Understanding Credibility of Adversarial Examples against Smart Grid: A Case Study for Voltage Stability Assessment

Qun Song
ERI@N, Interdisciplinary Graduate School
School of Computer Science and Engineering
Nanyang Technological University
Singapore
song0167@ntu.edu.sg

Rui Tan
School of Computer Science and Engineering
Nanyang Technological University
Singapore
tanrui@ntu.edu.sg

Chao Ren
ERI@N, Interdisciplinary Graduate School
School of Electrical and Electronic Engineering
Nanyang Technological University
Singapore
renc0003@ntu.edu.sg

Yan Xu
School of Electrical and Electronic Engineering
Nanyang Technological University
Singapore
xuyan@ntu.edu.sg

ABSTRACT

Stability assessment is an important task for maintaining reliable operations of power grids. With increased system complexity, deep learning-based stability assessment approaches are promising to address the shortfalls of the traditional time-domain simulation-based approaches. However, in the field of computer vision, the deep learning models are shown vulnerable to adversarial examples. Although this vulnerability has been noticed by the energy informatics research, the domain-specific analysis on the requirements imposed for implementing effective adversarial examples is still lacking. These attack requirements, albeit reasonable in computer vision tasks, can be too stringent in the context of power grids. In this paper, we systematically investigate the requirements and discuss the credibility of six representative adversarial example attacks for a case study of voltage stability assessment for the New England 10-machine 39-bus system. We show that (1) compromising the voltage traces of half of transmission system buses is a rule of thumb requirement; (2) the universal adversarial perturbations that are independent of the original clean voltage trajectory have the same credibility as the widely studied false data injection attacks on power grid state estimation, while other adversarial example attacks are less credible; (3) the universal perturbations can be effectively defended with strong adversarial training.

CCS CONCEPTS

• **Hardware → Smart grid**; • **Computing methodologies → Neural networks**; • **Security and privacy → Software and application security**.

e-Energy '21, June 28–July 2, 2021, Virtual Event, Italy

ACM ISBN 978-1-4503-8333-2/21/06…$15.00
https://doi.org/10.1145/3447555.3464859

KEYWORDS

Adversarial example, machine learning, cybersecurity, voltage stability assessment, smart grid

ACM Reference Format:
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 *The Twelfth ACM International Conference on Future Energy Systems (e-Energy '21), June 28–July 2, 2021, Virtual Event, Italy*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3447555.3464859

1 INTRODUCTION

Electric power grid is a critical cyber-physical system (CPS) that maintains reliable and economical generation, transmission, and distribution of electricity. It usually consists of the *generating stations* that convert the energy from other forms to electricity, the *transmission system* that carries the electric power from generating stations to load buses, and the *distribution systems* that distribute the electric power to the end customers. A control center monitors and manages the power grid to ensure the efficient and sustained operations [20]. By integrating modern information and communication technologies (ICTs), the traditional power grids are evolving into smart grids that possess improved sensing and control capabilities to deal with the new challenges caused by the increasing deployments of renewable energy, distributed generation, and demand response. Machine learning, as an ICT, has been considered and adopted for enhancing various grid capabilities such as load forecasting [34], fault detection [26], and automatic generation control [40].

The *deep neural networks* (DNNs) enabled by the advances of computing hardware acceleration have shown appealing efficiency in learning sophisticated patterns from big data. Thus, there are growing interests of applying deep learning to power grids [10, 26, 31, 34, 41]. However, the complex structures of DNNs engender vulnerabilities under adversarial settings. In this paper, we focus on the *adversarial example* threat [13]. It aims at misleading the DNN to yield wrong inference results by adding minute perturbations to the inputs. In particular, it is a specific type of the *false data injection*

(FDI) attack that has been widely studied under the context of power grid [27].

While DNNs can be used for various power grid operation tasks, this paper considers a representative task of online *voltage stability assessment* (VSA) to substantiate the evaluation and analysis. Losing stability can lead to widespread, catastrophic blackouts threatening people's properties and lives. Thus, maintaining stability is a fundamental requirement of any power system. The time-domain simulations used for offline VSA during the design of the power grid check the stability of the voltages at the transmission system buses under presumed disturbances. Although a high-fidelity system model can yield accurate VSA results, due to the power system complexities, the simulations are usually much slower than the evolution of the physical processes and ill-suited for online VSA. To develop online VSA capability that enables timely and proper reaction to a contingency, the grid operator can run extensive offline simulations under various disturbances and use the results to form a look-up table or train a machine learning model for online VSA on real-time voltage measurements [9]. Applying DNNs to better capture the inherent complexity of voltage dynamics and advance online VSA is an ongoing interest of the energy informatics research [16, 38, 42].

Wrong outputs of the DNN-based VSA can lead to catastrophic consequences. A false negative in detecting instability can cause missed or delayed activation of fault isolation, which can potentially result in widespread blackout; a false positive can cause unnecessary load shedding and thereby inconvenience and misery of the customers losing power. Thus, the cybersecurity risks faced by DNN-based VSA due to adversarial examples need to be understood. Various adversarial example construction algorithms have been proposed [3] and their effectiveness have been demonstrated in the safety-critical CPSes using computer vision (CV) to perceive the environment. For instance, an adversarial sticker pasted on road can mislead Tesla Autopilot to direct a car to the opposite lane [2]. However, the requirements for implementing these attacks, though reasonable in the CV tasks, may be too stringent in the context of VSA. For example, the attacks often require the original clean input to compute the malicious perturbation. In the CV-based lane recognition task, the camera's view of the road area as the clean input can be known *a priori* to the attacker and used to design the adversarial sticker. However, in VSA, obtaining the read access to all the transmission buses' real-time voltages for constructing effective adversarial examples can be a strong requirement. Coordinating the real-time eavesdropping and the data tampering in implementing certain attacks is a subtle task imposing high requirements on the attacker's resources and skills.

Therefore, simply transferring every worry from CV applications to DNN-based smart grid tasks without discrimination may hinder innovations. To the best of our knowledge, systematic studies on the credibility of adversarial example attacks with due discrimination on the requirements of implementing them in smart grids are still lacking. In this paper, we conduct a systematic case study to evaluate the effectiveness of various methods for constructing adversarial examples against VSA, which impose different requirements on (1) read access to the original clean voltage measurements, (2) write access to the voltage measurements, (3) knowledge about the DNN's internals, and (4) access to the DNN's training data. We also evaluate their effectiveness when the system defender adopts the prevailing countermeasures of *model hardening* and *input cleansing*. By relating the attack effectiveness with the attack requirement and also analyzing the difficulty/overhead of meeting the attack requirement, our evaluation results provide a comprehensive understanding on the credibility of the various adversarial example attacks on VSA.

From the case study, we summarize a methodology for evaluating the credibility of various types of adversarial example attacks on the DNN-based smart grid applications as follows. The methodology includes the following steps: (a) to investigate the individual attack model for each of the considered adversarial example attacks characterized by the minimal requirements needed to effectively mislead the DNN of the smart grid application; (b) to evaluate the credibility of the attacks through analyzing the feasibility of the requirements under the context of the considered smart grid application; and (c) to evaluate the effectiveness of prevailing countermeasures in protecting the smart grid application against the credible adversarial attacks.

The main contributions of this paper are summarized as follows:

- We study six adversarial example construction methods, i.e., FGSM [13], PGD [23], DeepFool [30], Carlini-Wagner [6], Universal Adversarial Perturbation (UAP) [29], and Universal Adversarial Network (UAN) [14]. We investigate the minimal requirement of implementing each of them to achieve effective attack on VSA.
- We show that tampering with the voltages of half of buses is a rule of thumb for adversarial examples to be effective. Moreover, the *universal adversarial example* attacks (i.e., UAP and UAN) that do not require read access of bus voltages have the same credibility as the FDI attack on grid state estimation [27] that has received much research attention. The other attacks are less credible because of their indispensable requirement on real-time bus voltage read access.
- We study the effectiveness of model hardening by adversarial training and input cleansing by APE-GAN [17] in defending each of the six adversarial example attacks. We show that the adversarial training using PGD adversarial examples can effectively protect the DNN-based VSA against the credible universal adversarial examples.

This paper is organized as follows. Section 2 reviews related work. Section 3 presents the background and preliminaries. Section 4 states the problem studied in this paper. Section 5 and Section 6 present the evaluation results on the attack requirement and defense effectiveness, respectively. Section 7 concludes this paper and discusses future work directions.

