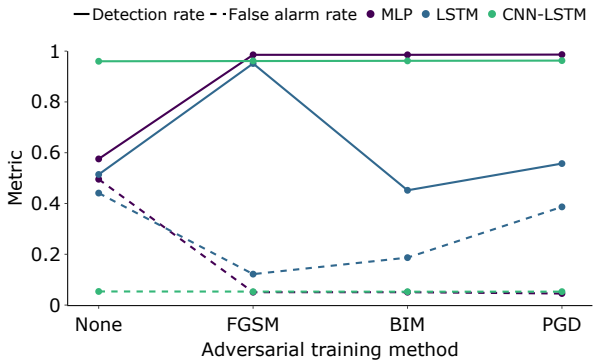

Fig. 4. Comparison of detection and false alarm rates across different adversarial training methods of increasing complexity for MLP, LSTM and CNN-LSTM models

Table 3. Detection rate (DR) and false alarm rate (FAR) for high frequency (>70%) and low wattage injection (<20 kW) attacks by prediction model and training type

Training Type	MLP		LSTM		CNN-LSTM	
	DR	FAR	DR	FAR	DR	FAR
Original	0.2843	0.8514	0.500	0.5224	0.6134	0.3588
FGSM	0.8918	0.0328	0.9419	0.0490	0.6312	0.4142
BIM	0.8911	0.0313	0.9285	0.0448	0.6300	0.4127
PGD	0.8983	0.0243	0.6913	0.4973	0.6620	0.3618

for MLP and LSTM, leading to a simple model like MLP to outperform a model like CNN-LSTM. Although there is a slight improvement in detection rates for CNN-LSTM, it results in an increased false alarm rate.

7.2 Model Adaptation to Attack Instances

When utilizing the original prediction models without adversarial training, we encounter a notable challenge in the anomaly detection process. Specifically, the performance of MLP and LSTM models is affected by their tendency to generalize excessively to unfamiliar data. Consequently, even in scenarios where an unrealistic amount of wattage is injected from a single household, the detection model fails to identify this anomaly. This is attributed to the prediction model adjusting to the perturbed input and generating predictions that align with the unrealistic input. Moreover, as the frequency of such injections increases, the prediction model begins to perceive them as the new norm, exacerbating the issue. While it is desirable for the prediction model to adapt to certain changes, such as seasonal fluctuations, accommodating unrealistic values presents a challenge. Achieving a balance between generalization and overfitting properties is crucial for effectively leveraging these complex time series models.

In addressing the issue of over-generalization as explained in Section 5.1, we examine how adversarial training influences the prediction model's response. Figure 5 illustrates the prediction responses of the original prediction model and the adversarially trained MLP models. Notably, a 12 kW injection occurs from 3:00 AM to 6:00 AM, representing a prolonged shift attack. We observe that the initially trained MLP model swiftly adjusts to this surge, albeit taking some time to stabilize. Consequently, the difference between the predicted and observed responses is low, resulting in a

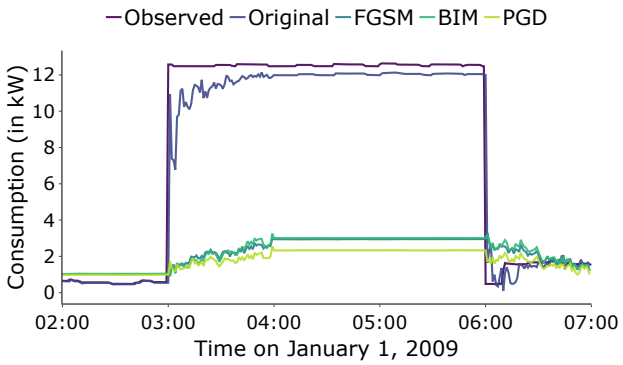

Fig. 5. Comparison of the prediction responses of the original and adversarially trained MLP models to a prolonged shift attack, injecting 12 kW, indicative of a high-wattage device, over a three-hour duration

low anomaly score. Thus, the threshold fails to flag this attack window as anomalous. However, after subjecting the MLP model to adversarial training using FGSM, BIM, and PGD, we observe resistance to immediate scaling to such high values. This behavior enables us to discern the gap between the predicted and observed consumption, facilitating the identification of this window as anomalous.

7.2.1 Prediction Range Comparison. To investigate the factors that contribute to the restriction in adjusting to larger injections and to determine if this behavior is consistent across different wattage injections, we analyze the prediction value range of both the original and adversarially trained models for MLP, LSTM, and CNN-LSTM. For MLP, we observe that the original model scales the values in response to increasing injected wattage ranging from 1 kW to 50 kW in the input data. The maximum prediction reaches up to 50 kW, aligning with the observed injections in the input data. Similarly, the original LSTM model exhibits a comparable behavior, although it scales less aggressively than MLP, reaching a maximum of approximately 14 kW. In contrast, CNN-LSTM does not exhibit significant adjustment without adversarial training, owing to CNN's inherent property of translation invariance. It reaches a maximum of around 8 kW when subjected to translations in the input. Upon subjecting all three models to adversarial training, we observe a compression in the prediction range to under 5 kW for both MLP and LSTM, while the range remains unchanged for CNN-LSTM. Adversarial training effectively constrains the prediction range, preventing over-generalization of the models to the input data.

