Energy and AI 17 (2024) 100381

Contents lists available at ScienceDirect

Energy and AI

journal homepage: www.elsevier.com/locate/egyai

# Evasive attacks against autoencoder-based cyberattack detection systems in power systems

Yew Meng Khaw [a,*], Amir Abiri Jahromi [b], Mohammadreza F.M. Arani [c], Deepa Kundur [a]

[a] Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada
[b] School of Electronic and Electrical Engineering, University of Leeds, Leeds, LS2 9JT, United Kingdom
[c] Department of Electrical, Computer & Biomedical Engineering, Toronto Metropolitan University (formerly Ryerson University), Toronto, ON M5B 2K3, Canada

## HIGHLIGHTS

- A novel algorithm is proposed for evasive attacks against autoencoder-based cyberattack detection systems (CDS) in smart grids.
- Effectiveness of the proposed evasive attack is demonstrated on an autoencoder-based CDS for transmission protective relays.
- Attacker with knowledge of the CDS model can successfully craft evasive samples to cause spurious protective relay operations.
- It is paramount to adequately assess the robustness of ML-based CDS prior to full-scale field implementation in smart grids.

## ARTICLE INFO

*Keywords:*
Cybersecurity
Adversarial attacks
Anomaly detection
Iterative-based methods
Substation automation

## ABSTRACT

The digital transformation process of power systems towards smart grids is resulting in improved reliability, efficiency and situational awareness at the expense of increased cybersecurity vulnerabilities. Given the availability of large volumes of smart grid data, machine learning-based methods are considered an effective way to improve cybersecurity posture. Despite the unquestionable merits of machine learning approaches for cybersecurity enhancement, they represent a component of the cyberattack surface that is vulnerable, in particular, to adversarial attacks. In this paper, we examine the robustness of autoencoder-based cyberattack detection systems in smart grids to adversarial attacks. A novel iterative-based method is first proposed to craft adversarial attack samples. Then, it is demonstrated that an attacker with white-box access to the autoencoder-based cyberattack detection systems can successfully craft evasive samples using the proposed method. The results indicate that naive initial adversarial seeds cannot be employed to craft successful adversarial attacks shedding insight on the complexity of designing adversarial attacks against autoencoder-based cyberattack detection systems in smart grids.

## 1. Introduction

### 1.1. Motivations

The number, frequency, and impact of cyberattacks have been escalating in a variety of domains over the past decade. Given the significance of critical infrastructures on societal welfare, cyberattacks on these systems have recently raised particular concern.

Power systems are arguably considered the most important critical infrastructure because most other critical infrastructures depend on the availability of electricity. A successful cyberattack against power systems can result in wide area, long duration electricity outages such as the one experienced by the Ukrainian power grid in 2015 [1,2] that can affect the safety of citizens and their economy. As such, the North American Electric Reliability Corporation (NERC) and other regulatory agencies have taken steps to ensure the cybersecurity of power systems, for example, by mandating various standards such as the critical infrastructure protection (CIP) standards [3]. As such, there is a pressing need by regulatory agencies and utilities alike to develop various approaches for the cybersecurity enhancement of smart grids.

Model-based and machine learning (ML)-based methods have been extensively investigated for cybersecurity enhancement of smart grids. ML-based methods for cyberattack detection are receiving extensive attention due to the increasing availability of large volumes of smart grid data. Despite the unquestionable benefits of ML-based approaches for smart grid cyberattack detection and mitigation, ML-based models

* Corresponding author.
*E-mail addresses:* ymkhaw@ece.utoronto.ca (Y.M. Khaw), a.abirijahromi@leeds.ac.uk (A.A. Jahromi), marani@torontomu.ca (M.F.M. Arani), dkundur@ece.utoronto.ca (D. Kundur).

https://doi.org/10.1016/j.egyai.2024.100381
Received 29 January 2024; Received in revised form 16 May 2024; Accepted 28 May 2024
Available online 4 June 2024

themselves can become the target of attacks. An experienced attacker with specific ML model knowledge can design and apply adversarial attacks that render ML-based attack detection ineffective.

### 1.2. Literature review

#### 1.2.1. Related literature on ML-based cyberattack detection systems in smart grids

ML-based cyberattack detection systems are extensively examined for different applications in smart grids. Supervised and unsupervised learning methods have been used in [4] to detect stealthy false data injection attacks against state estimators. In [5], ensemble learning algorithms using unsupervised and supervised classifiers have been proposed to detect stealthy false data injection attacks against state estimation. A semi-supervised learning approach based on mixture Gaussian distribution has been proposed in [6] for detecting false data injection attacks also targeting smart grid state estimation. Wavelet transform and deep neural network techniques have been employed in [7] to detect false data injection attacks against ac state estimation. A supervised ML-based method has been presented in [8] to detect cyberattacks targeting state estimators. A genetic algorithm-based method has been used in [8] for feature selection to enhance detection accuracy and reduce computational complexity.

A novel method based on margin setting algorithm has been presented in [9] to protect power systems against false data injection attacks. In [10], an artificial-intelligence-based algorithm has been introduced to detect compromised meters. An intrusion detection system using non-nested generalized exemplars and state extraction method has been proposed in [11] for wide area measurements. A combination of signature-based and deep learning methods have been employed in [12] to monitor and detect cyberattacks in transmission protection. In [13,14], a deep learning-based cyberattack detection system has been proposed for transmission line protection. The performance of different learning algorithms including supervised, semi-supervised, and online learning algorithms have been analysed in [15] for different attack scenarios. A conditional deep belief network-based method has been introduced in [16] for detecting the false data injection attacks in real-time. An unsupervised anomaly detection system using dynamic Bayesian networks and restricted Boltzmann machine has been proposed in [17] to differentiate an actual fault from a cyberattack in smart grids.

#### 1.2.2. Related literature on adversarial attacks against ML-based methods in smart grids

Various papers have investigated the adversarial attacks in the context of different applications in smart grids. In [18,19], the vulnerability of ML-based load forecasting model to adversarial attacks has been investigated. An adversarial attack against ML-based event analysis has been presented in [20]. In [21] a method has been proposed to generate adversarial signals to attack learning models in power systems.

The vulnerability of ML-based classifiers to adversarial attacks has been investigated in [22] for phasor measurement unit data. In [23], the vulnerabilities of deep learning-based energy theft detection systems to adversarial attacks has been investigated. A black-box adversarial attack construction algorithm for targeting ML-based models operating on smart meter data has been presented in [24]. An adversarial attack model has been presented in [25] to compromise dynamical controls of energy systems. In [26,27], machine learning have been used to craft false data injection attacks against power grid state estimation.