2 RELATED WORK

Applications of machine learning in power systems. In literature, machine learning-based approaches have been proposed for load [34] and price [31] forecasting, wind [31] and solar [41] power prediction, fault diagnosis [26] and FDI detection [10]. Deep reinforcement learning is considered for various power grid controls such as voltage control [39], frequency control [40], and emergency control [15]. Applying machine learning for VSA addresses the limitations of the conventional simulation-based VSA, i.e., poor

real-time performance [8] and scalability with respect to the power system size [19]. Machine learning algorithms used for VSA include extreme learning-based neural networks [38], recurrent neural network [16], and ensemble learning [42]. These studies focus on devising models that achieve good application performance and do not consider the potential cybersecurity risks caused by the use of machine learning models. This paper studies the adversarial example attacks against the machine learning model used for VSA and the countermeasures.

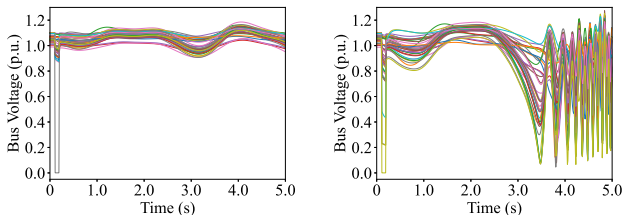
FDI attacks on power systems. The use of modern ICTs in power systems introduce cybersecurity concerns. The work [27] shows that FDI on power flow measurements can mislead state estimation and bypass the bad data detection mechanism at the control center. A further study [11] advances the attack by removing the requirement on the prior knowledge of the power grid topology. More studies show that FDI can be designed to mislead electricity market operations [37], energy routing processes [25], Optimal Power Flow (OPF) analysis [32], frequency control [35], centralized voltage control [24], and distributed voltage control [18]. In addition, the studies [18, 24, 35] consider the *optimal FDI* that schedules the FDI sequence to minimize the time left for the power grid to react [35], maximize the state estimation error [24] and voltage deviation [18]. Countermeasures against FDIs on power systems have also been studied, including attack detection [35] and mitigation [5]. The work [24] analyzes a joint detection-mitigation mechanism based on a Markov decision process formulation. The above studies consider the strategic planning of FDI. However, the targets of the FDI are not DNN-based.

Adversarial example attacks on power grid. Adversarial example attack is a specific form of FDI that aim at misleading DNN. A recent work [7] studies the impact of adversarial examples on the DNN-based load forecasting. Different from the impact analysis of a single adversarial example attack in [7], we perform a requirement analysis for six adversarial example attacks to investigate the conditions that the adversary needs to satisfy to launch effective adversarial example attacks based on a case-study application of VSA. The six attacks are representative ones frequently evaluated in literature [33]. Meanwhile, the construction of these attacks imposes distinct minimal requirement on the adversary. The minimal requirement provides insights into understanding the credibility of adversarial examples in the context of power systems. In addition, we investigate the effectiveness of prevailing countermeasures, while the existing work [7] does not consider defense.

3 BACKGROUND AND PRELIMINARIES

3.1 DNN-based Online Short-Term VSA

A stable power system can regain an equilibrium state after a disturbance [21]. It is essential to assess the power system stability against potential disturbances because loss of stability may result in loss of load in an area or tripping of transmission lines, leading to cascading failures and even widespread blackout [1]. Stability is often assessed in terms of *rotor angle*, *frequency*, and *voltage*. From the time scale of the post-contingency dynamics, short-term and long-term stability assessments concern time horizons of a few seconds and up to minutes, respectively. In this paper, we focus on short-term VSA, which classifies a seconds-long voltage trajectory



(a) Stable. **(b) Unstable.**

Figure 1: Example of stable and unstable situations.

that consists of the traces of the transmission buses' voltages into *stable* or *unstable*. Our study can be also extended to other types of stability assessment. Fig. 1 shows stable and unstable voltage trajectories over 5.0 seconds, which are caused by a fault occurring at 0.1 seconds followed by an automated fault clearance at 0.2 seconds. In the stable cases, the voltages of all buses restore to acceptable levels (less than 10% deviations from the nominal values). In the unstable cases, the voltages remain unacceptably far away from the nominal values or even collapse.

Offline VSA can be conducted using time-domain simulations with an extensive set of potential faults occur [20]. In contrast, online VSA, which needs to yield results based on real-time voltage measurements before a hard deadline for restorative actions should the system assessed unstable, faces two main challenges. First, the system operator often has limited or no information at run time regarding the occurred fault, which, however, are needed to bootstrap the time-domain simulation. Second, the time-domain simulation is often much slower than the state evolution of the power systems. To address these challenges, machine learning has been applied for online VSA [9, 16, 42]. Specifically, based on a training dataset $(\mathbf{X}, \mathbf{Y}) = [(\mathbf{x}_i, y_i), i = 1, ..., m]$, where $\mathbf{x}_i$ represents a post-fault voltage trajectory generated by an offline time-domain simulation and y_i represents the corresponding stability classification, a machine learning model $f(\mathbf{x}; \boldsymbol{\theta})$ with weights $\boldsymbol{\theta}$ can be trained to classify a voltage trajectory $\mathbf{x}$ at run time. With abundant training data, the well-trained model $f(\mathbf{x}; \boldsymbol{\theta})$ can handle a wide range of faults.

3.2 Adversarial Example Taxonomy

The study [3] provides a taxonomy of adversarial example construction methods, which is illustrated in Table 1. In term of applicable *scope* of the attack, *input-specific* means that a crafted perturbation is effective against a specific clean example, while *universal* means the perturbation is effective against many clean examples. In terms of the *computation* required, an adversarial example can be constructed by a *one-shot* computation step (e.g., by using a closed-form formula) or *iterative* computation that often involves a search process. In terms of the *knowledge* about the target DNN, the *white-box* methods require full knowledge about the DNN's internals including its architecture and weights, while the *black-box* methods only require the access to run the DNN without knowing its internals. Most effective methods require white-box knowledge. Some of them are still effective in the black-box setting by using a surrogate DNN. The adversary can query the black-box target

Table 1: Adversarial example construction methods.

Attack	Categorization [3]		
	Scope	Computation	Knowledge
FGSM [13]	Input-specific	One-shot	White/black-box
PGD [23]	Input-specific	Iterative	White/black-box
DF [30]	Input-specific	Iterative	White-box
CW [6]	Input-specific	Iterative	White-box
UAP [29]	Universal	Iterative	White/black-box
UAN [14]	Universal	Iterative	White/black-box

DNN with many input samples, train the surrogate DNN using the inputs and the target DNN's outputs, and then construct the adversarial examples against the surrogate DNN. In Table 1, we label such methods with "white/black-box."

As summarized in Table 1, this paper considers six representative adversarial example construction methods, i.e., Fast Gradient Sign Method (FGSM) [13], Project Gradient Descent (PGD) [23], DeepFool (DF) [30], Carlini and Wagner's method (CW) [6], Universal Adversarial Perturbation (UAP) [29], and Universal Adversarial Network (UAN) [14]. While Appendix A provides the formulations of these six construction methods, we describe their essences as follows. FGSM adds a one-step perturbation to the clean input sample. PGD performs multiple small-step FGSMs iteratively. DF finds the perturbation with the minimum distance from the clean input sample to the decision boundary. CW applies the Lagrangian relaxation to simplify the adversarial example construction problem (cf. Appendix A) to unconstrained optimization and then searches the solution. The above four methods are input-specific. The UAP and UAN are universal. UAP uses DF to find the perturbation for each clean input sample of the training dataset and accumulates the perturbations to form a single universal perturbation. UAN is a generative neural network that transforms a value randomly sampled from a distribution to a perturbation that can likely mislead the DNN.

3.3 Defenses against Adversarial Examples

Existing defenses can be divided into the *model hardening* and *input cleansing* categories. Model hardening modifies the target DNN to improve robustness against adversarial examples. *Adversarial training* is a model hardening technique that attempts to improve model robustness by including adversarial examples with their correct labels into the training dataset for model training. Existing studies [28] and the Competition on Adversarial Attacks and Defenses [36] show that adversarial training gives state-of-the-art performance on various benchmarks. The input cleansing defenses attempt to remove or disrupt the adversarial perturbations [3]. Compared with *ad hoc* approaches (e.g., data compression, foveation, and randomization [3]), APE-GAN [17] is a systematic input cleansing defense approach that aims to learn a manifold mapping from adversarial examples to original clean examples. The APE-GAN is trained under the generative adversarial network (GAN) setting [12]. With the help of a discriminator that aims to differentiate the clean input samples and outputs of the generator, the trained generator can cleanse the input adversarial example and output a benign counterpart. The detailed formulations of the adversarial training and APE-GAN can be found in Appendix A.

