Attacking Learning-based Models in Smart Grids: Current Challenges and New Frontiers

Gustavo Sánchez
KASTEL Security Research Labs,
Karlsruhe Institute of Technology
Eggenstein-Leopoldshafen, Germany
sanchez@kit.edu

Ghada Elbez
KASTEL Security Research Labs,
Karlsruhe Institute of Technology
Eggenstein-Leopoldshafen, Germany
ghada.elbez@kit.edu

Veit Hagenmeyer
KASTEL Security Research Labs,
Karlsruhe Institute of Technology
Eggenstein-Leopoldshafen, Germany
veit.hagenmeyer@kit.edu

ABSTRACT

Learning-based components applied to a plethora of use cases within smart grids are already a reality. These methods will undoubtedly play a key role in future energy systems. This paper addresses challenges in the field of adversarial attacks against learning-based models in the context of smart grids. We identify unexplored areas and potential improvements in current methodologies by categorizing attacks, and assessing their ability to be reproduced. Our survey showed a noticeable resistance to distributing experimental code. Additionally, we propose the integration of explainable artificial intelligence techniques into adversarial models. We carry out an initial experiment to showcase the possible effects of this integration, offering fresh perspectives on the behavior and vulnerabilities of learning-based models within smart grids. Our initial findings provide a basis for further investigation into adversarial attacks, with a special focus on use cases that affect electrical substation security. Finally, we outline the next steps of our research in this critical area.

CCS CONCEPTS

• **Security and privacy**; • **Hardware** → **Power and energy**; • **Computing methodologies** → **Artificial intelligence**;

KEYWORDS

Security, Smart Grid, Adversarial Machine Learning, IEC 61850.

ACM Reference Format:
Gustavo Sánchez, Ghada Elbez, and Veit Hagenmeyer. 2024. Attacking Learning-based Models in Smart Grids: Current Challenges and New Frontiers. In *The 15th ACM International Conference on Future and Sustainable Energy Systems (E-Energy '24), June 04–07, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3632775.3661984

1 INTRODUCTION

In the realm of critical infrastructure, the traditional approach of reactive security is often inadequate, because such measures are implemented after a security incident has already taken place [12]. Fortunately, many cybersecurity researchers and practitioners are dedicating their efforts to creating proactive cyber defense methods, in which future attack strategies are anticipated and these insights are incorporated into defense designs. As a result, research in offensive security has become crucial. This is particularly true for Smart Grids (SGs), where the integration of learning-based components introduces specific vulnerabilities [52]. These vulnerabilities must be thoroughly examined following the well-established "three golden rules" [7] of security in Machine Learning (ML): understanding the adversary, adopting a proactive stance, and implementing self-protection measures.

Attackers exploit weaknesses in learning models by conducting adversarial attacks. Such attacks, a refined form of False Data Injection (FDI), specifically target the susceptibilities of intelligent algorithms [38]. Adversarial attacks encompass a broad range of tactics, each with its own distinct characteristics and implications.

In light of this, eXplainable Artificial Intelligence (XAI) [5] methods have the potential to play an important role in fostering trust and clarity in algorithmic decisions in power systems. Despite their benefits, such as making intelligent systems more transparent and understandable to humans, the insights provided by XAI could be misused for nefarious purposes.

SGs represent controlled environments where data-driven techniques are increasingly being adopted as effective solutions for a variety of operational tasks. Notable applications include short-term load forecasting [23] and the detection of FDI [21]. Despite the prevalence of such studies, there remains a lack of detailed analysis concerning the resilience of these methods to adversarial attacks. Moreover, exploring adversarial attacks in SGs is not only about understanding attacker's strategies and goals, but also about developing and recommending effective defensive measures to counteract them.

Outside the power systems domain, attacks against learning-based methods is a wide research area. This field began to evolve with the influential work of Dalvi *et al.* [13] in 2004, which explored methods to circumvent learning-based email spam filters. More recent research has predominantly focused on adversarial perturbations in visual and auditory data, as seen in studies pertaining to image [9] and audio [10] domains.

Adversarial strategies against learning-based models vary considerably across different domains. Each domain presents unique challenges and requires domain-specific expertise and thorough analysis to understand the feasibility and impact of adversarial approaches [38]. This requires the development of customized strategies and assessments for every distinct area of application.

Contributions: In this paper, our goal is to enhance the community's comprehension of the challenges, research gaps, and future

E-Energy '24, June 04–07, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0480-2/24/06
https://doi.org/10.1145/3632775.3661984

directions concerning adversarial attacks, specifically those applicable to SGs. The main contributions can be summarized as follows.

- We categorize existing work on adversarial attacks against learning-based components in SGs (34 papers), and conduct a detailed assessment of the reproducibility of results.
- We show a correlation between the Confidentiality, Integrity and Availability (CIA) Triad and the Authentication, Authorization, and Accounting (AAA) security framework.
- Additionally, we explore the use of XAI as a potential tool for attackers, analyzing how it could be misused in the context of SGs to facilitate adversarial attacks.
- Lastly, we identify current research gaps and future challenges, paving the way for more robust and secure learning-based systems for SGs, with special emphasis on use cases related to electrical substations.

2 TAXONOMY OF ATTACKS

In this section, we describe the different types of adversarial attacks and their implications on SGs.

2.1 Threat Models

Firstly, it is important to understand the existing threat models of adversarial attacks.

