On Credibility of Adversarial Examples against Learning-Based Grid Voltage Stability Assessment

Qun Song, Rui Tan, Chao Ren, Yan Xu, Yang Lou, Jianping Wang, and Hoay Beng Gooi

Abstract—Voltage stability assessment is essential for maintaining reliable power grid operations. Stability assessment approaches using deep learning address the shortfalls of the traditional time-domain simulation-based approaches caused by increased system complexity. However, deep learning models are shown to be vulnerable to adversarial examples in the field of computer vision. While this vulnerability has been noticed by the power grid cybersecurity research, the domain-specific analysis on the requirements imposed upon effective attack implementation is still lacking. Although these attack requirements are usually reasonable in computer vision tasks, they can be stringent in the context of power grids. In this paper, we conduct a systematic investigation on the attack requirements and credibility of six representative adversarial example attacks based on a voltage stability assessment application for the New England 10-machine 39-bus power system. We show that (1) compromising about half the transmission system buses' voltage traces is a rule-of-thumb attack requirement; (2) the universal adversarial perturbations regardless of the original clean voltage trajectory possess the same credibility as the widely studied false data injection attacks on power grid state estimation, while the input-specific adversarial perturbations are less credible; (3) the prevailing strong adversarial training thwarts the universal perturbations but fails in defending certain input-specific perturbations. To advance defense to cope with both universal and input-specific adversarial examples, we propose a new approach that simultaneously estimates the predictive uncertainty of any given input of voltage trajectory and thwarts the attacks effectively.

Index Terms—Adversarial example, cybersecurity, neural networks, smart grid, voltage stability assessment

NOMENCLATURE

Acronyms

AMI	Advanced Metering Infrastructure
APT	Advanced persistent threat
CNN	Convolutional neural network
CPS	Cyber-physical system
CV	Computer vision
CW	Carlini and Wagner's method
DF	DeepFool
DNN	Deep neural network
DR	Detection rate
DSR	Defense success rate
FDI	False data injection
FGSM	Fast Gradient Sign Method
FLOPs	Floating-point operations
FPR	False positive rate
GPU	Graphics processing unit
ICTs	Information and communication technologies
MC	Monte Carlo
OPF	Optimal Power Flow
p.u.	Per unit
PGD	Projected Gradient Descent
ReLUs	Rectified linear units
ROC	Receiver operating characteristic
TPR	True positive rate
TR	Thwarting rate
UAN	Universal Adversarial Network
UAP	Universal Adversarial Perturbation
VSA	Voltage stability assessment

Symbols

ℓ_2	Euclidean norm
ϵ	Maximum perturbation intensity
γ	Uncertainty threshold
κ	CW's hyperparameter for perturbation intensity
$\boldsymbol{\delta}$	Adversarial perturbation
$\boldsymbol{\theta}$	Machine learning model weights
$\mathbf{a}$	FDI perturbation vector
$\mathbf{c}$	An arbitrary vector
$\mathbf{H}$	A constant matrix for state estimation
$\mathbf{M}$	Matrix to restrict the perturbed area
$\mathbf{x}'$	Adversarial example
$\mathbf{X}$	Training data generated by offline simulation
$\mathbf{x}$	Input voltage trajectory
$\mathbf{Y}$	Training data labels
a	An attack method
D	Distance metric
d	A defense method
dp	Dropout rate
f	Machine learning model
l	Number of attacked buses
N	Ensemble size
r	An attack requirement

- *Q. Song, R. Tan, C. Ren, Y. Xu, and H. B. Gooi are with the Nanyang Technological University (NTU), Singapore. Q. Song and C. Ren are with the Interdisciplinary Graduate School of NTU. R. Tan is with the School of Computer Science and Engineering of NTU. Y. Xu and H. B. Gooi are with the School of Electrical and Electronic Engineering of NTU. Y. Lou and J. Wang are with the Department of Computer Science, City University of Hong Kong.*

S_A	Set of attack methods
S_D	Set of defense methods
S_R	Union set of attack requirements
S_{R_i}	Individual set of attack requirements
y	Stability classification label

Subscripts

MLE	Maximum likelihood estimation
m	Number of training data samples
n	Number of attack requirements
p	Number of attack methods
q	Number of defense methods

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 efficient and sustained operations [1]. 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 [2]. Machine learning, as an ICT, has been considered and adopted for enhancing various grid capabilities such as load forecasting [3], fault diagnosis [4], solar power prediction [5], power grid controls [6].

The *deep neural networks* (DNNs), enabled by the advancements of hardware-based computing acceleration, have attracted growing interests in power grid applications [3]–[6] due to their appealing capabilities in extracting sophisticated patterns from big data. However, the complex structures of DNNs engender vulnerabilities under adversarial settings. In this paper, we focus on the threat of *adversarial example* [7], which adds minute crafted perturbations to clean inputs and misleads the DNN to yield wrong inference outputs. Adversarial example can be viewed as a specific type of the *false data injection* (FDI) that has been studied widely under the context of power grid [8].

DNNs can be applied in various power grid operation tasks. This paper considers a representative task of online *voltage stability assessment* (VSA). Maintaining stability is fundamental for any power system because losing stability may cause catastrophic blackouts that threaten people's properties and lives. During the power grid design phase, offline VSA conducts time-domain simulations to check the voltage stability of the grid when presumed disturbances are injected. Although a high-fidelity system model can yield accurate offline VSA outcomes, the simulations are often much slower than the evolution of the physical processes due to the power system complexities. Thus, time-domain simulations are ill-suited for online VSA. To develop online VSA for 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 with real-time voltage measurements [9], [10]. Applying DNNs to better capture the inherent complexity of voltage dynamics and advance online VSA is an ongoing interest of power grid operations [9], [10].

Wrong outputs of VSA can lead to catastrophic consequences. A false negative in detecting instability can cause missed or delayed activation of fault isolation, which may result in widespread blackout; a false positive may lead to unnecessary load shedding and thereby brings misery to the customers losing power. Thus, the cybersecurity risks faced by DNN-based VSA due to adversarial examples need to be understood. Various algorithms of constructing adversarial examples have been proposed [11]. The effects of adversarial examples have been demonstrated in the safety-critical CPSs that use computer vision (CV) for perception. For example, adversarial stickers pasted on road can mislead learning-based lane detection system of Tesla Autopilot [12]. However, the requirements for implementing these attacks, though reasonable in the CV tasks, can be too stringent in the context of VSA. For instance, the clean input is often needed to compute the malicious perturbation. In the CV-based lane recognition, the camera's view of the road as the clean input can be known *a priori* to the attacker and used to craft the adversarial sticker. However, in VSA, the requirement of obtaining real-time read access to all the transmission buses' voltages for constructing attacks can be very high. Coordinating the real-time eavesdropping and data tampering for implementing certain attacks imposes high requirements on the attacker's resources and skills.

Therefore, indiscriminately transferring the worry from CV to DNN-based smart grid applications may hinder innovations. This paper is motivated by the domain differences between CV and power grid applications in studying the vulnerability of adversarial examples. Existing studies mainly focus on investigating the threat of adversarial example attacks for the safety-critical cyber-physical systems that use CV for perception. However, the requirements and credibility of implementing adversarial attacks in the context of safety-critical power grid applications are not well understood. Thus, we conduct a domain-specific analysis to understand the credibility of adversarial example attacks against DNN-based power system applications. Specifically, we focus on the VSA application. To the best of our knowledge, systematic analysis on the credibility of adversarial example attacks with due discrimination on the requirements of implementing them in smart grids is still lacking. In this paper, we conduct a systematic study to evaluate the effectiveness of various adversarial example construction methods 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. By relating the attack effectiveness with the attack requirement and 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. We also evaluate the attack effectiveness when the system defender adopts the prevailing countermeasures of *model hardening* and *input cleansing*. From our evaluation,

although a class of model hardening techniques effectively defend the more credible universal attacks, they are not effective in counteracting certain input-specific attacks. While the input-specific attacks are viewed less credible from our analysis, their possibility cannot be completely ignored.

From the study, we summarize a methodology for evaluating the credibility of various types of adversarial example attacks on the DNN-based smart grid applications. The methodology includes: (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 the existing and/or new countermeasures in protecting the smart grid application against the adversarial example attacks.

The main contributions of this paper are as follows:

- We study six types of adversarial example, i.e., Fast Gradient Sign Method (FGSM) [7], Projected Gradient Descent (PGD) [13], DeepFool (DF) [14], Carlini and Wagner's method (CW) [15], Universal Adversarial Perturbation (UAP) [16], and Universal Adversarial Network (UAN) [17]. We investigate the minimal requirement of implementing each of them to achieve effective attack on VSA.
- We show tampering with the voltages of half buses is a rule of thumb for crafting effective adversarial examples. The *universal adversarial examples* (i.e., UAP and UAN) that do not require read access to bus voltages are as credible as the FDI on grid state estimation [8] that has been widely studied. The *input-specific adversarial examples* are less credible due to their indispensable requirement on real-time bus voltage read access.
- We study the effectiveness of the prevailing defenses of model hardening by adversarial training and input cleansing via APE-GAN [18] under each of the six adversarial example attacks. We show that the PGD adversarial training effectively protects the DNN-based VSA against the credible universal adversarial examples but fails to counteract certain input-specific attacks. To advance defense, we propose a new approach that simultaneously estimates the predictive uncertainty of any given input of voltage trajectory and thwarts both input-specific and universal adversarial example attacks effectively.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 contains preliminaries and background. Section 4 states the problem. Section 5 and Section 6 evaluate the attack requirement and existing defense effectiveness. Section 7 presents our proposed defense approach of joint attack detection and thwarting. Section 8 concludes this paper.

