GAN-GRID: A Novel Generative Attack on Smart Grid Stability Prediction

Emad Efatinasab[1(✉)], Alessandro Brighente[2], Mirco Rampazzo[1], Nahal Azadi[1], and Mauro Conti[2,3]

[1] Department of Information Engineering, University of Padova, Padua, Italy
emad.efatinasab@phd.unipd.it, mirco.rampazzo@unipd.it, nahal.azadi@studenti.unipd.it
[2] Department of Mathematics, University of Padova, Padua, Italy
{alessandro.brighente,mauro.conti}@unipd.it
[3] Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, Netherlands

Abstract. The smart grid represents a pivotal innovation in modernizing the electricity sector, offering an intelligent, digitalized energy network capable of optimizing energy delivery from source to consumer. It hence represents the backbone of the energy sector of a nation. Due to its central role, the availability of the smart grid is paramount and is hence necessary to have in-depth control of its operations and safety. To this aim, researchers developed multiple solutions to assess the smart grid's stability and guarantee that it operates in a safe state. Artificial intelligence and Machine learning algorithms have proven to be effective measures to accurately predict the smart grid's stability. Despite the presence of known adversarial attacks and potential solutions, currently, there exists no standardized measure to protect smart grids against this threat, leaving them open to new adversarial attacks.

In this paper, we propose GAN-GRID a novel adversarial attack targeting the stability prediction system of a smart grid tailored to real-world constraints. Our findings reveal that an adversary armed solely with the stability model's output, devoid of data or model knowledge, can craft data classified as stable with an Attack Success Rate (ASR) of 0.99. Also by manipulating authentic data and sensor values, the attacker can amplify grid issues, potentially undetected due to a compromised stability prediction system. These results underscore the imperative of fortifying smart grid security mechanisms against adversarial manipulation to uphold system stability and reliability.

1 Introduction

Smart Grid (SG) technology represents a modern electric power grid characterized by increased reliability, efficiency, sustainability, and bi-directional communication capabilities [31]. By integrating advanced hardware (such as phasor measurement units and smart meters) and advanced software solutions, SGs

J. Garcia-Alfaro et al. (Eds.): ESORICS 2024, LNCS 14982, pp. 374–393, 2024.
https://doi.org/10.1007/978-3-031-70879-4_19

provide safety and stability while concurrently reducing operational costs compared to previous energy distribution systems [22]. With the current urge to include renewable energy sources in the power market, the SG should be open to seamlessly including novel technologies together with their operations characteristics in terms of when they collect power, and how much power they can deliver. Accurately predicting renewable energy generation is crucial for ensuring the stable, efficient, and cost-effective operation of the power system [23]. This highlights the importance of employing advanced forecasting methods for anticipating fluctuations and maintaining the balance between electricity supply and demand, especially with sustainable energy sources. To achieve this, researchers have developed stability prediction systems as software components of the smart grid. These systems collect grid data and analyze historical trends to predict potential instability in the SG configuration, allowing for reconfiguration to ensure service availability. Machine Learning (ML) and Artificial Intelligence (AI) have proved to be a very efficient solution to this aim, with researchers proposing many different models with very good performance [2,8,13,30,35,45]. SGs represent the energy backbone of a nation and are hence among the critical infrastructures to be protected [17]. Indeed, critical infrastructures have been recently targets of many cyber attacks, as their disruption might significantly impact a whole country [12]. Several factors contribute to the vulnerabilities of the smart grid. High interconnection among devices and remote access points provide entry points for attackers, who can inject malicious data by compromising a single node. Additionally, the use of legacy systems, inherent system complexity, and lack of standardization make managing the SG challenging, particularly in terms of security [33]. Despite the investigation of authentication and access control mechanisms for securely collecting and managing data in SGs [6,24,36], SGs are nowadays still an easy target for cyberattacks [17].

While the successful integration of AI technologies shows that SGs are revolutionary in modernizing the electricity sector, they remain one of the most vulnerable points of SGs [22]. Indeed, a few studies [1,40] are assessing the susceptibility of AI-enabled stability prediction systems in SGs to adversarial attacks. The main idea behind these attacks is to inject maliciously crafted data into the smart grid network to deceive the AI-enabled stability prediction system, causing faults. This transforms potential adversarial attacks into false data injection attacks targeting the entire grid. Such attacks not only affect the stability prediction system but also disrupt interconnected systems that rely on accurate grid data. The attacker's ability to manipulate data distribution challenges grid operators who depend on accurate information for critical decisions. The risk escalates as manipulations may go unnoticed when the stability prediction model is compromised. This manipulation poses a significant risk as it could obscure any genuine instability within the grid, whether caused by the attacker or other factors. Up to now, all studies in the literature focus on state-of-the-art adversarial attacks, which however can be mitigated via state-of-the-art solutions. However, no proposal in the literature design attacks specifically for stability prediction systems leveraging mild assumptions related to the knowledge

of data and model parameters. This represents a fundamental need to address, as attacks on prediction systems may lead to severe malfunctioning, resulting in a lack of service and/or disruption of critical components of the infrastructure (e.g., due to overvoltage). SGs are part of a nation's critical infrastructures and need hence to be secured against these threats.

In this paper, we introduce GAN-GRID, a novel Adversarial attack using a Generative Adversarial Network (GAN) to generate grid-like data classified as stable by an ML-based stability prediction system. To the best of our knowledge, this is the first contribution proposing a new adversarial attack that requires minimal access to the real data and the model and demonstrates high success rates against stability prediction systems in SGs. Given the absence of openly available code for stability prediction systems in state-of-the-art papers, we first develop and test different ML and DL models specific to stability prediction tasks, achieving up to 0.999 accuracy. We then propose a novel adversarial model specifically targeting stability prediction systems. Starting from random data, our attack leverages a GAN optimized by reinforcement learning. When developing adversarial attacks, access to data and model specifics is crucial for creating effective adversarial samples that mislead the stability prediction system. Based on this consideration, we evaluate the vulnerability of these models to our attack in both a white box (i.e., access to model and data) and a grey box (i.e., access to model output) scenario, showcasing susceptibility even without access to authentic data or model details. The resulting injected data poses serious risks as it does not trigger any alarms regarding instability within the stability prediction system. Thus, other interconnected systems that rely on accurate grid data predictions could also be compromised. Our contributions can be summarized as follows.