7.2.2 Prediction Error Comparison. The prediction errors undergo a notable reduction for MLP and LSTM models. The root mean squared errors (RMSE) of predictions made by the original and adversarially trained models are presented individually in Table 4. Specifically, the MLP model exhibits an RMSE of 20.42 on attack data, reflecting its adjustment to anomalous instances. Following adversarial training, the RMSE experiences a substantial decrease to approximately 1, with BIM yielding the lowest error. Similarly, LSTM also undergoes a significant reduction in RMSE after adversarial training. However, for CNN-LSTM, there is no considerable reduction observed, with the RMSE remaining around the same value as that of the original model. Figure 6 depicts the box plot illustrating the differences between the predicted and observed values generated by both the original and adversarially trained models. After subjecting MLP and LSTM to adversarial training, we observe the distribution of differences being compressed closer to zero, indicating an improvement in prediction accuracy. However, CNN-LSTM has no significant changes in the

Table 4. Prediction error (RMSE) of the original and adversarially trained models on attack data

Model	Original	FGSM	BIM	PGD
MLP	20.42	0.9123	0.9121	1.1497
LSTM	8.0339	1.2076	0.8658	0.9630
CNN-LSTM	0.3285	0.3318	0.3316	0.3366

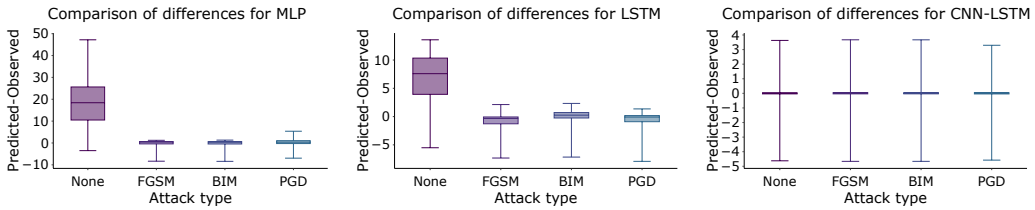

Fig. 6. Comparison of the differences between predicted and observed values between the different adversarial attack techniques

distribution of these differences. An interesting observation across all models is the tendency to under-predict after adversarial training. This suggests a preference for widening the gap between the observed and predicted values, facilitating the detection of certain attacks. However, this behavior introduces a new challenge of managing false positives, particularly when surges in consumption occur due to genuine customer behavior.

7.3 Adversarial Training using Predetermined Perturbations

Neural networks like LSTM and CNN-LSTM excel in time series prediction due to their ability to leverage historical data effectively. However, their computational complexity poses challenges, particularly in scenarios requiring real-time or edge computing applications where computational resources are limited. Adversarial training further compounds this challenge by increasing training and prediction times. As we transition towards edge computing, the demand for lightweight detection models grows, particularly in domains like power consumption monitoring, where smart meters operate autonomously within households. However, achieving adversarial robustness in such models often comes at the cost of increased computational overhead. To address this, we investigate into the efficacy of training models on predetermined random perturbations, contrasting it with the traditional approach of computing gradients during training. This hypothesis aims to shed light on how such an approach impacts detection performance compared to adversarially trained models.

7.3.1 Introducing Random Perturbations. To validate our hypothesis, we employ a methodology where we introduce random perturbations ranging from 4 kW to 13 kW to the last observation of the input data. This perturbation, applied to a selected proportion (5%) of the training dataset, creates a mixture of normal and adversarial instances. By training the model on this modified dataset, we eliminate the need for gradient monitoring during training, thereby accelerating the process. Once trained, we utilize the resulting prediction model for anomaly detection, evaluating its performance against traditional models trained with gradient-based methods. This streamlined approach aims to provide insights into the effectiveness of using predetermined perturbations for model training, particularly in scenarios where computational resources are constrained.

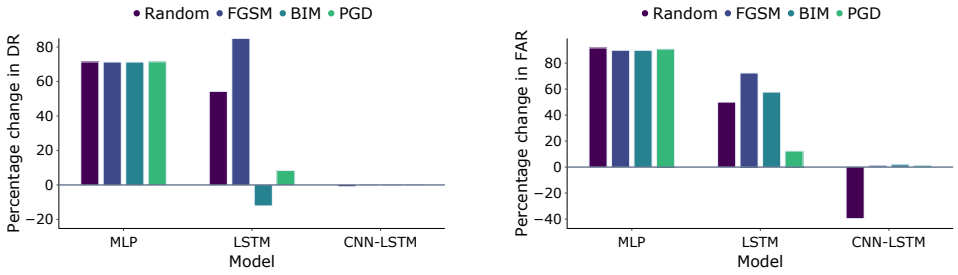

Fig. 7. Percentage improvement of the adversarially trained models over the original prediction model