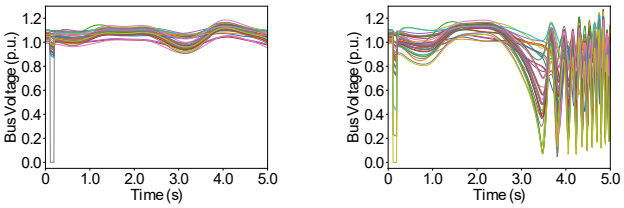
2 Related work

Machine learning applied in power systems. Machine learning-based approaches have been proposed in literature for load forecasting [3], solar power prediction [5], fault diagnosis [4], and attack detection [19]. Deep reinforcement learning is studied for various power grid controls including voltage, frequency, power, and emergency handling [6]. Applying machine learning addresses the shortfalls of the conventional simulation-based VSA, e.g., poor real-time performance and scalability with respect to the power system size. Various machine learning techniques such as extreme neural networks, ensemble learning, recurrent neural networks [9], and data augmentation [10] are used for VSA. In this paper, we focus on the adversarial example attacks against the machine learning models used for VSA.

FDI attacks on power systems. Modern ICTs adopted in power grids introduce cybersecurity concerns [20]. It is shown that FDI on power flow measurements can mislead state estimation and bypass the bad data detection [8]. Further studies show that FDI can be designed to mislead frequency control [21], voltage control [22], Optimal Power Flow (OPF) [23], and the Advanced Metering Infrastructure (AMI) [24]. The studies [21], [22] schedule *optimal FDI* that plans the FDI sequence to minimize the time left for reaction [21] or maximize the state estimation error [22]. The work [23] proposes a time-efficient framework to analyze the impact of FDI on the OPF. Countermeasures against FDIs on power systems have also been investigated, including both attack detection [21], [24] and mitigation [22], [25]. The above studies consider the strategic planning and mitigation of FDI. However, the targets of the FDI are not DNN-based.

Adversarial example attacks on power grid. Adversarial example is a specific form of FDI aiming to mislead DNN. The work [3] analyzes the impact of a specific type of adversarial examples on the DNN-based load forecasting. The work [9] studies the vulnerability of the machine learning models used for VSA under adversarial examples and evaluates the defense effectiveness of the existing adversarial training. Different from the previous studies, we perform a requirement investigation based on a VSA application for six representative adversarial examples that are frequently evaluated in literature [11] to analyze the conditions for effective attack launching. Meanwhile, the construction of these attacks imposes distinct minimal requirement on the adversary, which provides insights into understanding the credibility of adversarial examples in the context of power systems. Our prior work [26] analyzes the requirements and evaluates the prevailing countermeasures against adversarial examples. In this paper, we further propose a new defense approach that jointly detects and thwarts adversarial examples.

Countermeasures against adversarial examples. Countermeasures against adversarial examples are categorized into the *model hardening* and *input cleansing* methods. Model hardening improves robustness against adversarial examples by modifying the target DNN itself. *Adversarial training* is a model hardening method that modifies the training of the target DNN by including adversarial examples with their genuine labels. Adversarial training achieves state-of-the-art defense performance on various benchmarks as shown in existing research [27] and competition [28]. The input cleansing method eliminates the adversarial perturbations [11]. Different from *ad hoc* approaches (e.g., data randomization, compression, and foveation [11]), the sys-



(a) Stable. (b) Unstable.

Fig. 1: Example of stable and unstable situations.

tematic input cleansing defense of APE-GAN [18] aims to learn a manifold mapping from the adversarial input to clean input. The APE-GAN trains a discriminator and a generator simultaneously. The discriminator aims to differentiate between the clean input and the output of the generator, while the generator aims to cleanse the adversarial input and output its benign counterpart. The technical details of the adversarial training and APE-GAN are explained in the supplemental file. Recent defenses adopt ensembles to counteract adversarial examples [29], [30]. The work [30] trains multiple minimally overlapping models to detect adversarial examples. But the work [30] considers only a limited number of attack construction methods. The defense proposed in [29] performs input denoising as the first line of defense and then employs an ensemble to detect the adversarial examples that escape the input denoising based on kappa statistics. Different from the approach in [29], the defense approach proposed in this paper uses an ensemble to detect adversarial examples based on the predictive uncertainty. In addition, our approach applies majority vote to generate the final inference result that is robust to the adversarial example attacks, which is referred to as attack thwarting. The works in [29], [30] do not offer attack thwarting. Note that both existing approaches [29], [30] do not consider the universal attacks that are more credible in the context of power system applications according to our analysis, whereas our approach considers both the input-specific and universal attacks.

3 Background and Preliminaries

3.1 DNN-based Online Short-Term VSA

A stable power system has the capability to regain an equilibrium state after a disturbance [1]. Assessing the power system stability against possible disturbances is important because instability can lead to area load loss or transmission lines tripping, which may cause cascading failures and even widespread blackout. Stability is usually assessed concerning *rotor angle*, *frequency*, and *voltage*. According to the time scale of the post-contingency dynamics, short-term and long-term stability assessments cover horizons of several seconds and up to multiple minutes, respectively. We focus on the short-term VSA in this paper, which classifies the system into *stable* or *unstable* conditions based on a one-second trajectory that contains the voltage traces of the transmission buses. Note that our study can be extended to other forms of DNN-based stability assessment. Fig. 1 shows the voltage trajectories characterizing stable and unstable conditions over 5.0 seconds. In both trajectories, a fault occurs at 0.1 seconds followed by an automated fault clearance at 0.2 seconds. In stable conditions, all the bus voltages can restore to acceptable levels (e.g., less than 10% deviations from the nominal values). In unstable conditions, the bus voltages remain far away from the nominal values or even collapse.

Time-domain simulations simulating an extensive set of potential faults are conventionally considered for offline VSA [1]. Differently, online VSA assesses the system stability with a hard deadline based on real-time voltage measurements. Restorative actions will be taken if the system is assessed unstable. There are mainly two challenges for online VSA: (1) the system operator has limited/no information about the fault occurring at run time. However, the information is needed for bootstrapping the time-domain simulation; (2) the time-domain simulation is usually much slower than the evolution of the power system state. To address these challenges, machine learning has been adopted for online VSA [9], [10]. Specifically, a machine learning model $f(\mathbf{x};\boldsymbol{\theta})$ with weights $\boldsymbol{\theta}$ is trained to classify a voltage trajectory $\mathbf{x}$ at run time based on a training dataset $(\mathbf{X},\mathbf{Y}) = [(\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)]$, where $\mathbf{x}$ is the post-fault voltage trajectory generated by the offline time-domain simulation and y is the stability classification label. The model $f(\mathbf{x};\boldsymbol{\theta})$ trained with abundant training data can handle a wide range of faults.

TABLE 1: Adversarial example construction methods.

Attack	Categorization [11]		
	Scope	Computation	Knowledge
FGSM [7]	Input-specific	One-shot	White/black-box
PGD [13]	Input-specific	Iterative	White/black-box
DF [14]	Input-specific	Iterative	White-box
CW [15]	Input-specific	Iterative	White-box
UAP [16]	Universal	Iterative	White/black-box
UAN [17]	Universal	Iterative	White/black-box

3.2 Taxonomy of Adversarial Example

The taxonomy of adversarial examples is illustrated in Table 1, according to the categorization in [11]. In terms of applicable *scope* of the attack, *input-specific* means the adversarial perturbation is crafted to be effective against individual clean input sample, while *universal* means the perturbation is crafted to be effective on many clean examples. In terms of the *computation* required, the perturbation can either be generated by a *one-shot* computation (e.g., by using a closed-form formula) or an *iterative* search process. In terms of the *knowledge* about the target DNN, the *white-box* attacks need complete information of the DNN's internals, i.e., weights and architecture, while the *black-box* attacks only need the access to run the DNN without knowing its internals. Although many effective attacks require white-box knowledge, some of them are still effective under the black-box setting that uses a surrogate DNN to craft the adversarial examples. To train the surrogate DNN, the adversary may utilize the dataset obtained from querying the black-box target DNN using many input samples. In Table 1, such attacks are labeled as "white/black-box."

In this paper, we study six representative adversarial examples as shown in Table 1, i.e., FGSM [7], PGD [13],

DF [14], CW [15], UAP [16], and UAN [17]. The detailed formulations of these attacks are provided in the supplemental file. We briefly describe their essences here. FGSM computes a one-step perturbation using a closed-form formula. PGD iteratively performs mini-step FGSMs. DF finds the minimum perturbation added to the clean example to cross the approximated decision boundary. CW simplifies the optimization problem for crafting adversarial examples by applying the Lagrangian relaxation and then searches the solution. The aforementioned four attacks are input-specific. Then, we introduce the universal UAP and UAN attacks. UAP finds the DF perturbation for many clean examples and accumulates the perturbations to form a unified universal perturbation. UAN is a generative neural network that takes as input values randomly sampled from a distribution and outputs adversarial perturbations.