Attacker's Goal. In the context of SGs, comprehending the potential objectives of attackers is critical for ensuring robust system security. The CIA triad, a widely recognized model in information technology security, provides a framework for understanding these objectives. Reflecting on these three security violation grounds: *confidentiality* could be compromised when an attacker maliciously interacts with a learning algorithm with the objective of reverse-engineering it; *integrity* is compromised when performance is impacted without affecting normal operation; compromising *availability* refers to making the normal learning-based system functionalities unavailable to legitimate users.

Attacker's Knowledge. According to the information available to the attacker at the time of inception, there exist three main paradigms. *White box*: In the context of SGs, this represents an extreme, worst-case scenario, often linked to insider threats. It is critical for testing the resilience of learning-based systems against those who have full access to their inner workings. *Grey Box*: This paradigm reflects a more common scenario in real-world applications, where some system details might be obtained or publicly known. *Black Box*: Assessing the resilience of SGs against black box attacks is crucial, as it represents the common challenge of defending against external threats with minimal system information. In SGs, it is important to investigate all three categories, motivated by the fact that *security by obscurity* is not reasonable in this context, i.e., it is not good practice to expect security by code secrecy.

Attacker's Capability. Depending on the phase that an attacker influences the algorithm (i.e., during training or test time) there are different naming conventions to classify approaches. Identifying available transformations, preserving semantics, ensuring robustness to pre-processing and general plausibility are problem-space constraints [33] that represent major challenges to attackers. These constraints outline the boundaries within which attackers operate, and highlight the complex nature of securing learning-based systems in the critical infrastructure of SGs.

2.2 Types of Attack

The goal of attacks against integrity is to cause incorrect predictions that do not compromise normal system operation. These attacks, primarily in the form of adversarial examples (or evasion attacks) and data poisoning, pose a significant threat. Adversarial examples, introduced during the model's deployment phase, are engineered to mislead the SG's decision-making algorithms, which are critical for real-time monitoring and control. Data poisoning, targeting the model during its training phase, degrades the model's performance, potentially leading to situations such as flawed energy demand predictions or intrusion detection failures, which could have cascading effects on grid stability. Indeed, an attacker able to tamper with training data can rely on these manipulations to deceive the model with carefully crafted elements at inference time. Poisoning at test time, by targeting test-time adaptation methods that perform continuous fine-tuning of the target model, is also a possibility.

Attacks on SGs aiming to compromise availability are particularly disruptive. Sponge attacks, executed at the model's inference time, strain the computational resources, hindering the SG's ability to process legitimate operational data efficiently. These attacks have the potential to delay critical responses to grid conditions, among other effects. Indiscriminate poisoning, by corrupting training data, undermines the model's predictive accuracy, rendering it less effective in real-time applications. Sponge poisoning further exacerbates this by making the model resource-intensive, thus diminishing its operational efficiency and responsiveness in critical situations.

Confidentiality attacks in SGs involve sophisticated techniques to extract sensitive information, posing risks to both consumer privacy and proprietary system data. Model stealing attacks replicate the functionality of SG models, potentially revealing proprietary algorithms that are vital for grid security and efficiency. Model inversion attacks pose a direct threat to data privacy by extracting detailed patterns from the model's outputs. Likewise, membership inference attacks can further compromise data privacy by determining if specific data points were used in training models, revealing potentially sensitive behavior patterns. These attacks not only threaten the privacy of the SG's operational data but also risk violating consumer trust and regulatory compliance.

2.3 Mapping CIA and AAA

By mapping the attacker's capabilities and goals, we obtain a comprehensive overview of the different attacks against learning-based methods, as depicted in Table 1. Furthermore, the AAA framework addresses the main attributes of policy enforcement and access control to resources. AAA was designed to be applied in network security, but the principles provide a useful lens in the context of data governance within data-driven methods in SGs.

Authentication in ML security involves ensuring the legitimacy of SG data used for training and inference. *Authorization* relates to enforcing what data can influence the model and who can access the model's predictions and knowledge. *Accounting* involves monitoring and logging model access and usage, which is crucial

for detecting and responding to attacks. The integration of AAA is also included in Table 1.

Table 1: Categorization of attacks against learning-based methods in SGs according to the CIA Triad and AAA Framework. Legend: Authentication (Δ), Authorization (Ω), Accounting (Σ).

	Integrity	Availability	Confidentiality
Test Data	Evasion/ Adversarial examples, Test-time Poisoning (Δ)	Sponge Attack (Δ)	Model Extraction and Inversion, Membership Inference (Ω)(Σ)
Training Data	Poisoning (e.g., backdoors or trojans) (Δ) (Ω)	Indiscriminate Poisoning (i.e., DoS), Sponge Poisoning (Δ) (Ω)	Model Inversion with Poisoning (Δ) (Ω) (Σ)

For attacks that involve poisoning, an authorization step that addresses what data can influence the model would increase robustness. In relation to authentication, data governance measures should ensure that input data is legitimate with techniques such as validation and pre-processing, what would contribute towards increasing robustness against both evasion and poisoning attacks. Additionally, resource-aware authentication measures would contribute towards ensuring availability of learning-based models. Attacks against data confidentiality should be avoided by authorization policies that control access to the model's predictions and knowledge. Furthermore, these breaches are related to the system's accountability mechanisms; potential deficiencies in tracking and auditing access to and usage of the data facilitate attack success.