7.3.2 Performance Improvement. We employ the same set of test attack profiles to assess the performance of the model trained with random perturbations. Figure 7 illustrates the percentage improvement achieved by different adversarially trained models, including the one trained with random perturbations, over the original prediction model that lacked adversarial training. Our analysis reveals noteworthy enhancements in the detection rate of MLP with the random perturbation model, demonstrating competitive improvements in both detection and false alarm rates compared to traditional adversarially trained models. Similarly, the random perturbation model exhibits improvements with LSTM, albeit to a lesser extent compared to FGSM, which shows the highest improvement. These findings seem to align with random perturbations as an alternative training strategy, particularly in scenarios where traditional adversarial training may not yield significant enhancements. Interestingly, for CNN-LSTM, we observe a deterioration in performance when utilizing the random perturbation model, in contrast to the marginal improvements observed with other adversarially trained models.

While the observed improvements with the random perturbation model are intriguing, we delve deeper into its performance across the 2000 individual attack profiles to gain a comprehensive understanding. Upon closer examination, we find a consistent and significant improvement in performance across a majority of attack profiles. The detection rates remain consistently high, while the false alarm rates range from low to moderate for most profiles. However, certain areas present challenges for this model. For instance, the LSTM model struggles to detect high-frequency and high-intensity attacks, which are particularly aggressive. Surprisingly, even a 100 kW injection in approximately 80% of the data goes undetected, highlighting a notable limitation of the model. Similarly, CNN-LSTM encounters difficulties with attack profiles characterized by high frequency and low intensity, where it fails to detect obvious deviations in consumption patterns. These findings underscore the nuanced performance of the random perturbation model, demonstrating its possible effectiveness across a range of scenarios while also revealing areas that warrant further investigation and refinement.

8 Key Findings

In this section, we summarize the key findings from the results discussed in the previous section.

1. Increasing complexity of adversarial training does not guarantee better performance. This observation is evident in the case of the LSTM model that showed poorer performance with BIM and PGD compared to FGSM. Similarly, adversarial training did not consistently enhance performance, as seen with the LSTM model under BIM and PGD and the CNN-LSTM model under FGSM.

2. Simpler models can match the performance of a complex model through adversarial training. Without adversarial training, models struggled with detecting high-frequency and low-intensity

attacks. Through adversarial training, MLP showed significant improvement (DR: 89%, FAR: 30%) compared to CNN-LSTM (DR: 63%, FAR: 40%), highlighting the benefits of adversarial training on simpler models.

3. Controlling the adjustment property of a prediction model has a direct impact on the detection model's performance. The original models performed poorly without adversarial training due to their generalization properties, which caused them to overlook attack instances. Adversarial training aimed at limiting this adjustment improved anomaly detection by reducing prediction errors, constraining prediction ranges, and enhancing model robustness to input perturbations.

4. Effectiveness of an adversarial training method varies with prediction model's architecture. MLP and CNN-LSTM performed better with increased complexity from FGSM to PGD, while LSTM performed better with FGSM and worse with BIM. MLP also performed well with random perturbations, unlike CNN-LSTM, whose performance deteriorated compared to traditional adversarial techniques.

9 Discussion

In this section, we extend our evaluation to include additional datasets, a wider range of performance metrics, and ramp-based attacks to assess the consistency and generalizability of our approach. These analyses offer deeper insight into how adversarial training performs across varied conditions and highlight areas for further refinement. We also discuss the implications and limitations.

9.1 Assessment Across Multiple Datasets

To evaluate generalizability beyond the UCI dataset (Dataset 1 [16]), we assess model performance on two additional real-world datasets—Mexico (Dataset 2 [1]) and iFlex (Dataset 3 [17])—which vary in sampling rates, household behavior, and regional conditions. Detection rate (DR) and false alarm rate (FAR), averaged over all attack profiles, are summarized in Table 5.

Consistent with Dataset 1, adversarial training significantly improves MLP performance on Datasets 2 and 3, boosting detection rates while reducing false alarms. CNN-LSTM continues to exhibit strong performance even without adversarial training, showing minimal room for improvement. LSTM mirrors earlier findings where adversarial training occasionally degraded performance—particularly under BIM and PGD training. Similarly, while CNN-LSTM retains high detection rate, adversarial training slightly increases FAR in some cases, as previously observed.

Overall, performance trends are highly consistent across datasets, especially in terms of which models benefit from adversarial training. To statistically validate this, we apply the Kruskal–Wallis test on the AUC scores from Table 5. We choose the Kruskal–Wallis test as it is a non-parametric method suitable for comparing multiple independent groups without assuming normal distribution of the data. The result confirms no significant differences in performance across datasets ($H = 2.06$, $p = 0.36$), suggesting that adversarially trained models maintain robust and comparable detection performance across varied data conditions.

Table 5. Performance metrics under shift attacks for each model and training configuration across the three datasets (D1: UCI, D2: Mexico, D3: iFlex).