4 PROBLEM STATEMENT

4.1 System and Data Description

The power system considered in this case study is the 10-machine 39-bus New England system [4]. The system's single-line diagram can be found in Appendix B. We perform extensive time-domain simulations to generate voltage trajectories. In each simulation, a three-phase fault that lasts for a random time duration ranging from 0.1 to 0.3 seconds is injected to a randomly selected bus. The fault is cleared by a single or double transmission lines tripping, which simulates different topology change scenarios. Each voltage trajectory consists of the voltage traces of the 39 buses. The sampling rate is 100 samples per second. We generate 6,536 voltage trajectories from the simulations that cover a wide range of practical operating points of the power system. We divide the generated voltage trajectories into training, validation, and testing datasets with 4,536, 1,000, and 1,000 samples. Each sample is a 1×3900 data vector containing the 39 buses' voltage traces over a one-second duration after the clearance of the fault. For VSA, we use a convolutional neural network (CNN). The CNN has two convolutional layers with 128 1×5 filters followed by a 1×2 max pooling layer, two convolutional layers with 256 1×5 filters followed by a 1×2 max pooling layer, two fully connected layers with 512 rectified linear units (ReLUs) each, and a binary-class softmax layer. The validation accuracy of the trained CNN is 99.5%. Specifically, the empirically measured false positive rate and the false negative rate are 0% and 0.5%, respectively, in detecting the instability.

4.2 Threat Models and Research Problem

The general objective of the six adversarial example attacks is to mislead the target DNN, while minimizing the perturbation. As summarized in Table 1, the six attack construction methods have distinct features. Each of them imposes a distinct set of minimal requirements that need to be satisfied to render the attack effective. Thus, the six attack construction methods correspond to different threat models. The union set of their requirements contains the following four specific requirements.

(1) Read access to the clean voltage measurements: This is related to the applicable scope (i.e., input-specific or universal) of the adversarial example. An input-specific attack needs this read access. It cannot compute the adversarial perturbation until the whole clean voltage trajectory is obtained using this read access. In addition, the attack needs to add the perturbation to the voltage trajectory before it is fed to the DNN for VSA. Differently, the universal attacks do not need this access.

(2) Write access to the voltage measurements: A voltage trajectory consists of the voltage traces of all transmission system buses. The number of voltage traces that the adversary needs to tamper with is an important attack requirement aspect that is related to the cost and overhead of launching the attack. The full write access (i.e., being able to tamper with the voltage traces of all transmission system buses) apparently implies a strong and resourceful adversary.

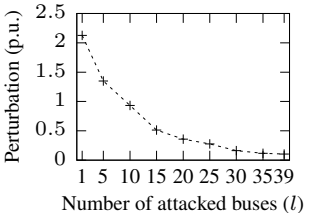


Figure 2: Per-bus average perturbation (Attack: DF).

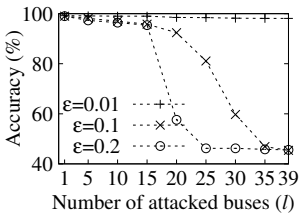


Figure 3: Accuracy vs. l and ϵ (Attack: PGD).

(3) Knowledge about DNN's internals: This is related to the white-/black-box features of the attack as summarized in Table 1.

(4) Access to DNN's training data: This specifies whether a considered adversarial example construction method needs the dataset used to train the target DNN, which can consist of labeled historical voltage trajectories.

In this paper, we inquire three issues. First, we investigate the minimal set of requirements that each attack needs to satisfy to mislead the VSA DNN. The results will depict precisely the threat models of the attack construction methods. Second, we analyze the credibility of the adversarial examples based on their minimal requirements. The different requirement aspects should be weighed differently, e.g., meeting the *real-time* requirement aspects (1) and (2) is often more difficult than meeting the *static* requirement aspects (3) and (4). Third, we evaluate whether the prevailing countermeasures can protect VSA against the credible attacks.

5 ATTACK REQUIREMENT INVESTIGATION

5.1 Attack Evaluation Settings

The effectiveness of attack is evaluated in terms of the accuracy of the target DNN on 1,000 perturbed samples from the test dataset described in Section 4.1. With an input-specific attack, an 1×3900 perturbation vector is computed for each specific test sample using the target DNN under the white-box setting or the surrogate DNN under the black-box setting. We set the surrogate DNN to have the same hyperparameters as the target DNN and train it from a random initialization using the training dataset. With UAP, we compute a fixed universal adversarial perturbation vector using 1,000 samples randomly selected from the training dataset and then apply it to all the 1,000 test samples. With UAN, we train the attack generator using 1,000 samples randomly selected from the training dataset. For each of the 1,000 test samples, the generator takes as input a Gaussian random vector and generates a 1×3900 adversarial perturbation vector. The generator's hyperparameters are adopted from [14].

5.1.1 Implementation of partial perturbation. As discussed in Section 4.2, the number of voltage traces that the adversary needs to tamper with is a key requirement. Now, we present our implementation of the adversarial example construction that only needs write access to l buses. The formulation of such *partial perturbation* is: $\boldsymbol{\delta}^* = \text{argmin}_{\boldsymbol{\delta}} D(\mathbf{x}, \mathbf{x}')$ subject to $f(\mathbf{x}'; \boldsymbol{\theta}) \neq y$ and only the input dimensions of $\mathbf{x}$ correspond to the l buses are modified. We use a mask $\mathbf{M}$, which is a matrix that has the same shape as the input of the target model and has unit values in the area corresponding to the l buses where adversarial perturbations are added and zero values in the area where no perturbation is added. $\mathbf{M}$ is used to restrict the area where the adversary can modify the measurements. For the one-step adversarial example, i.e., the FGSM attack, the computed adversarial perturbation is multiplied by $\mathbf{M}$ and added to each clean example. For the iterative adversarial example, the multiplication is performed at each iteration of the attack construction process. In the experiments, we set the value of l to be 1, 5, 10, 15, 20, 25, 30, and 39. For each setting of l, we perturb the fixed first l bus voltage measurements.

Table 2: Requirements for effective attacks against VSA.

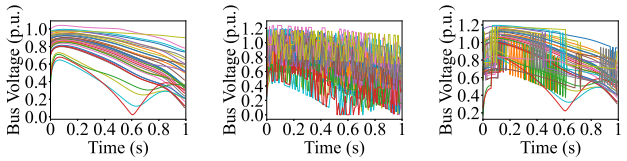
Attack	Minimal requirement			
	Access		Knowledge	
	Read	Write	DNN internal	Training data
FGSM [13]	Yes	Partial	Either	
PGD [23]	Yes	Partial	Either	
DF [30]	Yes	Partial	Yes	No
CW [6]	Yes	Full	Yes	No
UAP [29]	No	Partial	No	Yes
UAN [14]	No	Partial	No	Yes

5.1.2 Attack perturbation intensity settings. Intuitively, larger perturbations are more effective in misleading the target DNN. Thus, it is non-trivial to configure the attack perturbation intensity so that the comparison of attack effectiveness is fair. As DF finds the minimal perturbation needed to mislead the target DNN, we use DF to guide the settings of ϵ for FGSM, PGD, UAP, and UAN and κ for CW. Fig. 2 shows the per-bus average ℓ_2-norm of the perturbations found by DF versus the number of attacked buses (i.e., l). The intensity of the DF perturbation decreases with l because the needed perturbation intensity is larger to mislead the DNN when the perturbation is limited to fewer buses. When $1 \leq l \leq 25$, the average perturbation intensity is from 0.27 p.u. to 2.12 p.u., which is unreasonably large since the nominal bus voltage is 1 p.u.. Thus, we consider $30 \leq l \leq 39$ for DF such that the average perturbation intensity is at most 0.16 p.u..

Fig. 3 shows the DNN's accuracy versus l under various ϵ settings from 0.01 p.u. to 0.2 p.u.. The settings of 0.1 p.u. and 0.2 p.u. can render the attack effective when the adversary can tamper with sufficient bus voltage readings. We set $\epsilon = 0.2$ p.u. for FGSM, PGD, UAP, and UAN. For CW, when we set $\kappa = 0$, the average perturbation intensity is 0.18 p.u., which is similar to that of DF. In summary, by setting $\epsilon = 0.2$ p.u. and $\kappa = 0$, the comparison among the six attacks has a relatively fair basis. Note that the magnitude of the adversarial perturbation is a configurable parameter. In the VSA case study, we set the maximum allowed deviation from the nominal bus voltage to be ±0.2 p.u.. Under this setting, we evaluate the worse-case vulnerability of the VSA DNN.