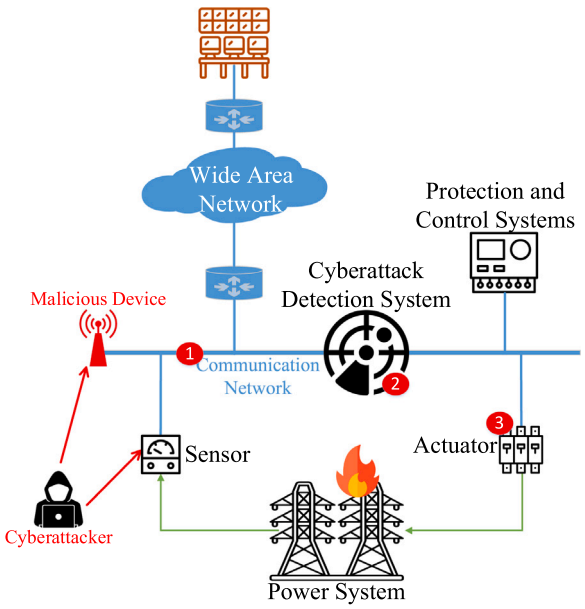


**Fig. 1.** High level representation of executing adversarial attacks against smart grids equipped with ML-based cyberattack detection systems.

#### 1.2.3. Related literature on adversarial attacks against cyberattack detection systems in smart grids

Recently, several papers have examined evasive attacks in the context of smart grid cybersecurity. While some references, [28,29], are focused on specific protocols, other references, [30–33], are application oriented. ML-based intrusion detection systems for IEC 60870-5-104 and Modbus protocols have been introduced in [28,29], respectively. Afterwards, the resilience of these intrusion detection systems have been evaluated against adversarial attacks. The vulnerability of deep learning-based intrusion detection systems against adversarial attacks has been examined in [30] in demand response applications. The adversarial attacks against intrusion detection systems for state estimators have been investigated in [31,32]. In [33], the adversarial attacks against deep reinforcement learning-based energy theft detectors have been explored.

### 1.3. Contributions

Considering the existing literature, there is a research gap on examining the robustness of machine learning-based cyberattack detection systems to adversarial attacks in smart grids. This paper builds on the available literature on adversarial attacks against machine learning to investigate the robustness of autoencoder-based cyberattack detection systems in smart grids. A novel iterative-based method is first proposed to craft adversarial attacks against an autoencoder-based unsupervised cyberattack detection systems. Afterwards, it is demonstrated that an attacker with white-box access to the autoencoder-based cyberattack detection systems can successfully craft evasive samples using the proposed method. We demonstrate that a naive initial adversarial seed cannot be used to craft successful adversarial attacks in smart grids using iterative-based methods.

The main contributions of this paper are as follows:

- For the first time, it is demonstrated that autoencoder-based cyberattack detection systems in digital substations are vulnerable to evasive attacks.
- A novel iterative-based algorithm is proposed to create evasive attacks against autoencoder-based cyberattack detection systems.
- It is demonstrated that the success of the proposed iterative-based algorithm does not depend on the availability of data from the targeted system.

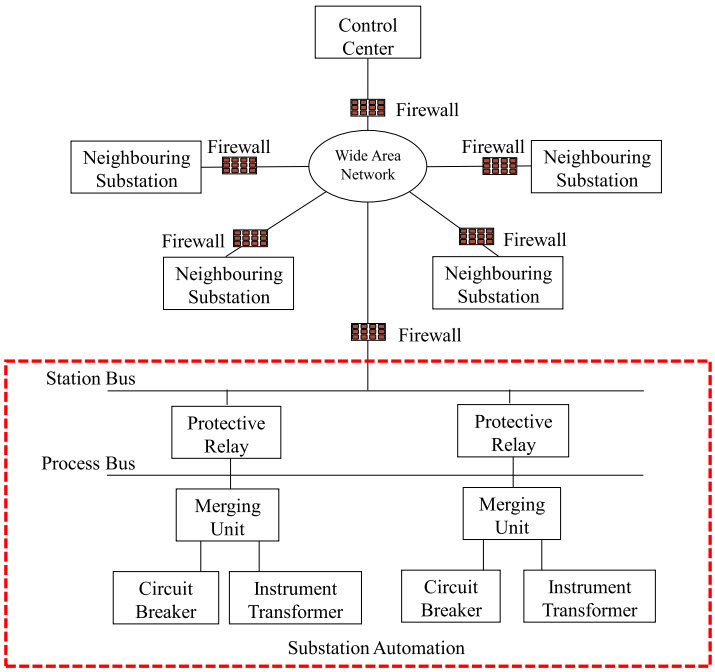


**Fig. 2.** A schematic diagram of the hierarchical structure of protection, control and monitoring systems in power systems.

### *1.4. Paper organization*

The remainder of this paper is organized as follows. Section 2 presents the basics of adversarial attacks against ML-based cyberattack detection systems in smart grids. The modelling of adversarial attacks is described in Section 3. Section 4 presents a novel iterative-based approach for generating adversarial attacks. The simulation results are provided in Section 5 before concluding the paper in Section 6.

## 2. Adversarial attacks against machine learning-based cyberattack detection systems

Interest and progress in the application of artificial intelligence (AI) and ML to cybersecurity of smart grids has increased in recent years. Consequently, there is a growing concern over these ML-based cybersecurity defences themselves becoming the target of attacks known as adversarial attacks. The objective of adversarial attacks is to deceive ML-based algorithms into making incorrect decisions by exploiting their vulnerabilities. Successful adversarial attacks against ML-based cybersecurity defences can prevent/delay attack detection or reduce their trustworthiness by increasing the number of misclassifications. As such, there is a need to evaluate the robustness of ML-based cybersecurity defences against such attacks.

Adversarial attacks can be conducted in training and/or testing phases. Poisoning attacks are performed in the training phase by injecting corrupting samples into the training set to achieve adversarial goals of attackers in the testing phase [34,35]. Evasive attacks are executed in the testing phase by deliberately adding subtle perturbation to the input data to mislead the ML model to make incorrect decisions [36,37]. The feasibility of evasive adversarial attacks is related to the characteristics of ML models as well as overfitting due to insufficient regularization. The focus of this paper is on the evasive adversarial attacks.

Evasive adversarial attacks can be classified into white-box or black-box attacks based on the extent of the attackers' knowledge of the machine learning system, which involves information about: (1) training dataset, (2) feature set, (3) learning algorithm, and (4) objective function and parameters/hyper-parameters during training. In the white-box evasive attack, the attacker has complete knowledge about the target system. In the black-box adversarial attack, the attacker has no knowledge about the ML algorithm and architecture including model parameters, feature set and training data, but can query the system and receive feedback to make educated conjectures about the model or to obtain training data that can subsequently be used to train a separate or substitute model.

