

ORIGINAL RESEARCH

Open access

White-Box Adversarial Threats to Autoencoder-Based Anomaly Detection in Smart Grid Cybersecurity

Ricardo Alves^{1*}, Bruno Costa², Helena Martins¹

Abstract

The shift of electrical power systems toward intelligent smart-grid architectures delivers better dependability, operational performance, and real-time visibility, yet simultaneously heightens exposure to cyber threats. Thanks to the massive amounts of data now generated in smart grids, machine learning techniques have emerged as a particularly promising avenue for bolstering defensive capabilities. While the strengths of machine-learning solutions for cybersecurity improvement are clear, these same models form part of the attack surface and remain especially susceptible to carefully crafted adversarial perturbations. This work analyzes the resilience of autoencoder-based intrusion detection mechanisms in smart grids against adversarial examples. We begin by presenting an original iterative procedure for synthesizing adversarial instances. Next, we demonstrate that a white-box adversary with full knowledge of the autoencoder detector can reliably produce stealthy, misclassifying samples using the proposed approach. Experimental results show that simply starting from basic adversarial seeds is insufficient to generate effective evasions, underscoring the considerable difficulty of constructing potent adversarial examples targeting autoencoder-based anomaly detectors deployed in smart-grid environments.

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

*Correspondence:

Ricardo Alves
ricardo.alves@ua.pt

¹ Department of Energy Analytics and Smart Systems, Faculty of Engineering, University of Aveiro, Aveiro, Portugal

² Department of Grid Optimization Technologies, Faculty of Engineering, University of Porto, Porto, Portugal

Introduction

Motivations

Over the last ten years, the scale, frequency, and severity of cyber intrusions have risen sharply across numerous fields. Because critical infrastructure underpins modern society, hostile cyber operations targeting these assets have attracted increased attention lately.

Among all critical infrastructures, power delivery networks stand out as arguably the single most vital, given that virtually every other essential service depends on a stable electricity supply. When cybercriminals succeed in compromising power grids, the outcome can be massive, extended blackouts—as seen during the well-documented 2015 assault on Ukraine's electricity network [1, 2]—with direct repercussions for human safety and national economies. In light of such risks, bodies such as the North American Electric Reliability Corporation (NERC) and various government regulators have introduced mandatory safeguards, notably the Critical Infrastructure Protection (CIP) framework [3]. This situation creates a strong and immediate demand—from both oversight agencies and grid operators—for innovative, robust methods to reinforce smart-grid cyber resilience.

Researchers have long examined both conventional model-driven techniques and modern data-driven machine-learning strategies to protect smart grids against cyber threats. In recent years, machine-learning approaches tailored to spotting malicious activity have gained particular momentum, fueled chiefly by the ever-growing repositories of high-resolution smart-grid measurements. Although ML-powered detection and response mechanisms clearly bring substantial value to smart-grid security, the models themselves can be deliberately subverted. A determined adversary with a detailed understanding of a specific machine-learning detector can engineer targeted adversarial manipulations that cause the system to fail to identify genuine attacks.

Literature review

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

A wide range of studies have explored machine learning-driven approaches for cyberattack detection across various smart grid applications. In Ref. [4], both supervised and unsupervised learning techniques were applied to identify stealthy false-data-injection attacks targeting state estimators. Reference [5] introduced ensemble learning frameworks that integrate unsupervised and supervised classifiers to counter covert false data injection attacks on state estimation processes. A semi-supervised strategy relying on Gaussian mixture models was developed in Ref. [6] specifically for identifying false data injection attacks aimed at smart grid state estimation. In Ref. [7], wavelet-based transformations combined with deep neural networks were used to detect false-data-injection attacks targeting AC state estimators.

A purely supervised machine learning technique was proposed in Ref. [8] for recognizing cyberattacks against state estimators. At the same time, the same study also employed a genetic algorithm for feature selection, thereby improving detection performance and reducing computational demands. Reference [9] presented an innovative protection scheme based on a margin-setting algorithm to defend power systems against false data injection attacks. In Ref. [10], an artificial intelligence algorithm was developed to identify tampered or compromised metering devices. An intrusion detection framework incorporating non-nested generalized exemplars and state extraction was proposed in Ref. [11] for monitoring wide-area measurement systems.

A hybrid approach merging signature-based detection with deep learning was applied in Ref. [12] to supervise and identify cyberattacks within transmission protection schemes. Deep learning models tailored for cyberattack detection in transmission line protection were proposed in Refs. [13] and [14]. Reference [15] conducted a comparative evaluation of multiple learning paradigms—supervised, semi-supervised, and online methods—across diverse attack scenarios. A conditional deep belief network was introduced in Ref. [16] to enable real-time detection of false data injection attacks. Finally, an unsupervised anomaly detection solution combining dynamic Bayesian networks with restricted Boltzmann machines was put forward in Ref. [17] to distinguish genuine physical faults from malicious cyberattacks in smart grid environments.

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

Numerous publications have analyzed the susceptibility of machine learning models to adversarial perturbations in various smart grid contexts. References [18] and [19] assessed how easily ML-based load forecasting models could be misled by adversarial inputs. An adversarial strategy targeting machine learning-driven event classification was described in Ref. [20]. In [21], a technique was developed to synthesize adversarial signals that can deceive learning-based models deployed in power systems.

The robustness of ML classifiers processing phasor measurement unit data was examined against adversarial examples in Ref. [22]. Reference [23] examined the vulnerability of deep learning systems designed for energy theft detection to adversarial manipulation. A black-box adversarial example generation algorithm for machine learning models operating on smart meter readings was introduced in Ref. [24]. In Ref. [25], an adversarial framework was proposed to undermine the dynamic control mechanisms of energy systems. References [26] and [27] demonstrated the use of machine learning techniques to construct sophisticated false data injection attacks against power grid state estimation procedures.

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

In recent years, a growing body of work has focused on evasion tactics within smart grid cybersecurity. Certain studies, such as Refs. [28] and [29], concentrate on protocol-specific defenses, whereas others—including Refs. [30–33]—adopt a more application-centric perspective. Intrusion detection systems tailored for the IEC 60870-5-104 and Modbus protocols were developed in Refs. [28] and [29], respectively, with subsequent assessments of their resilience to adversarial perturbations.

The susceptibility of deep learning-based intrusion detection systems to adversarial attacks was investigated in Ref. [30] within the domain of demand response services. References [31] and [32] explored adversarial threats targeting intrusion detection mechanisms protecting state estimators. In Ref. [33], researchers examined adversarial evasion techniques targeting deep reinforcement learning models used for detecting energy theft.

Contributions

Despite the extensive body of work already published, a notable research gap remains in evaluating the resilience of machine learning-driven cyberattack detection mechanisms to adversarial manipulations in smart grid environments. Drawing upon the existing studies on adversarial threats to machine learning models, this paper specifically examines the robustness of autoencoder-based cyberattack detection approaches deployed in smart grids. We begin by introducing an original iterative technique developed to generate adversarial examples targeting autoencoder-based unsupervised cyberattack detectors. Subsequently, we show that an adversary with white-box access to these autoencoder-based detection systems can reliably produce evasive samples using our proposed method. Furthermore, our analysis reveals that relying on simplistic initial adversarial seeds is inadequate for constructing adversarial examples in smart grids when using iterative generation strategies.

The primary contributions of this work can be summarized as follows:

- For the first time, this study establishes that autoencoder-based cyberattack detection systems implemented in digital substations are susceptible to evasion through adversarial perturbations.
- An innovative iterative algorithm is introduced for the purpose of synthesizing evasive attacks directed against autoencoder-based cyberattack detection frameworks.
- It is shown that the effectiveness of the proposed iterative algorithm remains independent of access to operational data from the victim system.

Adversarial Attacks Against Machine Learning-Based Cyberattack Detection Systems

In recent years, the adoption of artificial intelligence (AI) and machine learning (ML) techniques to strengthen cybersecurity in smart grids has seen substantial growth and advancement. This increased reliance on ML-based defensive tools has, in turn, sparked growing apprehension that these very systems could themselves become prime targets for a class of attacks known as adversarial attacks. The core purpose of adversarial attacks is to exploit inherent weaknesses in ML algorithms, causing them to produce erroneous outputs or decisions. When successful, such attacks can either obstruct or delay the identification of genuine cyber threats or erode confidence in the detection system by significantly elevating false-negative or false-positive rates. Consequently, rigorously assessing the robustness of ML-powered cybersecurity solutions against adversarial interference has become an essential research priority.

Adversarial attacks may occur during either the training stage or the inference (testing) stage. Poisoning attacks occur during the training phase, where maliciously crafted samples are inserted into the training dataset to compromise model

performance during subsequent deployment and testing [34, 35]. In contrast, evasive attacks are carried out at inference time by introducing imperceptible perturbations to input samples to induce misclassification or incorrect predictions by the ML model [36, 37]. The practicality of evasive adversarial attacks depends heavily on properties intrinsic to the ML model architecture, as well as on issues such as overfitting arising from inadequate regularization. This paper focuses exclusively on evasive adversarial attacks.