5.2 Attack Effectiveness and Requirements

The evaluation results are summarized in Table 2, which presents the minimal requirement of each of the six attack construction methods to effectively mislead the VSA DNN. The "read access"



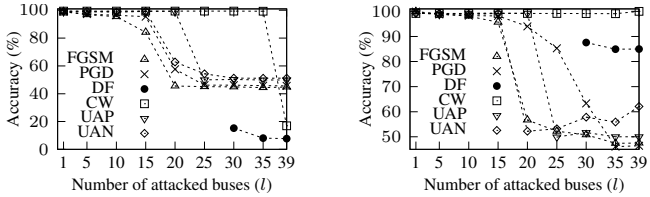
(a) Clean (unstable) (b) Random (c) Adversarial

Figure 4: Clean, randomly perturbed, and FGSM-perturbed bus voltage trajectories. The clean sample in (a) is classified as unstable; the randomly perturbed sample in (b) is correctly classified as unstable; the FGSM-perturbed sample in (c) is wrongly classified as stable.

column specifies whether the attack needs to obtain the clean voltage trajectory. The "write access" column specifies whether the attack needs to tamper with the voltage traces of all buses (full) or just a portion of them (partial) to be effective. The "DNN internal" and "training data" columns specify whether obtaining the DNN internal and a training dataset are needed, respectively.

5.2.1 Random perturbations versus adversarial examples. Fig. 4 shows a clean, unstable bus voltage trajectory, and its randomly perturbed and FGSM-perturbed counterparts. Each element of the random perturbation is randomly and independently sampled from the standard normal distribution and clipped to $[-0.2, 0.2]$ p.u.. The FGSM perturbations are applied to all bus voltage readings with 0.2 p.u. maximum perturbation intensity. In the presence of random perturbation, the DNN still achieves 99.3% test accuracy, which is just 0.2% lower compared with the accuracy on clean samples. However, in the presence of FGSM attack, the accuracy drops to 45.4%. These results show that, even if the adversarial is strong enough to compromise all bus voltage readings, they still need to apply intelligence to schedule the perturbation.

5.2.2 Effectiveness of input-specific adversarial examples. Fig. 5a shows the accuracy of the DNN in the presence of white-box attacks. When $1 \le l \le 15$, the effectiveness of the input-specific attacks (e.g., FGSM, PGD, DF, and CW) are limited. The lowest accuracy is 84.2%, which is caused by FGSM with $l = 15$. When $l = 20$, the accuracy of the DNN under FGSM attack drops to 45.5%. The accuracy under PGD attack drastically drops from 95.4% to 57.6% when l increases from 15 to 20. These results suggest that by tampering with about 50% of the bus voltage measurements, white-box input-specific adversarial examples can result in more than 50% VSA accuracy drops. When $30 \le l \le 39$, the DF attack is very effective. The DNN accuracies under DF attack are only 15.2% and 8.0% when l is 30 and 35, respectively. The effectiveness of CW is rather limited when only a portion of bus voltage measurements are under attack (i.e., $1 \le l < 39$). However, when CW can tamper with all bus voltage measurements (i.e., $l = 39$), the DNN accuracy drops to 15.5%. This suggests that, although CW is often considered the most effective adversarial example construction method in computer vision applications due to its optimization-based formulation [3], its effectiveness against VSA is conditioned on the write access to all input dimensions. In contrast, the gradient-based methods (i.e., FGSM, PGD, and DF) achieve non-negligible attack effectiveness



(a) White-box attacks (b) Black-box attacks

Figure 5: DNN accuracy in the presence of attack.

when partial input dimensions are under attack. These results are summarized in the "write access" column of Table 2.

Fig. 5b shows the results under black-box attacks. In the presence of DF and CW attacks, the DNN's accuracy remains at 84.9% and 99.6%, respectively. Thus, DF and CW are ineffective under the black-box setting. It is because the adversarial examples constructed by DF and CW overfit to the surrogate DNN and thus have reduced effectiveness against the target DNN. FGSM adversarial examples exhibit the highest transferability from the surrogate to the target DNN, compared with other input-specific attacks. Specifically, the DNN accuracy decreases from 95.7% to 56.7% when l increases from 15 to 20, which is similar to the results obtained under the white-box setting. PGD's effectiveness reduces when switching from the white-box to the black-box setting. However, the DNN accuracy still has a non-negligible drop from 63.3% to 46.1% when l increases from 30 to 35. The observed good transferability of FGSM and PGD adversarial examples from the surrogate DNN to the target DNN is because that they underfit the surrogate DNN. The above results suggest that preserving the confidentiality of the target model is a weak defense against FGSM and PGD. These results are summarized in the "DNN internal" column of Table 2. The "either" notes mean that, under the white-box setting, FGSM and PGD only require the DNN internal; under the black-box setting, FGSM and PGD require the training data to build the surrogate DNN.

5.2.3 Effectiveness of universal adversarial examples. Fig. 5a shows that, when $1 \le l \le 15$, the DNN accuracy remains at above 98.6% under white-box UAP and UAN attacks. White-box UAP decreases the DNN accuracy from 99.2% to 50.0% when l increases from 20 to 25. White-box UAN decreases the DNN accuracy from 99.2% to 62.8% when l increases from 15 to 20. These results show that the universal adversarial examples constructed in the white-box manner can decrease the target DNN's accuracy by up to 49.5% when only about 50% of bus voltage measurements are tampered with. Thus, in Table 2, UAP and UAN require partial "write access".

Fig. 5b shows that UAP and UAN under the black-box setting achieve similar attack effectiveness as under the white-box setting. Therefore, UAP and UAN do not require the target DNN internal to be effective, as summarized in the "DNN internal" column of Table 2. UAP and UAN start to take effect when l is larger than 25 and 20, respectively. Therefore, UAN is more effective than UAP. Moreover, under the black-box setting, universal adversarial example attacks constructed by UAP and UAN are more effective than the input-specific attacks. This suggests that UAN can effectively learn the distribution of the adversarial examples while avoiding overfitting to the surrogate DNN under the black-box setting.

Note that both UAP and UAN need a clean training dataset, as shown in Table 2, no matter whether they operate under the white-box or black-box settings.

5.3 Implications and Credibility Analysis

The key observations from the attack effectiveness evaluation results for VSA obtained in Section 5.2 are summarized as follows:

- Compared with adversarial examples, random perturbations are ineffective in misleading the VSA DNN.
- Except for CW attacks, all other adversarial example attacks can decrease the target DNN's accuracy by about 50% when tempering with only 50% of the input dimensions.
- CW and DF can be very effective, in that they can decrease the target DNN's accuracy to below 20% and 10%, respectively. However, they impose strong requirements such as read and write access to the voltage traces of many/all buses, as well as the DNN internal.
- Preserving the confidentiality of the DNN internal is a weak defense, because four attacks remain effective under the black-box setting.
- Universal adversarial example attacks are effective against VSA DNN under both the white-box and black-box settings.

In what follows, we discuss the implications of these results in the context of smart grids.

5.3.1 Static knowledge needed by attacker. DNN internal and training data are the static knowledge. From the last two columns of Table 2, each of the six attack methods needs at least one of them to be able to construct effective adversarial examples. However, as DNN internal and training data are static information, the adversary can obtain them in the scenario of advanced persistent threat (APT). The adversary may use social engineering against employees of the grid operator. Note that even if the adversary can only obtain a black-box VSA DNN (e.g., its binary executable), they can feed massive unlabeled input samples to the black-box DNN to obtain the corresponding labels, forming a training dataset for building a surrogate DNN. Then, the adversary can use FGSM, PGD, UAP, or UAN to construct effective adversarial examples. In summary, preserving the confidentiality of the static knowledge (i.e., DNN internal and training data) is a shaky defense under the APT scenario. Therefore, the weights of the last two columns of Table 5.2 are marginal in assessing the credibility of adversarial example attacks against VSA.

5.3.2 Implication of write access requirement. To launch adversarial example attack, the ability of tempering with the voltage traces of all or some buses is a must. We discuss the implication of our results from two facets.

Compromising half of buses is a rule of thumb: Our evaluation results show that some input-specific attacks, i.e., DF and CW, can nearly subvert VSA when all buses are under attack. However, because input-specific attacks are less credible as analyzed shortly in Section 5.3.3, the observed subversion is also less credible accordingly. Therefore, as shown in Fig. 5, the degradation of VSA DNN accuracy to about 50% by the universal adversarial examples is a more credible maximal attack effectiveness. Section 5.2.3 shows that UAN is more effective than UAP. With UAN, there is a significant drop of DNN accuracy when l increases from 15 to 20. When l increases further from 20, the further accuracy drops become less salient. Since the cost of the attack increases with l (which is discussed in the next paragraph), compromising half of the buses to obtain their write accesses is a rule of thumb for the adversary.