Model-Training	Precision			F1-score			Recall/Detection Rate			False Alarm Rate			AUC		
	D1	D2	D3	D1	D2	D3	D1	D2	D3	D1	D2	D3	D1	D2	D3
MLP-Original	0.524	0.524	0.525	0.507	0.51	0.501	0.553	0.56	0.541	0.484	0.485	0.498	0.559	0.557	0.553
MLP-FGSM	**0.854**	0.839	0.873	0.927	0.92	0.947	**0.986**	0.99	0.995	0.05	0.079	**0.048**	0.978	0.969	**0.984**
MLP-BIM	0.851	0.839	0.873	0.926	0.92	0.947	0.985	0.99	0.995	0.051	0.079	**0.048**	0.978	0.969	**0.984**
MLP-PGD	**0.854**	**0.842**	**0.875**	**0.928**	**0.922**	**0.949**	**0.986**	**0.991**	**0.996**	**0.046**	**0.076**	**0.048**	**0.98**	**0.972**	**0.984**
LSTM-Original	0.506	0.507	0.502	0.466	0.47	0.461	0.514	0.519	0.512	0.441	0.442	0.463	0.541	0.543	0.53
LSTM-FGSM	**0.785**	**0.787**	**0.791**	**0.869**	**0.874**	**0.873**	**0.951**	**0.953**	**0.954**	**0.122**	**0.126**	**0.146**	**0.938**	**0.936**	**0.943**
LSTM-BIM	0.586	0.594	0.604	0.472	0.482	0.514	0.452	0.457	0.508	0.187	0.19	0.22	0.637	0.638	0.652
LSTM-PGD	0.556	0.59	0.536	0.546	0.58	0.552	0.557	0.581	0.59	0.387	0.368	0.478	0.6	0.623	0.575
CNN-Original	0.879	0.873	0.894	0.935	0.933	0.952	**0.964**	**0.967**	**0.973**	0.059	0.069	**0.055**	0.969	0.966	**0.975**
CNN-FGSM	**0.91**	**0.914**	**0.905**	**0.956**	0.961	**0.956**	0.961	0.966	0.967	**0.053**	**0.063**	0.059	0.971	**0.969**	0.972
CNN-BIM	**0.91**	**0.914**	**0.905**	**0.956**	0.961	**0.956**	0.962	0.966	0.967	**0.053**	**0.063**	0.059	0.971	**0.969**	0.972
CNN-PGD	0.908	0.913	0.903	**0.956**	**0.962**	**0.956**	0.963	**0.967**	0.967	**0.053**	0.064	0.06	**0.973**	**0.969**	0.972

Table 6. Performance metrics under ramp attacks for each model and training configuration across the three datasets (D1: UCI, D2: Mexico, D3: iFlex).

Model-Training	Precision			F1-score			Recall/Detection Rate			False Alarm Rate			AUC		
	D1	D2	D3	D1	D2	D3	D1	D2	D3	D1	D2	D3	D1	D2	D3
MLP-Original	0.482	0.488	0.494	0.459	0.44	0.444	0.515	0.504	0.507	0.492	0.492	0.508	0.544	0.537	0.535
MLP-FGSM	0.844	0.836	0.876	0.808	0.756	0.805	0.796	0.778	0.805	0.055	0.088	0.051	0.917	0.896	0.94
MLP-BIM	0.844	0.836	0.876	0.808	0.756	0.805	0.796	0.778	0.805	0.055	0.088	0.051	0.917	0.896	0.94
MLP-PGD	**0.848**	**0.839**	**0.879**	**0.822**	**0.769**	**0.815**	**0.814**	**0.796**	**0.82**	**0.05**	**0.085**	**0.05**	**0.934**	**0.915**	**0.951**
LSTM-Original	0.499	0.506	0.501	0.472	0.455	0.449	0.524	0.514	0.518	0.455	0.457	0.478	0.538	0.533	0.524
LSTM-FGSM	**0.775**	**0.782**	**0.791**	**0.812**	**0.776**	**0.787**	**0.86**	**0.844**	**0.878**	**0.145**	**0.154**	**0.161**	**0.893**	**0.88**	**0.904**
LSTM-BIM	0.601	0.619	0.616	0.474	0.465	0.494	0.44	0.447	0.515	0.204	0.207	0.241	0.628	0.631	0.65
LSTM-PGD	0.566	0.609	0.548	0.526	0.535	0.519	0.524	0.542	0.589	0.33	0.31	0.443	0.613	0.635	0.592
CNN-Original	0.866	0.866	0.893	0.83	0.782	0.814	0.798	0.761	0.791	0.079	0.091	0.067	0.914	0.89	**0.925**
CNN-FGSM	**0.9**	**0.909**	**0.905**	**0.862**	**0.821**	**0.825**	**0.813**	**0.783**	0.796	**0.075**	0.087	0.075	**0.925**	**0.907**	0.923
CNN-BIM	**0.9**	**0.909**	**0.905**	**0.862**	**0.821**	**0.825**	**0.813**	**0.783**	0.796	**0.075**	0.087	0.075	**0.925**	**0.907**	0.923
CNN-PGD	0.896	0.905	0.903	0.859	0.817	0.824	0.81	0.78	**0.797**	**0.075**	0.09	**0.074**	0.924	0.902	0.924

Table 7. Average precision and Wilcoxon signed-rank p-values across 11 quarterly models. Statistically significant differences ($p < 0.05$) are marked with *.