4 Problem Statement

4.1 System and Data Description

We consider the New England 10-machine 39-bus system [31], which is a power grid model widely used in power system research. The system's single-line diagram can be found in the supplemental file. We perform extensive time-domain simulations to generate voltage trajectories using the commercial industry-standard software PSS/E [32]. PSS/E is a leading-edge electromechanical time-domain simulation tool designed for comprehensive assessment of dynamic behavior of power systems. Compared with existing voltage security assessment tools, PSS/E is more powerful and can be used to calculate security limits under specified criteria, contingencies, and transfer conditions for a rich set of models. Specifically, we consider composite load because high penetration of induction motor loads is the driving force for short-term voltage stability issues in today's power systems. We adopt the industry-standard composite load model "CLOD" [33] to model different load components including small motors, large motors, discharge lighting, transformer saturation, and voltage-dependent loads in the simulations of PSS/E software. 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 line tripping, which simulates different topology change scenarios. Each voltage trajectory consists of the voltage traces of the 39 buses. The sampling rate is 100Hz. In total, 6,536 voltage trajectories are generated, covering a wide range of practical system operating points. We divide the voltage trajectories into 4,536 training, 1,000 validation, and 1,000 testing samples. Each sample is a 1×3900 vector containing the 39 buses' voltage traces over a one-second duration after the clearance of the fault. We use a convolutional neural network (CNN) for VSA. The CNN has two convolutional layers with 128 1×5 filters followed by 1×2 max pooling, two convolutional layers with 256 1×5 filters followed by 1×2 max pooling, two dense layers with 512 rectified linear units (ReLUs) each, and a binary-class output layer. The trained CNN has an accuracy of 99.5% on the validation dataset. The empirically measured false positive rate and the false negative rate are 0% and 0.5%, respectively, in detecting the instability.

Algorithm 1: Credibility analysis of adversarial examples against DNN-based VSA.

Input: Training dataset $(\mathbf{X}, \mathbf{Y})$, VSA DNN $f(\cdot; \boldsymbol{\theta})$, set of attack methods $S_A = \{a_1, ..., a_p\}$, set of defense methods $S_D = \{d_1, ..., d_q\}$, union set of attack requirements $S_R = \{r_1, ..., r_n\}$, empty sets of individual attack requirements $S_{R_i} = \emptyset, i = 1, ..., p$

Output: Individual attack requirement sets $S_{R_i}, i = 1, ..., p$, credibility analysis results

```
for attack a_i ∈ S_A do
    Generate adversarial examples based on (X, Y) and f(·; θ);
    % Obtain minimal set of requirements for each attack.
    for requirement r_j ∈ S_R do
        Evaluate whether adversarial examples of a_i requires r_j to mislead f(·; θ);
        if a_i requires r_j then
            Add r_j to S_{R_i};
        else
            Continue;
    end
    Credibility analysis for a_i based on S_{R_i};
    for defense d_k ∈ S_D do
        Evaluate the defense effectiveness of d_k for a_i;
    end
end
```

4.2 Threat Models and Research Problem

Misleading the target DNN by adding minimized perturbation to the input is the general objective for the six adversarial example attacks. As shown in Table 1, the six attacks with different features impose distinct sets of minimal requirements 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: A voltage trajectory contains the voltage traces of all transmission buses. The read access of the voltage trajectory is related to the applicable scope of the adversarial example (i.e., input-specific or universal). The input-specific attacks need this read access since they require the whole clean voltage trajectory to craft attacks. In contrast, the universal attacks do not require this access.

(2) Write access to the voltage measurements: The number of voltage traces that the attacker needs to tamper with is an important requirement related to the overhead of launching attacks. The capability to tamper with all the voltage traces of transmission buses (i.e., full write access) apparently implies a strong and resourceful attacker.

(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 the dataset used to train the target DNN, e.g., the labeled historical voltage trajectories, is needed for effective attack construction.

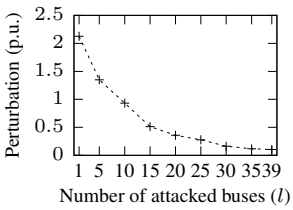


Fig. 2: Per-bus average perturbation (Attack: DF).

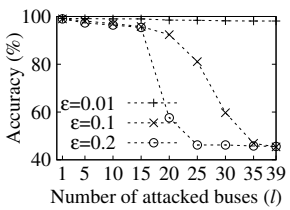


Fig. 3: Accuracy vs. l and ϵ (Attack: PGD).

TABLE 2: Requirements for effective attacks against VSA.

Attack	Minimal requirement			
	Access		Knowledge	
	Read	Write	DNN internal	Training data
FGSM [7]	Yes	Partial	Either	
PGD [13]	Yes	Partial	Either	
DF [14]	Yes	Partial	Yes	No
CW [15]	Yes	Full	Yes	No
UAP [16]	No	Partial	No	Yes
UAN [17]	No	Partial	No	Yes

This paper investigates the credibility of adversarial example attacks against the DNN-based VSA by inquiring three issues. First, we investigate the minimal set of requirements for each attack to mislead the VSA DNN, which precisely depicts the threat models for different attacks. Second, we analyze the credibility of the attacks based on their minimal requirements. The different requirement aspects should be weighed differently, e.g., meeting the *real-time* requirements (1) and (2) is often more difficult than meeting the *static* requirements (3) and (4). Finally, we evaluate the defense performance of prevailing countermeasures for VSA against the credible attacks and develop new countermeasures if the existing ones are found ineffective in certain cases. The credibility analysis procedure is summarized in Algorithm 1.

5 Attack Requirement Investigation

5.1 Attack Evaluation Settings

The adversarial examples are constructed based on the 1,000 test samples described in Section 4.1. For the input-specific attacks, a 1×3900 perturbation vector is computed for each test sample under the white-box setting against the target DNN or under the black-box setting against the surrogate DNN. The surrogate DNN has the same architecture and hyperparameters as the target DNN and is trained using the same training dataset at random initialization. The UAP attack computes a universal perturbation vector using 1,000 randomly selected training data samples. Then, the universal perturbation is applied to all the 1,000 test samples. We adopt the hyperparameters setting from [17] and train the UAN attack generator using 1,000 randomly selected training data samples. Then, we use the generator to generate a 1×3900 perturbation for each of the 1,000 test samples.

5.1.1 Partial perturbation implementation

We implement the adversarial example construction that only requires write access to a subset of l buses because the number of voltage traces that the attacker needs to tamper with is a key requirement, as discussed in Section 4.2. 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 to be attacked are perturbed. Specifically, a 1×3900 matrix $\mathbf{M}$, which has unit values in the dimensions of the l buses to be perturbed and zero values in the dimensions where no perturbation is added, is applied to restrict the area modified by the adversary. For one-shot attack, i.e., FGSM, each computed adversarial perturbation is multiplied by $\mathbf{M}$ and added to the clean example. For iterative attacks, we perform the multiplication at each step during the search process for attack construction. We set l to be 1, 5, 10, 15, 20, 25, 30, and 39 in the experiments. For each setting, we perturb the first l bus voltage measurements.

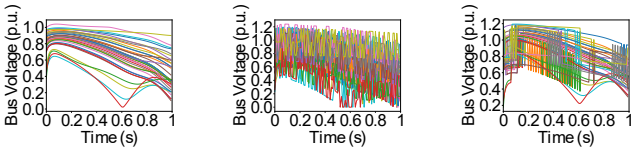
5.1.2 Attack perturbation intensity settings

Intuitively, adversarial examples with larger intensity of perturbation can mislead the target DNN more effectively. Thus, for fair comparison of attack effectiveness, it is non-trivial to configure the attack perturbation intensity. We use DF to guide the settings of ϵ for FGSM, PGD, UAP, and UAN and κ for CW, as DF automatically finds the minimal adversarial perturbations to mislead the target DNN. Note that ϵ is the maximum perturbation intensity, and κ is a hyperparameter controlling perturbation intensity. Please refer to the supplemental file for the definition of these two hyperparameters. From Fig. 2, the per-bus average intensity of the ℓ_2 perturbations found by DF decreases with the number of attacked buses (i.e., l). This is because, when the perturbations are restricted to fewer buses, the needed perturbation intensity is larger to mislead the DNN. The average perturbation intensity ranges from 0.27 per unit (p.u.) to 2.12 p.u. when $1 \leq l \leq 25$. Since the nominal bus voltage is 1 p.u., these values are unacceptably large. Thus, we consider DF with $30 \leq l \leq 39$ such that the average perturbation intensity is at most 0.16 p.u..