However, these measures are not always considered in SGs and, therefore, adversaries take advantage.

3 STATE-OF-THE-ART IN SMART GRIDS

In this section we present our survey, with a focus on works that made their data available. The complete analysis of the state-of-the-art of scientific research in adversarial attacks within SGs can be consulted in Table 2. The reproducibility factors are adapted from [31].

Reproducibility of Results. We attempted to evaluate the consistency of the claims made in the state-of-the-art by reproducing their results. In the repository provided in [11] there are files missing, what prevented us from checking the validity of their results (e.g., *data_all.csv* in [11]). In [27] there are missing modules (e.g., *cleverhans_copy.utils*), as well as missing instructions on the execution order and dissimilarity between files (e.g., *attacks.py* and *Attacks.py*). Code instructions are present for [43] and [29] (while partially described in [11]). Additionally, errors occur when trying to import a component from a library that has been relocated or renamed in a newer version; this would be solved by specifying library version requirements. Adversarial attack code is provided in [11, 27, 29, 44]. Works [1, 6, 15–17, 20, 25, 34, 34, 43, 49, 53, 54] provide pseudo-code of their proposed attack(s) but not an implementation. In the studies [36] and [29], it is mentioned that the code and artifacts would be released upon request. Upon obtaining access to the resources from [29] and subsequent execution, it was observed that their claims were consistent. On the contrary, we contacted the corresponding author of [36], who was unable to make data available.

Authors in [43] do not provide in-depth information about target models due to relying on previous work; we consider this practice acceptable as far as the referenced work is reproducible. In the case of [43], it was possible to execute the model used [50] as a target for adversarial attacks only after several code modifications for compatibility (*keras*, *tensorflow* and *matplotlib* scripts), as the library versions used were not specified. Due to the library issues, results were slightly different although the same random seed was used. In [44], there are files missing (e.g., *20191014universal_pert_5000_1_2.0.npy* and *confusion_matrix_SAA.npy*), and dependencies issues (e.g., *cannot import name 'cast' from partially*

Table 2: Survey of existing scientific work. Legend: ●Fully met, ◐Partially, ○Missing, - Not Applicable.

Ref	Reproducibility										Attack			Other		
	Target Model	Hyper-parameters	Training Info	Dataset Available	Data Split Info	Source Code	Code Instructions	Adv. Attack Code	Code Works	Claims Consistent	Integrity	Availability	Confidentiality	Use of XAI	White/Black/Grey	Publication Year
[25]	●	○	◐	○	○	○	-	◐	-	-	●	○	○	○	W	'19
[11]	●	●	●	●	●	●	◐	●	○	-	●	○	●	○	W/B	'19
[53]	●	●	●	○	●	○	-	◐	-	-	●	○	○	○	W	'19
[37]	●	◐	○	●	○	○	-	-	-	-	●	○	○	○	W	'20
[27]	●	●	●	●	●	●	○	●	○	-	●	○	●	○	W/B	'20
[15]	●	●	●	●	-	○	-	◐	-	-	●	○	○	○	G	'21
[42]	●	●	●	●	●	○	-	○	-	-	○	●	○	○	B	'21
[40]	●	●	●	●	●	○	-	○	-	-	●	○	○	○	W	'21
[38]	●	●	●	●	●	○	-	○	-	-	●	○	○	○	W/B	'21
[19]	●	◐	◐	○	◐	○	-	○	-	-	●	○	○	○	W/B	'21
[36]	●	●	◐	○	○	○	-	○	-	-	●	○	○	○	G	'21
[17]	●	●	●	◐	●	○	-	◐	-	-	●	○	○	○	W/B	'21
[45]	●	●	●	○	●	○	-	◐	-	-	●	○	○	○	W	'21
[32]	●	●	●	●	●	○	-	○	-	-	●	○	○	○	W/B	'21
[34]	●	●	●	○	○	○	-	◐	-	-	●	○	○	○	W/B	'22
[44]	●	●	●	●	●	●	○	●	○	-	●	○	●	○	B	'22
[6]	●	●	●	●	●	○	-	◐	-	-	●	○	○	○	W/B	'22
[49]	●	●	●	◐	●	○	-	◐	-	-	●	○	●	○	W/B	'22
[43]	●	●	●	●	◐	●	●	◐	○	-	●	○	○	○	G	'22
[41]	●	●	●	●	●	○	-	○	-	-	●	○	○	○	W/B/G	'22
[35]	●	◐	◐	○	○	○	-	○	-	-	●	○	○	○	B	'22
[48]	●	●	●	●	●	○	-	○	-	-	●	○	○	○	B	'22
[18]	●	○	●	●	●	●	●	○	○	-	●	○	○	○	W	'22
[8]	●	●	●	●	○	○	-	○	-	-	●	○	○	○	W	'23
[16]	●	●	●	◐	○	○	-	◐	-	-	●	○	○	○	B	'23
[39]	●	●	●	○	●	○	-	○	-	-	○	●	○	○	B	'23
[3]	●	◐	●	○	○	○	-	○	-	-	●	○	○	○	W	'23
[20]	●	●	●	○	○	○	-	◐	-	-	●	○	○	○	G	'23
[54]	●	●	●	●	○	○	-	◐	-	-	●	○	○	○	G	'23
[1]	●	◐	●	◐	●	○	-	◐	-	-	●	○	○	○	B	'23
[30]	●	○	●	○	○	○	-	○	-	-	●	○	○	○	W	'23
[29]	●	●	●	●	●	●	●	●	●	●	●	○	○	●	W	'23
[46]	●	●	●	○	○	○	-	○	-	-	●	○	○	○	G	'23
[51]	●	●	●	●	○	○	-	○	-	-	●	○	○	○	B	'24

initialized module 'keras.src.backend'). In [18], after following the specified execution instructions by their cited source [2], pulling data from the cloud failed.