Model	Average Precision	Wilcoxon p-value (vs. Original)
MLP (Original)	0.4825	–
MLP (FGSM)	0.5992	0.00098*
MLP (BIM)	0.5992	0.00098*
MLP (PGD)	0.5361	0.00293*
LSTM (Original)	0.5888	–
LSTM (FGSM)	0.7673	0.00293*
LSTM (BIM)	0.7502	0.00488*
LSTM (PGD)	0.6537	0.24023
CNN-LSTM (Original)	0.8923	–
CNN-LSTM (FGSM/BIM/PGD)	0.8913	0.76465

9.2 Performance Evaluation Using Additional Metrics

Beyond detection rate (DR) and false alarm rate (FAR), we report additional evaluation metrics—precision, F1-score, and AUC in Table 5 to better assess the trade-offs between true and false positives in anomaly detection. These findings reinforce earlier observations that adversarial training enhances performance—especially for simpler models—without degrading precision. Additionally, we perform a sliding window evaluation across 11 sequential models to verify the consistency of improvements from adversarial training. Each model is trained on a three-month interval and tested on the following month using the UCI dataset (Dataset 1) spanning three years. We calculate precision on each test set and compare the distributions between the original and adversarially trained models using the Wilcoxon signed-rank test—a paired, non-parametric test suited for time-evolving metrics. Table 7 presents the average precision values across the eleven models, along with p-values indicating statistically significant improvements for all adversarially trained models compared to the original. For MLP, improvements are observed across all adversarial training methods. Similarly, LSTM shows consistent gains except in the case of the PGD-trained model, which does not show a statistically significant improvement. For CNN-LSTM, changes are minimal, aligning with its already strong baseline performance. These findings align with the performance trends previously observed when using the entire dataset.

9.3 Robustness Against Ramp-Based Attacks

To evaluate robustness against more gradual, time-continuous anomalies, we assess model performance under ramp attacks. Unlike shift attacks that add a constant perturbation, ramp attacks introduce incremental additions across a continuous window, simulating stealthier, more realistic manipulation scenarios. Table 6 shows precision, F1-score, recall, false alarm rate and AUC under ramp attacks for all models and training configurations across the three datasets.

Adversarial training improves detection performance compared to the original models in most cases. However, when comparing ramp performance to the corresponding shift results (in Table 5), we observe noticeable drops in recall and F1-score across most models, especially under adversarial training. Precision, in contrast, remains relatively stable, indicating the models are missing more anomalies but not increasing false positives.

MLP exhibits the largest overall degradation under ramp attacks, with MLP-FGSM showing F1-score drops ranging from 0.1189 to 0.1644 across datasets. CNN-LSTM also experiences significant

Table 8. Recall difference (Shift − Ramp) across proportion ranges for selected cases

Model/Training/Data	0–30%	35–75%	80–100%
CNN-Original-D1	0.01	0.16	0.30
LSTM-FGSM-D2	0.01	0.08	0.18
MLP-BIM-D3	0.02	0.17	0.31
MLP-PGD-D2	0.02	0.19	0.33

recall and F1 declines, though its precision remains largely unaffected. LSTM models show mixed results—original versions remain stable, but adversarial variants vary in effectiveness.

To understand why ramp attacks reduce recall, we analyze selected model–dataset combinations across three injection proportion ranges (0–30%, 35–75%, and 80–100%). Table 8 shows that recall differences are minimal at low frequencies ($\leq$0.02) but increase substantially with higher injection proportions, reaching up to 0.33. This is due to smaller stepwise perturbations at higher frequencies, which are harder for the model to distinguish from normal fluctuations. These findings highlight the challenge of detecting high-frequency ramp attacks and emphasize the need to evaluate detection robustness not only by attack type, but also by injection intensity and frequency.

9.4 Implication and Limitations

By fortifying the prediction models against adversarial attacks, the overall security of anomaly detection systems in power grids can be significantly improved. This ensures that the systems are more resilient to demand manipulation attempts by malicious actors, thereby safeguarding critical infrastructure. Before adversarial training, models had difficulty in detecting low wattage and more frequent attacks (types 2 and 3 in Figure 2) that either lead to increased operating costs or even a generator failure due to frequency instability. By limiting the models from adjusting to attack data, we are able to increase the gap (or the difference) between the predicted and observed consumption values when there is an ongoing manipulation attack. This gap is then eventually captured by the anomaly score, leading to higher scores that are above the threshold, thus aiding in detecting such high frequency attacks. The false alarm rates are also significantly reduced because of the adjusting property being handled during adversarial training. Perturbing only the last input time step enables detection of various types of attacks, ranging from point to more continuous forms.