Evasive adversarial attacks are commonly divided into white-box and black-box categories, depending on the degree of knowledge the adversary possesses regarding the target ML system. This knowledge may encompass: (1) the composition of the training dataset, (2) the set of input features, (3) the specific learning algorithm employed, and (4) the objective function together with its associated parameters and hyperparameters. In a white-box evasive attack scenario, the adversary enjoys full visibility into all aspects of the targeted ML system. Conversely, in a black-box adversarial attack, the adversary lacks any direct insight into the model's algorithm, architecture, parameters, feature representation, or training data; however, the attacker is typically permitted to submit queries to the system and observe its responses, enabling inference about the underlying model or the collection of sufficient data to construct a surrogate model for further exploitation.

Adversarial attacks were initially explored within the fields of image processing and computer vision [38–40]. In those contexts, the goal is to apply the smallest possible alterations to legitimate inputs, leading to incorrect classifications while ensuring the changes remain undetectable to human observers. Although substantial research exists on adversarial attacks in image-related and vision applications, the growing integration of ML-based techniques into industrial control systems (ICS) security challenges has recently spurred research on adversarial threats in this domain. Notably, well-known adversarial generation techniques such as the fast gradient sign method (FGSM) and the basic iterative method (BIM) have proven largely ineffective when applied to industrial control systems and smart grids, as thoroughly analyzed in Refs. [41–44]. Nevertheless, only a limited number of studies have addressed the specific consequences of adversarial attacks on ML-driven cyberattack detection mechanisms in smart grids [45].

Several distinct constraints must be satisfied to effectively execute attacks on ICSs and smart grids, rendering adversarial efforts in these environments significantly more difficult, as depicted in **Figure 1**. Firstly, modern ICSs and smart grids inherently depend on communication networks, requiring attackers to circumvent any intrusion detection systems that scrutinize network packets and traffic patterns for irregularities. Additionally, the crafted malicious packet must successfully reach the intended recipient device—for instance, it must include the correct destination media access control (MAC) address to reach the targeted endpoint. Secondly, adversarial examples must successfully bypass ML-based cyberattack detection systems that analyze the content of communication payloads. For example, adversaries might need to replicate realistic physical behaviors or stay within permissible operational bounds to avoid triggering ML-based detectors in smart grids. Thirdly, the malicious content delivered via the communication packet must produce the intended physical consequence, such as the incorrect activation of a protective relay that leads to unnecessary tripping and thereby achieves the attacker's objectives. The latter two requirements are central to developing potent and viable adversarial samples.

Both iterative optimization approaches and generative adversarial networks can serve as means to construct adversarial attacks. Iterative-based techniques primarily aim to progressively modify an original sample to maximize its likelihood of evading ML-based cyberattack detectors while simultaneously inducing the targeted physical outcome. The primary difficulty with iterative methods is formulating an effective objective function that reliably produces high-quality adversarial instances. By comparison, the key obstacle for generative adversarial networks involves securing sufficient training data for model development.

Existing research has largely overlooked the distinctive features of smart grids relative to other types of ICSs. Power systems encompass a vast array of substations and interconnected elements distributed across extensive geographic regions. Furthermore, critical protection and control actions must occur within extremely brief time windows. These attributes demand the deployment of distributed intrusion detection systems that operate using only a handful of localized

measurements (such as sensor data) within a single substation. In contrast, much of the current literature on adversarial attacks against intrusion detection systems has concentrated on water distribution networks and water treatment facilities. Those systems typically aggregate data from all sensors and actuators at notably low sampling rates, which complicates the execution of successful attacks against them. Moreover, attackers often need to alter most sensor values to achieve their goals, making adversarial implementation highly challenging, if not impractical [46]. Although certain foundational concepts from these domains may be partially applicable to tasks such as state estimation in power systems, they cannot be directly transferred to protection and control functions without substantial adaptation.

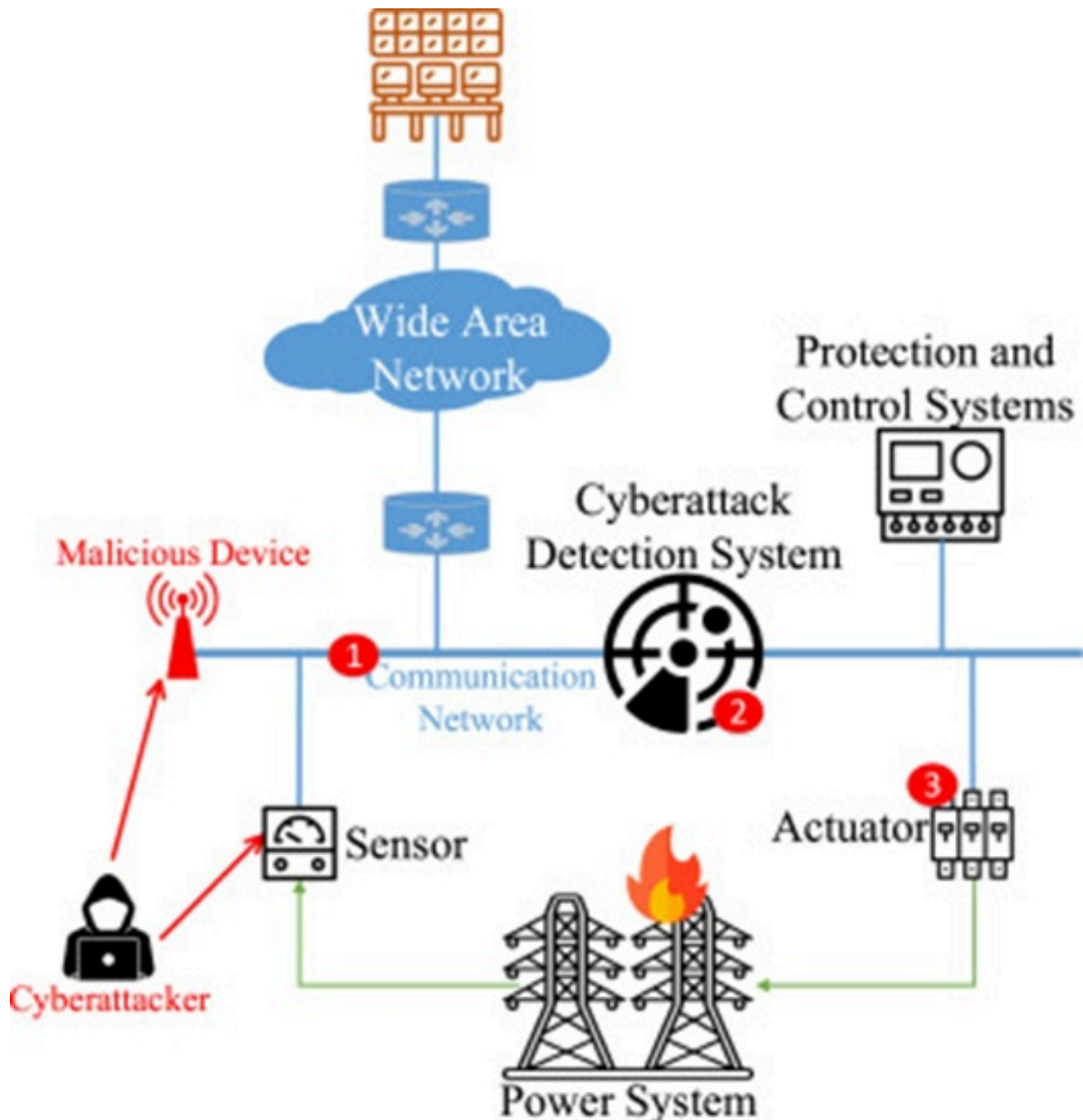


Figure 1. High-level illustration of the process for launching adversarial attacks on smart grids that incorporate machine learning-based cyberattack detection mechanisms.

Threat Model

Protection, control, and monitoring systems within power networks exhibit a hierarchical organization, as shown in [Figure 2](#), and typically function across varying time scales. In general, lower-level protection and control components are engineered to process high-resolution time-stamped data and respond more rapidly than their counterparts operating at higher hierarchical levels. The present study concentrates on sensors and actuators situated at these lower tiers of the power system, particularly within substation protection schemes. A key characteristic of these lower-level communications is that encryption cannot be applied due to the stringent time-critical nature of the functions they support. Furthermore, the actuators at this level serve as the initial line of defense for the power system. Consequently, a successful cyberattack on these sensors and actuators can cause severe disruptions across the power infrastructure. These constraints have motivated numerous researchers to implement AI-driven intrusion detection solutions, notably autoencoders [\[14\]](#). Unsupervised and semi-supervised autoencoder-based detection approaches outperform many other AI techniques in this context, owing to the plentiful supply of labeled normal (non-anomalous) data and the extreme scarcity of genuine cyberattack samples in smart grid environments.