Several papers [6, 11, 27, 29, 38, 40–42, 44] provide comprehensive details about the training stage of learning-based methods. On the other hand, papers [18, 25, 30] do not provide information about training hyperparameters. Papers [1, 3, 19, 35, 37] provide partial hyperparameter information.

Attack Types. We observed that most published papers investigate attacks compromising the integrity of models, revealing a potential research gap in studying adversarial efforts against the availability and confidentiality of learning-based models.

Authors in [42] and [39] deal with attacks against availability. In [42], it is a reported an analysis on the impact of indiscriminate data poisoning attacks, and how to detect them within the topic of electricity theft. By significantly lowering the accuracy and reliability of the learning models, these attacks have the potential to render the system ineffective for its intended purpose, essentially making the service unavailable or less available to its users. In [39], we observe a similar approach, where authors investigate indiscriminate poisoning through different injection levels in the context of FDI detection. Works [11, 27, 44, 49] compromise confidentiality via the use of surrogate models. In these cases, the objective is to enhance integrity attacks by testing more realistic attacks scenarios (i.e., that do not require access to the inner workings of the models). This is achieved via targeting a substitute model, and then transferring the attacks. However, these approaches assume that attackers have access to data for training testbed models similar to the real targets.

In terms of the amount of information accessible to the attackers, authors in [6, 11, 17, 19, 27, 32, 34, 38, 49] provide threat models based on both white and black box scenarios; most notably, [41] present white, black and grey box attack approaches. In the rest of the papers, the focus is solely on a single scenario, i.e., either white (10), black (8) or grey (6) box.

Learning-based methods. The vast majority of papers leverage ML and Deep Learning (DL) algorithms, and use tabular data for training. We encountered a number of papers investigating adversarial attacks against Reinforcement Learning (RL) models in SG use cases [2, 16, 18, 32, 35, 45, 46, 48]. In RL, an agent learns to make decisions by interacting with an environment to achieve a goal; in Table 2, for RL papers the column *Dataset Available* relates to the availability of this whole environment. Furthermore, in [51], authors present an investigation that focuses on attacks against computer vision applied to SGs, such as object recognition and defect detection tasks. Another outlier is presented in [15], where authors employ Natural Language Processing (NLP) and describe a sentence-level text adversarial attack algorithm, evaluated in the context of a SG based on industrial Internet-of-Things.

Other Findings. Only one of the surveyed papers consider XAI in their methodology. Authors in [29] leverage XAI to identify the two most relevant features used by a learning-based Intrusion Detection System (IDS). By focusing their adversarial efforts on adding perturbations exclusively to these features, they attempt to optimize their evasion capabilities against the IDS.

4 ADVERSARIAL EXPLAINABILITY

XAI techniques are currently used in domains such as computer vision [24] to detect the presence of adversarial directions in images. For instance, saliency maps [47] have been long used for detecting adversarial perturbations in the ML literature. However, we envision the use of XAI for crafting more powerful adversarial attacks against learning-based models based on heterogeneous, tabular data from SGs. This section presents an initial experiment and analysis of how XAI can be utilized from an adversarial perspective.

Experimental Setup. We explore an attacker's potential to analyze a target model, either directly (in a white-box approach) or through a surrogate model (black-box approach) due to the transferability of attacks [14]. We use the code and dataset from [29]. The dataset is from Modbus TCP traffic, used for intrusion detection in a electrical substation testbed [22]. We focus on the proposed Random Forest (RF) model, chosen for its complexity compared to linear SVMs.

SHAP Summary Plot. We employ SHapley Additive exPlanations (SHAP) [26] to produce a summary plot (see Figure 1). This type of plot is used to show the contribution of each feature to the output of the model. The plot uses a *bee swarm* style to display the density of the points, avoiding overlaps in order to see each point clearly. Each row represents a feature from the dataset. Features are ranked by their importance, which can be inferred by the spread and color intensity of the points. The X-axis represents the SHAP value for each feature. This value indicates the impact of a feature on the model's output. A higher absolute SHAP value means a higher impact on the model output. Points placed to the right of the vertical line (zero impact) indicate a positive impact on the model output, while points to the left indicate a negative impact. The color of the points represents the value of the feature (not the SHAP value). In Figure 1, *High* feature values are colored in pink and *Low* feature values are colored in blue. This means that high values of a feature tend to push the model output higher if the SHAP value is positive or lower if the SHAP value is negative. The spread of the points along the X-axis shows the distribution of the impacts each feature has across the data. A wide spread means the feature has varied effects depending on the context (other feature values in the vector).