9.4.1 Efficiency in Model Development. For a simple neural network model like MLP, we observe a significant improvement in the overall detection rate across different profiles by about 70% and a reduction in the false alarm rate by about 90% for all adversarial training methods. Similarly, we observe an improvement in the LSTM model when using the FGSM training method. However, the adversarial training method demands more time and resources during the prediction and training processes. The similarity in performance between randomly trained and adversarially trained models suggests that complex adversarial training techniques may not always be necessary to achieve robust anomaly detection, especially for simpler neural network models. This makes anomaly detection more accessible to a wide range of applications, including edge detection, which demands cost-effective solutions. Table 9 presents the memory usage and prediction time for each model. MLP is the most efficient, using the least memory and delivering predictions 50% and 30% faster than LSTM and CNN-LSTM, respectively, in batch mode. Its lightweight nature makes it ideal for real-time, edge computing applications.

9.4.2 Limitations. While adversarial training improves resilience against many attack profiles, it is still limited in its ability to adaptively steer the model's internal behavior. The process is

Table 9. Run-time memory of a loaded model and time for a single prediction

Model	Memory when loaded	Prediction time
MLP	0.067 MB	0.014 s
LSTM	9.909 MB	0.017 s
CNN-LSTM	12.240 MB	0.019 s

externally driven focusing on perturbing inputs rather than being guided by a more controlled adjustment of the model's decision boundaries or temporal reasoning strategies. Both ramp and shift attacks at high frequencies with small magnitude injections remain especially challenging–their subtlety often makes them nearly indistinguishable from normal fluctuations in power usage. This can lead to a consistent under-detection of such attacks, especially in models that rely heavily on local deviations. To better handle these scenarios, future models should be designed to reason over temporal windows more globally, capturing cumulative changes rather than reacting only to instantaneous shifts. For example, a model could learn to assess whether the total increase across a short time span exceeds a meaningful threshold, even if individual steps are minor. In addition, developing simulation pipelines or generative models to produce more realistic attack sequences could provide richer adversarial examples for training and testing. Finally, integrating this detection framework into a real-time system introduces further challenges—especially around early detection. Subtle attacks may evolve slowly yet cause significant deviation, highlighting the need for models that balance sensitivity, specificity, and rapid response—an important direction for future work.

10 Conclusion

In conclusion, this paper has explored the effectiveness of adversarially robust prediction models for anomaly detection in power consumption data. Through the integration of modified adversarial training techniques, including the Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), and Projected Gradient Descent (PGD), we have demonstrated the potential to enhance the resilience of anomaly detection models against adversarial attacks, by controlling the over-generalization property of the neural network prediction models. Our findings underscore the importance of incorporating adversarial training with a specific objective within the development of anomaly detection models for power grids. By exposing the model to carefully crafted adversarial examples during training, we can improve its ability to identify and mitigate anomalous patterns, thereby bolstering the security and reliability of critical infrastructure. Furthermore, our study highlights the need for ongoing research and development efforts to advance the state-of-the-art in adversarially robust anomaly detection methods, particularly in the context of modern power grid environments. Moving forward, future work should focus on refining adversarial training and exploring novel approaches to enhance the robustness of anomaly detection models against evolving cyber threats, by considering the over-generalization problem. Additionally, there is a need for comprehensive evaluation frameworks to assess the effectiveness and scalability of adversarially trained models in real-world deployment scenarios. By addressing these challenges, we can contribute to the development of more secure and resilient smart grid systems capable of withstanding adversarial attacks.

References

[1] Baldemar Aguirre-Fraire, Jessica Beltrán, and Valeria Soto. 2024. Household energy consumption enriched with weather data in northeast of Mexico. Data set. https://doi.org/10.17632/tvhygj8rgg.1

[2] Sheila Alemany and Niki Pissinou. 2021. The Dilemma Between Data Transformations and Adversarial Robustness for Time Series Application Systems. arXiv:2006.10885 [cs.LG] https://arxiv.org/abs/2006.10885

[3] Moayad Aloqaily, Burak Kantarci, and Hussein T. Mouftah. 2017. Trusted Third Party for service management in vehicular clouds. In *13th International Wireless Communications and Mobile Computing Conference*. 928–933.

[4] Taha Belkhouja and Janardhan Rao Doppa. 2020. Analyzing Deep Learning for Time-Series Data Through Adversarial Lens in Mobile and IoT Applications. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 39 (2020), 3190–3201.

[5] Taha Belkhouja and Janardhan Rao Doppa. 2022. Adversarial framework with certified robustness for time-series domain via statistical features. *Journal of Artificial Intelligence Research* 73 (2022), 1435–1471.

[6] Gowtham Bellala, Manish Marwah, Martin Arlitt, Geoff Lyon, and Cullen E. Bash. 2011. Towards an understanding of campus-scale power consumption. In *Proceedings of the Third ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings (BuildSys '11)*. Association for Computing Machinery, 73–78.

[7] Liron Bergman, Niv Cohen, and Yedid Hoshen. 2020. Deep Nearest Neighbor Anomaly Detection. arXiv:2002.10445 [cs.LG]

[8] C Chahla, Hichem Snoussi, L Merghem, and M Esseghir. 2020. A deep learning approach for anomaly detection and prediction in power consumption data. *Energy Efficiency* 13, 8 (2020), 1633–1651.