The attacker's goal is to orchestrate a covert attack that slips past autoencoder-based cyberattack detection mechanisms in smart grids while compelling one or more actuators to malfunction in a manner that produces the desired physical outcome. We posit that the adversary can access and tamper with a selected subset of sensors or their data streams. Two distinct attack vectors are examined in the context of IEC 61850-compliant substations. In the first scenario, the adversary is assumed to have either physical access to compromise transformer instruments or merging units directly or sufficient privileges to alter their configuration parameters, thereby enabling false data injection attacks. In the second scenario, the adversary can intercept and manipulate the communication flows emanating from sensors—for instance, by inserting a rogue device into the network to carry out man-in-the-middle (MITM) and false data injection (FDI) operations. Specifically, the attacker interferes with sampled-value (SV) packets transmitted from the merging unit to protective relays and substitutes them with fabricated SV packets that embed synthetic fault signatures generated via the proposed adversarial technique. This deception induces erroneous operation of the protective relay. We further assume that the attacker has white-box access to the autoencoder-based cyberattack detection system, with complete knowledge of its architecture and learned parameters. Such an adversary could, for example, be a disaffected insider with legitimate internal access to the detection system, or an external actor previously engaged in the system's design, testing, or deployment phases.

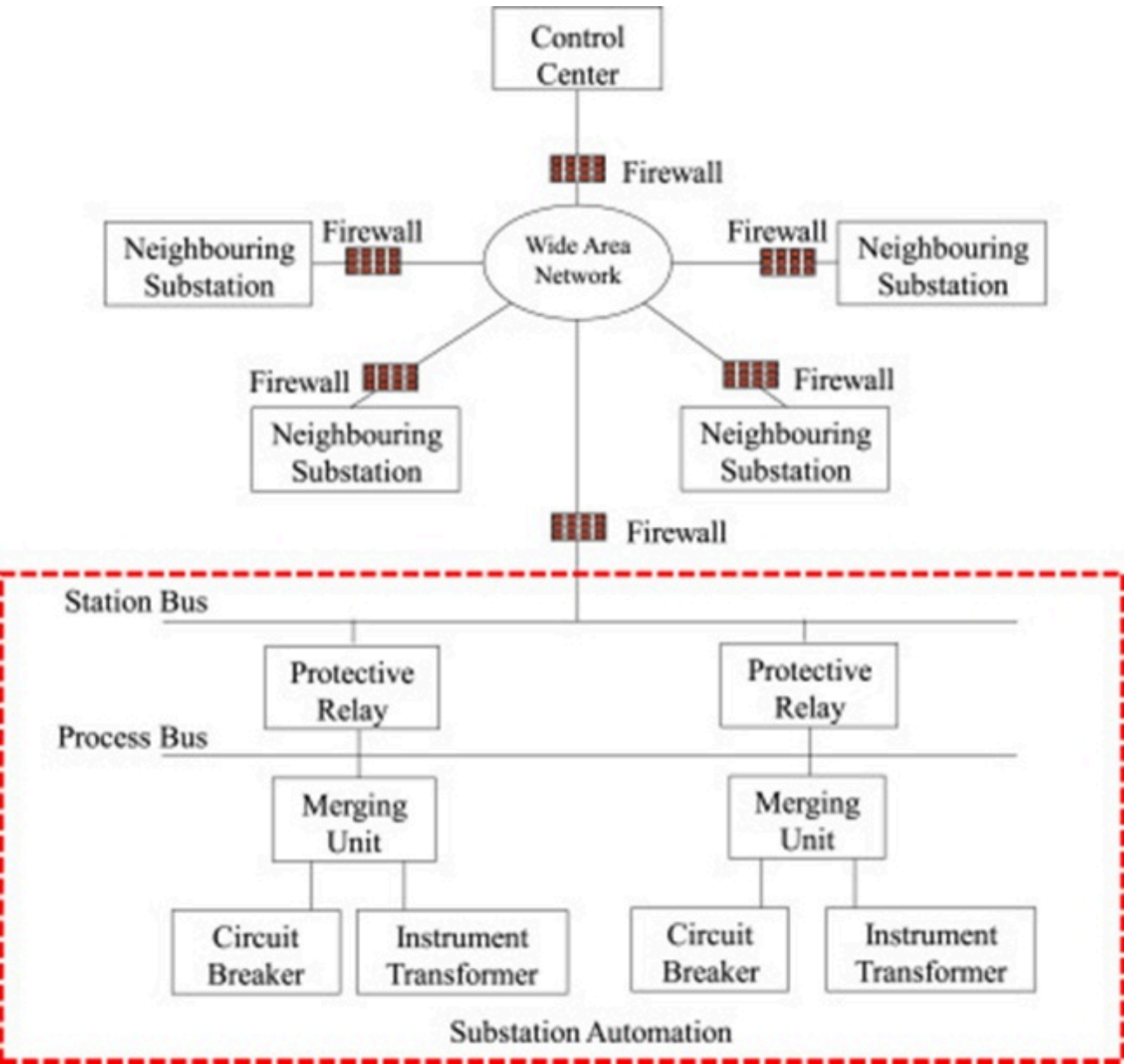


Figure 2. Schematic overview illustrating the hierarchical architecture of protection, control, and monitoring functions within power systems.

Proposed Evasive Attack Algorithm

We first outline a general sequence of steps an attacker might follow to mount an effective attack against smart grids protected by autoencoder-based cyberattack detection systems. Initially, the adversary selects a specific target within the smart grid infrastructure. For instance, protective relays—such as instantaneous overcurrent relays—may serve as the intended victim of the evasion effort. Next, the attacker determines the precise properties the malicious data must exhibit to induce misoperation of the chosen target. In the case of an instantaneous overcurrent relay, injecting current measurements that exceed the relay's pickup threshold would trigger an unintended trip. The core difficulty for the attacker is constructing an evasive sample that simultaneously bypasses the substation's cyberattack detection system and provokes the desired misbehavior in the relay. Although the instantaneous overcurrent relay example is used here for

clarity, the same principles can readily be adapted to other relay types or to additional protection and control elements within substations.

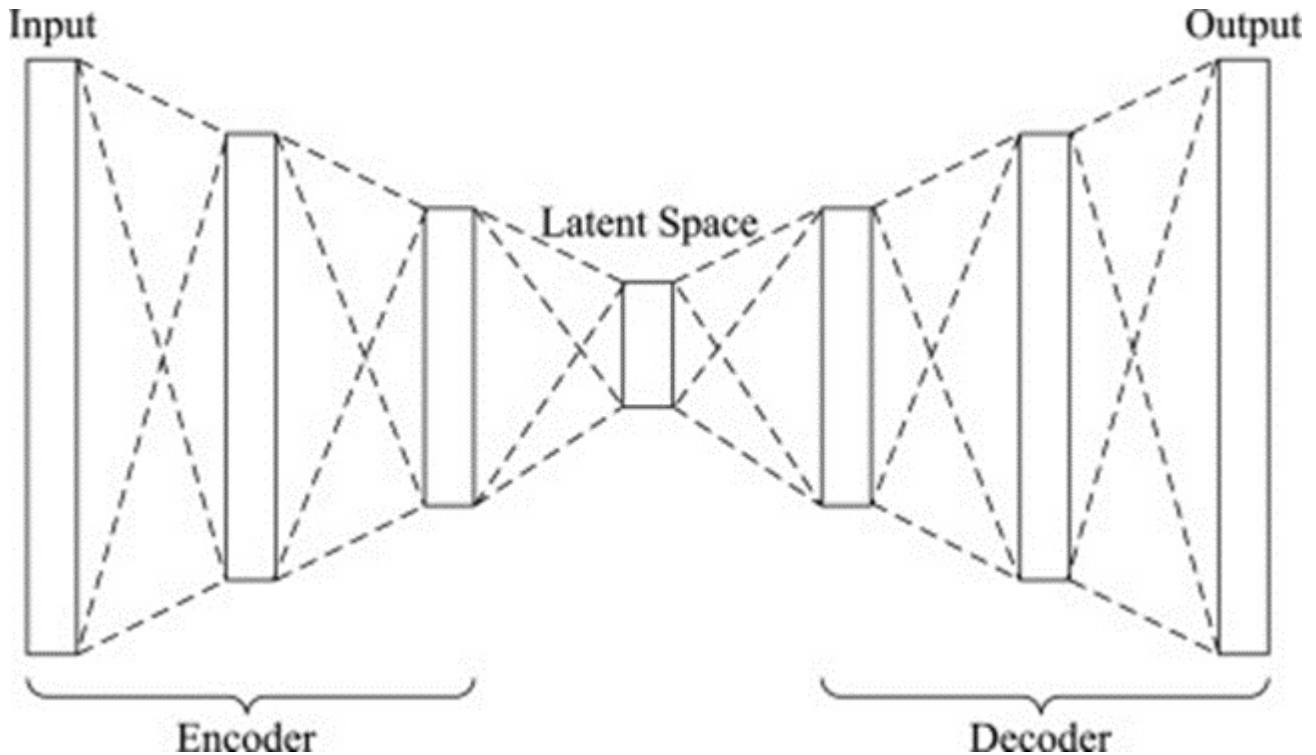


Figure 3. Architecture of a standard autoencoder, including the encoder component, decoder component, and the latent representation space.