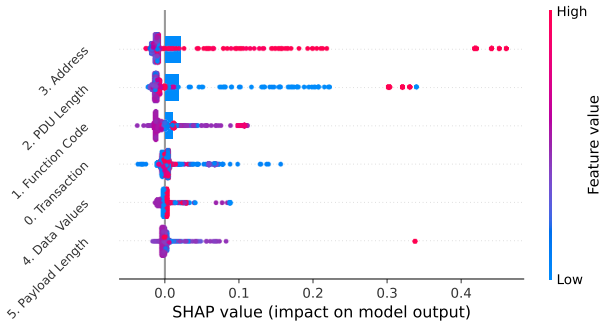

Figure 1: SHAP summary of the RF model from [29].

Adversarial Insights. The SHAP summary plot in Figure 1 provides insights that can be exploited to craft adversarial examples.

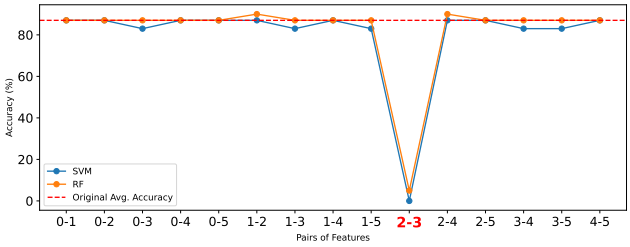

Figure 2: IDS accuracy under adversarial attack targeting different pairs of features.

By determining the direction to alter the most relevant feature's values, an attacker can optimally steer predictions. In our experiment, we observe that including higher values of *Modbus Address* would lower prediction success. Furthermore, we see that *Payload Length* is not as relevant, and should remain unchanged. This is because higher values of addresses are linked to the learned profile of the system under normal operation. Therefore, an adversary would include increased values of modbus addresses deliberately to maximize evasion probability and stealthiness. The SHAP values give a sense of how much a feature should be changed. Features with a wider spread of SHAP values might be more sensitive to changes, and even small perturbations could lead to significant impacts on the output. To empirically demonstrate this, we applied the same perturbation ($\varepsilon = 2.7$, as per the original paper) to all possible pairs of features. As can be seen in Figure 2, the most impactful manipulation corresponds to features with index 2 and 3 (i.e., *Modbus Address* and *PDU Length*), which are the most important—and therefore, vulnerable— ones according to SHAP. The feature that correspond to each index can be consulted in Figure 1 (on the Y-axis labels).

In summary, the attacker can leverage insights on natural feature variability to make subtle, hard-to-detect changes, thus crafting stealthy attacks that can potentially bypass SG security. Nevertheless, the trade-off between impact and stealthiness of the attack must be addressed. In [29], authors attempt to address this by ensuring that perturbations remain within realistic bounds, that is, setting minimum/maximum limits for each feature. This is accomplished by identifying the smallest/largest values for each feature in the original dataset and ensuring that the values of the adversarial examples do not surpass these boundaries. Consequently, these constraints ensure that the adversarial modifications maintain feature values within a practical range. In practice, it may occur that in order to keep an attack feasible, a combination of features need to be perturbed to mislead the predictor, not only the most important one(s). However, authors do not show a realistic implementation (i.e., end-to-end exploit in the problem space) of the evasion attack, limiting their study to the feature space.

Through XAI, it is also possible for an attacker to identify spurious correlations [4] in learning models. These occur when unrelated data artifacts mistakenly guide the model in classifying tasks, leading it to rely on irrelevant patterns rather than addressing the actual problem. For instance, in a network intrusion detection scenario, if most attacks in the training data come from a specific network region, the model might wrongly focus on identifying attacks based on IP ranges rather than the true nature of the attacks. This issue is a ML pitfall [4] exploitable by sophisticated attackers.

5 DISCUSSION AND EXTENDED WORK

After our analysis, in this section we present challenges and research gaps that we aim to cover in the future.

5.1 Reproducibility Analysis

We found that only 6 out of the 34 papers surveyed provided source code, constituting approximately 18% of the publications; this is approximately half compared to the baseline [31], which measures reproducibility of general ML papers in Tier 1 security conferences, where authors found that 39% of the papers provided code. A lesson we learn from this: not sharing the code used in experiments makes it harder for future researchers to build upon or compare their work with published techniques. Moreover, starting from scratch to develop complex pre-processing tasks, new analytical methods, and sophisticated system designs is a challenging task. For instance, we managed to run further experiments on the code provided by [29], adding to their investigation. Therefore, making code available not only enhances reproducibility but also supports further innovation and development. From the source codes provided, only one worked out-of-box. This problem also exists in the top tier ML security literature [31], where 82% of the papers with code required installing further packages, changing paths or directory structures, or fixing errors that appear. In our survey, we identified a lack of code instructions (e.g., a detailed ReadMe file) in approximately 40% of the papers with source code, increasing the difficulty of executing the implementations. Furthermore, 40% of the source codes did not include the corresponding adversarial attack code, being in most cases only reported via pseudo-code as a paper figure.

5.2 Confidentiality and Availability

In our survey, we observe a major interest in attacks compromising the integrity of learning-based models. Confidentiality and availability attacks are underrepresented in this body of work, but are equally —or even more— important in the context of critical infrastructure. As a potential reason behind the focus on integrity, particularly through adversarial examples, we consider the existence of immediate and tangible consequences for performance and safety. The initial discovery and exploration of adversarial examples (e.g., in general computer vision) and their effects on integrity in real world scenarios (e.g., in autonomous driving) opened up a new research frontier. As scholars dove into the complexities of these vulnerabilities, a momentum built up around this line of inquiry, leading to a concentration of effort and resources in understanding and mitigating integrity attacks across many different domains, including SGs security. In 2018, Biggio and Roli [7] published a categorization of a decades worth of research in attacks against ML, where one can notice that availability attacks via test data, and confidentiality attacks via training data were still undiscovered.