Attack implementation: There are three possible ways to implement the attack. (1) An adversary within the enterprise network of the power grid control center can compromise the measurements of all buses. However, this strong adversary is ill-motivated, because they should subvert the VSA results directly. (2) An adversary compromises the communication links from the buses to the control center. To launch such attack, on one hand, the adversary must have the capability to intercept the network transmission of the clean voltage trajectory on the communication paths, e.g., on a router, in order to transmit the maliciously perturbed voltage trajectory to the control center without causing suspicion. On the other hand, the adversary needs the capability to breach the cryptographic protection. The adversary may have obtained the master keys of the compromised links, which represents a strong adversary as well. Exploiting zero-day vulnerability of the cryptographic protection (e.g., OpenSSL's Heartbleed bug) does not require the master key. However, the availability of such zero-day vulnerabilities is opportunistic and obtaining them is often costly. (3) An adversary manipulates the analog sensors by using remote electromagnetic inferences, which have been demonstrated feasible in [22]. However, such sensor reading manipulation attack is delicate and requires extensive skills. Through the above discussions, the attacks on the communication links and the analog sensors, though requiring significant investment and expertise, have certain credibility and cannot be complacently ignored.

5.3.3 Implication of read access. We separately discuss the implications of the input-specific and universal attacks in the context of VSA, which require full and no read access to the clean input.

Input-specific attacks: Since the adversary cannot construct the input-specific adversarial example until the whole voltage trajectory is obtained, the sensor reading manipulation by electromagnetic interference discussed in Section 5.3.2 is not applicable. Therefore, the adversary has to compromise the communication links from all buses to the control center, which represents a high overhead. The full read access requirement renders the input-specific attacks sophisticated, resource- and skill-demanding.

Universal attacks: Since the universal examples are independent of the real-time clean examples, they can be implemented by either the sensor reading manipulation by electromagnetic interference or compromising the communication links. Thus, the delicate interception required by the input-specific attacks is not a must. Note that the widely studied FDI attack against the power grid state estimation [27] is also a universal attack. Specifically, the perturbation to the power flow vector is given by $\mathbf{a} = \mathbf{Hc}$, where $\mathbf{H}$ is a constant matrix for state estimation and $\mathbf{c}$ is an arbitrary vector. Therefore, $\mathbf{H}$ is a static knowledge regarding the power grid that the adversary should obtain and the perturbation $\mathbf{a}$ is independent of the real-time power flow state of the power grid. Given the same nature of the universal adversary example attacks and the state estimation FDI attack studied in [27], they have the same

credibility that has substantially concerned the relevant research communities.

5.3.4 Summary. From the above analysis, the universal adversarial example attacks pose credible threats against VSA. Between UAP and UAN, the latter is more effective according to our evaluation. If the UAN adversary can compromise the voltage traces of more than half of the buses, devastating effects on VSA will be generated. While we should not expel the possibility of the input-specific adversarial example attacks, they are less credible and their results presented in this section help us understand the attack effectiveness more comprehensively.

6 EVALUATION OF DEFENSE EFFECTIVENESS

6.1 Defense Evaluation Settings

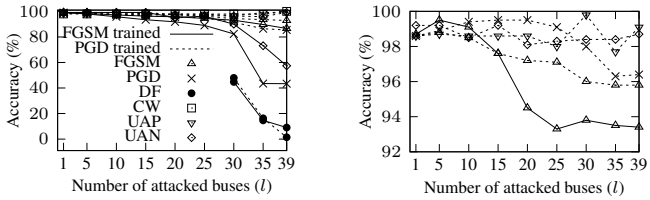
We employ adversarial training, APE-GAN, or the combination of them as the defense, which are explained as follows.

6.1.1 Setting of adversarial training. We consider two variants of adversarial training called *FGSM adversarial training* [13] and *PGD adversarial training* [28], which add 1,000 adversarial examples crafted by FGSM or PGD, respectively, into the training dataset. The adversarial training samples are crafted using validation samples with $\epsilon = 0.2$ p.u.. The DNN hardened by FGSM achieves 99.2% accuracy on clean test samples and 98.6% on FGSM adversarial examples constructed from clean test samples. The DNN hardened by PGD achieves 98.6% accuracy on clean test samples and 98.2% on PGD adversarial examples constructed from clean test samples. These results show that the hardening is effective against the considered attack method. To evaluate the effectiveness of the defense, we consider both white-box and black-box attacks. In the white-box setting, the adversary can access the hardened DNN's internals. Therefore, this white-box setting follows the Kerckhoffs's principle, in which the enemy knows the system including its defense mechanism. In the black-box setting, the adversary cannot access the internals of the hardened DNN and constructs the adversarial examples based on the surrogate DNN that is trained by the adversary using the clean training data from random initialization. Note that the adversary does not apply adversarial training to harden the surrogate DNN.

6.1.2 Setting of APE-GAN. APE-GAN is trained using the approach presented in [17] with all clean training samples and 1,000 FGSM adversarial examples constructed based on the VSA DNN. The trained APE-GAN is used before the VSA DNN to cleanse the input samples. To evaluate the effectiveness of the defense, we also consider white-box and black-box attacks. Under the white-box setting, the adversary constructs the adversarial examples using the target DNN. Note that this white-box attack construction does not consider APE-GAN, because how to construct adversarial perturbations that can bypass APE-GAN is still an open issue. Under the black-box setting (which is not considered in the paper proposing APE-GAN [17]), the adversary constructs the adversarial examples based on the surrogate DNN.

6.1.3 Combination of PGD adversarial training and APE-GAN. The first combination scheme is to combine a single PGD hardened DNN and the APE-GAN. During the training, we first apply PGD adversarial training to harden the DNN and then train the APE-GAN using clean training samples and 1,000 FGSM adversarial examples constructed based on the hardened DNN. During the inference, the input is first cleansed by the APE-GAN and then fed to the hardened DNN. To evaluate this defense, we consider both white-box and black-box attacks. The white-box adversarial examples are crafted using the target PGD-hardened DNN; the black-box adversarial examples are crafted using the surrogate DNN that is not hardened by adversarial training.

The second scheme is to combine multiple PGD-hardened DNNs and APE-GAN. The decision fusion among the multiple DNNs is as follows. Given an input, if the percentage of the majority of the outputs of N PGD-hardened DNNs is greater than a threshold T_s, the input is considered clean and the majority is yielded as the final result; otherwise, the input is considered adversarial. For the input classified adversarial, we use APE-GAN to cleanse the input, feed the cleansed input to the N PGD-hardened DNNs, and use the majority of the DNNs' outputs as the final classification result. We set $N = 5$ and $T_s = 100\%$ in the evaluation. These settings will be explained shortly in Section 6.2.3. This second scheme is only evaluated with black-box attacks generated by the surrogate DNN as in the first scheme. Note that how to construct adversarial examples under the white-box setting against the structure that fuses multiple DNNs' outputs is still an open issue.



(a) White-box attacks **(b) Black-box attacks**

Figure 6: VSA accuracy in the presence of adversarial training defense. The legends of (b) are the same as (a).

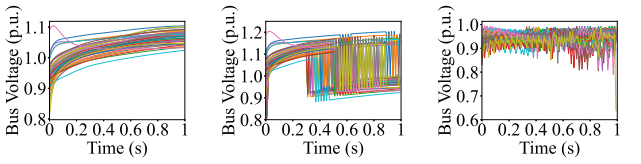
6.2 Defense Effectiveness Results

This section presents the defense effectiveness results, which are summarized in Table 3. We say an attack is effective if it can decrease the accuracy of the target DNN to 80% and below; we say a defense is effective if it can restore the accuracy of the target DNN to 80% and above in the presence of an effective attack. From Section 5.2, the black-box DF and CW attacks are not effective. Thus, we do not consider these two attacks in this section.

6.2.1 Effectiveness of adversarial training. Fig. 6a and Fig. 6b show the VSA DNN's accuracy in the presence of adversarial training defense versus l under the white-box and black-box attack settings. First, we analyze the defense effectiveness against input-specific attacks. In Fig. 6a, FGSM adversarial training is not effective against white-box PGD attack when $l = 35$ and $l = 39$. But PGD adversarial training is effective against white-box PGD attack. This suggests that PGD adversarial training is more effective than FGSM adversarial training. Both FGSM and PGD adversarial training defenses are not effective against white-box DF attack when $30 \leq l \leq 39$.

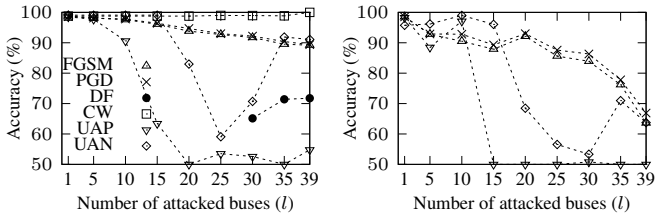
Table 3: Summary of defense effectiveness. "✓" and "✗" represent effective and ineffective defenses. "White" and "Black" refer to "White-box" and "Black-box" attacks. "N.A." for the black-box DF and CW attacks means these attacks are not effective and thus not used to evaluate defense effective. "\" means the results are not available because the attack construction approach is still an open issue.