- We propose a novel realistic threat model that reflects a real-world scenario of an attack on a stability prediction system that has not been discussed before in literature.
- We propose **GAN-GRID**, a novel class of adversarial attacks to stability prediction systems To the best of our knowledge, we are the first to develop such attacks in this context.
- We propose and evaluate several stability prediction models to determine which are the most effective for stability prediction applications. Our evaluation together with our open-source code, provides a reference for future studies on stability prediction models and their security.
- We evaluate our system and attacks on the Electrical Grid Stability Simulated Dataset. We show an accuracy of up to 0.999 for our stability prediction models. Also, our attack was able to deceive the stability prediction models to classify the generated data as stable with an Attack Success Rate (ASR) of up to 0.99. Notably, it outperformed other attacks in both ASR and the level of access required to execute the attack.
- We make the code of our systems, attacks, and the dataset available at: https://github.com/emadef1/GAN_GRID/. Thanks to our code, we foster research on this subject providing a common baseline for future evaluation and developments.

2 Related Work

In this section, we present related works on stability prediction systems and their security. In particular, we review existing stability prediction methodologies in Sect. 2.1, while we review currently available attacks to these systems in Sect. 2.2.

2.1 AI and ML for SG Stability

In this context, AI has emerged as one of the most transformative and impactful technologies for the effective management of power grids and SGs [5]. These cutting-edge AI techniques offer powerful and promising solutions for the stability analysis and control of SGs, attracting increasing interest and attention from researchers and practitioners alike [37]. For instance Önder et al. [35] introduced five distinct cascade methodologies, encompassing pre-processing, training, testing division, and classification stages within the stability estimation procedure for SGs. Bashir et al. [5] utilized a range of state-of-the-art ML algorithms, including Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Logistic Regression, Naive Bayes, Neural Networks, and Decision Tree classifiers, to forecast SG stability. Gorzałczany et al. [19] tackle the challenge of transparent and precise prediction of decentralized SG control stability by leveraging a knowledge-based data-mining methodology, specifically a fuzzy rule-based classifier. Their approach utilizes multi-objective evolutionary optimization algorithms to enhance the balance between interpretability and accuracy within the classification system. An improved model is introduced in [43], harnessing the capabilities of explainable AI and feature engineering for predicting SG stability. Notably, this study adopts a symmetrical approach by addressing the problem from both classification and regression perspectives. Dewangan et al. [13] have presented a new and enhanced genetic algorithm (GA)-based extreme learning machine (ELM) model for forecasting the stability of SG. They explore the outcomes of this model and compare them with those of other modern AI and DL models for comprehensive analysis. Furthermore, there is a growing emphasis on the utilization of Recurrent Neural Networks (RNNs) such as Long Short-Term Memory Network (LSTM) and Gated Recurrent Unit (GRU) in the literature [2,45]. Their widespread adoption underscores their effectiveness in capturing temporal dependencies and modeling sequential data, thus enhancing the accuracy and reliability of stability prediction systems in SG environments. Convolutional Neural Networks (CNNs) are emerging as a popular choice in stability prediction research within SGs, evidenced by their recurrent application in the literature [11,20,38].

2.2 Adversarial Attacks

Ahmadian et al. [1] introduced a False Data Injection Attacks (FDIA) utilizing a GAN architecture. In this model, the attacker assumes the role of the generative network, while the Energy System Operator (ESO) acts as the discriminative network. By formulating an optimization problem, the attacker generates deceptive data that evades detection by the power system state estimator.

Li et al. [27] illustrate the susceptibility of well-established ML models used for detecting energy theft to adversarial attacks. Specifically, they develop an approach for generating adversarial measurements, allowing attackers to report significantly reduced power consumption to utility companies, effectively evading detection by the ML-based energy theft detection systems. Chenet al. [10] endeavor to tackle security concerns surrounding ML applications within power systems. They highlight that the majority of ML algorithms currently proposed for power systems exhibit vulnerability to adversarial examples, which are inputs deliberately crafted with malicious intent. As discussed in the literature, ML/DL models are frequently employed as stability prediction systems, yet they are vulnerable to adversarial attacks, an issue often overlooked in previous research [21].

3 System and Threat Model

System Model. In an operational scenario devoid of active threats targeting system disruption, the stability prediction system receives input data from the SG infrastructure, i.e., different sensors and Phasor Measurement Unit (PMU) measurements from different points across the grid. The stability prediction model is designed to analyze grid conditions and determine whether stability is maintained or compromised. Thus, this system focuses solely on stability prediction, which entails discerning whether the grid is stable or unstable (binary classification task). To this aim, it uses ML and/or AI algorithms to discern whether, based on the current observations, the SG will be stable or not in the near future. Before deployment, the stability prediction model undergoes training using uncorrupted data to ensure accurate and reliable predictions within the operational environment.

Threat Model. The attacker's goal is to inject fraudulent data into the grid's stream, covertly aiming to manipulate the stability model's classification. To this aim, the attacker might exploit known or new vulnerabilities to gain remote access [9,41]. The primary goal is to deceive the stability prediction system into classifying the injected packets as belonging to the stable class. One potential real-world case of an attacker compromising the stability prediction system is during peak demand times when actual grid conditions become unstable. For instance, heat waves can significantly impact power system operations by increasing peak loads and reducing transmission and generation capacity [25]. The uncertainty and variability of wind and solar generation can pose challenges for grid operators, requiring additional actions to balance the system [7]. During these peak demand times, the unstable conditions strain the grid. The stability prediction system, misled by adversarial data injected by the attacker, fails to initiate preventive measures such as load shedding or switching to backup generators. This failure is critical because these measures are designed to alleviate the strain on the grid by reducing demand or supplementing supply. Without these interventions, the grid remains under excessive load, causing transformers, generators, and other critical components to fail. A local outage in one part of

the grid causes a chain reaction, leading to widespread blackouts. In a blackout, access to critical services like telecommunications, transportation, and medical assistance is also compromised [16]. We define two scenarios based on the attacker's knowledge of the SG's data and of the stability prediction model.