Consider the autoencoder structure depicted in **Figure 3**. The encoder function, denoted f , maps the original input x into a compressed latent representation z , whose dimensionality is generally lower than that of the input. The decoder function, denoted g , then attempts to reconstruct the input from this latent code. Building on this architecture, we introduce an iterative procedure for generating evasive attacks that can both deceive autoencoder-based cyberattack detectors and activate undesired responses in smart grid protection or control devices, formalized as follows:

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

Where x_a represents the resulting adversarial (evasive) attack sample.

The first term in **Eq. 1** guarantees a low reconstruction error, enabling the crafted evasive sample to successfully evade the cyberattack detection mechanism. The second term ensures that the generated evasive sample induces misoperation in the targeted relay. Various functional forms can be adopted to realize $h(x)$. The term $h(x)$ in (1) is derived directly from the white-box representation of the autoencoder, whereas $h(x)$ must be carefully formulated according to the operational logic of the specific protective relay or control device under attack. It should be emphasized that the behavior and decision logic of protective relays and control systems in power networks are thoroughly standardized and well-documented in publicly available resources. This accessibility allows attackers to obtain sufficiently accurate details about these systems, facilitating the design of an appropriate $h(x)$ term in **Eq. 1**.

Taking the instantaneous overcurrent protective relay as a concrete example, the relay compares incoming current measurements from the current transformer against a predefined threshold to identify faults and initiate circuit breaker

tripping. Accordingly, the attacker can define $h(x) = \text{ReLU}(\theta - \max(x))$, where the threshold parameter θ is set to a value greater than or equal to the relay's pickup setting. Here, ReLU denotes the rectified linear unit activation function. Publicly disclosed specifications regarding current transformer ratings and typical settings for instantaneous overcurrent relays make it straightforward for adversaries to select a suitable value for θ . Moreover, incorporating the ReLU function into $h(x)$ not only ensures activation of the protective relay once the threshold is exceeded but also prevents any conflict or trade-off between the first and second terms in Eq. 1 after the relay's tripping condition has been satisfied.

```

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

```

Algorithm 1. Crafting the evasive attack sample

The optimization task formulated in Eq. 1 is addressed by Algorithm 1, implemented in TensorFlow. In step 3 of Algorithm 1, x_f serves as the initial adversarial seed. During steps 5–7, the procedure iteratively adjusts the evasive sample to simultaneously reduce its reconstruction error, expressed as $\|g(f(x_i)) - x_i\|_2^2$, and minimize the value of $h(x)$ so that the sample reliably triggers the targeted relay. This iterative perturbation process continues until the reconstruction error reaches an acceptably low level.

Simulation Results

This section evaluates the effectiveness of the proposed evasive attack strategy when applied against autoencoder-based cyberattack detection systems protecting protective relays. We begin by outlining the nature of cyberattacks directed at protective relays. We then describe the design and implementation of an autoencoder-based cyberattack detection system tailored for protective relays. Subsequently, we present evidence that the proposed evasive attack can circumvent the autoencoder-based detection mechanism. All simulations and experiments were performed on a system equipped with an Intel i7-9700K processor and an NVIDIA RTX 2080 graphics card. The deep learning components of the cyberattack detection system were developed using Keras with TensorFlow serving as the backend. Algorithm 1 was also implemented in TensorFlow to generate the adversarial (evasive) attack samples.

Cyberattacks against protective relays

Power systems rely on a variety of protective relays that analyze measurements, such as current and voltage, to identify abnormal operating conditions. These abnormal states are typically detected by applying combinations of magnitude thresholds and time durations. For instance, overcurrent relays identify faults when the measured current exceeds a predefined threshold and persists for a specified period. Consequently, in the absence of cyberattack-detection mechanisms, a cyberattacker can deploy numerous strategies to achieve a successful compromise. Notably, the attacker does not need to faithfully reproduce realistic fault waveforms; any injected signal that violates the relay's threshold and timing criteria is sufficient to induce misoperation, irrespective of its waveform characteristics. For example, the adversary

could modify the configuration of a current measurement device or inject artificially elevated current values directly into the substation's local area network, thereby activating the overcurrent relay.

In this study, we utilize the IEEE Power System Relaying Committee (PSRC) D6 benchmark test system to illustrate cyberattacks targeting protective relays [47]. As shown in **Figure 4**, the system comprises three substations that connect a generating plant to the broader grid via four transmission lines. These lines are protected by a variety of protective relays, including distance, overcurrent, and differential protection schemes. The benchmark system is modeled using a co-simulation environment that integrates Riverbed Modeler with the OPAL-RT HYPERSIM real-time simulator.

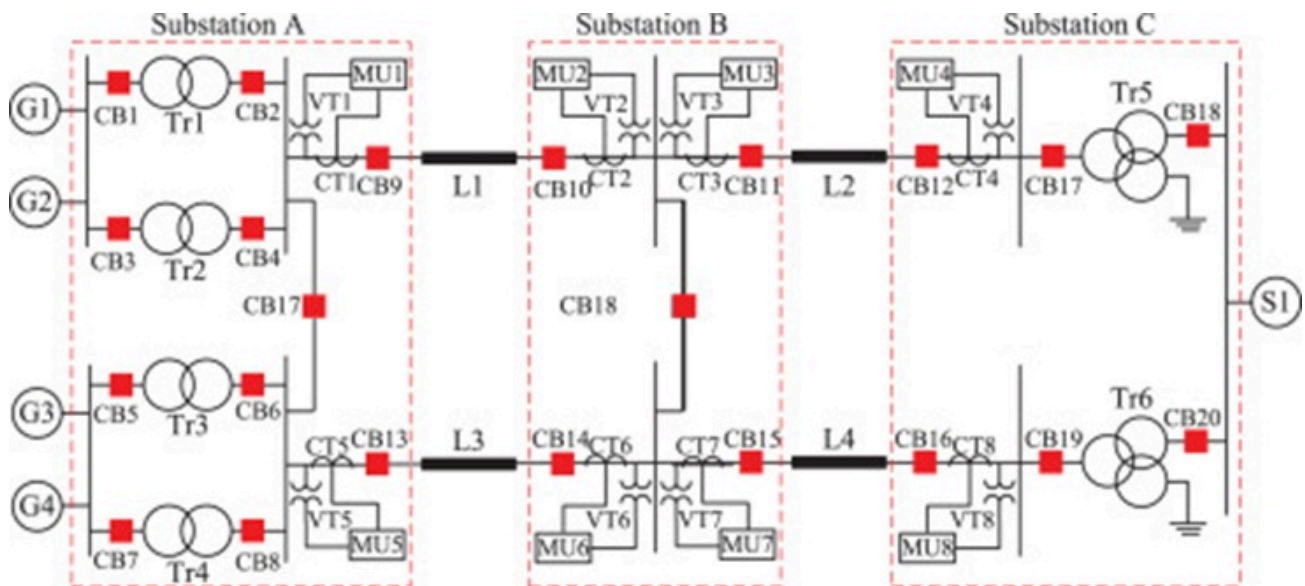


Figure 4. The IEEE PSRC D6 benchmark test system.

To highlight the susceptibility of protective relays without cyberattack detection, we implemented two distinct attack scenarios in this section. In the first scenario, we altered the current transformer settings to provoke misoperation of the overcurrent relay protecting a transmission line, as depicted in **Figure 5**. In the second scenario, we injected high-magnitude current values through the substation local area network, resulting in overcurrent relay malfunction, as illustrated in **Figure 6**. Both attacks successfully caused the corresponding transmission line to trip.