Adversarial attacks were first investigated in the image processing and computer vision domains [38–40]. The objective here is to minimally perturb benign data to cause misclassification while keeping the modifications imperceptible to the human eye. Although there is a large body of literature on adversarial attacks in image and computer vision domains, the increased attention of ML-based models to industrial control system (ICS) cybersecurity problems has resulted in recent adversarial attack publications in this application area. It is worth noting that famous methods for generating adversarial attacks such as fast gradient sign method (FGSM) and basic iteration method (BIM) are ineffective in industrial control systems and smart grids as extensively discussed in [41–44]. Yet, few papers have investigated the implications of adversarial attacks for ML-based cyberattack detection systems in smart grids [45].

There are several requirements for successful implementation of attacks against ICSs and smart grids that make adversarial attacks in these fields more challenging as illustrated in Fig. 1. First, communication networks are innate to modern ICSs and smart grids. As such, attackers must bypass any intrusion detection systems that monitor communication packets and communication traffic for abnormal behaviours. Moreover, the communication packet must reach the desired device. For example, the communication packet containing a malicious payload must have the correct destination media access control (MAC) address to reach the desired target device. Second, adversarial attacks need to evade ML-based cyberattack detection systems that monitor the payload of communication packets. For instance, the attackers may need to mimic laws of physics or acceptable ranges to evade ML-based cyberattack detection systems in smart grids. Third, the malicious payload of communication packets should cause the desired physical impact, *e.g.* maloperation of a protective relay resulting in tripping operations to achieve the desired impact of an attacker. The second and third aspects are related to designing effective and successful adversarial samples.

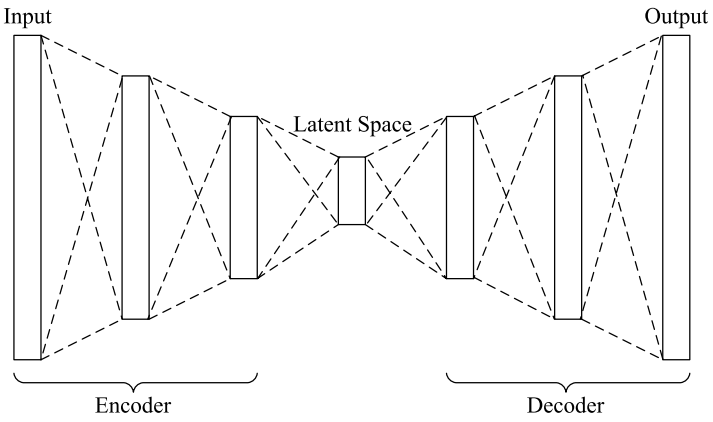


**Fig. 3.** Structure of a typical autoencoder consisting of the encoder, decoder and the latent space.

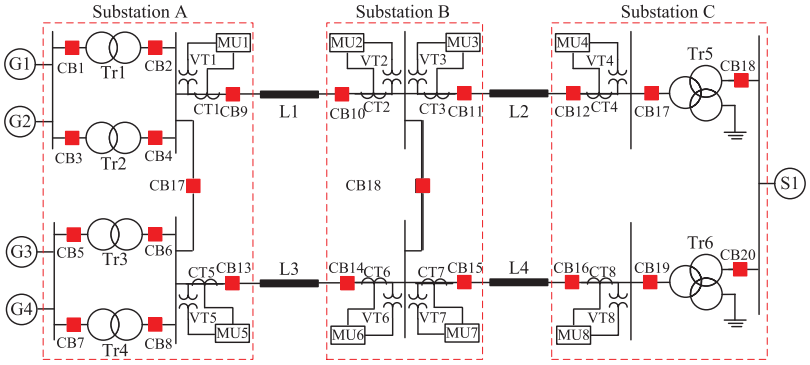


**Fig. 4.** The IEEE PSRC D6 benchmark test system.

Both iterative-based and generative adversarial networks can be used to devise adversarial attacks. The main objective of the iterative-based methods is to transform an initial sample iteratively to maximize the probability of the sample evading the ML-based cyberattack detection systems and also causing the desired physical impact. The main challenge in the iterative-based methods is to design an objective function that can successfully create adversarial samples to a high degree. In contrast, the main challenge in generative adversarial networks is the availability of data to train the models.

The present literature has neglected the unique characteristics of smart grids in contrast to other ICSs. Power systems consist of a large number of substations and components that are spread over a wide geographical area. Moreover, essential protection and control actions need to be executed within a short period of time. These characteristics necessitate distributed intrusion systems relying on a few local measurements (sensor readings) in one substation. This is while the adversarial attacks against intrusion detection systems studied in the present literature usually have focused on water distribution and water treatment systems. These water distribution and water treatment systems collect all sensors and actuators data with very low sampling rates which makes execution of successful attacks against these ICSs more difficult. Moreover, the attacker should manipulate most of the sensor readings in order to execute a successful attack which makes the implementation of adversarial attacks difficult, if not impossible [46]. While the underlying principles of these approaches can be generalized to some extent to applications like state estimation in power systems, they cannot be simply extended to protection and control systems.

## 3. Threat model

The protection, control and monitoring systems in power systems have a hierarchical structure as illustrated in Fig. 2 which commonly operate in different time scales. In general, the protection and control systems on the lower system levels are designed to function based on data with higher time resolution and act faster compared to protection and control systems on the higher system levels. In this paper, we focus on the sensors and actuators at the lower power system levels like protection systems in substations. It is worth noting that no encryption can be used for communication between these lower level systems because of the time critical functions involved. Moreover, the actuators at this level form the first line of defence in power systems. As such, successful cyberattacks against these sensors and actuators can result in serious consequences for power systems. These characteristics persuade many researchers to adopt AI-based intrusion detection system such as autoencoders [14]. Unsupervised or semi-supervised autoencoder-based cyberattack detection systems are superior to other AI-based methods because of the abundance of labeled non-anomalous datasets and rarity of cyberattack datasets in smart grids.

The objective of the attacker is to devise a stealthy attack which can evade autoencoder-based cyberattack detection systems in smart grids and force an actuator or a set of actuators to misoperate and cause the intended physical impact. We assume that the attacker can access and compromise a particular set of sensors or data coming from sensors. Two attack scenarios against IEC 61850-based substations are considered here. In the first scenario, we assume that the attacker has the ability to either compromise the transformer instruments/merging units physically or manipulate their settings to perform the false data injection attacks. In the second scenario, we assume that the attacker has the ability to compromise and control the communication traffic coming from the sensors, for example, by connecting a malicious device to the communication network to perform man-in-the-middle (MITM) and false data injection (FDI) attacks. The attacker disrupts the sampled value (SV) packets from the merging unit to the protective relays and sends falsified SV packets which contain the synthetic fault data