Fig. 3 shows the accuracy of DNN under PGD attack versus l when ϵ is from 0.01 p.u. to 0.2 p.u.. With $\epsilon = 0.1$ p.u. and 0.2 p.u., the attack is effective when sufficient bus voltage readings are tempered with. We set $\epsilon = 0.2$ p.u. for FGSM, PGD, UAP, and UAN and $\kappa = 0$ for CW. Note that the average perturbation intensity for CW with $\kappa = 0$ is 0.18 p.u.. In a word, the settings of $\epsilon = 0.2$ p.u. and $\kappa = 0$ enable fair comparison among the six attacks. Note that we can configure the adversarial perturbation magnitude. In this paper, the maximum deviation from the nominal bus voltage is set at ± 0.2 p.u.. We evaluate the worst-case vulnerability of the VSA DNN under this setting.

5.2 Attack Effectiveness and Requirements

Table 2 summarizes the results from our evaluation. The results indicate the minimal requirement for each of the six attacks to fool the VSA DNN. The column of "read access" states the necessity for the adversary to obtain the clean voltage trajectory. The column of "write access" describes whether the adversary needs to perturb all the bus voltage traces (full) or just a portion of them (partial) for the attack



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

Fig. 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.

to be effective. The "DNN internal" and "training data" columns specify whether the DNN internal and a training dataset are needed, respectively.

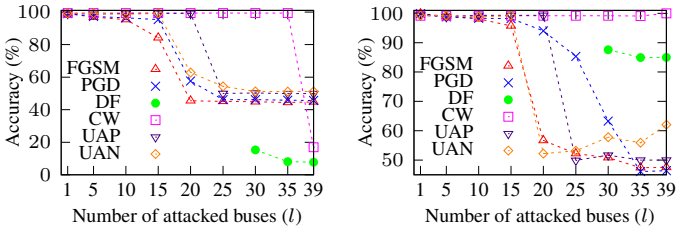
5.2.1 Random perturbations versus adversarial examples

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 with maximum intensity at 0.2 p.u. are applied to all buses. The DNN achieves 99.3% test accuracy under the random perturbation, which is only 0.2% lower than that on clean samples. However, in the presence of FGSM attack, the accuracy drops to 45.4%. This shows that, even if the adversary is able to compromise all bus voltage readings, they still need to apply intelligence to plan the perturbation.

5.2.2 Effectiveness of input-specific adversarial examples

Fig. 5a presents the VSA accuracies under white-box attacks. The input-specific FGSM, PGD, DF, and CW are not effective when $1 \leq l \leq 15$, where FGSM with $l = 15$ causes the lowest accuracy of 84.2%. The DNN accuracies drop to 45.5% and 57.6% under the FGSM and PGD attacks with $l = 20$, respectively, which suggest that input-specific adversarial examples under white-box setting can decrease VSA accuracy by more than 50% through tampering with about half the bus voltage measurements. The DF attacks with $30 \leq l \leq 39$ are very effective, which cause only 15.2% and 8.0% VSA accuracies when $l = 30$ and 35, respectively. When $1 \leq l < 39$, the CW attack is not effective. When CW can temper with all buses, the DNN accuracy drops to 15.5%. This suggests that, although the optimization-based CW is often regarded as one of the most powerful attacks in CV applications [11], its effectiveness against VSA is conditioned on the write access to all input dimensions. In contrast, the gradient-based FGSM, PGD, and DF attacks achieve non-negligible attack effectiveness when partial input dimensions are under attack. We summarize the results in the "write access" column of Table 2.

Fig. 5b presents the VSA accuracies under black-box attacks. We can see that the VSA accuracies remain at 84.9% and 99.6%, respectively, under the DF and CW attacks, which implies the ineffectiveness of the black-box DF and CW attacks. The reduced attack effectiveness against the black-box target DNN is because the DF and CW adversarial examples overfit the surrogate DNN used for constructing



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

Fig. 5: DNN accuracy in the presence of attack.

them. Among all the input-specific attacks, FGSM shows the strongest transferability from the surrogate DNN to the target DNN. When l increases from 15 to 20 for FGSM, the DNN accuracy drops from 95.7% to 56.7%, which is similar to the results obtained under the white-box setting. Attack effectiveness of PGD reduces in the black-box setting. However, when l increases from 30 to 35, the DNN accuracy still drastically drops from 63.3% to 46.1%. The FGSM and PGD exhibit good transferabilities because they are less overfit to the surrogate DNN. These results imply that keeping the target model confidential is a weak defense against FGSM and PGD. We summarize the results in the "DNN internal" column of Table 2. The "either" note means the attack requires only the DNN internal under the white-box setting or only the training data for building the surrogate DNN under the black-box setting.

5.2.3 Effectiveness of universal adversarial examples

As shown in Fig. 5a, the VSA accuracies remain at above 98.6% under white-box UAP and UAN with $1 \leq l \leq 15$. The white-box universal adversarial examples can decrease the VSA accuracy by up to 49.5% when tempering with only half the buses. Specifically, when l increases from 20 to 25, VSA accuracy drops from 99.2% to 50.0% under white-box UAP. When l increases from 15 to 20, DNN accuracy drops from 99.2% to 62.8% under white-box UAN. Thus, in Table 2, UAP and UAN require partial "write access".

From Fig. 5b, black-box UAP and UAN exhibit similar attack effectiveness as under the white-box setting, which indicates the target DNN internal is not a must for UAP and UAN to take effect. This is summarized in the "DNN internal" column of Table 2. When $l \geq 25$ and 20, UAP and UAN are effective, respectively, showing that UAP is slightly less effective than UAN. Meanwhile, black-box universal UAP and UAN attacks are more effective than all the black-box input-specific attacks. This indicates that universal attacks effectively capture the distribution of the adversarial examples while avoiding overfitting to the surrogate DNN under the black-box setting. Note that crafting both white-box and black-box universal attacks requires a clean training dataset, as summarized in Table 2.

5.3 Implications and Credibility Analysis

The key observations from the evaluation results in Section 5.2 are summarized as follows:

- Except CW and DF attacks, all other adversarial example attacks can decrease the target DNN's accuracy by around 50% when tempering with about 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%. However, they impose strong requirements such as read and write access to the voltage traces of many/all buses, and 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

Static knowledge required by the adversary contains training data and DNN internal. From Table 2, each of the six attacks requires at least one of them for effective attack construction. However, since training data and DNN internals are static information, it is not difficult for the adversary to obtain them under the scenario of advanced persistent threat (APT), e.g., conducting social engineering against employees of the grid operator. Even if the attacker 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 craft the effective FGSM, PGD, UAP, or UAN adversarial examples. In summary, under the APT scenario, preserving the confidentiality of the static knowledge is a shaky defense. Therefore, the weights of the last two columns of Table 5.2 are marginal in the attack credibility assessment against VSA.

5.3.2 Implication of write access requirement

The adversary must have the capability to modify the voltage traces of all or some buses, in order to launch adversarial example attack. We discuss the implication of our results from two facets.

Compromising half the buses is a rule of thumb: Our evaluation shows that certain input-specific attacks such as DF and CW can almost subvert VSA when all buses are tempered with. However, as shortly analyzed in Section 5.3.3 that input-specific attacks are less credible, the subversion is also less credible accordingly. Thus, the degradation of VSA accuracy to about 50% caused by the universal attacks as shown in Fig. 5 is a more credible maximal attack effectiveness. Section 5.2.3 shows that UAN is more effective than UAP. A significant drop of DNN accuracy occurs under UAN when l increases from 15 to 20. When l continues to increase from 20, the further accuracy drops are less salient. Since the cost of the attack increases with l (which is discussed in the next paragraph), compromising half the buses to obtain their write accesses is a rule of thumb for the adversary.

Attack implementation: There are three possible ways for attack implementation. (1) An adversary within the enterprise network of the power grid control center can compromise the measurements of all buses. This strong adversary, however, is ill-motivated because it should directly subvert the VSA results. (2) An adversary compromises the communication links from the buses to the control center. On one hand, interception of the network transmission of the clean voltage trajectory on the communication paths, e.g., on a router, is required to transmit the maliciously perturbed voltage trajectory to the control center without causing suspicion. On the other hand, the cryptographic protection needs to be breached. For instance, the attacker may have obtained the master keys of the compromised links, which, however, implies a strong adversary. 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 [34]. 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 input-specific adversarial examples cannot be generated until obtaining the whole voltage trajectory, it is not applicable for the adversary to conduct the sensor reading manipulation by electromagnetic interference as discussed in Section 5.3.2. As a result, the adversary must compromise the communication links from all buses to the control center, which incurs a high overhead. The requirement of full read access renders the input-specific attacks resource- and skill-demanding.

Universal attacks: The universal adversarial examples are independent of the real-time clean input samples. Therefore, they can be implemented either through sensor reading manipulation by electromagnetic interference or by compromising the communication links. The sophisticated interception required by the input-specific attacks is not required. Note that the widely studied FDI attack against the power grid state estimation [8] is also a universal attack. To be more specific, the perturbation vector added to the power flow measurement is given by $\mathbf{a} = \mathbf{H}\mathbf{c}$, where $\mathbf{H}$ is a constant matrix for state estimation and $\mathbf{c}$ is an arbitrary vector. That is to say, the perturbation $\mathbf{a}$ is independent of the power grid real-time power flow state and only the piece of static knowledge of the system (i.e., $\mathbf{H}$) should be obtained by the adversary. Given the same nature of the universal adversary example attacks and the state estimation FDI attack studied in [8], they have the same credibility that has substantially concerned the relevant research communities.