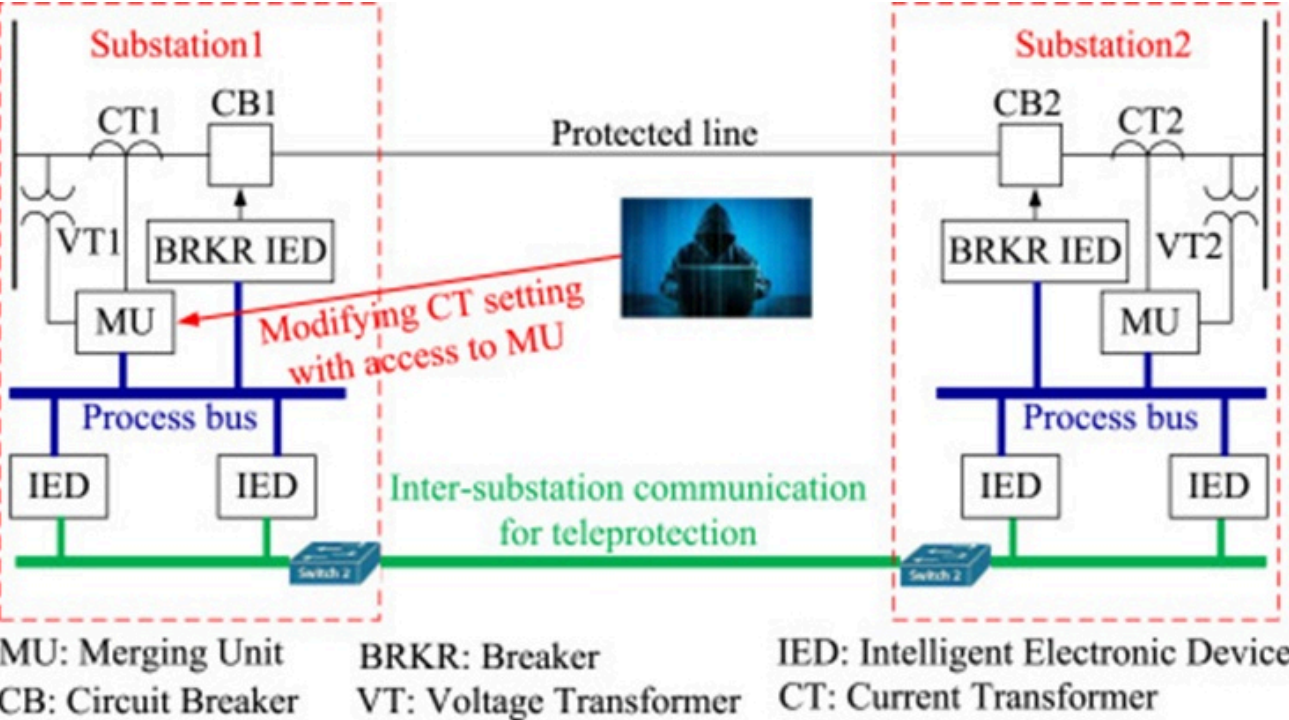


Figure 5. Malicious alteration of current transformer parameters via the merging unit interface.

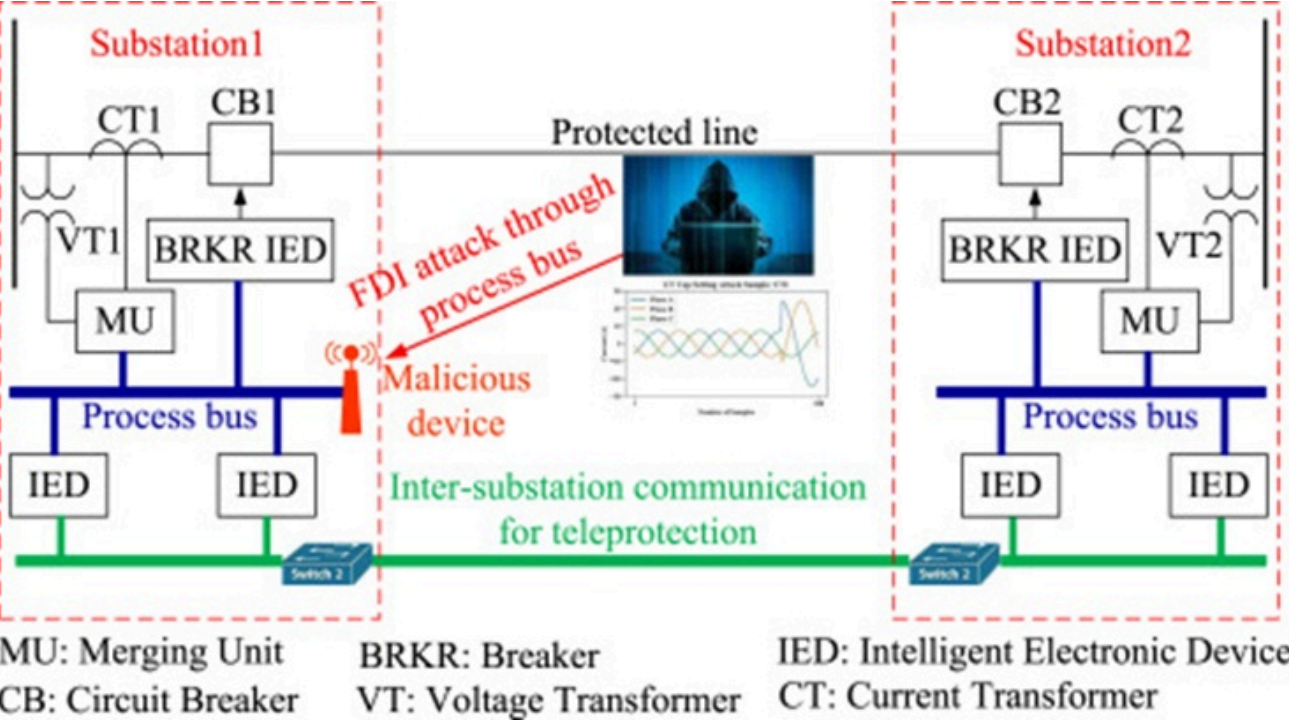


Figure 6. Execution of a false data injection attack across the process bus.

The increasing frequency and sophistication of cyberattacks on power systems in recent years have emphasized the critical need for effective cyberattack detection solutions.

Autoencoder-based cyberattack detection system for protective relays

The elusive, constantly evolving nature of modern cyberattacks has driven the adoption of machine learning-based detection systems in substations to counter these threats. Among these approaches, unsupervised autoencoder-based cyberattack detection systems stand out for their ability to identify previously unseen (zero-day) attacks. Additionally, they do not require an ongoing supply of evolving attack samples for training purposes.

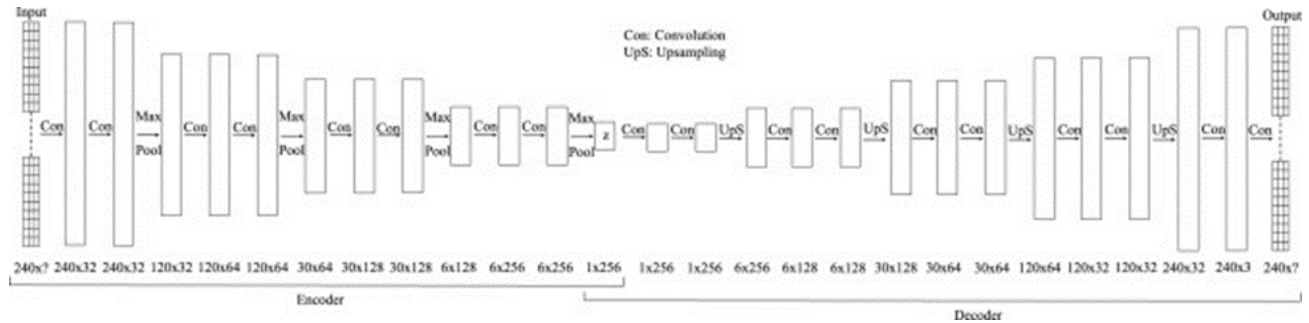


Figure 7. Architectural overview of the autoencoder-based cyberattack detection framework.

In this work, we implement a one-dimensional convolutional autoencoder originally proposed in [14], with its structure outlined in Figure 7. The model employs a convolutional filter size of 10, a stride of 1, and ReLU activation functions throughout. The autoencoder processes time-series sequences of current measurements corresponding to the overcurrent protective relay and raises an alarm upon detecting anomalous (cyberattack-related) behavior. Input data are presented to the detection system in fixed-size windows containing 240 consecutive measurement samples.

$$precision = \frac{\# True \quad Positive}{\# True \quad Positive + \# False \quad Positive} \quad (2)$$

$$recall = \frac{\# True \quad Positive}{\# True \quad Positive + \# False \quad Negative} \quad (3)$$

The performance of the cyberattack detection system is evaluated using standard precision and recall metrics. Instances of cyberattacks correctly identified by the system are counted as True Positives. Normal (benign) measurements incorrectly flagged as cyberattacks constitute False Positives. Undetected cyberattacks are classified as False Negatives, while correctly identified normal measurements are True Negatives. Let # denote the count of each category. Precision is then defined as the proportion of all attack predictions made by the model that are actually correct. Recall represents the proportion of genuine attacks that the system successfully detects (i.e., “recalls” as attacks).

The implemented autoencoder-based cyberattack detection system achieved perfect performance—100% precision and 100% recall—against the attack scenarios. Therefore, an adversary seeking to bypass this detection mechanism must either supply authentic historical fault recordings from the specific protection zone of the relay or resort to adversarial perturbation techniques. Acquiring genuine historical fault data is often impractical, however, since no fault event may have ever occurred within the relay’s protected zone despite the attacker’s access to measurement streams. The following section demonstrates how evasive attack algorithms enable the adversary to circumvent autoencoder-based cyberattack detection systems.

Evasive attacks

In this section, we assess the effectiveness of evasive attacks in circumventing autoencoder-based cyberattack detection systems protecting protective relays. The iterative algorithm for generating evasive attacks is applied here to construct those adversarial samples. The attacker requires an initial seed, denoted x_f , to initiate the crafting process using the proposed algorithm.