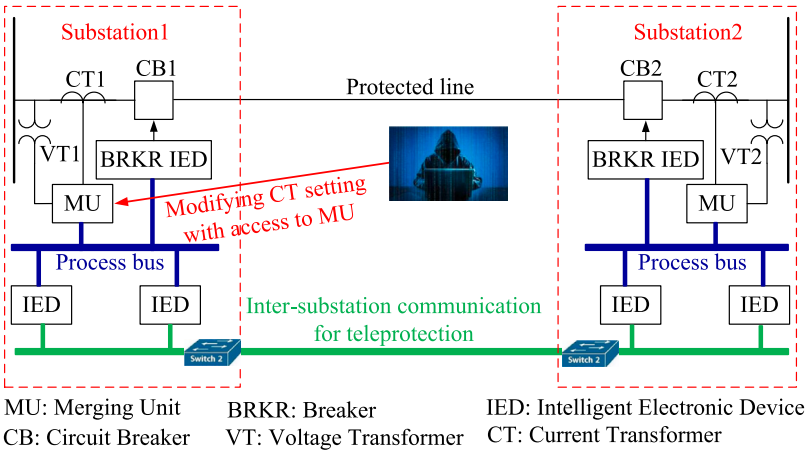


**Fig. 5.** Malicious modification of the current transformer setting through the merging unit.

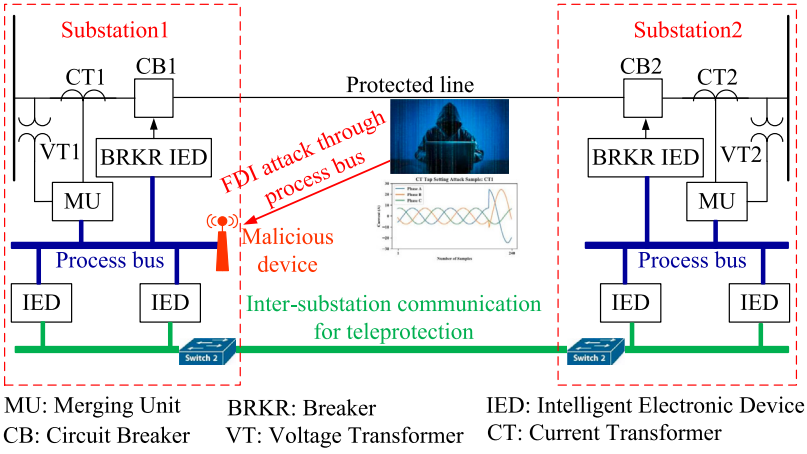


**Fig. 6.** False data injection attack through the process bus.

crafted using the proposed adversarial attack algorithm. This causes the protective relay to misoperate. We also assume that the attacker has white-box access to the autoencoder-based cyberattack detection systems and knows the model architecture and parameters. The attacker may be a disgruntled internal employee with white-box access to the cyberattack detection system. The attacker may also be external to the organization, but has been involved with the development, evaluation and installation of the cyberattack detection system.

## 4. Proposed evasive attack algorithm

We start by identifying a roadmap that an attacker can take to execute a successful attack against smart grids with autoencoder-based cyberattack detection systems. First, the attacker chooses a target in the smart grid for the attack. For example, protective relays such as instantaneous overcurrent relays can be considered as the targets of the evasive attack. The attacker should then identify the characteristics of the attack data that can force the target to misoperate. For an instantaneous overcurrent relay, an attacker can force a misoperation by injecting current measurements with magnitude above the instantaneous overcurrent relay setting. The challenge for the attacker is to create an evasive attack sample which can both bypass the cyberattack detection system in a substation and force the relay to misoperate. The example of instantaneous overcurrent protective relays discussed here can be easily extended to other types of protective relays or other protection and control systems in substations.

Consider an autoencoder illustrated in Fig. 3. The encoder, $f$, encodes the input, $x$, to a latent space, $z$, with dimensions typically smaller than the input space. The decoder, $g$, then reconstructs the input from the latent space. We propose an iterative algorithm for devising evasive attacks to bypass the autoencoder-based cyberattack detection systems and trigger a protection or control system in smart grids as follows:

$$x_a = \underset{x}{\operatorname{argmin}} \|g(f(x)) - x\|_2^2 + h(x) \tag{1}$$

where $x_a$ denotes the crafted evasive attack sample.

The first term in (1) ensures that the reconstruction error is small and allows the evasive attack sample to bypass the cyberattack detection system. The second term ensures that the crafted evasive attack sample can force the target relay to misoperate. Different functions can be used to implement $h(x)$. The $g(f(x))$ term in (1) is taken from the whitebox model of the autoencoder and $h(x)$ should be designed based on the logic of the targeted protective relay or control system. It is worth noting that the logic of the protective relays or control systems in power systems are well-documented and standardized. Thus, it is convenient for attackers to access roughly accurate information about these systems to design the term $h(x)$ in (1).

Considering the example of the instantaneous overcurrent protective relay, the current measurements received from the current transformer are compared with a threshold by the instantaneous protective relay to detect the fault and trip the circuit breaker. Therefore, the attacker can select $h(x) = ReLU(\theta - max(x))$ where $\theta$ must be larger than or equal to the relay setting. $ReLU$ represents rectified linear unit function. The accessible information in the public domain about the ratings of current transformers and settings of the instantaneous

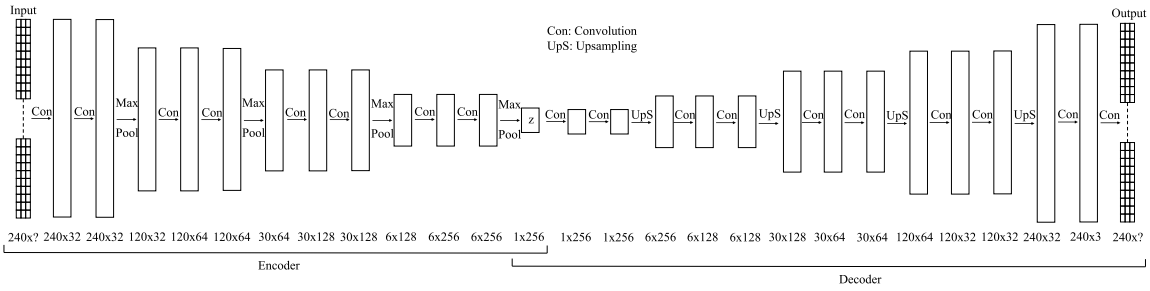


**Fig. 7.** The architecture of the autoencoder-based cyberattack detection system.

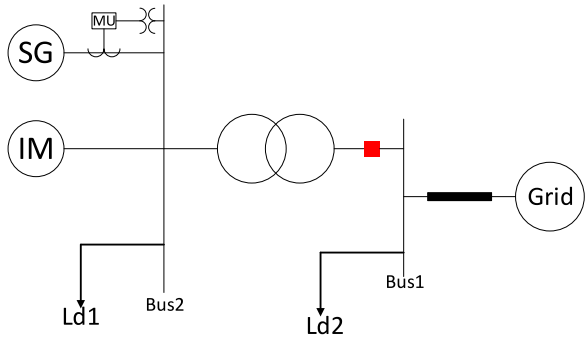


**Fig. 8.** Simulation-based distribution system test case.

overcurrent protective relays make the selection of appropriate value for $\theta$ convenient for attackers. Furthermore, the choice of a $ReLU$ function in $h(x)$ not only guarantees that the protective relay will be triggered but also eliminates the competition between the first and second terms in (1) once the relay threshold is met.