- *White-box Scenario*: In this scenario, the attacker possesses comprehensive access to both the data employed in testing the model and detailed information regarding the model's architecture and parameters. This advantageous position provides the attacker with ample opportunities to exploit vulnerabilities in the system. By leveraging this intelligence, the attacker can meticulously craft powerful adversarial samples aimed at deceiving the model. Additionally, having access to the model weights enables the adversary to fine-tune the attack parameters offline, enhancing the effectiveness and sophistication of their attacks.
- *Gray-box Scenario*: In real-world contexts, scenarios where adversaries successfully infiltrate systems to compromise stability prediction models through unauthorized access to data and trained models are rare. Various defense strategies outlined in the literature empower real-world systems to integrate countermeasures aimed at deterring direct breaches [14,32,42]. It is also suggested by [39] that while we shouldn't dismiss the potential for input-specific adversarial attacks, they are generally considered less plausible as attacks against SG stability assessment systems. In a more realistic scenario, termed a grey-box setting, attackers can only access the trained models' output without obtaining data from the grid or accessing the model architecture and training details. However, it's crucial to note that attackers may possess knowledge of the features used by the stability prediction model for the development of adversarial attacks, which could be inferred from widely available literature or through interactions with the model itself. In this grey-box scenario, the attacker preemptively uses the model output to train the generator of a GAN. By leveraging the stability prediction model as an oracle, the attacker can train a neural network using its feedback.

4 GAN-GRID: Our Proposed Adversarial Attack

We now discuss the attacks that we employ against stability prediction systems in SGs. In Sect. 4.1 we describe our proposed methodology to generate adversarial samples in a greybox setting. In Sect. 4.2 we then present common whitebox adversarial approaches that represent a baseline for comparison with our proposed attack model.

4.1 GAN-GRID Model

In this section, we describe the workflow of GAN-GRID as depicted in Fig. 1. In our scenario, the attacker gains access to the stability prediction model response without direct access to the underlying data. This is akin to a modified GAN

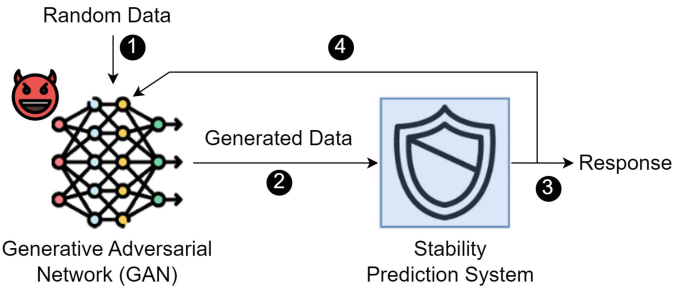


Fig. 1. GAN-GRID Attack Workflow.

training process, where the attacker utilizes the legitimate model to train the generator. The attacker starts by providing input to the GAN randomly sampled data ❶. The output of the GAN network is then distributed in the SG grid network ❷. We optimize our generator model in conventional GAN training to outsmart a fixed discriminator, represented by the stability prediction system, rather than training both components iteratively. By leveraging the stability prediction model as an oracle ❸, the attacker trains a neural network capable of generating fraudulent samples, even from random data. Our generative network leverages discriminator feedback ❹, provided by the stability prediction system's output, for optimization and loss computation. This feedback guides the generator model in producing fraudulent data that can trick the stability prediction system. To address the challenges of convergence and navigating local minima in the large search space, we use reinforcement learning to improve the generator's learning procedure. This strategic choice allows for more efficient exploration of the search space and adaptation in response to feedback and rewards received during the learning process. By employing exploration and exploitation strategies, the generator can strike a balance between trying new approaches and leveraging existing knowledge to identify promising search spaces. Through the integration of reinforcement learning techniques, our approach transcends the limitations typically associated with traditional optimization methods.

The reinforcement learning process involves several key parameters, including the maximum episode length, discount factor denoted as γ, the number of episodes, and the learning rate represented by α. These parameters govern the update mechanism for the generator's latent input using reinforcement learning. The training loop operates across episodes, where each episode begins by initializing the latent input parameters and the episode reward. Within each episode, the generator generates a sample based on the latent input. This generated sample undergoes evaluation by the stability prediction model, which provides predictions against randomly generated target labels for comparison. The reward is computed as the mean accuracy of the predictions matching the targets. To update the latent input using reinforcement learning, the temporal difference error (td_{error}) is calculated as the difference between the reward and the cumulative episode reward. The reward reflects the agent's performance in an

episode, offering immediate feedback on its decisions. Conversely, the cumulative episode reward signifies the total reward gathered throughout an entire episode, bounded by the maximum number of steps or actions allowed in the reinforcement learning process. By calculating the td_{error} as the difference between the reward and the cumulative episode reward, we capture the discrepancy between the immediate feedback received and the overall performance over an extended period. Subsequently, the latent input is updated by incorporating a scaled noise term to introduce randomness and facilitate exploration. The scaling factor for the noise term is determined by α, td_{error}, and the γ factor raised to the power of the current step. Mathematically, the update equation for the latent input is expressed as:

$$latent\_input = \alpha \cdot td_{error} \cdot \gamma^{step} \cdot latent\_input. \tag{1}$$

This scaling factor influences the magnitude of the noise added to the latent input, potentially increasing or decreasing the level of exploration based on the td_{error}'s magnitude.

Scaling the noise with td_{error} enables dynamic exploration adjustment during training. Higher td_{error} yields larger scaling factors, increasing exploration and randomness in latent input updates. Conversely, lower td_{error} results in smaller scaling factors, decreasing exploration and increasing exploitation as the agent refines estimates and converges towards better solutions, reducing randomness in latent input updates. This mechanism allows the generator to adapt its latent input based on the reward signal, facilitating the exploration of diverse latent space regions. As the agent learns from experience, the future rewards' impact on the scaling factor diminishes, allowing the agent to prioritize immediate feedback for policy optimization. After each episode, the generator updates using the final latent input. The stability prediction model assesses the generator's output, generating a target label tensor for loss calculation. Binary cross-entropy loss computes the generator's loss, and parameters are updated via backward propagation. Upon completing the specified number of episodes, the trained generator is returned, capable of producing deceptive data without knowing the real data distribution. This updating mechanism enables the generator to adapt its latent input according to the received reward signal, allowing it to explore diverse regions within the latent space. As the agent gains more experience and learns from previous steps, the influence of future rewards on the scaling factor decreases, allowing the agent to focus more on optimizing its policy based on immediate feedback. Following each episode, the generator undergoes an update using the final latent input. Once the designated number of episodes is completed, the trained generator is returned, equipped with the capacity to generate deceptive data effectively even without a glance at real data distribution. We use the Leaky ReLU activation function [44] to prevent the dying ReLU problem and to improve gradient flow, which in turn helps stabilize the training of our generator. The generator architecture is composed of 5 feed-forward layers of 128, 256, 512, 64, and 12 units respectively.