Defense \ Attack	Input-specific attacks								Universal attacks			
	FGSM		PGD		DF		CW		UAP		UAN	
	White	Black	White	Black	White	Black	White	Black	White	Black	White	Black
❶ FGSM adv training	✓	✓	✗	✓	✗	N.A.	✓	N.A.	✓	✓	✗	✓
❷ PGD adv training	✓	✓	✓	✓	✗	N.A.	✓	N.A.	✓	✓	✓	✓
❸ APE-GAN	✓	✗	✓	✗	✗	N.A.	✓	N.A.	✗	✗	✗	✗
❹ APE-GAN+PGD adv training	✓	✓	✓	✓	✗	N.A.	✓	N.A.	✗	✗	✓	✓
❺ APE-GAN+N PGD adv training	\	✓	\	✓	\	N.A.	\	N.A.	\	✓	\	✓



(a) Clean (stable) **(b) Adversarial** **(c) APE-GAN cleansed**

Figure 7: Clean, FGSM-perturbed, and APE-GAN cleansed bus voltage trajectories. The clean sample in (a) is classified as stable; the FGSM-perturbed sample in (b) is wrongly classified as unstable; the APE-GAN cleansed sample in (c) is correctly classified as stable.



(a) White-box attacks **(b) Black-box attacks**

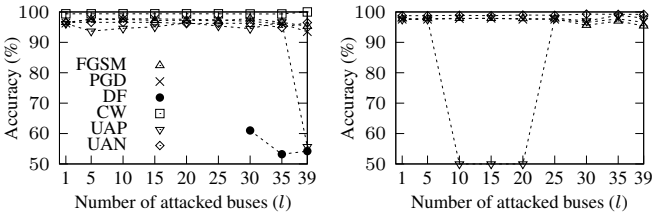
Figure 8: VSA accuracy when APE-GAN defends attacks.

From Fig. 6b, adversarial training is effective against all effective black-box input-specific attacks. Then, we analyze the defense effectiveness against universal attacks. As shown in Fig. 6, PGD adversarial training is effective against both the white-box and black-box universal attacks. The results observed from Fig. 6b are summarized in the two rows of Table 3 headed by ❶ and ❷.

6.2.2 Effectiveness of APE-GAN. Fig. 7 shows a clean, stable bus voltage trajectory, its FGSM-perturbed counterpart, and the output of APE-GAN when the input is the aforementioned FGSM-perturbed trajectory. Fig. 8a and Fig. 8b show the VSA DNN's accuracy after the input is cleansed by APE-GAN under the setting of white-box and black-box attack. First, we analyze the defense effectiveness against input-specific attacks. Under the white-box setting, APE-GAN is effective against all input-specific attacks except DF. Under the black-box setting, APE-GAN is ineffective against FGSM and PGD attacks. Thus, APE-GAN is more effective against white-box attacks. This is because APE-GAN is designed to eliminate the adversarial perturbations crafted by white-box adversary based on the target DNN [17]. Then, we analyze APE-GAN's defense effectiveness against universal attacks. From Fig. 8a, when l is 15 and 20, 36.1% and 49.5% of the white-box UAP adversarial examples bypass APE-GAN. When l is 25, 40.4% of the white-box UAN adversarial examples bypass APE-GAN. As shown in Fig. 8b, APE-GAN performs worse against black-box universal attacks. In summary, APE-GAN is ineffective against universal attacks. The results observed from Fig. 8 are summarized in the row of Table 3 headed by ❸.

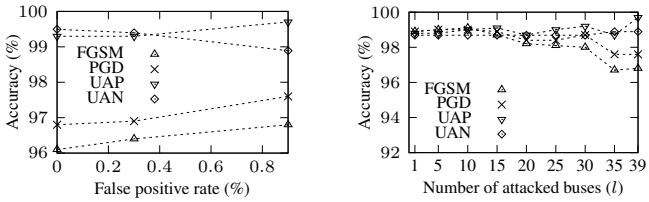
6.2.3 Effectiveness of adversarial training combined with APE-GAN. From the results in Section 6.2.1, PGD adversarial training is more effective than FGSM adversarial training. Therefore, we attempt to combine PGD adversarial training with APE-GAN for better defense effectiveness.

Fig. 9 shows the VSA DNN's accuracy when a single hardened DNN is combined with APE-GAN (i.e., the first scheme discussed in Section 6.1.3). As both PGD adversarial training and APE-GAN are ineffective against the white-box DF, the combination of them is also ineffective against white-box DF. In addition, the combination has deteriorated defense effectiveness against UAP attacks under certain settings (i.e., when $l = 39$ under the white-box attack setting and $10 \leq l \leq 20$ under the black-box attack setting), compared with the sole PGD adversarial training. This suggests that the preprocessing performed by APE-GAN may reduce the effectiveness of the DNN hardened by adversarial training in counteracting certain adversarial examples. Non-monotonicity of accuracy versus l can be observed in Fig. 8a, Fig. 8b, and Fig. 9b. This is because the output of APE-GAN is unpredictable and in some cases may disturb the input samples and decrease the accuracy. Meanwhile, since we only consider one random combination of l buses to be compromised from all 39 buses, the randomness may also contribute to the non-monotonicity. We do not consider all combinations of the l compromised buses because otherwise the number of experiments to generate one point in the figures will be huge. For example, to choose 10 buses from 39 ones, there are $\binom{39}{10} = 635,745,396$ possible combinations. The results observed from Fig. 9 are summarized in the row of Table 3 headed by ❹.



(a) White-box attacks **(b) Black-box attacks**

Figure 9: Defense effectiveness of combining PGD adversarial training and APE-GAN against various attacks.



(a) Accuracy vs. FPR (l = 39) **(b) Accuracy vs. l (T_s = 100%)**

Figure 10: Defense effectiveness of combining 5 PGD-hardened DNNs and APE-GAN against black-box attacks.

Next, we evaluate the second combination scheme discussed in Section 6.1.3. From our extensive evaluation, the settings for N beyond 5 do not bring substantial benefit in terms of the VSA accuracy in the absence and presence of attacks. The detailed results are omitted here due to space constraints. Thus, we set $N = 5$. Under the majority fusion rule, with a higher T_s setting, the decision fusion among the N PGD-hardened DNNs will have higher false positive rate (FPR), i.e., more clean inputs will be classified adversarial. As a result, more clean inputs will be processed by APE-GAN and the second round of DNN executions and decision fusion will be triggered, which represents higher computation overhead. Thus, we evaluate the overall VSA accuracy achieved by the second combination scheme versus FPR, which is shown in Fig. 10a. The points on a curve are obtained by varying T_s within 60%, 80%, and 100%. For all attacks except UAN, accuracy increases with FPR, presenting a trade-off between defense effectiveness and computation overhead. Although the accuracy decreases with FPR when the system is subject to UAN attack, when T_s = 100%, the accuracy is 98.9%, higher than that achieved by a single PGD-hardened DNN (i.e., 98.7%). Thus, we set T_s = 100% to achieve the best defense effectiveness. Under the setting of T_s = 100%, Fig. 10b shows the VSA accuracy versus l under various black-box attacks. The VSA accuracy is always above 96.7%. In particular, when l = 39, the accuracy ranges from 96.8% to 99.7%, higher than the counterparts in Fig. 6b and Fig. 8b. This suggests that fusing multiple hardened DNNs can improve defense effectiveness. Observations from Fig. 10 are summarized in the row of Table 3 headed by ❺.

6.3 Implication of Results

From Table 3, PGD adversarial training is effective against all attacks except for the white-box DF attack. From our analysis in Section 5.3, the input-specific attacks are less credible compared with the universal attacks. Therefore, the shortfall of PGD adversarial training in addressing white-box DF attack is mitigated. As FGSM adversarial training is ineffective against white-box UAN which is a universal attack, it is less desirable than PGD adversarial training. The scheme that combines N PGD-hardened DNNs and APE-GAN is a potential competitor of PGD adversarial training. However, under the white-box setting, how to construct meaningful adversarial examples against the decision fusion-based structure using the principle of the six attack methods is still an open issue. Although begging sophisticated attacks is exorbitant, the uncertainty lying in the effectiveness of the combination scheme against white-box attacks cannot be ignored. In addition, the combination scheme incurs higher run-time overhead (at least N times) compared with PGD adversarial training. Based on the above considerations, PGD adversarial training is still preferred.

After jointly considering the attack credibility analysis presented in Section 5.3, the PGD adversarial training should be applied to protect DNN-based VSA against the universal adversarial examples that generate non-negligible concerns. From the results in Fig. 6, when PGD adversarial training is applied, the universal attacks (UAP or UAN) can cause at most 3.6% accuracy drop under any setting for l, compared with the case without attack.