**Algorithm 1** Crafting The Evasive Attack Sample

1: $\tau$ is the learning rate, $\gamma$ is the momentum term
2: $meansquarederror(MSE)_{target}$ is the desired reconstruction error
3: $x_0 \leftarrow x_f$
4: **while** $||g(f(x_i)) - x_i||_2^2 > MSE_{target}$ **do**
5: $\quad L_i = ||g(f(x_i)) - x_i||_2^2 + h(x)$
6: $\quad v_{i+1} = \gamma v_i + (1-\gamma)\nabla_{x_i} L_i$
7: $\quad x_{i+1} = x_i - \tau v_i$
8: **end while**

The optimization problem in (1) can be solved using Algorithm 1 which is implemented using TensorFlow. $x_f$ in step 3 of Algorithm 1 is the adversarial seed. In steps 5–7, the algorithm perturbs the evasive sample to minimize its reconstruction error, $\|g(f(x_i)) - x_i\|_2^2$, while also minimizing the term $h(x)$ to ensure that the evasive sample triggers the relay. This perturbation of the evasive sample is repeated iteratively until the desired reconstruction error is achieved.

## 5. Simulation results

In this section, we investigate the performance of the proposed evasive attack against autoencoder-based cyberattack detection systems for protective relays. We first describe the cyberattacks against protective relays. Next, we explain an autoencoder-based cyberattack detection systems for protective relays. Afterwards, the capability of the proposed evasive attack to bypass the autoencoder-based cyberattack detection systems is demonstrated. All the experiments are conducted on a computer with i7-9700K CPU and RTX2080 GPU. The deep learning model for the cyberattack detection system is implemented with Keras with a Tensorflow backend. Algorithm 1 is implemented in Tensorflow to craft the evasive attack samples.

### 5.1. Cyberattacks against protective relays

Power systems are commonly protected by various protective relays which use measurements like current and/or voltage to identify abnormal conditions. The abnormal conditions are generally recognized by considering a combination of thresholds and time limits. For example, the overcurrent relays detect the abnormal conditions when the current level goes beyond a certain threshold and for a certain duration of time. As such, a cyberattacker can employ many different methods to execute a successful attack in the absence of cyberattack detection systems. Moreover, the attacker does not even need to replicate waveforms of abnormal power system conditions. This is because any signal that violate the threshold and time settings of a protective relay is sufficient to force the relay to misoperate regardless of the waveform shape. For example, the attacker can change the setting of the current measurement instrument or inject any false current measurements with high magnitude through the substation local area network to trigger the overcurrent relay.

We employ the IEEE power system relaying committee (PSRC) D6 benchmark test system in this paper to demonstrate the cyberattacks against protective relays [47]. The test system comprises of three substations connecting a power plant to the rest of the power system through four transmission lines as illustrated in Fig. 4. The transmission lines are protected by various protective relays including distance, overcurrent and differential protection. The test system is implemented using a co-simulation platform including Riverbed Modeler and OPAL-RT HYPERSIM simulator.

We implemented two cyberattacks in this section to demonstrate the vulnerability of the protective relays in power systems in the absence of cyberattack detection systems. In the first attack, we changed the current transformer settings to force the overcurrent relay of the transmission line to misoperate as illustrated in Fig. 5. In the second attack scenario, we injected currents with high magnitude through the substation local area network to cause overcurrent relay malfunction as illustrated in Fig. 6. Both attack scenarios resulted in the transmission line tripping.

The growing number of cyberattacks against power systems in recent years has underlined the importance of cyberattack detection systems.

### 5.2. Autoencoder-based cyberattack detection system for protective relays

The stealthy and polymorphic nature of cyberattacks motivated the development of machine learning-based cyberattack detection systems in substations to counteract these attacks. The unsupervised autoencoder-based cyberattack detection systems are superior to other machine learning-based methods as they can detect zero-day attacks. Moreover, autoencoder-based cyberattack detection systems do not need continually evolving cyberattack data for training.

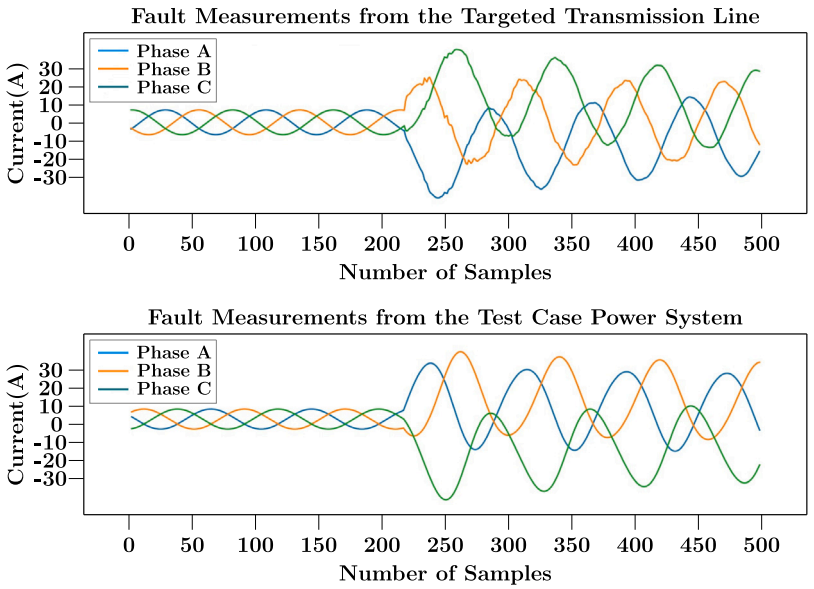


**Fig. 9.** The fault measurements from both the targeted power system (top) and the test case power system employed to generate initial attack seed (bottom).

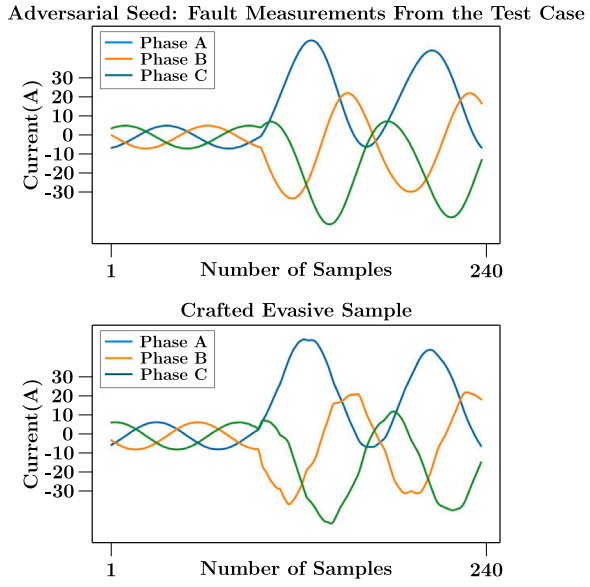


**Fig. 10.** A window of initial adversarial seed (top) and crafted evasive attack using Algorithm 1 (bottom).

In this paper, we adopt a 1-dimensional convolutional based autoencoder from [14] as illustrated in Fig. 7. The autoencoder has a convolution filter size of 10, convolutional stride length of 1 and used ReLU as the activation function. The autoencoder receives time series of current measurements as input for the overcurrent protective relay and generates an alarm when it detects a cyberattack. Data are fed to the autoencoder-based cyberattack detection system in windows with the size of 240 measurement samples.