4.2 Reference Whitebox Attacks

In a white-box threat model, the adversary is equipped with complete knowledge of both the data utilized and the trained model itself. Consequently, we undertake an examination of notable adversarial attacks to unveil vulnerabilities inherent in these models. Notice that this setting represents the most advantageous one for the attacker. Consequently, since this has been widely studied in the literature, we leverage well-studied and understood attacks as a reference to evaluate GAN-GRID that leverages a less advantageous graybox setting. Our attention is directed toward specific attacks that have been emphasized in the literature due to their significance and effectiveness in uncovering weaknesses within ML models. However, it is important to note that many well-known attacks have not been tested or implemented in libraries for binary classification problems. This constraint posed challenges in identifying and selecting appropriate attack methodologies.

- *Fast Gradient Sign Method (FGSM):* FGSM efficiently generates adversarial examples by leveraging the gradient sign of the loss function. Renowned for its computational efficiency, FGSM serves as a fundamental benchmark for assessing model robustness [18].
- *Basic Iterative Method (BIM):* BIM builds upon FGSM by iteratively applying small perturbations at each step, thereby enhancing the attack's potency. This iterative approach offers insights into the cumulative effects of perturbations, shedding light on nuanced aspects of model robustness [26].
- *Projected Gradient Descent (PGD):* PGD adopts an iterative optimization strategy similar to BIM, but distinguishes itself by incorporating a projection step to confine perturbations within a predefined constraint set. This distinctive feature enables PGD to craft highly potent adversarial examples, facilitating thorough examination of model robustness under rigorous conditions. [29].
- *Random noise:* This custom implementation of random noise attack strategy utilizes a method of introducing random noise to generate adversarial instances aimed at undermining our models. The attack introduces random perturbations drawn from a normal distribution to the original samples. Each input sample undergoes multiple iterations of perturbation, guided by the user-defined epsilon (ϵ) parameter, representing the strength of each attack and the extent of perturbation introduced. In the context of adversarial attacks, the epsilon (ϵ) parameter controls the magnitude of the perturbation added to the input data. A larger epsilon value means a stronger attack, as it allows for greater deviation from the original data, potentially leading to more noticeable changes. Conversely, a smaller epsilon value results in subtler perturbations, which might be harder to detect but could still be effective in misleading the model. Following perturbation, the samples are subjected to the models classification process. If the resulting accuracy is lower than the original predictions, signifying successful deception, the perturbed sample replaces the original in the set of adversarial examples. This iterative process continues until either a successful adversarial instance is identified or

the maximum number of perturbation attempts, specified by the number of samples parameter, is exhausted. We opted for a sample size of 50 to minimize computational burden.

5 Grid Stability Prediction

In this section, we thoroughly explore models developed specifically for stability prediction. Despite the presence of a vast literature that proposes models for stability prediction, we explore new models for stability prediction in response to a critical concern. While some models in the literature may perform satisfactorily, their reproducibility is a significant limitation. Indeed, the lack of sufficient information about the model architecture and hyperparameters or the lack of their open-source code prevents accurate replication of these models. Therefore, we resort to creating state-of-the-art-based stability prediction models to test the effectiveness of our devised attack. To ensure a thorough and complete analysis, we employ both classical ML (Sect. 5.1) and DL (Sect. 5.2) models. This dual approach helps us understand their performance and vulnerability to attacks comprehensively, drawing robust conclusions about stability prediction efficacy and security against potential threats.

5.1 ML Model Design

In our ML model implementation, we consider classical ML algorithms such as Decision Trees, Extra Trees, XGBoost, KNN, Light Gradient-Boosting Machine (LGBM), and Random Forest. After thorough training and comparison experiments with other algorithms (see Sect. 6.2 for details), we selected the XGBoost architecture. Following hyper-parameter tuning, XGBoost emerged as the optimal choice for our stability prediction system due to its superior performance and lightweight nature. This efficiency ensures swift data processing and model evaluation, making it well-suited for real-time prediction tasks and enhancing the responsiveness and reliability of our system.

5.2 DL Model Design

To ensure practicality and efficiency, we engineered our DL stability prediction model to be streamlined, minimizing computational demands while maximizing effectiveness. This design philosophy aligns with our goal of creating a robust yet resource-efficient system. Our stability prediction model employs a one-layer Bi-directional LSTM architecture with 220 neurons to capture temporal dependencies in both forward and backward directions within the time sequence. To prevent overfitting, we introduced a dropout layer with a 0.5 dropout rate during training. Following the dropout layer, the LSTM layer's output passes through a Linear layer with 440 neurons, activating an element-wise sigmoid function. The deliberate choice of LSTMs aims to capture potential causal relationships between data points. For model optimization, we use the Binary Cross-Entropy

loss function, a standard metric for binary classification tasks. The training process utilizes the Adam optimizer with a learning rate of 1×10^{-3} for efficient gradient descent. We structure training iterations into 10 epochs to balance duration and performance.