7 CONCLUSION AND FUTURE WORK

This paper analyzed the requirement and credibility of six adversarial example attacks on the voltage stability assessment. We showed that effective adversarial example attacks need to compromise the voltage traces of at least half of the transmission system buses. The universal adversarial examples pose similar credibility as the widely studied false data injection on power grid state estimation. In addition, we found that the model hardening using an adversarial training approach can effectively counteract the universal adversarial examples. The credibility analysis methodology adopted in this paper can also be applied to other types of adversarial example attacks and power grid applications.

For future work, beyond the open issues pointed in this paper, it is also interesting to follow the approach of ensemble learning to generate multiple deep models hardened by adversarial training using different types of adversarial examples and then fuse their results to output the final result. In addition, how to construct adversarial examples under the white-box setting against the ensemble and the corresponding attack effectiveness results will improve our understanding on the cybersecurity of the deep learning-based voltage stability assessment.

ACKNOWLEDGMENTS

The authors wish to thank the anonymous reviewers and shepherd Dr. Nilanjan Banerjee for providing valuable feedback on this work. This research is supported by the National Research Foundation, Singapore and National University of Singapore through its National Satellite of Excellence in Trustworthy Software Systems (NSOE-TSS) office under the Trustworthy Computing for Secure Smart Nation Grant (TCSSNG) award no. NSOE-TSS2020-01.

REFERENCES

[1] [n.d.]. Final Report on the August 14, 2003 Blackout in the United States and Canada: Causes and Recommendations. https://www.energy.gov/sites/prod/files/oeprod/DocumentsandMedia/BlackoutFinal-Web.pdf. Accessed: 2020-12-11.

[2] [n.d.]. Tencent Keen Security Lab: Experimental Security Research of Tesla Autopilot. https://keenlab.tencent.com/en/2019/03/29/Tencent-Keen-Security-Lab-Experimental-Security-Research-of-Tesla-Autopilot/.

[3] Naveed Akhtar and Ajmal Mian. 2018. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access* 6 (2018), 14410–14430.

[4] T Athay, R Podmore, and S Virmani. 1979. A practical method for the direct analysis of transient stability. *IEEE Transactions on Power Apparatus and Systems* 2 (1979), 573–584.

[5] Suzhi Bi and Ying Jun Zhang. 2014. Graphical methods for defense against false-data injection attacks on power system state estimation. *Transactions on Smart Grid* 5, 3 (2014), 1216–1227.

[6] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *Proceedings of the IEEE Symposium on Security and Privacy*.

[7] Yize Chen, Yushi Tan, and Baosen Zhang. 2019. Exploiting vulnerabilities of load forecasting through adversarial attacks. In *Proceedings of the ACM International Conference on Future Energy Systems*.

[8] Sambarta Dasgupta, Magesh Paramasivam, Umesh Vaidya, and Venkataramana Ajjarapu. 2013. Real-time monitoring of short-term voltage stability using PMU data. *Transactions on Power Systems* 28, 4 (2013), 3702–3711.

[9] Zhao Yang Dong, Yan Xu, Pei Zhang, and Kit Po Wong. 2013. Using IS to assess an electric power system's real-time stability. *IEEE Intelligent Systems* 28, 4 (2013), 60–66.

[10] Mohammad Esmalifalak, Lanchao Liu, Nam Nguyen, Rong Zheng, and Zhu Han. 2014. Detecting stealthy false data injection using machine learning in smart grid. *IEEE Systems Journal* 11, 3 (2014), 1644–1652.

[11] Mohammad Esmalifalak, Huy Nguyen, Rong Zheng, and Zhu Han. 2011. Stealth false data injection using independent component analysis in smart grid. In *Proceedings of the IEEE International Conference on Smart Grid Communications*.

[12] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks. In *Proceedings of the International Conference on Neural Information Processing Systems*.

[13] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *Proceedings of the International Conference on Learning Representations*.

[14] Jamie Hayes and George Danezis. 2018. Learning universal adversarial perturbations with generative models. In *Proceedings of the IEEE Symposium on Security and Privacy Workshop*.

[15] Qiuhua Huang, Renke Huang, Weituo Hao, Jie Tan, Rui Fan, and Zhenyu Huang. 2019. Adaptive power system emergency control using deep reinforcement learning. *Transactions on Smart Grid* 11, 2 (2019), 1171–1182.

[16] JQ James, David J Hill, Albert YS Lam, Jiatao Gu, and Victor OK Li. 2017. Intelligent time-adaptive transient stability assessment system. *Transactions on Power Systems* 33, 1 (2017), 1049–1058.

[17] Guoqing Jin, Shiwei Shen, Dongming Zhang, Feng Dai, and Yongdong Zhang. 2019. Ape-gan: Adversarial perturbation elimination with gan. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.

[18] Peizhong Ju and Xiaojun Lin. 2018. Adversarial attacks to distributed voltage control in power distribution networks with DERs. In *Proceedings of the International Conference on Future Energy Systems*.

[19] K Kawabe and K Tanaka. 2014. Analytical method for short-term voltage stability using the stability boundary in the PV plane. *Transactions on Power Systems* 29, 6 (2014), 3041–3047.

[20] Prabha Kundur, Neal J Balu, and Mark G Lauby. 1994. *Power system stability and control*. Vol. 7. McGraw-hill New York.

[21] Prabha Kundur, John Paserba, Venkat Ajjarapu, Göran Andersson, Anjan Bose, Claudio Canizares, Nikos Hatziargyriou, David Hill, Alex Stankovic, Carson Taylor, et al. 2004. Definition and classification of power system stability IEEE/CIGRE joint task force on stability terms and definitions. *Transactions on Power Systems* 19, 3 (2004), 1387–1401.

[22] Denis Foo Kune, John Backes, Shane S Clark, Daniel Kramer, Matthew Reynolds, Kevin Fu, Yongdae Kim, and Wenyuan Xu. 2013. Ghost talk: Mitigating EMI signal injection attacks against analog sensors. In *Proceedings of the IEEE Symposium on Security and Privacy*.

[23] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2017. Adversarial examples in the physical world. In *Proceedings of the International Conference on Learning Representations Workshop*.

[24] Subhash Lakshminarayana, Teo Zhan Teng, David KY Yau, and Rui Tan. 2017. Optimal attack against cyber-physical control systems with reactive attack mitigation. In *Proceedings of the International Conference on Future Energy Systems*.

[25] Jie Lin, Wei Yu, Xinyu Yang, Guobin Xu, and Wei Zhao. 2012. On false data injection attacks against distributed energy routing in smart grid. In *Proceedings of the International Conference on Cyber-Physical Systems*. IEEE.

[26] WY Liu, BP Tang, JG Han, XN Lu, NN Hu, and ZZ He. 2015. The structure healthy condition monitoring and fault diagnosis methods in wind turbines: A review. *Renewable and Sustainable Energy Reviews* 44 (2015), 466–472.

[27] Yao Liu, Peng Ning, and Michael K Reiter. 2009. False data injection attacks against state estimation in electric power grids. In *Proceedings of the ACM Conference on Computer and Communications Security*.

[28] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the International Conference on Learning Representations*.

[29] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2017. Universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[30] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[31] Michael Negnevitsky, Paras Mandal, and Anurag K Srivastava. 2009. Machine learning applications for load, price and wind power prediction in power systems. In *Proceedings of the International Conference on Intelligent System Applications to Power Systems*. IEEE.

[32] Mohammad Ashiqur Rahman, Ehab Al-Shaer, and Rajesh G Kavasseri. 2014. A formal model for verifying the impact of stealthy attacks on optimal power flow in power grids. In *Proceedings of the International Conference on Cyber-Physical Systems*. IEEE.

[33] Kui Ren, Tianhang Zheng, Zhan Qin, and Xue Liu. 2020. Adversarial attacks and defenses in deep learning. *Engineering* (2020).

[34] Seunghyoung Ryu, Jaekoo Noh, and Hongseok Kim. 2017. Deep neural network based demand side short term load forecasting. *Energies* 10, 1 (2017), 3.

[35] Rui Tan, Hoang Hai Nguyen, Eddy YS Foo, Xinshu Dong, David KY Yau, Zbigniew Kalbarczyk, Ravishankar K Iyer, and Hoay Beng Gooi. 2016. Optimal false data injection attack against automatic generation control in power grids. In *Proceedings of the International Conference on Cyber-Physical Systems*. IEEE.

[36] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. 2019. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[37] Le Xie, Yilin Mo, and Bruno Sinopoli. 2011. Integrity data attacks in power market operations. *Transactions on Smart Grid* 2, 4 (2011), 659–666.

[38] Yan Xu, Zhao Yang Dong, Jun Hua Zhao, Pei Zhang, and Kit Po Wong. 2012. A reliable intelligent system for real-time dynamic security assessment of power systems. *Transactions on Power Systems* 27, 3 (2012), 1253–1263.