Nonetheless, the rapid proliferation of new approaches against confidentiality and availability observed in the broad artificial intelligence community will undoubtedly affect the future of attacks

tailored to learning models in SGs. The arms race between developing more sophisticated attacks and defenses contributes to this dynamic and rapidly evolving field of study.

5.3 Focus on Electrical Substations

We observe a lack of investigations in the subtopic of adversarial attacks against electrical substation-related learning-based methods. Electrical substations are pivotal components of the power grid, acting as nodes where transmission lines are connected, transformed, and distributed to various consumers. As part of a nation's critical infrastructure, ensuring their security is vital to prevent disruptions that could have wide-ranging consequences on other sectors such as healthcare, finance, transportation, and water supply.

In consequence, we propose to explore the identified research challenges within the KASTEL Security lab [22]. Our experimental environment consists of three key subsystems [28]: a microgrid, a transmission/distribution substation, and a Software-Defined Network. The transmission/distribution substation is structured into three layers: the station level, equipped with a substation automation system and a human-machine interface for overarching control and surveillance; the bay level, which includes devices for control and protection; and the process level, designed with a test set that simulates the actual physical processes. Both physical and virtual elements are integrated into these subsystems.

5.4 Challenges

To the best of our knowledge, research on these scenarios within this contextual framework has not been conducted despite their high potential and importance:

Evasion in the Problem Space. Developing realistic and practical implementations of proof-of-concepts for adversarial examples in intrusion detection use cases related to electrical substations. Specifically, understanding how evasion attacks targeting the problem space of electrical substation-related systems compromise their operational integrity, and what advanced mitigation strategies can be developed to safeguard learning-based components.

Poisoning Training Data. When developers build datasets for training, it is in their best interest to avoid miss-labeling (if supervised) and/or pollution of the normal profile (if unsupervised). Additionally, a malicious actor could purposely inject malicious data to compromise the model's performance. The goal here is to maximize classification error[1] by injecting poisoning samples into the training set. Furthermore, these malicious data points can be tailored to make the model overfit with the objective of facilitating model inversion.

Sponge Attacks. In SGs, the availability of systems is the most critical security aspect. Most models that use learning to detect events and control systems are not installed directly on the components found in substations. Instead, field data is gathered and consolidated at a utility data center, where there are sufficient computing resources to process the data, train models, and run applications that make use of these models. Nevertheless, electrical substations can be considered resource-constrained. Future research is warranted to increase our understanding on attacker capabilities when it comes to compromising the availability of learning-based models via attacks that soak up resources.

Model Extraction. If an attacker obtains data used to train a given model, it would be possible to train surrogate models known to be highly similar to the original. This situation allows an attacker to generate transferable adversarial attacks. Further investigation is needed to better understand the impact of model extraction attacks against learning-based components using data from electrical substations, and what are the potential risks to the confidentiality of sensitive information contained in proprietary models.

The objective of proactively testing learning-based models is to eventually increase resilience against these attacks in SGs . We envision the application of the following techniques to further protect electrical substations:

Adversarial Training. Incorporating adversarially generated examples into the training phase of models helps them recognize and counteract sophisticated attack patterns. By exposing the model to these malicious inputs during training, it becomes better equipped to identify similar threats during operational use, enhancing its defense capabilities against real-world adversarial attacks in the SG environment.

Feature Removal. XAI methods can identify vulnerable features in SG datasets, which could be strategically removed or altered to strengthen the model's security (i.e., an adversary-aware feature selection step). This might result in a performance trade-off, but enhancing security in critical SG operations could outweigh the loss in precision, especially in high-stakes scenarios like grid stability and outage prevention.

Expert Knowledge. Each use case exhibits potential defense directions that are tightly related to a given subdomain. A throughout understanding of subdomain technicalities (e.g., IEC 61859 communication protocols) is critical to understand vulnerabilities prone to be exploited by adversaries.

6 CONCLUSION AND OUTLOOK

This paper categorizes and evaluates work on attacks against learning-based models in smart grids, with a focus on reproducibility. Additionally, we propose exploring the explainability and interpretability of data-driven models to uncover vulnerabilities and develop countermeasures. We elaborate on the reasons behind the current limitations and challenges, with the objective of providing further insights to fill the identified research gaps. Our findings serve as a roadmap for the research community to develop stronger and more secure learning-based systems in smart grids, particularly in use cases related to electrical substations. As future work, apart from addressing the indicated challenges, we plan to evolve this survey into a more detailed systematic literature review. This will include exploring additional aspects, such as identifying the most frequently used learning models and defensive mechanisms.

ACKNOWLEDGMENTS

This work was supported by funding from the topic Engineering Secure Systems of the Helmholtz Association (HGF) and by KASTEL Security Research Labs (structure 46.23.02).

[1]The attacker's goal might extend beyond causing misclassification and include any type of misprediction. This is important because decision making and control applications mostly rely on regression models rather than classification models.

REFERENCES

[1] Afia Afrin and Omid Ardakanian. 2023. Adversarial Attacks on Machine Learning-Based State Estimation in Power Distribution Systems. In *eEnergy '23*.