6 Evaluation

We now delve into the evaluation of the attack and baseline stability prediction systems. As metrics, we use accuracy and F1 score to evaluate the models and Attack Success Rate (ASR) to evaluate attacks, defined as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}, \tag{2}$$

$$F1 = \frac{2TP}{2TP + FP + FN}, \tag{3}$$

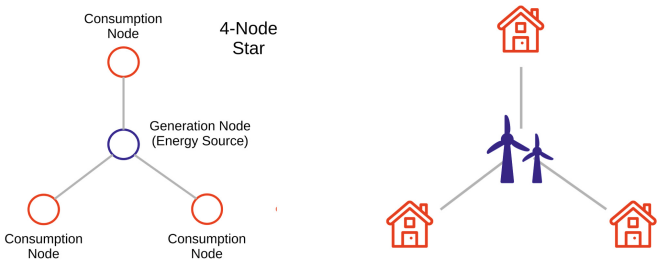
$$ASR = \frac{\#\text{ malicious batches fooling the stability prediction}}{\#\text{ malicious batches sent}}. \tag{4}$$

6.1 Dataset

The dataset utilized for evaluating our systems originates from an augmented version of the *Electrical Grid Stability Simulated Dataset* obtained from the University of California (UCI) Machine Learning Repository [3]. Initially containing 10,000 samples, this dataset contains simulation outcomes regarding grid stability for a reference 4-node star network, as depicted in Fig. 2a. Also a real-world example of such architecture can be seen in Fig. 2b. By augmentation, the dataset expanded to 60,000 samples, leveraging the grid's inherent symmetry and increasing the dataset sixfold. It comprises 12 primary predictive features and two dependent variables, offering insights into grid stability dynamics. To manage the dataset effectively, we used a robust windowing technique, segmenting it into predefined-size segments. Each window was created iteratively by traversing the data with a step size equal to half of the window size, set at 16 for our dataset. Additionally, we partitioned the dataset into training (75%), validation (5%), and test (20%) subsets. Preprocessing steps focused on normalization to prepare the dataset for prediction models effectively.

6.2 Baseline Evaluation

During the evaluation phase, we assess the performance of our stability prediction systems. We first utilize the training data to train both ML and LSTM models. Subsequently, we evaluate the efficacy of our stability prediction system on the test set. The results of our evaluation are noteworthy. The Best ML model, i.e., XGBoost, achieves a mean accuracy of 0.994 ± 0.001, while the



(a) Topology: 4-node star architecture

(b) Real-world example of the topology

Fig. 2. Laboratory setup for the real attack experimentation.

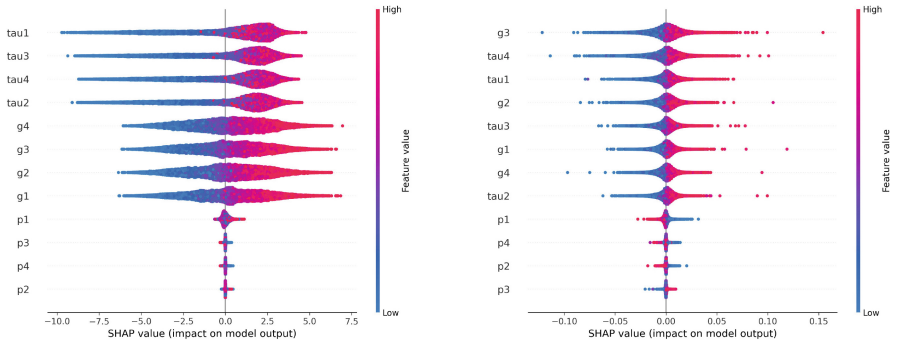
DL model demonstrates even higher accuracy, reaching 0.999 ± 0.001 for the stability prediction task. A comprehensive presentation of results is provided in Table 1.

Table 1. ML and DL Models Performance Metrics

Model	Performance Metrics	
	Accuracy	F1 Score
LSTM	0.999	0.999
XGBoost	0.994	0.994
LGBM	0.97	0.97
Decision Tree	0.974	0.974
Extra Trees	0.991	0.991
KNN	0.875	0.874
Random Forest	0.988	0.988

Feature Importance. To discern the most influential features employed by both DL and ML models, we employ Explainable Artificial Intelligence (XAI) techniques. Specifically, we leverage SHapley Additive exPlanations (SHAP) [28], recognized for its model-agnostic nature and robust interpretability. SHAP allows us to quantify the contribution of each feature to the model's predictions, offering insights into the underlying decision-making process. We use SHAP Gradient Explainer for interpreting the LSTM model and SHAP Tree Explainer for the XGBoost model, with results depicted in Fig. 3a and 3b. The analysis indicates varying feature importance between XGBoost and LSTM models. In XGBoost, participant reaction time (tau[x]) is primary, followed by price

elasticity coefficients (gamma). Nominal power consumption or production features (p[x]) have less impact. In contrast, the LSTM model prioritizes price elasticity coefficients and then participant reaction time. However, both models consider nominal power consumption or production features less critical in decision-making processes. This observation aligns with findings from the literature, where Erdem et al. [15] utilized Layer-Wise Relevance Propagation (LRP) to determine relevance scores for each input, thereby confirming the diminished importance of nominal power consumption or production features in decision-making processes.



(a) Shap values for XGBoost.

(b) Shap values for LSTM.

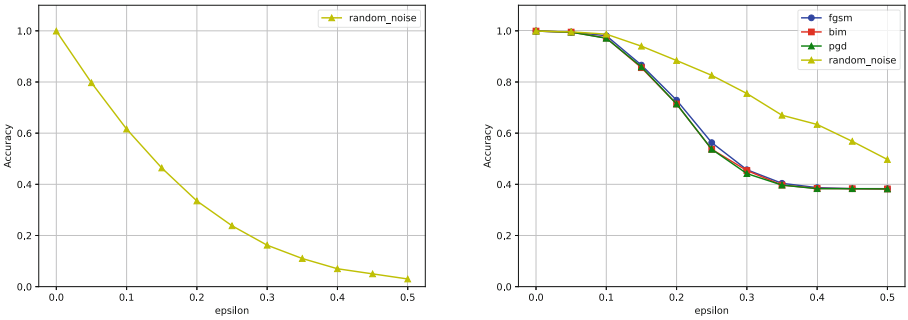
Fig. 3. Shap values.

6.3 Attack Evaluation

We proceed to assess our attacks against the stability prediction models, dividing the evaluation into the scenarios outlined in the threat model in Sect. 3: white-box attacks and the GAN-GRID attack.