We use a simulation-based test scenario to generate the seed required by [Algorithm 1](#) for crafting evasive attacks. That said, alternative methods could equally be employed to obtain such a seed. For this analysis, we assume the attacker possesses only minimal knowledge of the victim system. Specifically, the attacker is aware only that the target is a generator substation, lacking any details on the generator's technical specifications, the substation layout, or the network's topology and impedance values.

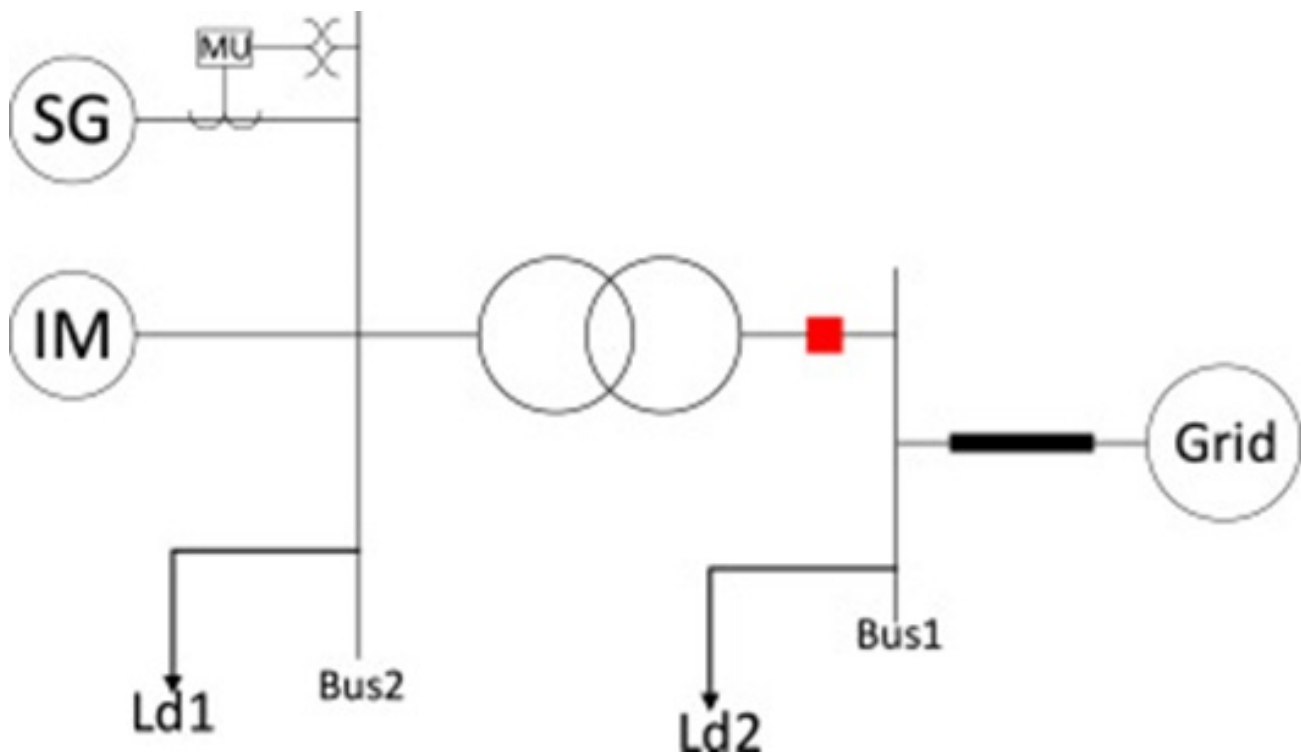


Figure 8. Test case of the simulation-based distribution system.

The simulation test case leveraged by the attacker in our experiments is depicted in [Figure 8](#). This scenario is derived from a demonstration model built in MATLAB Simulink by G. Sybille and T. Zabaoui. Within the test case, a 3.125-MVA synchronous generator (SG) supplies a 2400-V distribution network that powers a 2250-hp induction motor (IM) and a resistive load (Ld1). The synchronous generator is fitted with a non-reheat thermal turbine, governor, automatic generation control (AGC), and an AC1A excitation system. The 2400-V distribution feeder is further interconnected to a 25-kV grid via a 6 MVA step-up transformer, with the 25-kV network represented as an ideal voltage source in series with an R-L impedance. A side-by-side comparison of fault measurements from both the actual target power system and the simulation-based test case used to derive the initial attack seed is provided in [Figure 9](#).

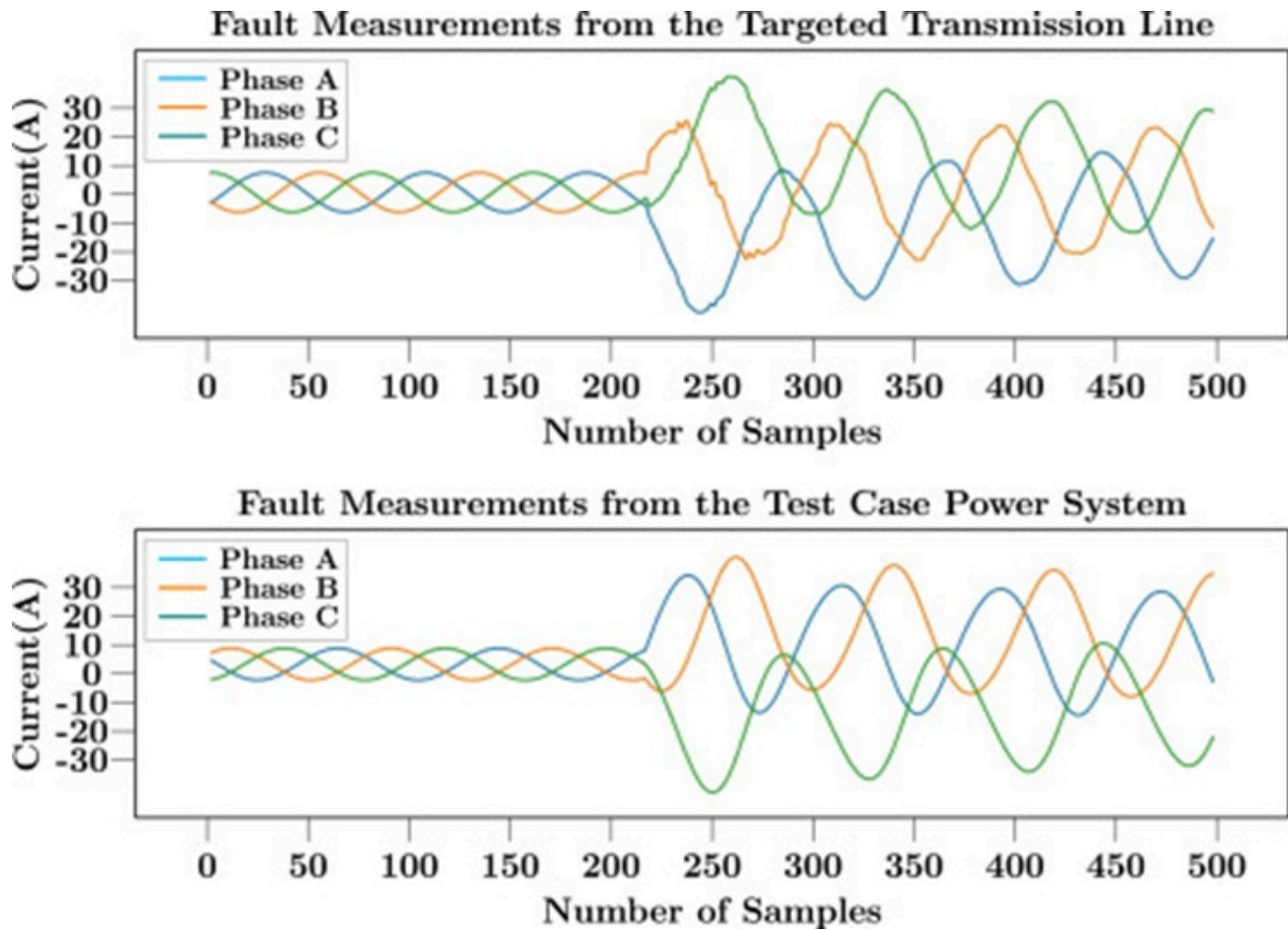


Figure 9. Fault measurements recorded from the actual targeted power system (top) alongside those from the test-case system used to create the initial attack seed (bottom).

It is important to stress that the simulation test case employed by the attacker to create the initial seed differs substantially from the actual target system shown in [Figure 4](#) in multiple respects. First, the generator ratings and parameters vary between the two systems. Second, the exciter, governor, and turbine configurations differ. Third, the voltage levels and network topologies differ. Due to these mismatches, the raw initial seed data generated by this test case cannot evade the autoencoder-based cyberattack detection system trained on the target system in [Figure 4](#) unless adversarial perturbation techniques are applied.

Using [Algorithm 1](#), the attacker can synthesize artificial evasive samples designed to mimic legitimate measurements from the target substation. Note that the cyberattack detection system processes inputs in fixed windows of 240 measurement samples. Consequently, each output generated by [Algorithm 1](#) consists of exactly 240 samples. Provided the mean squared error (MSE) target in [Algorithm 1](#) is set below the detector's established threshold, each of these standalone evasive windows is assured to bypass detection. An example segment of such evasive data appears in [Figure 10](#).