[9] Jui-Sheng Chou and Abdi Suryadinata Telaga. 2014. Real-time detection of anomalous power consumption. *Renewable and Sustainable Energy Reviews* 33 (2014), 400–411.

[10] Adrian Dabrowski, Johanna Ullrich, and Edgar R Weippl. 2017. Grid shock: Coordinated load-changing attacks on power grids: The non-smart power grid is vulnerable to cyber attacks as well. In *33rd Annual Computer Security Applications Conference*. Association for Computing Machinery, 303–314.

[11] Islam Debicha, Thibault Debatty, Jean-Michel Dricot, and Wim Mees. 2021. Adversarial Training for Deep Learning-based Intrusion Detection Systems. arXiv:2104.09852 [cs.CR]

[12] U.S.-Canada Power System Outage Task Force. 2004. Final Report on the August 14, 2003 Blackout in the United States and Canada: Causes and Recommendations. https://www.energy.gov/sites/prod/files/oeprod/DocumentsandMedia/BlackoutFinal-Web.pdf.

[13] Honghao Gao, Binyang Qiu, Ramon J Duran Barroso, Walayat Hussain, Yueshen Xu, and Xinheng Wang. 2022. Tsmae: a novel anomaly detection approach for internet of things time series data using memory-augmented autoencoder. *IEEE Transactions on network science and engineering* 10, 5 (2022), 2978–2990.

[14] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. 2019. Memorizing Normality to Detect Anomaly: Memory-augmented Deep Autoencoder for Unsupervised Anomaly Detection. arXiv:1904.02639 [cs.CV]

[15] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. arXiv:1412.6572 [stat.ML] https://arxiv.org/abs/1412.6572

[16] Georges Hebrail and Alice Berard. 2012. Individual household electric power consumption. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C58K54.

[17] M. Hofmann and T. Siebenbrunner. 2023. A rich dataset of hourly residential electricity consumption data and survey answers from the iFlex dynamic pricing experiment. Data set. https://doi.org/10.5281/zenodo.8248802

[18] Zahra Jadidi, Shantanu Pal, Nithesh Nayak, Arawinkumaar Selvakkumar, Chih-Chia Chang, Maedeh Beheshti, and Alireza Jolfaei. 2022. Security of machine learning-based anomaly detection in cyber physical systems. In *International Conference on Computer Communications and Networks*. 1–7.

[19] Yifan Jia, Jingyi Wang, Christopher M Poskitt, Sudipta Chattopadhyay, Jun Sun, and Yuqi Chen. 2021. Adversarial attacks and mitigation for anomaly detectors of cyber-physical systems. *International Journal of Critical Infrastructure Protection* 34 (2021), 100452.

[20] Fazle Karim, Somshubra Majumdar, and Houshang Darabi. 2020. Adversarial attacks on time series. *IEEE transactions on pattern analysis and machine intelligence* 43, 10 (2020), 3309–3320.

[21] Tae-Young Kim and Sung-Bae Cho. 2018. Predicting the household power consumption using CNN-LSTM hybrid networks. In *International Conference on Intelligent Data Engineering and Automated Learning*. 481–490.

[22] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2017. Adversarial examples in the physical world. arXiv:1607.02533 [cs.CV]

[23] Jiao Li, Yang Liu, Tao Chen, Zhen Xiao, Zhenjiang Li, and Jianping Wang. 2020. Adversarial attacks and defenses on cyber–physical systems: A survey. *IEEE Internet of Things Journal* 7, 6 (2020), 5103–5115.

[24] Jiangnan Li, Yingyuan Yang, Jinyuan Stella Sun, Kevin Tomsovic, and Hairong Qi. 2021. Conaml: Constrained adversarial machine learning for cyber-physical systems. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*. 52–66.

[25] Srinidhi Madabhushi and Rinku Dewri. 2022. On the Impact of Model Tolerance in Power Grid Anomaly Detection Systems. In *International Conference on Information Systems Security*. 220–234.

[26] Srinidhi Madabhushi and Rinku Dewri. 2023. A survey of anomaly detection methods for power grids. *International Journal of Information Security* 22, 6 (2023), 1799–1832.

[27] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2019. Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv:1706.06083

[28] Gautam Raj Mode and Khaza Anuarul Hoque. 2020. Crafting adversarial examples for deep learning based prognostics. In *19th IEEE International Conference on Machine Learning and Applications*. 467–472.

[29] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. DeepFool: a simple and accurate method to fool deep neural networks. arXiv:1511.04599

[30] John Mulo, Pu Tian, Adamu Hussaini, Hengshuo Liang, and Wei Yu. 2023. Towards an adversarial machine learning framework in cyber-physical systems. In *2023 IEEE/ACIS 21st International Conference on Software Engineering Research, Management and Applications (SERA)*. IEEE, 138–143.