White-Box Evaluation. In this section, we thoroughly assess the effectiveness of white-box attacks, as detailed in Sect. 4.2. To execute these attacks, we utilize the Adversarial Robustness Toolbox (ART) library [34], probing the baseline systems to evaluate the susceptibility of our models without incorporating any countermeasures or defenses. In classical ML models, characterized by non-differentiable architectures such as decision trees or ensemble methods, applying white-box adversarial attacks like FGSM, BIM, and PGD is not straightforward due to the absence of easily obtainable gradients. Unlike DL models, which readily provide gradients, classical ML models often lack this accessibility, rendering the application of gradient-based attacks impractical or challenging. This challenge extends beyond just accessibility; it also pertains to fundamental differences in architecture and the methods employed in classical ML compared to

DL. These classical ML techniques often rely on discrete decisions and non-linear transformations, making the computation and propagation of gradients inherently difficult. Additionally, the library implementations of these attacks do not offer built-in support for ML classifiers. As a result, we do not employ these three attacks against our classical ML model. Instead, we utilize our proposed random noise-based attack tailored for XGBoost, to explore potential vulnerabilities and assess robustness. The attacks are conducted with varying epsilon values, representing the strength of each attack and the extent of perturbation introduced. Specifically, we explore epsilon values ranging from 0.05 to 0.50. The outcomes of these attacks across different models are visually depicted in Fig. 4. The XGBoost model is more susceptible to the same random noise attack compared to the LSTM model. Moreover, it is noteworthy that the FGSM, BIM, and PGD methods exhibit nearly identical performance, surpassing that of random noise. Also, increasing the epsilon value beyond 0.5 does not provide any significant advantage.



(a) Model accuracy vs epsilon for XGBoost. (b) Model accuracy vs epsilon for LSTM.

Fig. 4. Model's accuracy at varying epsilon values on the white-box attacks.

GAN-GRID Evaluation. In our attack evaluation, we utilize a generator model optimized through reinforcement learning, leveraging the output of our stability prediction systems as surrogate data. Our aim is to generate data classified as stable by the prediction system, without access to actual data or model architecture and training details. We train the generator against both XGBoost and LSTM models, with negligible training time per episode, even on CPU (1 s). After training, we synthesize data from noise using the generator, matching the number of batches in the test set. We subsequently evaluate this generated data against the stability prediction models. Results show an ASR of 0.99 ± 0.01 % for the attack against both models. This highlights the vulnerability of these models to our attack, as our generator can converge to a data distribution classified as stable without access to real data. During our experiments, we conducted multiple training iterations with the generator to determine the

mean convergence episode and the required time and number of data batches for classification by the surrogate model, ensuring generator convergence. For the LSTM model, convergence typically occurs after 15 episodes of training, requiring approximately 60 batches of data to be sent for classification. This process takes roughly 16 min. With the XGBoost model, convergence is achieved after about 5 episodes of training, necessitating around 20 batches of data and taking approximately 6 min. These results underscore stability prediction models' vulnerability to sophisticated attacks, even with limited access to data or models, emphasizing the need for enhanced robustness and security in critical systems. The DL model takes longer than the ML model to process. In our simulations, data collection for the stability prediction model happens every 16 s, with the model requiring the same amount of time to receive data and generate predictions.

In light of our discussion regarding the potential manipulation of authentic data and sensor values by malicious actors, we undertake an analysis to investigate the ramifications of the grid infrastructure. Our objective is to shed light on the capacity for manipulative actions to introduce distortions that could

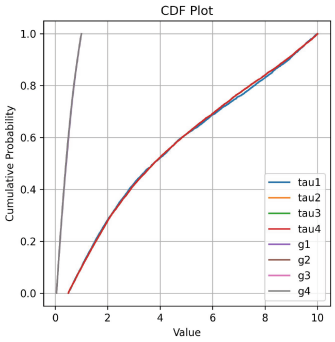


Fig. 5. Cumulative distribution of Real data

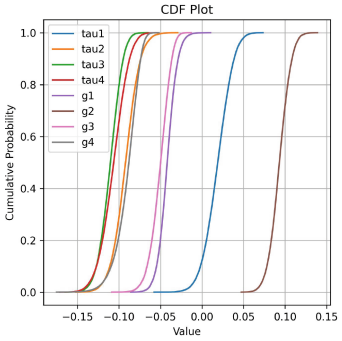


Fig. 6. Cumulative distribution of data generated for the ML model

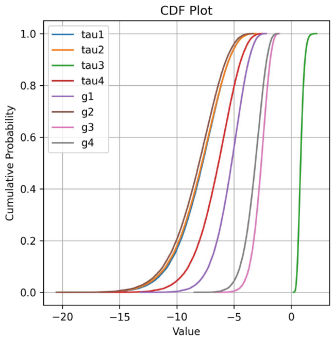


Fig. 7. Cumulative distribution of data generated for the DL model

exacerbate existing grid issues while evading detection due to compromised stability systems. The outcomes, depicted in Figs. 5, 6, and 7, reveal a significant discrepancy in the distribution patterns of relevant features (according to SHAP analysis in Sect. 6.2), leaning towards smaller values compared to authentic data. These changes have the potential to cause significant problems within the grid infrastructure. Skewed distributions of relevant features towards smaller values can trigger operational challenges within the grid. For instance, such skewness might lead to underestimation of power demand, causing inadequate resource allocation and grid instability during peak demand periods. This situation can also lead to overvoltage, frequency deviations, and heightened stress on grid components, potentially resulting in equipment failures, service disruptions, and compromised grid reliability. Based on these results, we recommend implementing defensive measures such as adversarial training, which is one of the most effective approaches against adversarial attacks [4]. Additionally, the use of anomaly detection systems, which have demonstrated good results in other smart grid applications [14], can potentially enhance the security of AI-enabled stability prediction systems.

Summary. The Table 2 summarizes the success rates of the outlined attacks. It is evident that white-box attacks demand extensive access to both the model and data, as discussed in our threat model in Sect. 3. However, this scenario is often not feasible in real-world settings. On the contrary, the GAN-GRID attack merely requires access to the model's output, significantly reducing the required level of access. Moreover, in terms of ASR, the GAN-GRID outperforms all other attacks. Additionally, we can estimate the potential time required for an attacker to employ the GAN-GRID attack in a real scenario, further highlighting its efficiency and effectiveness.