5.3.4 Summary

Based on 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 attacker compromises the voltage traces of more than half the buses, devastating effects can be

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 effectiveness.

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.	✗	✗	✓	✓

generated on VSA. Meanwhile, the possibility of launching the input-specific adversarial example attacks should not be expelled. 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 evaluate the defense performance of adversarial training, APE-GAN, and the combination of them as the defense.

Setting of adversarial training. We consider two variants of adversarial training called *FGSM adversarial training* [7] and *PGD adversarial training* [27], which add 1,000 FGSM or PGD adversarial examples with genuine labels, respectively, into the training dataset. The added samples for adversarial training are generated based on the validation dataset with ϵ set to 0.2 p.u.. The FGSM-hardened DNN achieves 99.2% accuracy on clean test samples and 98.6% on FGSM adversarial examples generated from clean test samples. The PGD-hardened DNN achieves 98.6% accuracy on clean test samples and 98.2% on PGD adversarial examples crafted from clean test samples. These results show that the hardening is effective against the considered attack method. To evaluate the defense performance, we consider both white-box setting, where the attacker can access the internals of the hardened DNN, and black-box setting, where the attack cannot access the hardened DNN's internals. Thus, under the white-box setting, the defense follows the Kerckhoffs's principle to assume an enemy knowing the system including its defense mechanism. Under the black-box setting, the adversary crafts the attacks based on the surrogate DNN trained using the obtained clean training data. Note that the adversarial training is not applied by the adversary to harden the surrogate DNN.

Setting of APE-GAN. We follow the procedure in [18] to train the APE-GAN using the clean training dataset and 1,000 FGSM adversarial examples crafted against the target DNN. The APE-GAN is first applied to cleanse the input samples before the inferencing by VSA DNN. To evaluate the defense performance, we consider both settings of white-box and black-box. In the white-box setting, the adversarial examples are generated against the target DNN. Note that we do not consider APE-GAN in this white-box setting, since the generation of adversarial perturbations aiming to bypass APE-GAN is still an open issue. In the black-box setting (which is not mentioned in the APE-GAN paper [18]), the adversarial examples are constructed based on the surrogate DNN.

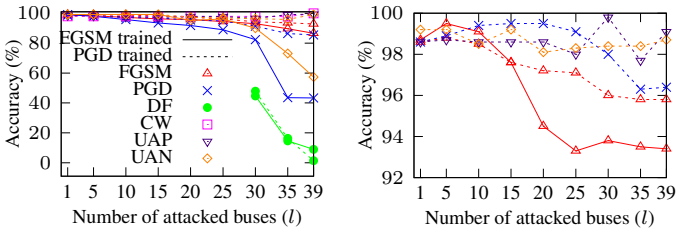
Combination of adversarial training and APE-GAN. The results that will be shortly presented in Section 6.2 show that PGD adversarial training is more effective than FGSM adversarial training. As a result, we consider the combination of PGD adversarial training and APE-GAN as defense for better performance. During the training, we first apply PGD adversarial training to harden the DNN. Then, we train the APE-GAN using the clean training dataset and 1,000 FGSM adversarial examples crafted based on the PGD-hardened DNN. During the testing, the APE-GAN is firstly applied to cleanse the input and then the PGD-hardened DNN is used to make the inference. We evaluate the defense performance of this combination scheme under both white-box and black-box attacks. In particular, we use the target PGD-hardened DNN to generate white-box adversarial examples; and we use the surrogate DNN that is not hardened by adversarial training to generate the black-box adversarial examples.

6.2 Defense Effectiveness Results

Table 3 summarizes the defense performance. An "effective" attack means that it can decrease the accuracy of the target DNN to 80% and below; an "effective" defense means that it can maintain the accuracy of the target DNN at above 80%, in the presence of an effective attack. From Section 5.2, the black-box DF and CW are not effective attacks. Thus, these two attacks are considered in the evaluation of the defense performance in this section.

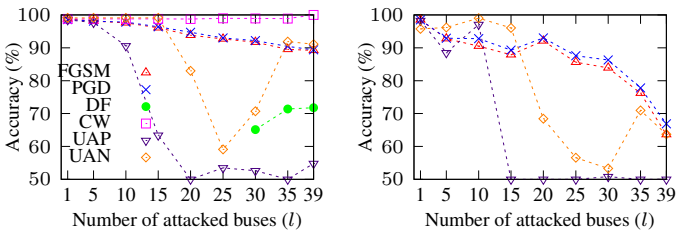
Effectiveness of adversarial training. Fig. 6 presents the accuracy of the VSA DNN that is hardened by adversarial training versus l under the white-box and black-box attacks. We first analyze the defense performance of adversarial training under input-specific attacks. From Fig. 6a, under the white-box PGD with $l = 35$ and 39, FGSM adversarial training is not effective. However, under the white-box PGD attack, PGD adversarial training is effective. This result indicates that PGD adversarial training is more effective in protecting the VSA DNN than the FGSM adversarial training. Under the white-box DF with $30 \leq l \leq 39$, both defenses of FGSM and PGD adversarial training are not effective. Adversarial training is effective against all effective black-box input-specific attacks, as shown in Fig. 6b. We then analyze the defense performance under universal adversarial examples. As shown in Fig. 6, PGD adversarial training effectively protects the target DNN against both the white-box and black-box universal attacks. The observations from Fig. 6 are summarized in the top two rows of Table 3 headed by ❶ and ❷.

Effectiveness of APE-GAN. Fig. 7 shows the VSA DNN's accuracy versus l when APE-GAN is applied to



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

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

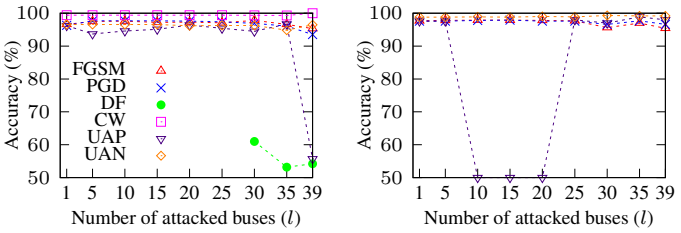


(a) White-box attacks (b) Black-box attacks

Fig. 7: VSA accuracy when APE-GAN defends attacks.

cleanse the input under both the white-box and black-box settings. We first evaluate the defense effectiveness of APE-GAN under the input-specific attacks. Under the white-box setting, APE-GAN is only ineffective against the input-specific DF attack. Under the black-box setting, APE-GAN is ineffective against both attacks of FGSM and PGD. We can see that the defense of APE-GAN is more effective under white-box attacks. This is caused by the design of APE-GAN with the goal of removing the adversarial perturbations generated by the adversary based on the white-box target DNN [18]. We then evaluate the defense effectiveness of APE-GAN under universal attacks. As shown in Fig. 7a, when $l = 15$ and 20, white-box UAP adversarial examples can bypass the APE-GAN with probabilities of 36.1% and 49.5%. When $l = 25$, 40.4% of the white-box UAN adversarial examples bypass APE-GAN. From Fig. 7b, APE-GAN achieves poor defense effectiveness against universal attacks in black-box setting. In summary, APE-GAN cannot defend against universal adversarial examples. The observations obtained from Fig. 7 are summarized in the row of Table 3 headed by ❸.

Effectiveness of adversarial training combined with APE-GAN. Fig. 8 plots the VSA DNN's accuracy when APE-GAN is combined with a PGD-hardened DNN (i.e., the combination scheme presented in Section 6.1). As none of the PGD adversarial training and APE-GAN are effective against the white-box DF, combining them is also ineffective to counteract with the white-box DF. Moreover, the combination has deteriorated the defense performance for UAP attacks under certain settings (i.e., when $l = 39$ under the white-box setting and $10 \leq l \leq 20$ under black-box setting), compared with applying PGD adversarial training solely. This suggests that the APE-GAN pre-processing may worsen the defense effectiveness of the adversarially hardened DNN in defending against certain attacks. Non-monotonicity of DNN accuracy versus l can be observed in



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

Fig. 8: Defense effectiveness of combining PGD adversarial training and APE-GAN against various attacks.

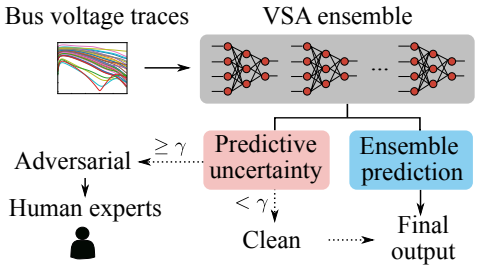


Fig. 9: Approach overview. This paper considers four variants of VSA ensemble, i.e., MC-dropout DNN units, MC-dropout DNN, DNN unit ensemble, and DNN ensemble.

Fig. 7a, Fig. 7b, and Fig. 8b. This is because the cleansing of APE-GAN is unpredictable and may sometimes disrupt the input samples and thus 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 for the reason that the number of experiments to generate one point in the figures will be huge. For instance, to choose 10 buses from 39 ones, there are $\binom{39}{10} = 635,745,396$ possible combinations. The results observed from Fig. 8 are summarized in the row of Table 3 headed by ❹.