[39] Yinliang Xu, Wei Zhang, Wenxin Liu, and Frank Ferrese. 2012. Multiagent-based reinforcement learning for optimal reactive power dispatch. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, 6 (2012), 1742–1751.

[40] Tao Yu, Bin Zhou, Ka Wing Chan, Liang Chen, and Bo Yang. 2011. Stochastic Optimal Relaxed Automatic Generation Control in Non-Markov Environment Based on Multi-Step $Q(\lambda)$ Learning. *Transactions on Power Systems* 26, 3 (2011), 1272–1282.

[41] Jianwu Zeng and Wei Qiao. 2013. Short-term solar power prediction using a support vector machine. *Renewable Energy* 52 (2013), 118–127.

[42] Yuchen Zhang, Yan Xu, Zhao Yang Dong, and Rui Zhang. 2018. A hierarchical self-adaptive data-analytics method for real-time power system short-term voltage stability assessment. *Transactions on Industrial Informatics* 15, 1 (2018), 74–84.

A ADVERSARIAL EXAMPLE AND DEFENSE FORMULATIONS

Consider a classifier $f(\cdot;\boldsymbol{\theta})$ with weights $\boldsymbol{\theta}$ that classifies an intact input $\mathbf{x}$ as y, i.e., $f(\mathbf{x};\boldsymbol{\theta}) = y$. An adversarial example $\mathbf{x}' = \mathbf{x} + \boldsymbol{\delta}$, where $\boldsymbol{\delta}$ is a perturbation, is classified as $y' \neq y$ by the classifier. To reduce the chance of being detected, the adversary aims at minimizing the distance between $\mathbf{x}$ and $\mathbf{x}'$, which is denoted by $D(\mathbf{x}, \mathbf{x}')$. Thus, the attack construction can be formulated as a constrained optimization problem: $\boldsymbol{\delta}^* = \operatorname{argmin}_{\boldsymbol{\delta}} D(\mathbf{x}, \mathbf{x}')$ subject to $f(\mathbf{x}';\boldsymbol{\theta}) \neq y$. As this problem is difficult, existing studies propose various heuristic solutions and the crafted perturbations may not always meet the constraint $f(\mathbf{x}';\boldsymbol{\theta}) \neq y$. The effectiveness of an adversarial example construction method is often characterized by the empirical rate of yielding $f(\mathbf{x}';\boldsymbol{\theta}) \neq y$. If the input $\mathbf{x}$ is a time series, an additional constraint regarding the autocorrelation of $\mathbf{x}'$ can be integrated into the above formulation, such that the attack

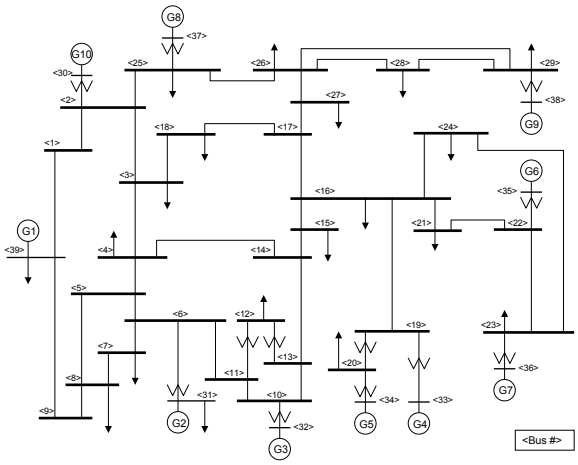


Figure 11: New England 10-machine 39-bus power system.

traces will follow certain patterns and be smooth. In this paper, we do not enforce the autocorrelation constraint.

The FGSM [13] is a representative one-step attack construction method. The FGSM adversarial example is calculated by: $\mathbf{x}' = \mathbf{x} + \epsilon \cdot sign(\nabla J(\boldsymbol{\theta}, \mathbf{x}, y))$, where $sign(\cdot)$ is the sign function and $J(\cdot)$ is the loss function of the model. The PGD [23] attack performs small-sized FGSM for multiple iterations. At iteration i, the PGD adversarial example is updated by: $\mathbf{x}_i = \Pi_\epsilon(\mathbf{x}_{i-1} + \alpha sign(\nabla J(\boldsymbol{\theta}, \mathbf{x}, y)))$, where $\Pi_\epsilon(\cdot)$ projects $\mathbf{x}_i$ in the $\epsilon - L_p (p = 1, 2, \infty)$ neighbor of $\mathbf{x}$ and α determines the step size. The DF [30] aims to find the minimum perturbation that takes the clean input to boundary of the region deciding the classifier's prediction of the input. For non-linear general classifier, at each iteration i, the DF adversarial example is accumulated by a small vector which is calculated by $\boldsymbol{\delta}_i^* = \text{argmin}_{\boldsymbol{\delta}_i} \|\boldsymbol{\delta}_i\|_2$ subject to $\hat{k}(\mathbf{x}_i) + \nabla\hat{k}(\mathbf{x}_i)^T \boldsymbol{\delta}_i = 0$, where $\hat{k}(\cdot)$ is the estimated classifier derived by linearizing the decision boundaries of the original classifier $f(\cdot; \boldsymbol{\theta})$. The CW [6] formulates the attack construction as: $\boldsymbol{\delta}^* = \text{argmin}_{\boldsymbol{\delta}} D(\mathbf{x}, \mathbf{x}')$ subject to $L(\mathbf{x}') \leq 0$, where $f(\mathbf{x}'; \boldsymbol{\theta}) \neq y$ if and only if $L(\mathbf{x}') \leq 0$. Then, the Lagrangian relaxation is applied to simplify the problem as: $\boldsymbol{\delta}^* = \text{argmin}_{\boldsymbol{\delta}} D(\mathbf{x}, \mathbf{x}') + c \cdot L(\mathbf{x}')$, where c is a constant weight for combining the two minimization objectives. The specific form of the function $L(\mathbf{x}')$ is $L(\mathbf{x}') = \mathbf{max}\{Z(\mathbf{x}')_y - \mathbf{max}_{y_i \neq y}\{Z(\mathbf{x}')_{y_i}\}, -\kappa\}$, where $Z(\cdot)$ represents the logits output of the classifier $f(\cdot; \boldsymbol{\theta})$ and κ controls the strength of the adversarial example. The UAP attack [29] is generated by finding the adversarial perturbation for each of the data samples from a training set using the DF algorithm and accumulating these perturbations to form a universal adversarial perturbation. The UAN attack [14] learns a generative model $G(\cdot; \phi)$ with parameters ϕ that can take as input a random vector z sampled from normal distribution and output a universal adversarial perturbation. The loss function for training $G(\cdot; \phi)$ is $\mathbf{max}\{Z(\mathbf{x}')_y - \mathbf{max}_{y_i \neq y}\{Z(\mathbf{x}')_{y_i}\}, -\kappa\} + c \cdot D(\mathbf{x}, \mathbf{x}')$.

The adversarial training [28] follows the idea of robust optimization and formulates a min-max problem to find the robust model parameters $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \max_{D(\mathbf{x}, \mathbf{x}') \leq \epsilon} J(\boldsymbol{\theta}, \mathbf{x}', y)$. The formulation is viewed as the composition of an inner maximization problem aiming to find the effective adversarial examples and an outer minimization problem minimizing the adversarial loss given by the inner attack problem. The loss function for adversarial training can be $J_{adv}(\boldsymbol{\theta}, \mathbf{x}, y) = \alpha J(\boldsymbol{\theta}, \mathbf{x}, y) + (1 - \alpha) J(\boldsymbol{\theta}, \mathbf{x}', y)$, where α balances the loss on benign and adversarial examples and $\mathbf{x}'$ can be computed by FGSM or PGD attack. For APE-GAN [17], the loss function for training the discriminator is $-\log D(\mathbf{x}; \boldsymbol{\theta}_D) + \log D(G(\mathbf{x}'; \boldsymbol{\theta}_G); \boldsymbol{\theta}_D)$. The loss function for the generator contains two parts. The first part is a pixel-wise mean square error loss $\frac{1}{W}\frac{1}{H}\sum_{i=1}^{W}\sum_{i=1}^{H}(\mathbf{x}_{i,j} - G(\mathbf{x}'_{i,j}; \boldsymbol{\theta}_G))^2$, where W and H represent the width and height of the input. The second part is the adversarial loss function $1 - \log D(G(\mathbf{x}'; \boldsymbol{\theta}_G); \boldsymbol{\theta}_D)$. The discriminator and generator are trained together under the GAN setting.

B POWER SYSTEM SINGLE-LINE DIAGRAM

The single-line diagram of the power system considered in the VSA case study is shown in Fig. 11.