Table 2. Comparison of Model Performance Under Adversarial Attacks ($\epsilon = 0.5$)

Model	Accuracy					
	Baseline	GAN-GRID	FGSM	BIM	PGD	Random noise
LSTM	0.999	0.01	0.383	0.382	0.381	0.497
XGBoost	0.994	0.01	–	–	–	0.038

7 Conclusions

Our study emphasizes the critical need to strengthen SG security mechanisms to defend against adversarial manipulation and maintain system stability and reliability. Using advanced ML algorithms, including XGBoost and LSTM-based DL models, we explore stability prediction using the Electrical Grid Stability Simulated dataset. Through rigorous experimentation, we achieved high predictive performance. However, our findings reveal the vulnerability of SG stability prediction systems to our novel attack, even with limited information, achieving an ASR of 0.99 outperforming other attack methods. We also demonstrated that by injecting the data generated by our attack, adversary can exacerbate grid issues without triggering alarms in compromised stability prediction systems. These results underscore the importance of enhancing resilience against cyberattacks in SG environments to ensure the ongoing integrity and efficiency of modernized electricity networks.

Future Work. In future research, there is potential to refine the GAN-GRID attack to improve its effectiveness and success rate while reducing deployment time. This could entail exploring various generator architectures, optimization techniques, and injection strategies to optimize the attack process. Furthermore, a primary focus will be on developing defenses against GAN-GRID attacks and investigating potential countermeasures. Additionally, examining poisoning attacks could offer valuable insights into the resilience of stability prediction systems. By establishing a new system and threat model that accounts for this type of attack, we aim to identify vulnerabilities within the models and strengthen their security posture. Moreover, in addition to addressing GAN-GRID attacks in stability prediction systems, future research may entail evaluating the impact of these attacks on other ML-based systems, such as fault prediction systems in SGs.

References

1. Ahmadian, S., Malki, H., Han, Z.: Cyber attacks on smart energy grids using generative adverserial networks. In: 2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 942–946 (2018). https://doi.org/10.1109/GlobalSIP.2018.8646424

2. Alazab, M., Khan, S., Krishnan, S.S.R., Pham, Q.V., Reddy, M.P.K., Gadekallu, T.R.: A multidirectional LSTM model for predicting the stability of a smart grid. IEEE Access **8**, 85454–85463 (2020). https://doi.org/10.1109/ACCESS.2020.2991067
3. Arzamasov, V.: Electrical grid stability simulated data. UCI Mach. Learn. Repository (2018). https://doi.org/10.24432/C5PG66
4. Bai, T., Luo, J., Zhao, J., Wen, B., Wang, Q.: Recent advances in adversarial training for adversarial robustness. In: Zhou, Z.H. (ed.) Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, pp. 4312–4321. International Joint Conferences on Artificial Intelligence Organization (2021). https://doi.org/10.24963/ijcai.2021/591
5. Bashir, A.K., et al.: Comparative analysis of machine learning algorithms for prediction of smart grid stability†. Int. Trans. Electr. Energy Syst. **31**(9), e12706 (2021). https://doi.org/10.1002/2050-7038.12706
6. Bera, B., Saha, S., Das, A.K., Vasilakos, A.V.: Designing blockchain-based access control protocol in IoT-enabled smart-grid system. IEEE Internet Things J. **8**(7), 5744–5761 (2021). https://doi.org/10.1109/JIOT.2020.3030308
7. Bird, L., Milligan, M., Lew, D.: Integrating variable renewable energy: challenges and solutions. Tech. rep., National Renewable Energy Lab.(NREL), Golden, CO (United States) (2013)
8. Breviglieri, P., Erdem, T., Eken, S.: Predicting smart grid stability with optimized deep models. SN Comput. Sci. **2**, 1–12 (2021)
9. Chen, T.M., Abu-Nimeh, S.: Lessons from stuxnet. Computer **44**(4), 91–93 (2011)
10. Chen, Y., Tan, Y., Deka, D.: Is machine learning in power systems vulnerable? In: 2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), pp. 1–6 (2018). https://doi.org/10.1109/SmartGridComm.2018.8587547
11. Ciaramella, G., Martinelli, F., Mercaldo, F., Santone, A.: Explainable deep learning for smart grid stability detection. In: 2023 IEEE International Conference on Big Data, pp. 6131–6137 (2023). https://doi.org/10.1109/BigData59044.2023.10386170
12. CISA: The attack on colonial pipeline: what we've learned and what we've done over the past two years. https://www.cisa.gov/news-events/news/attack-colonial-pipeline-what-weve-learned-what-weve-done-over-past-two-years. Accessed 20 Apr 2024
13. Dewangan, F., Biswal, M., Patnaik, B., Hasan, S., Mishra, M.: Chapter five - smart grid stability prediction using genetic algorithm-based extreme learning machine. In: Bansal, R.C., Mishra, M., Sood, Y.R. (eds.) Electric Power Systems Resiliency, pp. 149–163. Academic Press (2022). https://doi.org/10.1016/B978-0-323-85536-5.00011-4
14. Efatinasab, E., Marchiori, F., Brighente, A., Rampazzo, M., Conti, M.: FaultGuard: a generative approach to resilient fault prediction in smart electrical grids. arXiv preprint arXiv:2403.17494 (2024)
15. Erdem, T., Eken, S.: Layer-wise relevance propagation for smart-grid stability prediction. In: Djeddi, C., Siddiqi, I., Jamil, A., Ali Hameed, A., Kucuk, İ (eds.) MedPRAI 2021. CCIS, vol. 1543, pp. 315–328. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-04112-9_24
16. Federation of American Scientists: Grid failure and extreme heat (2024). https://fas.org/publication/grid-failure-extreme-heat/