The precision and recall metrics are employed to measure the performance of the cyberattack detection system.

$$precision = \frac{\#True\ Positive}{\#True\ Positive + \#False\ Positive} \tag{2}$$

$$recall = \frac{\#True\ Positive}{\#True\ Positive + \#False\ Negative} \tag{3}$$

Cyberattacks that are correctly detected by the cyberattack detection system represent True Positive. Measurements with normal behaviour that are incorrectly classified as a cyberattack represent False Positive. Cyberattacks that are not detected by the cyberattack detection system represent False Negative. Measurements with normal behaviour that are correctly classified as legitimate measurements represent True Negative. # represents the count of each event. Therefore, precision is the fraction of attack classifications made by the cyberattack detection model that is correct. Recall is the fraction of actual attacks that are "recalled", *i.e.*, correctly classified as attacks by the cyberattack detection system.

The autoencoder-based cyberattack detection system was able to identify the attacks presented in Section 5.1 with 100% precision and 100% recall. As such, the attacker can only bypass the cyberattack detection system either by injecting genuine historical fault data from the location of the protective relay or using adversarial attacks. It is worth noting that obtaining historical fault data is difficult as a fault may have never occurred on the zone of protection of the relay while the attacker has gained access to these measurement data. In the next section, we demonstrate how the attacker can use evasive attack algorithms to bypass autoencoder-based cyberattack detection systems.

### 5.3. Evasive attacks

In this section, we investigate the ability of evasive attacks to bypass the autoencoder-based cyberattack detection systems for protective relays. The iterative algorithm proposed in Section 4 for evasive attacks is employed here to craft the evasive attacks. The attacker needs a seed, $x_f$, to craft evasive attacks using the proposed algorithm as cited in Section 4.

We consider a simulation-based test case to generate the seed for Algorithm 1 to craft evasive attacks. Nevertheless, any other approach can be adopted to generate the seed. We assume that the attacker has limited information about the target system. For example, we assume that the attacker only knows that the target substation is a generator substation. Nevertheless, the attacker does not have any information about the generator parameters, the configuration of the substation or the topology and impedances of the network.

The simulation-based test case used by the attacker in our study is illustrated in Fig. 8. The test case is based on a demo test system in MATLAB Simulink developed by G. Sybille and T. Zabaiou. In the test case, a 3.125-MVA synchronous generator, SG, is connected to a 2400-V distribution system which feed a 2250-hp induction motor, IM, and a resistive load, Ld1. The synchronous generator is equipped with a

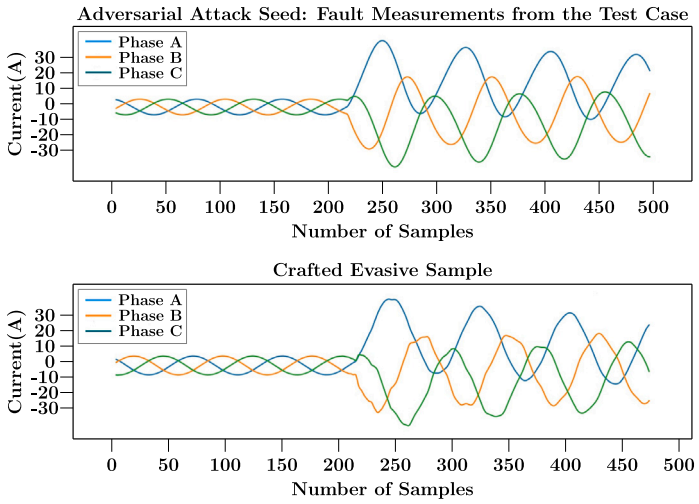


**Fig. 11.** The initial adversarial seed (top) and the joined crafted evasive attack using the proposed method (bottom).

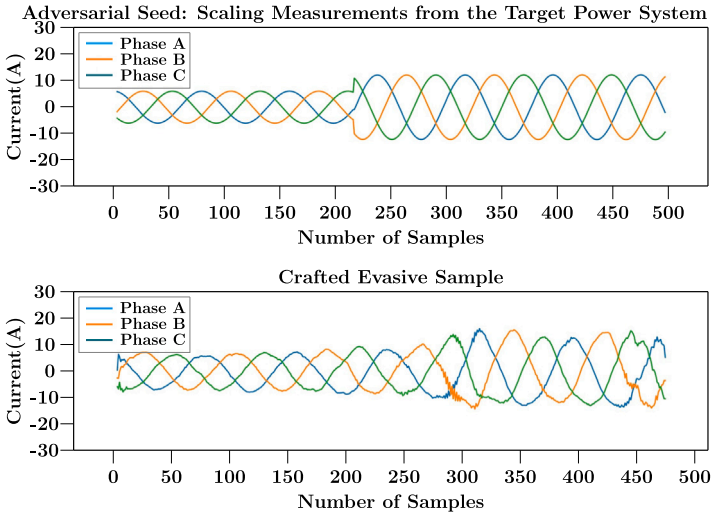


**Fig. 12.** The scaled fault-free measurements as the initial adversarial seed (top) and the joined crafted evasive attack (bottom).

non-reheat thermal turbine, governor, automatic generation controller (AGC) and an AC1 A excitation system. The 2400-V distribution system is also connected to a 25-kV network through a 6 MVA transformer. The 25-kV network is modelled as an ideal voltage source behind an R-L impedance. Fault measurements from both the targeted power system and the test case power system employed to generate the initial attack seed is shown in Fig. 9 for comparison.

We emphasize the simulation-based test case used by the attacker to generate the initial attack seed is different from the target system illustrated in Fig. 4 in several aspects. First, the rating and parameters of the generator in the two systems are different. Second, the generator exciter, governor, and turbine of the generator in the two systems are different. Third, the networks voltage level and topology are different. Because of these differences, the initial attack seed data generated using this test case cannot bypass the autoencoder-based cyberattack detection system developed for the target system illustrated in Fig. 4 without using adversarial attack algorithms.