6.3 Implication of Results

From Table 3, PGD adversarial training is effective against all attacks except the white-box DF attack. Thus, the PGD adversarial training can be applied to protect DNN-based VSA against the more credible universal adversarial examples that generate non-negligible concerns. However, although it is exorbitant to craft sophisticated input-specific attacks, the possibility of launching such attacks cannot be totally ignored. The following section presents a new defense approach that jointly detects and thwarts both universal and input-specific attacks.

7 Joint Attack Detection and Thwarting

The results in Section 6 show that the state-of-the-art defenses cannot fully thwart adversarial example attacks. Moreover, it is essential to detect adversarial examples in the power system, such that these samples can be flagged and passed to human experts for further analysis. In this section, we propose a new defense approach to counteract both the input-specific and universal adversarial example attacks.

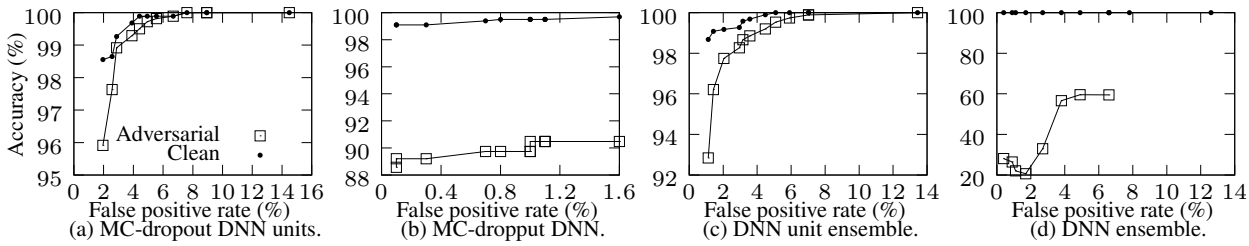


Fig. 10: VSA ensembles' accuracies on clean and adversarial samples that do not trigger the attack detector.

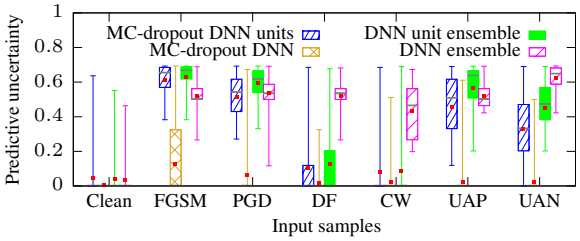


Fig. 11: Predictive uncertainty values. Gray bar represents median; red dot represents mean; box represents 25th and 75th percentiles; whiskers represent minimum and maximum. The same applies to all the box plots in this paper.

The approach consists of (1) an attack detection module that measures the predictive uncertainty of any given input to detect adversarial examples; and (2) an attack thwarting module that aims to generate the genuine label for the input.

7.1 Approach Overview

Fig. 9 illustrates the run-time workflow of our approach. Given as input the voltage trajectory of all buses' voltage traces, each DNN of the VSA ensemble generates a prediction. The design of VSA ensemble will be presented in Section 7.1.2. Based on the multiple predictions generated by the VSA ensemble, a predictive uncertainty and an ensemble prediction for the input are generated together. Then, the predictive uncertainty is compared with a predefined uncertainty threshold. If the predictive uncertainty is smaller than the threshold, the input is considered clean and the ensemble prediction is the final output. Otherwise, the input is considered adversarial. In this case, the ensemble prediction for this input is invalid and the input will be passed to human expert for further analysis.

7.1.1 Preliminary on predictive uncertainty estimation

Predictive uncertainty estimation measures the predictive uncertainty of DNN [35], which is essential for safety-critical decision making in autonomous systems. Bayesian deep learning is commonly applied to model predictive uncertainty [35]. Given the training dataset $(\mathbf{X}, \mathbf{Y}) = [(\mathbf{x}_i, y_i), i = 1, ..., m]$, as opposed to learning a point estimate of the model parameters (denoted by $\hat{\boldsymbol{\theta}}_{\text{MLE}}$) via maximum likelihood estimation by minimizing the negative log-likelihood $-\sum_{i=1}^{m} \log p(y_i|\mathbf{x}_i, \boldsymbol{\theta})$, Bayesian deep learning infers the probability distribution over the model parameters under the Bayesian inference framework. Specifically, the posterior distribution over the model parameters is modeled by $p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{Y}|\mathbf{X})}$. Then, the predictive posterior distribution given a new input sample $\mathbf{x}^*$ is obtained by marginalizing out the estimated posterior distribution over the model parameters: $p(y^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y}) = \int p(y^*|\mathbf{x}^*, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y})d\boldsymbol{\theta}$. However, in practice, it is intractable to marginalize over the whole parameter space for complex DNNs due to the vast dimensionality. Among the various approximations made to tackle this issue, the Monte Carlo dropout [36] (MC-dropout) and deep ensembles [37] are two approaches widely adopted. In particular, MC-dropout activates dropout layers during both the training and testing processes, which can be viewed as performing variational inference with Bernoulli distributed random variables. The predictive uncertainty is given by performing multiple stochastic forward passes on the same input. Deep ensembles approach trains multiple models with random initialization and shuffled training data. The obtained ensemble is used to produce uncertainty estimation. The existing studies of uncertainty estimation mainly focus on evaluating the attack detection performance of adversarial examples. In this paper, we propose an approach that jointly performs attack detection based on predictive uncertainty and attack thwarting using majority vote.

7.1.2 VSA ensemble design and setting

In our experiments, we consider two basic neural network architectures for the DNNs in the VSA ensemble. The first architecture is the one described in Section 4.1, which we refer to as *VSA DNN*. The second architecture is the *VSA DNN units*. Specifically, each VSA DNN unit is trained using the voltage traces of a single bus only. We investigate this architecture to evaluate whether the DNN units trained using non-overlapping parts of the input data are more diverse and generate more distinct predictions for adversarial examples. Each of the VSA DNN units has two convolutional layers with 32 and 64 1×5 filters, two fully connected layers with 128 neurons each, and one binary-class softmax layer. We then combine the two architectures with the two uncertainty estimation methods mentioned in Section 7.1.1, i.e., MC-dropout and deep ensemble. Specifically, we consider four variants of VSA ensemble as follows.

1) **MC-dropout DNN units** are trained and tested with dropout applied before every weight layer. The ensemble has 39 VSA DNN units corresponding to all buses. For each input, the ensemble parameters are different due to the randomness of dropout.
2) **MC-dropout DNN** trains a VSA DNN with dropout applied before every layer. The ensemble has N

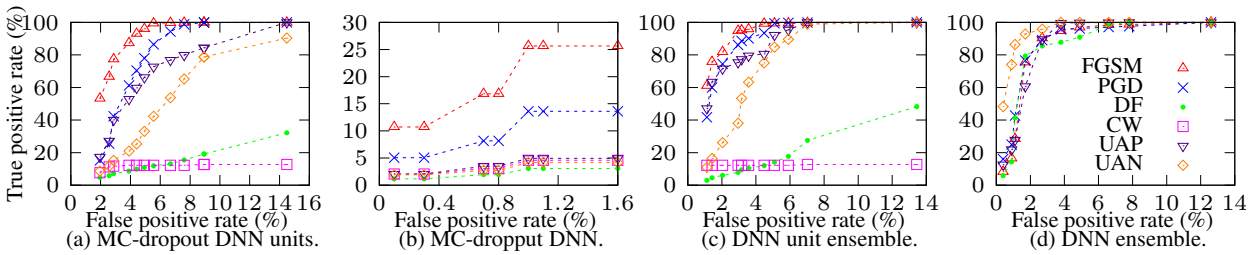


Fig. 12: ROCs of VSA ensembles.

VSA DNNs sampled during testing with dropout activated.

3) **DNN unit ensemble** has 39 VSA DNN units trained from random initialization and with shuffled training data. It can be viewed as the MC-dropout DNN units with a zero dropout rate. The ensemble parameters during testing are static.
4) **DNN ensemble** consists of N VSA DNNs trained with shuffled training data at random initialization.

The dropout rate for the MC-dropout DNN and MC-dropout DNN units is denoted by dp.

7.1.3 Attack detection

At run time, each input sample consisting of the voltage traces of all buses is sent to the VSA ensemble. Based on the multiple predictions generated by each of the DNN in the VSA ensemble, the predictive uncertainty is calculated. In this paper, the predictive uncertainty is defined as the entropy of the multiple predictions. The predictive uncertainty is then compared with a pre-defined uncertainty threshold γ to decide whether the input is clean or adversarial. The γ can be configured to achieve the desired trade-off between the unnecessary overhead incurred to human experts (which is characterized by the false positive rate, FPR) and the attack detection performance (which is characterized by the true positive rate, TPR).

7.1.4 Attack thwarting

The attack thwarting aims to give the genuine labels of the adversarial examples that are not flagged by the attack detection. The attack thwarting performance is evaluated by the accuracy on the adversarial examples that are not detected by the VSA ensemble. In our approach, if an input is not detected as adversarial, the ensemble prediction for this input is computed by applying majority vote on the predictions generated by all members of the VSA ensemble, i.e., the predicted label that has the most occurrence is yield as the final output.