17. Forbes: 3 alarming threats to the U.S. energy grid - Cyber, physical, and existential events. https://www.forbes.com/sites/chuckbrooks/2023/02/15/3-alarming-threats-to-the-us-energy-grid--cyber-physical-and-existential-events/. Accessed 20 Apr 2024
18. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv:1412.6572 (2015)
19. Gorzałczany, M.B., Piekoszewski, J., Rudziński, F.: A modern data-mining approach based on genetically optimized fuzzy systems for interpretable and accurate smart-grid stability prediction. Energies **13**(10), 2559 (2020). https://doi.org/10.3390/en13102559
20. Gupta, A., Gurrala, G., Sastry, P.S.: An online power system stability monitoring system using convolutional neural networks. IEEE Trans. Power Syst. **34**(2), 864–872 (2019). https://doi.org/10.1109/TPWRS.2018.2872505
21. Hao, J., Tao, Y.: Adversarial attacks on deep learning models in smart grids. Energy Rep. **8**, 123–129 (2022). https://doi.org/10.1016/j.egyr.2021.11.026, https://www.sciencedirect.com/science/article/pii/S2352484721011707, 2021 6th International Conference on Clean Energy and Power Generation Technology
22. He, Y., Mendis, G.J., Wei, J.: Real-time detection of false data injection attacks in smart grid: a deep learning-based intelligent mechanism. IEEE Trans. Smart Grid **8**(5), 2505–2516 (2017). https://doi.org/10.1109/TSG.2017.2703842
23. Jiao, J.: Application and prospect of artificial intelligence in smart grid. IOP Conf. Ser. Earth Environ. Sci. **510**(2), 022012 (2020). https://doi.org/10.1088/1755-1315/510/2/022012
24. Jung, M., Hofer, T., Döbelt, S., Kienesberger, G., Judex, F., Kastner, W.: Access control for a smart grid SOA. In: 2012 International Conference for Internet Technology and Secured Transactions, pp. 281–287 (2012)
25. Ke, X., Wu, D., Rice, J., Kintner-Meyer, M., Lu, N.: Quantifying impacts of heat waves on power grid operation. Appl. Energy **183**, 504–512 (2016). https://doi.org/10.1016/j.apenergy.2016.08.188
26. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. arXiv:1607.02533 (2017)
27. Li, J., Yang, Y., Sun, J.S.: SearchFromFree: adversarial measurements for machine learning-based energy theft detection. In: 2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), pp. 1–6 (2020). https://doi.org/10.1109/SmartGridComm47815.2020.9303013
28. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS 2017, pp. 4768–4777 (2017)
29. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv:1706.06083 (2019)
30. Massaoudi, M., Abu-Rub, H., Refaat, S.S., Chihi, I., Oueslati, F.S.: Accurate smart-grid stability forecasting based on deep learning: Point and interval estimation method. In: 2021 IEEE Kansas Power and Energy Conference (KPEC), pp. 1–6 (2021). https://doi.org/10.1109/KPEC51835.2021.9446196
31. Muqeet, H.A., Liaqat, R., Jamil, M., Khan, A.A.: A state-of-the-art review of smart energy systems and their management in a smart grid environment. Energies **16**(1), 472 (2023). https://doi.org/10.3390/en16010472
32. Musleh, A.S., Chen, G., Dong, Z.Y.: A survey on the detection algorithms for false data injection attacks in smart grids. IEEE Trans. Smart Grid **11**(3), 2218–2234 (2020). https://doi.org/10.1109/TSG.2019.2949998

33. Nafees, M.N., Saxena, N., Cardenas, A., Grijalva, S., Burnap, P.: Smart grid cyber-physical situational awareness of complex operational technology attacks: a review. ACM Comput. Surv. **55**(10), 1–36 (2023)
34. Nicolae, M.I., et al.: Adversarial robustness toolbox v1. 0.0. arXiv preprint arXiv:1807.01069 (2018)
35. Önder, M., Dogan, M.U., Polat, K.: Classification of smart grid stability prediction using cascade machine learning methods and the internet of things in smart grid. Neural Comput. Appl. **35**, 17851–17869 (2023). https://doi.org/10.1007/s00521-023-08605-x
36. Saxena, N., Choi, B.J.: State of the art authentication, access control, and secure integration in smart grid. Energies **8**(10), 11883–11915 (2015). https://doi.org/10.3390/en81011883
37. Shi, Z., et al.: Artificial intelligence techniques for stability analysis and control in smart grids: methodologies, applications, challenges and future directions. Appl. Energy **278**, 115733 (2020). https://doi.org/10.1016/j.apenergy.2020.115733
38. Shi, Z., et al.: Convolutional neural network-based power system transient stability assessment and instability mode prediction. Appl. Energy **263**, 114586 (2020). https://doi.org/10.1016/j.apenergy.2020.114586
39. Song, Q., Tan, R., Ren, C., Xu, Y.: Understanding credibility of adversarial examples against smart grid: a case study for voltage stability assessment. In: Proceedings of the Twelfth ACM International Conference on Future Energy Systems, e-Energy 2021, pp. 95–106. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3447555.3464859
40. Song, Q., et al.: On credibility of adversarial examples against learning-based grid voltage stability assessment. IEEE Trans. Dependable Secure Comput. **21**(2), 585–599 (2024). https://doi.org/10.1109/TDSC.2022.3213012
41. Sullivan, J.E., Kamensky, D.: How cyber-attacks in Ukraine show the vulnerability of the us power grid. Electr. J. **30**(3), 30–35 (2017)
42. Tounsi, W.: Cyber deception, the ultimate piece of a defensive strategy - proof of concept. In: 2022 6th Cyber Security in Networking Conference (CSNet), pp. 1–5 (2022). https://doi.org/10.1109/CSNet56116.2022.9955605
43. Ucar, F.: A comprehensive analysis of smart grid stability prediction along with explainable artificial intelligence. Symmetry **15**(2), 289 (2023). https://doi.org/10.3390/sym15020289
44. Xu, J., Li, Z., Du, B., Zhang, M., Liu, J.: Reluplex made more practical: leaky ReLU. In: 2020 IEEE Symposium on Computers and Communications (ISCC), pp. 1–7. IEEE (2020)
45. Zhang, Y., Zhang, H., Zhang, J., Li, L., Zheng, Z.: Power grid stability prediction model based on BiLSTM with attention. In: ISEEIE 2021, 2021 International Symposium on Electrical, Electronics and Information Engineering, pp. 344–349. Association for Computing Machinery (2021). https://doi.org/10.1145/3459104.3459160