The attacker can craft synthetic evasive samples using Algorithm 1 to replicate the measurements in the target substation. Recall that the cyberattack detection system has an input window of 240 measurement samples. As such, each crafted evasive output from Algorithm 1 contains 240 samples. These individually crafted evasive windows are guaranteed to evade the detector as long as we set mean squared error (MSE) target to a value less than the detection threshold in Algorithm 1. A window of evasive data is shown in Fig. 10.

However, in reality, an attacker has to feed a continuous series of measurement data to the overcurrent relay instead of a disjoint sequence of windows of data. As such, the attacker has to join the crafted evasive windows of data into a single final evasive sample. The method of generating the final evasive sample by joining the first few measurements of each window and discarding the measurements that overlap with the next evasive window does not work. This is because joining several windows of evasive data which are crafted independently introduces discontinuities in the final crafted evasive sample. These discontinuities can be detected by the cyberattack detection system as an attack.

To address this challenge, the stride of the windows of evasive data should be reduced to minimize the discontinuities introduced in the final joined evasive sample. When crafting the evasive samples, a window stride of 1 produced the results with the minimum discontinuity in our simulations. We performed gradient averaging during each update iteration of the evasive samples to further increase the success rate of the final joined evasive sample and reduce its reconstruction error. In Step 7 of Algorithm 1, every data point is updated using the average

gradient computed across all windows that contain that data point. As such, we are using the average for the sections where the evasive windows overlap.

The evasive attack crafted using the seed produced by the test power system in Fig. 8, Algorithm 1 and techniques explained above was able to evade the autoencoder-based cyberattack detection system with 100% success rate. An example of this joined evasive sample is illustrated in Fig. 11. The results here proved that the proposed Algorithm 1 can be used successfully to craft evasive attacks against autoencoder-based cyberattack detection systems in substations.

It is worth noting that the attacker cannot craft evasive attacks using naive seeds. To demonstrate this notion, the evasive attack seed is generated by using scaled fault-free measurement data from the targeted power system similar to the case illustrated in Fig. 5 in Section 5.1. Although Algorithm 1 was able to reduce the reconstruction error, it was unable to craft a successful evasive attack as illustrated in Fig. 12.

Training ML models with synthetically generated evasive samples has been proposed in the literature as an effective way to counteract adversarial attacks against classifier models. For example, training datasets are augmented with adversarial examples in [48] to provide protection against adversarial attacks on the classifiers. However, such an approach may not work in an unsupervised learning setting and additional adaptations are needed as the unsupervised model does not use labeled datasets.

## 6. Conclusion

This paper investigated the vulnerability of autoencoder-based cyberattack detection systems in digital substations to adversarial attacks. A novel iterative-based algorithm is proposed to craft evasive attacks against autoencoder-based cyberattack detection systems. It is demonstrated that an attacker with white-box access to the autoencoder-based cyberattack detection system and limited knowledge about the target substation can craft an evasive attack which can bypass the cyberattack detection system and cause a physical impact. It is worth noting that the same attacker cannot craft evasive attacks using naive seeds. This highlighted the importance of keeping both the information about substations, and model architecture and parameters of autoencoder-based cyberattack detection systems confidential.

## CRediT authorship contribution statement

**Yew Meng Khaw:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Amir Abiri Jahromi:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Mohammadreza F.M. Arani:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Deepa Kundur:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The data that has been used is confidential.

## Acknowledgment

The authors wish to thank the NSERC Strategic Partnerships Grants for Projects program, the NSERC Discovery Grants program, the Fonds Nature et technologies – Québec (FRQNT) Postdoctoral Awards program and the NSERC Postdoctoral Fellowships program for funding this research.

## References

[1] Liang G, Weller SR, Zhao J, Luo F, Dong ZY. The 2015 ukraine blackout: Implications for false data injection attacks. IEEE Trans Power Syst 2016;32(4):3317–8.
[2] Industrial control systems cyber emergency response team (ICS-CERT) cyberattack against ukrainian critical infrastructure. 2016, (accessed 29 August 2019).
[3] North American electric reliability corporation (NERC) critical infrastructure protection (CIP) reliability standards. 2019, [Online]. Available: http://www.nerc.com. (accessed 29 August 2019).
[4] Esmalifalak M, Liu L, Nguyen N, Zheng R, Han Z. Detecting stealthy false data injection using machine learning in smart grid. IEEE Syst J 2017;11(3):1644–52.
[5] Ashrafuzzaman M, Das S, Chakhchoukh Y, Shiva S, Sheldon FT. Detecting stealthy false data injection attacks in the smart grid using ensemble-based machine learning. Comput Secur 2020;97.
[6] Foroutan SA, Salmasi FR. Detection of false data injection attacks against state estimation in smart grids based on a mixture Gaussian distribution learning method. IET Cyber Phys Syst Theory Appl 2017;2(4):161–71.
[7] Yu JJQ, Hou Y, Li VOK. Online false data injection attack detection with wavelet transform and deep neural networks. IEEE Trans Ind Inform 2018;14(7):3271–80.
[8] Ahmed S, Lee Y, Hyun S, Koo I. Feature selection–based detection of covert cyber deception assaults in smart grid communications networks using machine learning. IEEE Access 2018;6:27518–29.
[9] Wang Y, Amin MM, Fu J, Moussa HB. A novel data analytical approach for false data injection cyber–physical attack mitigation in smart grids. IEEE Access 2017;5:26022–33.
[10] Khanna K, Panigrahi BK, Joshi A. AI-based approach to identify compromised meters in data integrity attacks on smart grid. IET Gener Trans Distrib 2018;12(5):1052–66.
[11] Adhikari U, Morris TH, Pan S. Applying non-nested generalized exemplars classification for cyber-power event and intrusion detection. IEEE Trans Smart Grid 2018;9(5):3928–41.
[12] Ahmed A, Krishnan VVG, Armina Foroutan S, Touhiduzzaman Md, Rublein C, Srivastava A, et al. Cyber physical security analytics for anomalies in transmission protection systems. IEEE Trans Ind Appl 2019;55(6):6313–23.
[13] Khaw YM, Abiri Jahromi A, Arani MFM, Kundur D, Sanner S, Kassouf M. Preventing false tripping cyberattacks against distance relays: a deep learning approach. In: Proc. IEEE inter. conf. comm. cont. and comput. tech. for smart grids (smartGridComm). Beijing, China; 2019, p. 1–6.
[14] Khaw YM, Abiri Jahromi A, Arani MFM, Sanner S, Kundur D, Kassouf M. A deep learning-based cyberattack detection system for transmission protective relays. IEEE Trans Smart Grid 2021;12(3):2554–65.
[15] Ozay M, Esnaola I, Yarman Vural FT, Kulkarni SR, Poor HV. Machine learning methods for attack detection in the smart grid. IEEE Trans Neural Netw Learn Syst 2016;27(8):1773–86.
[16] He Y, Mendis GJ, Wei J. Real-time detection of false data injection attacks in smart grid: A deep learning-based intelligent mechanism. IEEE Trans Smart Grid 2017;8(5):2505–16.
[17] Karimipour H, Dehghantanha A, Parizi RM, Choo KR, Leung H. A deep and scalable unsupervised machine learning system for cyber-attack detection in large-scale smart grids. IEEE Access 2019;7:80778–88.
[18] Chen Y, Tan Y, Zhang B. Exploiting vulnerabilities of load forecasting through adversarial attacks. In: Proceedings of the tenth ACM e-energy. ACM; 2019, p. 1–11.
[19] Zhou X, Li Y, Barreto CA, Li J, Volgyesi P, Neema H, et al. Evaluating resilience of grid load predictions under stealthy adversarial attacks. In: 2019 resilience week. RWS, 2019, p. 206–12.
[20] Niazazari I, Livani H. Attack on grid event cause analysis: An adversarial machine learning approach. In: 2020 IEEE power & energy society innovative smart grid technologies conference. ISGT, 2020, p. 1–5.
[21] Tian J, Li T, Shang F, Cao K, Li J, Ozay M. Adaptive normalized attacks for learning adversarial attacks and defenses in power systems. In: 2019 IEEE international conference on communications, control, and computing technologies for smart grids (smartGridComm). 2019, p. 1–6.