7.2 Performance Evaluation

The evaluation is conducted using the four variants of VSA ensemble described in Section 7.1.2. We set $dp = 0.2$ and $N = 40$.

7.2.1 Performance in absence of attack

This section evaluates the accuracy of the four variants of VSA ensemble on the clean test samples. Fig. 10 shows the accuracies of the clean samples that do not trigger the adversarial example detector versus the FPR. The FPR is the percentage of the clean samples that are detected as adversarial. We can see that all variants of VSA ensemble except the MC-dropout DNN can achieve accuracies of 100% on clean examples when the FPR is larger than or equal to 7.6%. This is because the models sampled from MC-dropout DNN are less diverse and tend to generate similar predictions. This lack of diversity of the MC-dropout DNN also results in the lower values of predictive uncertainty as shown in Fig. 11 and poor attack detection and thwarting performance presented in Section 7.2.2.

TABLE 4: Defense performance measured by defense success rate/detection rate/thwarting rate (%) when false positive rate $\leq 8\%$. VSA ensemble variants denoted by (1)-(4) are explained in Section 7.1.2.

VSA ensemble variants			
(1)	(2)	(3)	(4)
100/73.8/26.2	91.5/11.2/80.3	99.9/89.6/10.3	99.7/99.2/0.5

7.2.2 Performance in presence of attack

In this section, we evaluate the attack detection and the attack thwarting performance of our approach, respectively. The adversarial examples in our experiments are crafted using a surrogate VSA DNN, as described in Section 5.1. We only consider the adversarial examples that can successfully mislead the surrogate VSA DNN.

First, we evaluate the attack detection performance. Fig. 12 shows the receiver operating characteristic (ROC) curves of the four VSA ensemble variants under different adversarial example attacks. Different points in a curve are obtained by varying the uncertainty threshold γ from 0.1 to 0.65 with a step size of 0.05. The FPR indicates the unnecessary overhead incurred to human experts in performing analysis on the clean examples. The TPR shows the effectiveness of the attack detection. A higher ROC curve means a better trade-off between the unnecessary overhead and attack detection effectiveness. From Fig. 12, DNN ensemble achieves the highest ROC curves under different attacks. Specifically, when the FPR is higher than 6.6%, almost all the adversarial examples are detected. This can be inferred from the values of predictive uncertainty as shown in Fig. 11. In Fig. 11, the predictive uncertainty of different adversarial examples given by the DNN ensemble are always higher than zero. Thus, all adversarial examples can be detected using a lower uncertainty threshold γ, which, however, increases the FPR. Differently, the remaining three variants of VSA ensemble produce zero predictive uncertainty values

for the adversarial examples. Therefore, these adversarial examples cannot be detected no matter how low the uncertainty threshold γ is set. In the following experiments, we will show that these undetected adversarial examples can be rectified by the attack thwarting module using the MC-dropout DNN units and DNN unit ensemble.

Next, we evaluate the attack thwarting performance. Fig. 10 presents the accuracy on the undetected adversarial examples versus the FPR of the attack detection. A higher curve indicates a better trade-off between the unnecessary overhead and the attack thwarting performance. Under certain settings, all adversarial examples are detected and thus are not shown in Fig. 10. We can see that the MC-dropout DNN units achieve the best performance. When the FPR is 7.6%, the accuracy on the undetected adversarial examples is 100%. In comparison, when the accuracy is 100% for the DNN unit ensemble, the FPR is 13.4%.

7.2.3 Computation performance

This section investigates the computation overhead and execution latency of our approach. The experiments are conducted on our computing server. The server has 10-core Intel Core i9-7900X 3.30GHz CPU and runs Ubuntu 18.04. The server also has four NVIDIA GeForce RTX 2080 Ti 11GB graphics processing units (GPUs). All our codes are implemented using Python 3. For VSA ensemble variants (1), (3), and (4), we evenly distribute all the VSA DNNs or VSA DNN units to the four GPUs for parallel inference execution. The VSA ensemble variant (2) is replicated on each GPU. For parallel inference execution, the ensemble has $\frac{N}{4}$ VSA DNNs sampled on each GPU. To compare our approach with the adversarial training, we run the single adversarially trained VSA DNN on a single GPU. Since we focus on short-term VSA in this paper, we only measure the computation performance of inferencing on the input sample with batch size = 1.

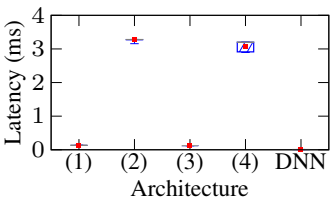
Table. 13a summarizes the number of parameters and the floating-point operations (FLOPs) for the four VSA ensemble variants as well as the single adversarially trained VSA DNN. Fig. 13b presents the execution latency for different VSA DNN architectures. From the results, we can see that the VSA ensemble variants (1) and (3) achieve less than 1 ms latency, which is comparable with running the single adversarially hardened VSA DNN. The variants (2) and (4) have higher latency of around 3 ms, which is intuitive because the two variants consist of more parameters and require more FLOPs when inferencing. However, this latency is still relatively low given that the time duration for the input sample is one second.

7.3 Summary of Results

We summarize the defense performance of our approach by three metrics: **Defense success rate (DSR)** measures the rate of detecting or correctly classifying the undetected adversarial examples; **Detection rate (DR)** measures the percentage of adversarial examples being detected; and **Thwarting rate (TR)** measures the percentage of the adversarial examples that are undetected but correctly classified by the attack thwarting. Note that DSR = DR + TR. Table. 4 summarizes

Arch	Params	FLOPs
(1)	16	53
(2)	128	56,073
(3)	16	53
(4)	5,114	56,073
DNN	128	1,402

(a) The parameter number and FLOPs in 10^6 for different VSA DNN architectures.



(b) Execution latency versus different VSA DNN architectures.

Fig. 13: Computation performance evaluation results. (1)-(4) denote the VSA ensemble variants explained in Section 7.1.2. "DNN" represents the single adversarially trained VSA DNN.

the results. We can see that our approach using the MC-dropout DNN units can achieve 100% of DSR with a relatively low FPR of less than 8%. Besides, our measurement in Section 7.2.3 shows running the MC-dropout DNN units achieves inference latency that is comparable with running the single adversarially harden DNN. Thus, we recommend to use the MC-dropout DNN units as the VSA ensemble in counteracting adversarial examples on VSA.

8 Conclusion

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 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 more credible universal adversarial examples but fail in thwarting certain less credible input-specific adversarial examples. Since the possibility of such attack cannot be completely ignored, we propose an approach using an ensemble to jointly detect adversarial examples based on predictive uncertainty and thwart adversarial examples using majority vote. The evaluation shows our approach using the MC-dropout DNN units can effectively defend against all the adversarial examples. Specifically, compared with the PGD adversarial training that achieves accuracy as low as 1.4% under certain less credible attack, our approach using the MC-dropout DNN units can achieve 100% defense success rates for all the six attacks, including both the more credible and less credible attacks. Meanwhile, the average execution latency for the MC-dropout DNN units is less than 1 ms, which is comparable with that of the adversarially harden DNN. The credibility analysis methodology adopted in this paper can also be applied to other types of adversarial example attacks and power grid applications.

Acknowledgment

Rui Tan's work was 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. Hoay Beng Gooi's work was supported by the Department of the Navy, Office of Naval Research Global under ONRG Award N62909-19-1-2037.

REFERENCES

[1] P. Kundur, N. J. Balu, and M. G. Lauby, *Power system stability and control*. McGraw-hill New York, 1994, vol. 7.

[2] M. S. Mahmoud, H. M. Khalid, and M. M. Hamdan, *Cyberphysical Infrastructures in Power Systems: Architectures and Vulnerabilities*. Academic Press, 2021.

[3] Y. Chen, Y. Tan, and B. Zhang, "Exploiting vulnerabilities of load forecasting through adversarial attacks," in *ACM e-Energy*, 2019.

[4] W. Liu, B. Tang, J. Han, X. Lu, N. Hu, and Z. He, "The structure healthy condition monitoring and fault diagnosis methods in wind turbines: A review," *Renewable and Sustainable Energy Reviews*, vol. 44, 2015.

[5] D. W. Van der Meer, J. Widén, and J. Munkhammar, "Review on probabilistic forecasting of photovoltaic power production and electricity consumption," *Renewable and Sustainable Energy Reviews*, vol. 81, 2018.

[6] Z. Zhang, D. Zhang, and R. C. Qiu, "Deep reinforcement learning for power system applications: An overview," *CSEE Journal of Power and Energy Systems*, vol. 6, 2019.

[7] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *ICLR*, 2015.

[8] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," in *CCS*, 2009.

[9] C. Ren, X. Du, Y. Xu, Q. Song, Y. Liu, and R. Tan, "Vulnerability analysis, robustness verification, and mitigation strategy for machine learning-based power system stability assessment model under adversarial examples," *IEEE Trans. Smart Grid*, 2021.

[10] Y. Li, M. Zhang, and C. Chen, "A deep-learning intelligent system incorporating data augmentation for short-term voltage stability assessment of power systems," *Applied Energy*, vol. 308, p. 118347, 2022.