[2] Utkarsha Agwan, Lucas Spangher, William Arnold, Tarang Srivastava, Kameshwar Poolla, and Costas J Spanos. 2021. Pricing in prosumer aggregations using reinforcement learning. In *eEnergy '21*.

[3] Carmelo Ardito, Yashar Deldjoo, Tommaso Di Noia, Eugenio Di Sciascio, Fatemeh Nazary, and Giovanni Servedio. 2023. Machine-learned Adversarial Attacks against Fault Prediction Systems in Smart Electrical Grids. *arXiv* (2023).

[4] Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallaro, and Konrad Rieck. 2022. Dos and don'ts of machine learning in computer security. In *USENIX Security '22*.

[5] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* (2020).

[6] Shameek Bhattacharjee, Mohammad Jaminur Islam, and Sahar Abedzadeh. 2022. Robust anomaly based attack detection in smart grids under data poisoning attacks. In *CPSS '22*.

[7] Battista Biggio and Fabio Roli. 2018. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition* (2018).

[8] Atef H Bondok, Mohamed Mahmoud, Mahmoud M Badr, Mostafa M Fouda, Mohamed Abdallah, and Maazen Alsabaan. 2023. Novel Evasion Attacks against Adversarial Training Defense for Smart Grid Federated Learning. *Access* (2023).

[9] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *SP '17*.

[10] Nicholas Carlini and David Wagner. 2018. Audio adversarial examples: Targeted attacks on speech-to-text. In *SPW '18*.

[11] Yize Chen, Yushi Tan, and Baosen Zhang. 2019. Exploiting vulnerabilities of load forecasting through adversarial attacks. In *eEnergy '19*.

[12] Richard Colbaugh and Kristin Glass. 2011. Proactive defense for evolving cyber threats. In *ISI '11*.

[13] Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, and Deepak Verma. 2004. Adversarial classification. In *KDD '04*.

[14] Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli. 2019. Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In *USENIX Security '19*.

[15] Jialiang Dong, Zhitao Guan, Longfei Wu, Xiaojiang Du, and Mohsen Guizani. 2021. A sentence-level text adversarial attack algorithm against IIoT based smart grid. *Computer Networks* (2021).

[16] Ahmed T El-Toukhy, Mohamed MEA Mahmoud, Atef H Bondok, Mostafa M Fouda, and Maazen Alsabaan. 2023. Countering Evasion Attacks for Smart Grid Reinforcement Learning-based Detectors. *IEEE Access* (2023).

[17] Zhang Guihai and Biplab Sikdar. 2021. Adversarial machine learning against false data injection attack detection for smart grid demand response. In *SmartGridComm'21*.

[18] Sam Gunn, Doseok Jang, Orr Paradise, Lucas Spangher, and Costas J Spanos. 2022. Adversarial poisoning attacks on reinforcement learning-driven energy pricing. In *BuildSys '22*.

[19] Kian Hamedani, Lingjia Liu, Jithin Jagannath, and Yang Yi. 2021. Adversarial classification of the attacks on smart grids using game theory and deep learning. In *WiseML @ WiSec '21*.

[20] Rong Huang and Yuancheng Li. 2023. Adversarial Attack Mitigation Strategy for Machine Learning-Based Network Attack Detection Model in Power System. *IEEE Transactions on Smart Grid* (2023).

[21] JQ James, Yunhe Hou, and Victor OK Li. 2018. Online false data injection attack detection with wavelet transform and deep neural networks. *IEEE Transactions on Industrial Informatics* (2018).

[22] KASTEL. 2023. KASTEL - Security and Privacy for Future Energy Systems. Available at https://www.kastel.kit.edu/english/energie.php(accessed20/03/2024).

[23] Weicong Kong, Zhao Yang Dong, Youwei Jia, David J Hill, Yan Xu, and Yuan Zhang. 2017. Short-term residential load forecasting based on LSTM recurrent neural network. *IEEE transactions on smart grid* (2017).

[24] Aditya Kuppa and Nhien-An Le-Khac. 2021. Adversarial XAI methods in cybersecurity. *IEEE transactions on information forensics and security* (2021).

[25] Tian Liu and Tao Shu. 2019. Adversarial false data injection attack against nonlinear ac state estimation with ann in smart grid. In *SecureComm '19*.

[26] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *NIPS '17*.

[27] Gautam Raj Mode and Khaza Anuarul Hoque. 2020. Adversarial Examples in Deep Learning for Multivariate Time Series Regression. *arXiv:2009.11911* (2020).

[28] Aneeqa Mumrez, Muhammad M Roomi, Heng Chuan Tan, Daisuke Mashima, Ghada Elbez, and Veit Hagenmeyer. 2023. Comparative Study on Smart Grid Security Testbeds Using MITRE ATT&CK Matrix. In *SmartGridComm '23*.

[29] Aneeqa Mumrez, Gustavo Sánchez, Ghada Elbez, and Veit Hagenmeyer. 2023. On Evasion of Machine Learning-based Intrusion Detection in Smart Grids. In *SmartGridComm '23*.

[30] Fatemeh Nazary, Yashar Deldjoo, Tommaso Di Noia, Carmelo Ardito, and Eugenio Di Sciascio. 2023. Smart Electrical grids Under the Lens of Adversarial Attacks. (2023).