[22] Biswal M, Misra S, Tayeen AS. Black box attack on machine learning Assisted Wide Area monitoring and protection systems. In: 2020 IEEE power & energy society innovative smart grid technologies conference. ISGT, 2020, p. 1–5.
[23] Li J, Yang Y, Sun JS. Exploiting vulnerabilities of deep learning-based energy theft detection in ami through adversarial attacks. 2020.
[24] Wang J, Srikantha P. Stealthy Black-Box Attacks on deep learning non-intrusive load monitoring models. IEEE Trans Smart Grid 2021;12(4):3479–92.
[25] Sabounchi M, Wei-Kocsis J. A practical adversarial attack on contingency detection of smart energy systems. 2021.
[26] Mohammadpourfard M, Ghanaatpishe F, Mohammadi M, Lakshminarayana S, Pechenizkiy M. Generation of false data injection attacks using conditional generative adversarial networks. In: 2020 IEEE PES innovative smart grid technologies europe (ISGT-europe). 2020, p. 41–5.
[27] Liu T, Shu T. On the security of ANN-based AC state estimation in smart grid. Comput Secur 2021;105.
[28] Teryak H, Albaseer A, Abdallah M, Al-Kuwari S, Qaraqe M. Double-edged defense: Thwarting cyber attacks and adversarial machine learning in IEC 60870-5-104 smart grids. IEEE Open J Ind Electron Soc 2023;4:629–42.
[29] Mumrez A, Sánchez G, Elbez G, Hagenmeyer V. On evasion of machine learning-based intrusion detection in smart grids. In: 2023 IEEE international conference on communications, control, and computing technologies for smart grids (smartGridComm). Glasgow, United Kingdom; 2023, p. 1–7.
[30] Guihai Z, Sikdar B. Adversarial machine learning against false data injection attack detection for smart grid demand response. In: 2021 IEEE international conference on communications, control, and computing technologies for smart grids (smartGridComm). Aachen, Germany; 2021, p. 352–7.
[31] Sayghe A, Zhao J, Konstantinou C. Evasion attacks with adversarial deep learning against power system state estimation. In: 2020 IEEE power & energy society general meeting. PESGM, Montreal, QC, Canada; 2020, p. 1–5.
[32] Li J, Yang Y, Sun JS, Tomsovic K, Qi H. Towards adversarial-resilient deep neural networks for false data injection attack detection in power grids. In: 2023 32nd international conference on computer communications and networks. ICCCN, Honolulu, HI, USA; 2023, p. 1–10.
[33] El-Toukhy AT, Mahmoud MMEA, Bondok AH, Fouda MM, Alsabaan M. Countering evasion attacks for smart grid reinforcement learning-based detectors. IEEE Access 2023;11:97373–90.
[34] Rubinstein BIP, Nelson B, Huang L, Joseph AD, Lau S, Rao S, Taft N, Tygar JD. ANTITODE: Understanding and defending against poisoning of anomaly detectors. In: Proc. th ACM SIGCOMm conference on internet measurement, association for computing machinery. 2009.
[35] Shafahi A, Huang WR, Najibi M, Suciu O, Studer C, Dumitras T, Goldstein T. Poison frogs! Targeted clean-label poisoning attacks on neural networks. In: Proc. 32nd conference on neural information processing systems. NIPS, Montreal; 2018.
[36] Xu W, Qi Y, Evans D. Automatically evading classifiers. In: Network and distributed system security symposium. NDSS, San Diego; 2016.
[37] Biggio B, Corona I, Maiorca D, Nelson B, Šrndić N, Laskov P, Giacinto G, Roli F. Evasion attacks against machine learning at test time. Mach Learn Knowl Discov Databases 2013;8190:387–402.
[38] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow IJ, Fergus R. Intriguing properties of neural networks. 2013.
[39] Moosavi-Dezfooli S, Fawzi A, Frossard P. DeepFool: a simple and accurate method to fool deep neural networks. 2015.
[40] Papernot N, McDaniel P, Jha S, Fredrikson M, Berkay Celik and Z, Swami A. The limitations of deep learning in adversarial settings. 2015.
[41] Gomez ALP, Maimo LF, Celdran AH, Clemente FJG, Cleary F. Crafting adversarial samples for anomaly detectors in industrial control systems. Procedia Comput Sci 2021;184:573–80.
[42] Kong X, Ge Z. Adversarial attacks on neural-network-based soft sensors: Directly attack output. IEEE Trans Ind Inf 2022;18(4):2443–51.
[43] Liu Y, Xu L, Yang S, Zhao D, Li X. Adversarial sample attacks and defenses based on LSTM-ED in industrial control systems. Comput Secur 2024;140.
[44] Jia Y, Wang J, Poskitt CM, Chattopadhyay S, Sun J, Chen Y. Adversarial attacks and mitigation for anomaly detectors of cyber–physical systems. Int J Crit Infrastruct Prot 2021;34.
[45] Anthi E, Williams L, Rhode M, Burnap P, Wedgbury A. Adversarial attacks on machine learning cybersecurity defences in industrial control systems. J Inf Secur Appl 2021;58.
[46] Kravchik M, Shabtai A. Efficient cyber attack detection in industrial control systems using lightweight neural networks and pca. IEEE Trans Dependable Secure Comput 2021.
[47] Gras H, Mahseredjian J, Rutovic E, Karaagac U, Haddadi A, Saad O, Kocar I, El-Akoum A. A new hierarchical approach for modeling protection systems in EMT–type software. In: Proc. international conference on power system transients. Seoul, Republic of Korea; 2017.
[48] Goodfellow I, Schlens J, Szegedy C. Explaining and harnessing adversarial examples. 2015, arXiv arXiv:1412.6572v3.