In practice, however, the attacker must supply the overcurrent relay with a continuous stream of measurement values rather than isolated, non-overlapping windows. This necessitates concatenating the individually crafted evasive windows into one cohesive final evasive trace. A naive concatenation approach—retaining only the initial non-overlapping portion of each window and discarding overlapping segments—turns out to be ineffective. When independently generated evasive windows are stitched together, abrupt discontinuities emerge in the resulting composite sample, which the cyberattack detection system can readily identify as anomalous.

To overcome this issue, the window stride used during evasive sample generation should be minimized to reduce visible discontinuities in the final concatenated trace. In our experiments, adopting a stride of 1 yielded the smoothest transitions with the least discontinuity. Additionally, we incorporated gradient averaging across overlapping regions during each iteration of the evasive sample update process, thereby improving the overall success rate of the joined evasive sample and further lowering its reconstruction error. Specifically, in Step 7 of [Algorithm 1](#), each data point receives an update based on the averaged gradient across all windows that include it, effectively smoothing transitions across overlapping areas.

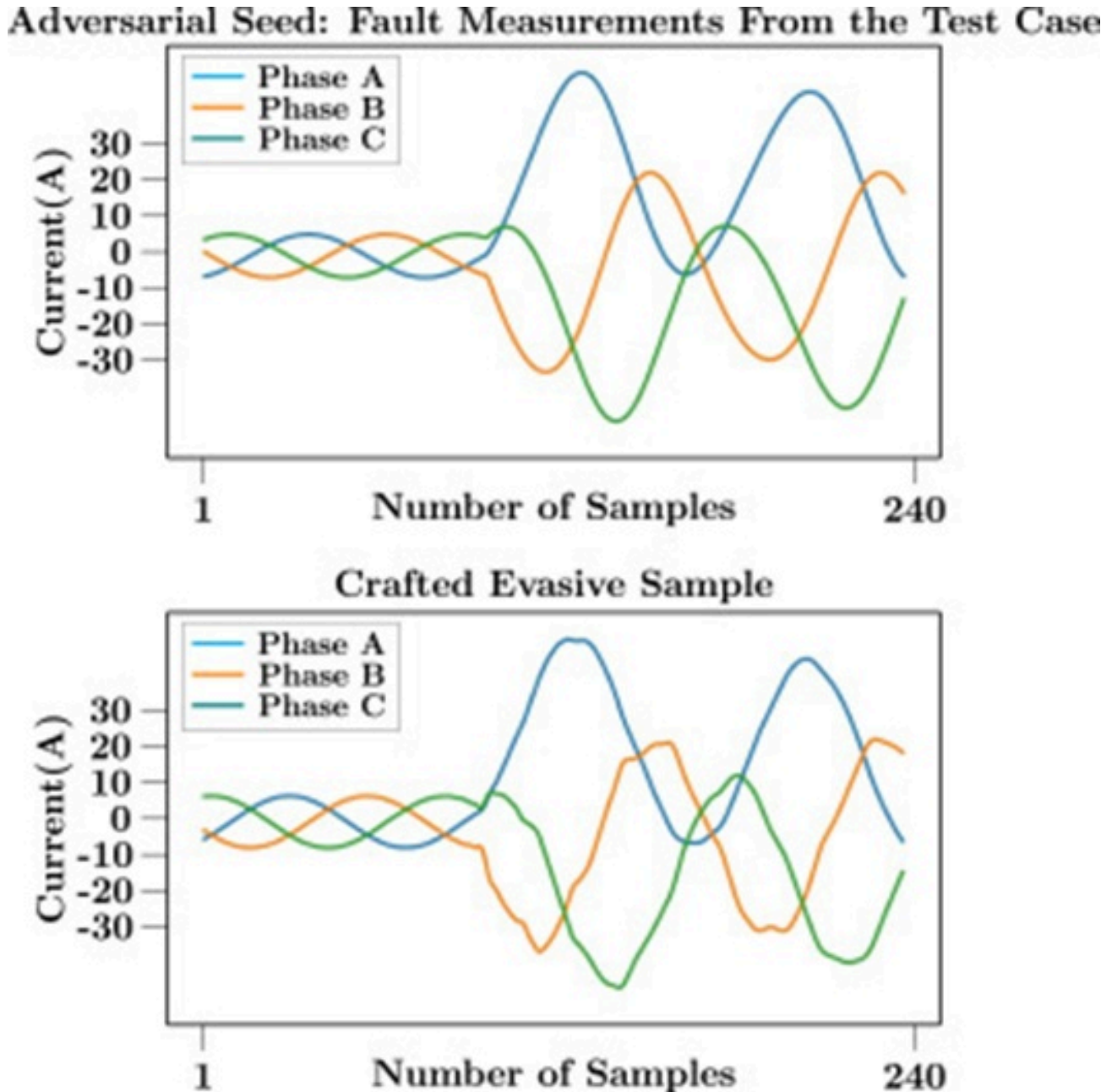


Figure 10. A segment of the initial adversarial seed (top) and the resulting evasive attack sample generated via [Algorithm 1](#) (bottom).

The evasive attack constructed from the seed generated by the test system in [Figure 8](#), processed through [Algorithm 1](#) and the aforementioned refinement techniques, achieved a 100% evasion success rate against the autoencoder-based

cyberattack detection system. An illustrative example of the final joined evasive sample is presented in **Figure 11**. These findings confirm that the proposed **Algorithm 1** can reliably generate evasive attacks capable of defeating autoencoder-based cyberattack detectors in substation environments.

Importantly, evasive attacks cannot be successfully crafted from naive initial seeds. To illustrate this limitation, we generated an adversarial seed using scaled fault-free measurement data from the target power system—analogue to the scenario shown in **Figure 5**. Although **Algorithm 1** reduced the reconstruction error, it failed to produce a viable evasive attack, as demonstrated in **Figure 12**.

Some prior work has suggested that augmenting training datasets with synthetically generated evasive (adversarial) samples can effectively defend classifier-based models against adversarial threats—for instance, by incorporating adversarial examples during training, as explored in Ref. [48]. Nevertheless, such defensive strategies are less straightforward in unsupervised learning contexts and typically require further modifications, since unsupervised models do not rely on labeled training data.

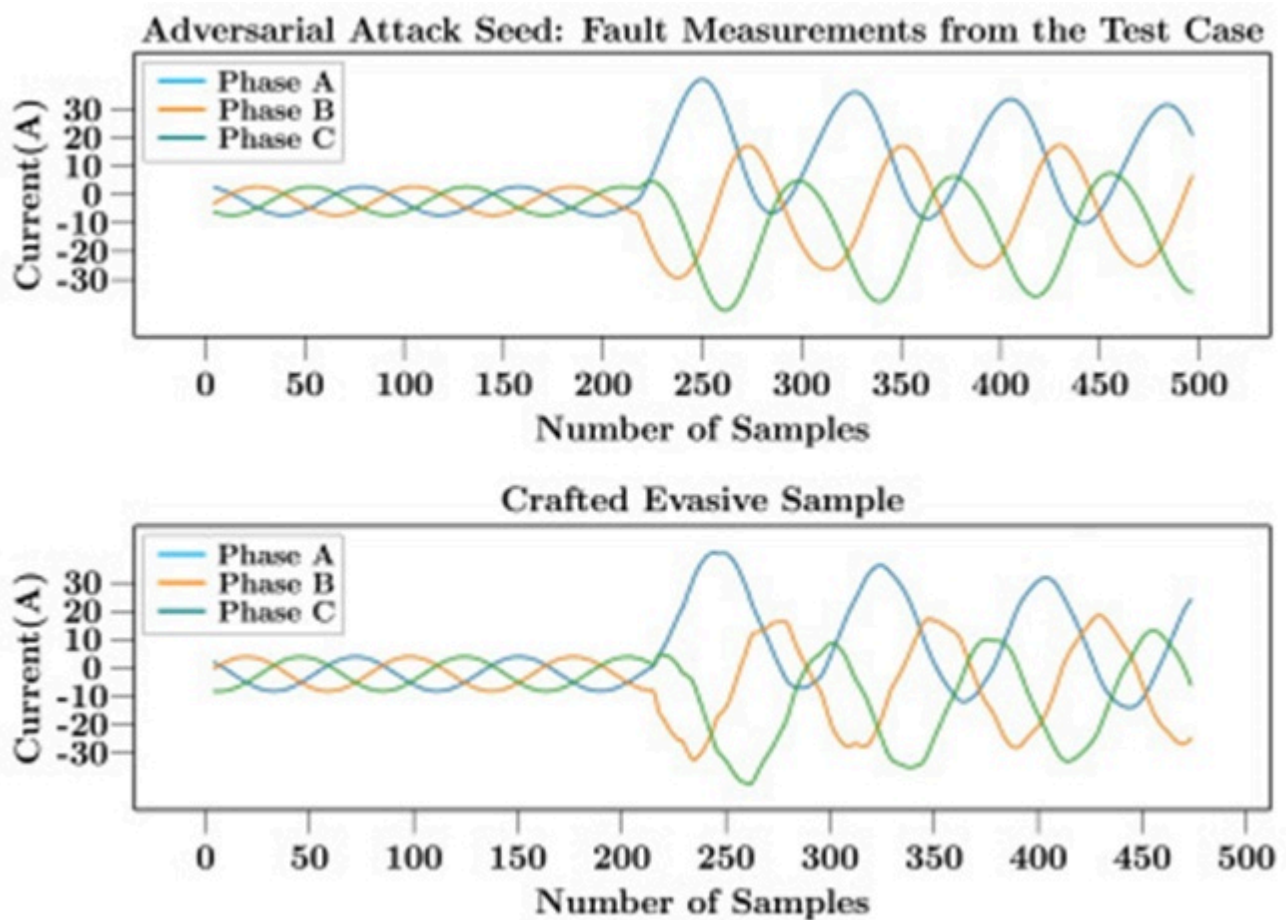


Figure 11. The original adversarial seed (top) and the concatenated evasive attack sample produced by the proposed method (bottom).