[31] Daniel Olszewski, Allison Lu, Carson Stillman, Kevin Warren, Cole Kitroser, Alejandro Pascual, Divyajyoti Ukirde, Kevin Butler, and Patrick Traynor. 2023. " Get in Researchers; We're Measuring Reproducibility": A Reproducibility Study of Machine Learning Papers in Tier 1 Security Conferences. In *CCS '23*.

[32] Alexander Pan, Yongkyun Lee, Huan Zhang, Yize Chen, and Yuanyuan Shi. 2021. Improving robustness of reinforcement learning for power system control with adversarial training. *RL4RL @ ICML '21* (2021).

[33] Fabio Pierazzi, Feargus Pendlebury, Jacopo Cortellazzi, and Lorenzo Cavallaro. 2020. Intriguing properties of adversarial ml attacks in the problem space. In *SP*.

[34] Chao Ren, Xiaoning Du, Yan Xu, Qun Song, Yang Liu, and Rui Tan. 2021. Vulnerability analysis, robustness verification, and mitigation strategy for machine learning-based power system stability assessment model under adversarial examples. *IEEE Transactions on Smart Grid* (2021).

[35] Moein Sabounchi and Jin Wei-Kocsis. 2022. A practical adversarial attack on contingency detection of smart energy systems. In *ISGT '22*.

[36] Everton Jose Santana, Ricardo Petri Silva, Bruno Bogaz Zarpelão, and Sylvio Barbon Junior. 2021. Detecting and mitigating adversarial examples in regression tasks: a photovoltaic power generation forecasting case study. *Information* (2021).

[37] Ali Sayghe, Junbo Zhao, and Charalambos Konstantinou. 2020. Evasion attacks with adversarial deep learning against power system state estimation. In *PESGM'20*.

[38] Qun Song, Rui Tan, Chao Ren, and Yan Xu. 2021. Understanding credibility of adversarial examples against smart grid: A case study for voltage stability assessment. In *eEnergy '21*.

[39] Abdulrahman Takiddin, Muhammad Ismail, Rachad Atat, Katherine R Davis, and Erchin Serpedin. 2023. Robust Graph Autoencoder-Based Detection of False Data Injection Attacks Against Data Poisoning in Smart Grids. *IEEE Transactions on Artificial Intelligence* (2023).

[40] Abdulrahman Takiddin, Muhammad Ismail, and Erchin Serpedin. 2021. Robust detection of electricity theft against evasion attacks in smart grids. In *ICC '21*.

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

[42] Abdulrahman Takiddin, Muhammad Ismail, Usman Zafar, and Erchin Serpedin. 2020. Robust electricity theft detection against data poisoning attacks in smart grids. *IEEE Transactions on Smart Grid* (2020).

[43] Jiwei Tian, Buhong Wang, Jing Li, and Charalambos Konstantinou. 2022. Adversarial attack and defense methods for neural network based state estimation in smart grid. *IET Renewable Power Generation* (2022).

[44] Jiwei Tian, Buhong Wang, Jing Li, and Zhen Wang. 2021. Adversarial attacks and defense for CNN based power quality recognition in smart grid. *IEEE Transactions on Network Science and Engineering* (2021).

[45] Zhiqiang Wan, Hepeng Li, Hang Shuai, Yan Lindsay Sun, and Haibo He. 2021. Adversarial attack for deep reinforcement learning based demand response. In *PESGM'21*.

[46] Yu Wang and Bikash Pal. 2023. Destabilizing attack and robust defense for inverter-based microgrids by adversarial deep reinforcement learning. *IEEE Transactions on Smart Grid* (2023).

[47] Dian Ang Yap, Joyce Xu, and Vinay Uday Prabhu. 2019. On detecting adversarial inputs with entropy of saliency maps. *CV-COPS @ CVPR '19* (2019).

[48] Lanting Zeng, Dawei Qiu, and Mingyang Sun. 2022. Resilience enhancement of multi-agent reinforcement learning-based demand response against adversarial attacks. *Applied Energy* (2022).

[49] Guihai Zhang and Biplab Sikdar. 2022. Ensemble and Transfer Adversarial Attack on Smart Grid Demand-Response Mechanisms. In *SmartGridComm '22*.

[50] Liang Zhang, Gang Wang, and Georgios B Giannakis. 2019. Real-time power system state estimation and forecasting via deep unrolled neural networks. *IEEE Transactions on Signal Processing* (2019).

[51] Yu Zhang, Chao Huo, Huifeng Bai, and Ganghong Zhang. 2023. Adversarial Defense Based on Mimic Defense and Reinforcement Learning for Power Vision Task in Smart Grid. In *ACCES '23*.

[52] Zhenyong Zhang, Mengxiang Liu, Mingyang Sun, Ruilong Deng, Peng Cheng, Dusit Niyato, Mo-Yuen Chow, and Jiming Chen. 2024. Vulnerability of Machine Learning Approaches Applied in IoT-Based Smart Grid: A Review. *IEEE Internet of Things Journal* (2024).

[53] Xingyu Zhou, Yi Li, Carlos A Barreto, Jiani Li, Peter Volgyesi, Himanshu Neema, and Xenofon Koutsoukos. 2019. Evaluating resilience of grid load predictions under stealthy adversarial attacks. In *RWS '19*.

[54] Yanxu Zhu, Hong Wen, Runhui Zhao, Yixin Jiang, Qiang Liu, and Peng Zhang. 2023. Research on Data Poisoning Attack against Smart Grid Cyber–Physical System Based on Edge Computing. *Sensors* (2023).