[31] Felix O Olowononi, Danda B Rawat, and Chunmei Liu. 2020. Resilient machine learning for networked cyber physical systems: A survey for machine learning security to securing machine learning for CPS. *IEEE Communications Surveys & Tutorials* 23, 1 (2020), 524–552.

[32] Gururaghav Raman, Jimmy Chih-Hsien Peng, and Talal Rahwan. 2019. Manipulating residents' behavior to attack the urban power distribution system. *IEEE Transactions on Industrial Informatics* 15, 10 (2019), 5575–5587.

[33] Md Mamunur Rashid, Joarder Kamruzzaman, Mohammad Mehedi Hassan, Tasadduq Imam, Santoso Wibowo, Steven Gordon, and Giancarlo Fortino. 2022. Adversarial training for deep learning-based cyberattack detection in IoT-based smart city applications. *Computers & Security* 120 (2022), 102783.

[34] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. 2022. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14318–14328.

[35] Zakir Ahmad Sheikh, Yashwant Singh, Pradeep Kumar Singh, and Paulo J Sequeira Gonçalves. 2023. Defending the defender: adversarial learning based defending strategy for learning based security methods in Cyber-physical systems (CPS). *Sensors* 23, 12 (2023), 5459.

[36] Shoaib Ahmed Siddiqui, Andreas Dengel, and Sheraz Ahmed. 2020. Benchmarking adversarial attacks and defenses for time-series data. In *International Conference on Neural Information Processing*. Springer, 544–554.

[37] Saleh Soltan, Prateek Mittal, and H. Vincent Poor. 2018. BlackIoT: IoT Botnet of High Wattage Devices Can Disrupt the Power Grid. In *27th USENIX Security Symposium*.

[38] Junho Song, Keonwoo Kim, Jeonglyul Oh, and Sungzoon Cho. 2023. Memto: Memory-guided transformer for multivariate time series anomaly detection. *Advances in Neural Information Processing Systems* 36 (2023), 57947–57963.

[39] Zhe Sun and Jinguo Li. 2022. Anomaly detection for CPS via memory-augmented reconstruction and time series prediction. In *2022 IEEE 19th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*. IEEE, 530–536.

[40] Abdulrahman Takiddin, Muhammad Ismail, and Erchin Serpedin. 2022. Robust data-driven detection of electricity theft adversarial evasion attacks in smart grids. *IEEE Transactions on Smart Grid* 14, 1 (2022), 663–676.

[41] Kai Liang Tan, Yasaman Esfandiari, Xian Yeow Lee, Soumik Sarkar, et al. 2020. Robustifying reinforcement learning agents via action space adversarial training. In *2020 American control conference (ACC)*. IEEE, 3959–3964.

[42] Jiwei Tian, Buhong Wang, Jing Li, and Zhen Wang. 2022. Adversarial Attacks and Defense for CNN Based Power Quality Recognition in Smart Grid. *IEEE Transactions on Network Science and Engineering* 9, 2 (2022), 807–819.

[43] Kebei Wang and Manimaran Govindarasu. 2024. FGSM-based Synthetic Data Generation Technique and Application to Anomaly Detection in Smart Grid. In *2024 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, 1–5.

[44] Pengyuan Wang and Manimaran Govindarasu. 2019. *Cyber-Physical Anomaly Detection for Power Grid with Machine Learning*. Springer International Publishing, 31–49.

[45] Shunyao Wang, Ryan K. L. Ko, Guangdong Bai, Naipeng Dong, Taejun Choi, and Yanjun Zhang. 2024. Evasion Attack and Defense on Machine Learning Models in Cyber-Physical Systems: A Survey. *Commun. Surveys Tuts.* 26, 2 (2024).

[46] Xiaohui Wang, Ting Zhao, He Liu, and Rong He. 2019. Power consumption predicting and anomaly detection based on long short-term memory neural network. In *4th International Conference on Cloud Computing and Big Data Analysis*.

[47] Leon Wu, Gail Kaiser, Cynthia Rudin, David Waltz, Roger Anderson, Albert Boulanger, Ansaf Salleb-Aouissi, Haimonti Dutta, and Manoj Pooleery. 2011. Evaluating machine learning for improving power grid reliability. In *Proceedings of ICML 2011 Workshop on Machine Learning for Global Challenge*.

[48] Tao Wu, Xuechun Wang, Shaojie Qiao, Xingping Xian, Yanbing Liu, and Liang Zhang. 2022. Small perturbations are enough: Adversarial attacks on time series prediction. *Information Sciences* 587 (2022), 794–812.

[49] Aidong Xu, Xuechun Wang, Yunan Zhang, Tao Wu, and Xingping Xian. 2022. Adversarial Attacks on Deep Neural Networks for Time Series Prediction. In *2021 10th International Conference on Internet Computing for Science and Engineering*. Association for Computing Machinery, 8–14.

[50] Haiqi Zhu, Shaohui Liu, and Feng Jiang. 2022. Adversarial training of LSTM-ED based anomaly detection for complex time-series in cyber-physical-social systems. *Pattern Recognition Letters* 164 (2022), 132–139.