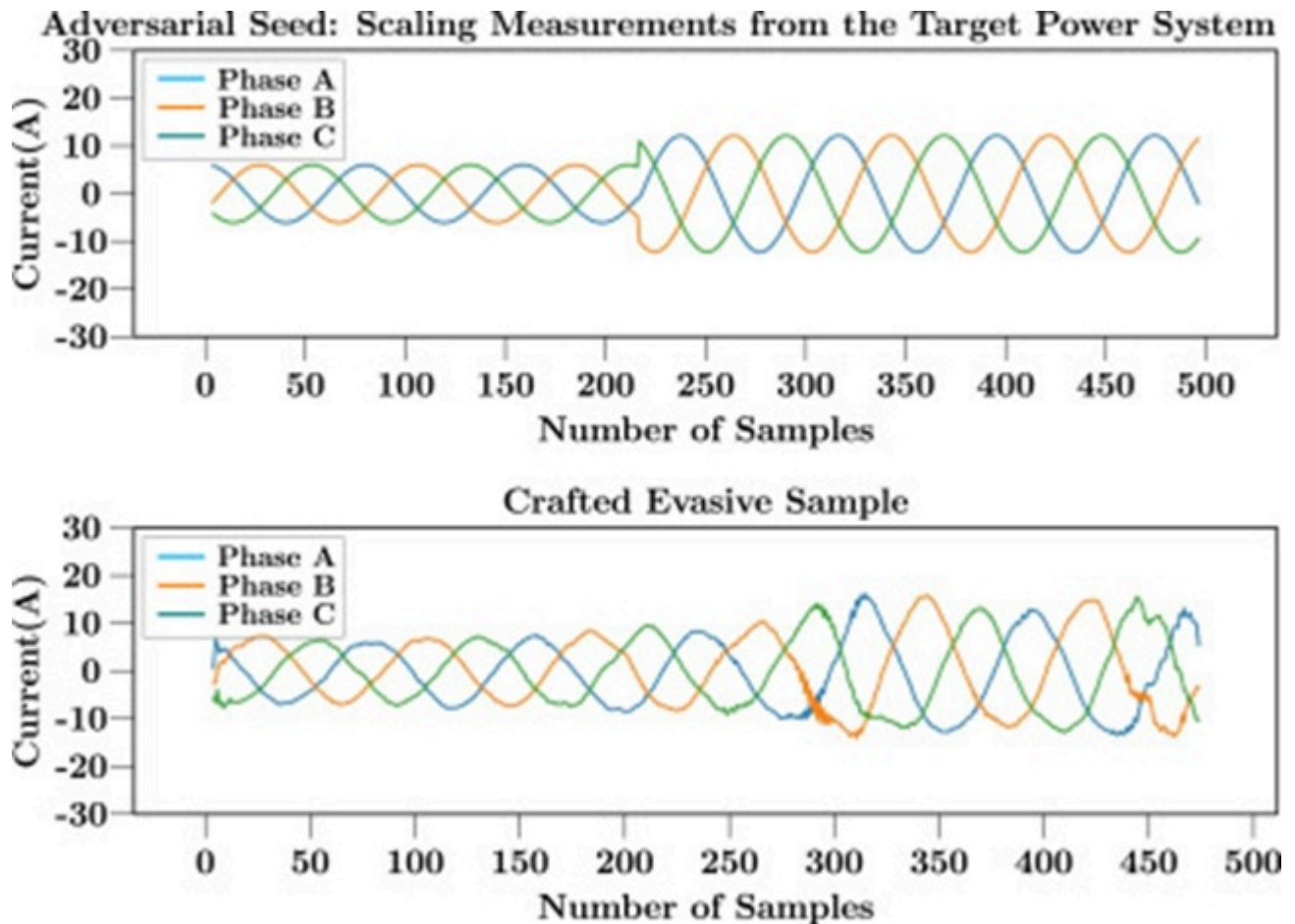


Figure 12. Scaled fault-free measurements serving as the initial adversarial seed (top) and the resulting concatenated evasive attack (bottom).

Conclusion

This study examined the susceptibility of autoencoder-based cyberattack detection systems deployed in digital substations to adversarial attacks. We introduce an original iterative algorithm specifically designed to generate evasive attacks targeting autoencoder-based cyberattack detection mechanisms. Our results show that an adversary equipped with white-box access to the autoencoder-based detection system—combined with only partial knowledge of the target substation—can successfully produce an evasive sample that both evades detection and induces a tangible physical effect. By contrast, the same adversary cannot achieve evasion when using naive initial seeds. These observations underscore the critical need to protect both substation-specific operational details and the architectural and parametric configuration of autoencoder-based cyberattack detection systems from unauthorized disclosure.

Acknowledgements

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.

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 26 Mar 2024

Revised: 08 May 2024

Accepted: 20 Jun 2024

Published online: 20 September 2024

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- 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.
- Industrial Control Systems Cyber Emergency Response Team (ICS-CERT) Cyberattack Against Ukrainian Critical Infrastructure. 2016.
- North American Electric Reliability Corporation (NERC) Critical Infrastructure Protection (CIP) Reliability Standards. 2019.
- 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.
- 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;97:101994.
- 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.
- 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.
- 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.
- 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.

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.

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.

Ahmed A, Krishnan VVG, Foroutan SA, Touhiduzzaman M, 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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

Li J, Yang Y, Sun JS. Exploiting vulnerabilities of deep learning-based energy theft detection in AMI through adversarial attacks. 2020.

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.

Sabounchi M, Wei-Kocsis J. A practical adversarial attack on contingency detection of smart energy systems. 2021.

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.

Liu T, Shu T. On the security of ANN-based AC state estimation in smart grid. *Comput Secur*. 2021;105:102253.

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.

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.

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.

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.

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.

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.

Rubinstein BIP, Nelson B, Huang L, Joseph AD, Lau S, Rao S, et al. ANTITODE: Understanding and defending against poisoning of anomaly detectors. In: Proc th ACM SIGCOMM Conference on Internet Measurement, Association for Computing Machinery. 2009.

Shafahi A, Huang WR, Najibi M, Suci O, Studer C, Dumitras T, et al. Poison frogs! Targeted clean-label poisoning attacks on neural networks. In: Proc 32nd conference on neural information processing systems (NIPS). Montreal; 2018.

Xu W, Qi Y, Evans D. Automatically evading classifiers. In: Network and distributed system security symposium (NDSS). San Diego; 2016.

Biggio B, Corona I, Maiorca D, Nelson B, Šrđić N, Laskov P, et al. Evasion attacks against machine learning at test time. *Mach Learn Knowl Discov Databases.* 2013;8190:387-402.

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow IJ, et al. Intriguing properties of neural networks. 2013.

Moosavi-Dezfooli S, Fawzi A, Frossard P. DeepFool: a simple and accurate method to fool deep neural networks. 2015.

Papernot N, McDaniel P, Jha S, Fredrikson M, Berkay Celik Z, Swami A. The limitations of deep learning in adversarial settings. 2015.

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.

Kong X, Ge Z. Adversarial attacks on neural-network-based soft sensors: Directly attack output. *IEEE Trans Ind Inf.* 2022;18(4):2443-51.

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;140:103786.

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;34:100452.

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;58:102717.

Kravchik M, Shabtai A. Efficient cyber attack detection in industrial control systems using lightweight neural networks and PCA. *IEEE Trans Dependable Secure Comput.* 2021;19(5):3179-97.

Gras H, Mahseredjian J, Rutovic E, Karaagac U, Haddadi A, Saad O, et al. 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.

Goodfellow I, Schlenz J, Szegedy C. Explaining and harnessing adversarial examples. *arXiv [Preprint]*. 2015. Available from: [arXiv:1412.6572v3](https://arxiv.org/abs/1412.6572v3).