[11] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, 2018.

[12] "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/.

[13] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *ICLR Workshop*, 2017.

[14] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *CVPR*, 2016.

[15] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE S&P (Oakland)*, 2017.

[16] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *CVPR*, 2017.

[17] J. Hayes and G. Danezis, "Learning universal adversarial perturbations with generative models," in *IEEE S&P (Oakland) Workshop*, 2018.

[18] G. Jin, S. Shen, D. Zhang, F. Dai, and Y. Zhang, "Ape-gan: Adversarial perturbation elimination with gan," in *ICASSP*, 2019.

[19] U. Inayat, M. F. Zia, S. Mahmood, H. M. Khalid, and M. Benbouzid, "Learning-based methods for cyber attacks detection in iot systems: A survey on methods, analysis, and future prospects," *Electronics*, vol. 11, no. 9, p. 1502, 2022.

[20] S. Ashraf, M. H. Shawon, H. M. Khalid, and S. Muyeen, "Denial-of-service attack on iec 61850-based substation automation system: A crucial cyber threat towards smart substation pathways," *Sensors*, vol. 21, no. 19, p. 6415, 2021.

[21] R. Tan, H. H. Nguyen, E. Y. Foo, X. Dong, D. K. Yau, Z. Kalbarczyk, R. K. Iyer, and H. B. Gooi, "Optimal false data injection attack against automatic generation control in power grids," in *IEEE ICCPS*, 2016.

[22] S. Lakshminarayana, T. Z. Teng, D. K. Yau, and R. Tan, "Optimal attack against cyber-physical control systems with reactive attack mitigation," in *ACM e-Energy*, 2017.

[23] M. A. Rahman and A. Datta, "Impact of stealthy attacks on optimal power flow: A simulink-driven formal analysis," *IEEE Trans. Dependable Secure Comput.*, vol. 17, 2018.

[24] S. Bhattacharjee and S. K. Das, "Detection and forensics against stealthy data falsification in smart metering infrastructure," *IEEE Trans. Dependable Secure Comput.*, vol. 18, 2018.

[25] A. S. Musleh, H. M. Khalid, S. Muyeen, and A. Al-Durra, "A prediction algorithm to enhance grid resilience toward cyber attacks in wamcs applications," *IEEE Syst J*, vol. 13, 2017.

[26] Q. Song, R. Tan, C. Ren, and Y. Xu, "Understanding credibility of adversarial examples against smart grid: A case study for voltage stability assessment," in *ACM e-Energy*, 2021.

[27] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.

[28] C. Xie, Y. Wu, L. v. d. Maaten, A. L. Yuille, and K. He, "Feature denoising for improving adversarial robustness," in *CVPR*, 2019.

[29] W. Wei and L. Liu, "Robust deep learning ensemble against deception," *IEEE Trans. Dependable Secure Comput.*, vol. 18, 2020.

[30] M. Javaheripi, M. Samragh, B. D. Rouhani, T. Javidi, and F. Koushanfar, "Curtail: Characterizing and thwarting adversarial deep learning," *IEEE Trans. Dependable Secure Comput.*, vol. 18, 2020.

[31] T. Athay, R. Podmore, and S. Virmani, "A practical method for the direct analysis of transient stability," *IEEE Trans. Power Apparatus and Systems*, no. 2, 1979.

[32] P. Siemens, "Pss/e 33.0 program application guide: Volume ii," *Siemens PTI: Schenectady, NY, USA*, 2011.

[33] H. Renmu, M. Jin, and D. J. Hill, "Composite load modeling via measurement approach," *IEEE Trans. Power Syst.*, vol. 21, 2006.

[34] D. F. Kune, J. Backes, S. S. Clark, D. Kramer, M. Reynolds, K. Fu, Y. Kim, and W. Xu, "Ghost talk: Mitigating emi signal injection attacks against analog sensors," in *IEEE S&P (Oakland)*, 2013.

[35] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *NIPS*, vol. 30, 2017.

[36] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *ICML*, 2016.

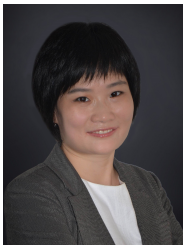
[37] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *NIPS*, vol. 30, 2017.

Qun Song received the Ph.D. degree in computer science from Nanyang Technological University, Singapore in 2022 and the B.Eng. degree in computer science from Nankai University, China in 2018. Her research interests include resilient sensing, edge computing, and cyber-physical systems. She is the recipient of IPSN'21 Best Artifact Award Runner-Up and NTU SCALE Best Demo Award. She serves on the technical program committees (TPCs) of 2023 ACM e-Energy and 2022 SenSys Shadow Program. She is a Member of IEEE.

Rui Tan is an Associate Professor at School of Computer Science and Engineering, Nanyang Technological University, Singapore. Previously, he was a Research Scientist (2012-2015) and a Senior Research Scientist (2015) at Advanced Digital Sciences Center, a Singapore-based research center of University of Illinois at Urbana-Champaign, and a postdoctoral Research Associate (2010-2012) at Michigan State University. He received the Ph.D. (2010) degree in computer science from City University of Hong Kong, the B.S. (2004) and M.S. (2007) degrees from Shanghai Jiao Tong University. His research interests include cyber-physical systems, sensor networks, and pervasive computing systems. He is the recipient of ICCPS'22 Best Paper Award Finalist, SenSys'21 Best Paper Award Runner-Up, IPSN'21 Best Artifact Award Runner-Up, IPSN'17 and CPSR-SG'17 Best Paper Awards, IPSN'14 Best Paper Award Runner-Up, PerCom'13 Mark Weiser Best Paper Award Finalist, and CityU Outstanding Academic Performance Award. He is currently serving as an Associate Editor of the ACM Transactions on Sensor Networks. He also serves frequently on the technical program committees (TPCs) of various international conferences related to his research areas, such as SenSys, IPSN, and IoTDI. He received the Distinguished TPC Member recognition thrice from INFOCOM in 2017, 2020, and 2022. He is a Senior Member of IEEE.

Chao Ren received the B.E. degree from the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China, in July 2017, and the Ph.D. degree from Interdisciplinary Graduate School, Nanyang Technological University, Singapore, in March 2022. Currently, he is a Research Fellow in School of Computer Science and Engineering, and conducts postdoctoral research with Wallenberg-NTU Presidential Postdoctoral Fellowship at Nanyang Technological University, Singapore. His research interests include adversarial machine learning, data-analytics, security assessment, and their applications to smart grid. He won several programming contest awards, including the Champion of Chinese Software Cup, NeurIPS Competition, and Golden Award of International College Student "Internet+" Competition. He is a Member of IEEE.

Yan Xu received the B.E. and M.E degrees in electrical engineering from South China University of Technology, Guangzhou, China in 2008 and 2011, respectively, and the Ph.D. degree in electrical engineering from The University of Newcastle, NSW, Australia, in 2013, all in electrical engineering. He is currently an Associate Professor at School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), and a Cluster Director at Energy Research Institute @ NTU (ERI@N), Singapore. Previously, he held The University of Sydney Postdoctoral Fellowship in Australia. His research interests include power system stability and control, microgrid, and data-analytics for smart grid applications. Dr Xu is an Editor for IEEE TRANSACTIONS ON SMART GRID, IEEE TRANSACTIONS ON POWER SYSTEMS, CSEE Journal of Power and Energy Systems, and an Associate Editor for IET Generation, Transmission & Distribution. He is a Senior Member of IEEE.



Jianping Wang received the B.S. and M.S. degrees in computer science from Nankai University, Tianjin, China, in 1996 and 1999, respectively, and the Ph.D. degree in computer science from the University of Texas at Dallas, Richardson, TX, USA, in 2003. She is currently a Professor at the Department of Computer Science, City University of Hong Kong, Hong Kong. Her research interests include autonomous driving, dependable networking, wireless networking, and cloud computing. She is a Senior Member of IEEE.

Yang Lou received the B.S. degree in Computer Science from the City University of Hong Kong, Hong Kong, in 2021. He is currently pursuing the Ph.D. degree in Computer Science from the City University of Hong Kong, Hong Kong. His research interests include autonomous driving security, uncertainty estimation, and sensor security.

Hoay Beng Gooi (LSM'20) received his Ph.D. degree from Ohio State University, Columbus, Ohio in 1983. He worked as Assistant Professor at Lafayette College, Easton, Pennsylvania during 1983-85 and Senior Engineer at Control Data - Energy Management System Division, Plymouth, Minnesota before joining Nanyang Technological University (NTU) in 1991, Singapore. He is an Associate Professor with the School of Electrical and Electronic Engineering. During 2008–14, he was Deputy Head of Power Engineering Division. He has been Co-Director of SP Group-NTU Joint Lab since 2020 and Chairman, LMAG, IEEE Singapore since 2021. He received Outstanding Associate Editor Award in 2021 for his contribution towards IEEE Transactions on Power Systems. He is a registered professional engineer in Pennsylvania, USA, and Singapore and serves in the Fundamentals of Engineering Examination (Electrical) sub-committee, Professional Engineers Board Singapore. His current research interests include microgrid energy management systems, energy storage, condition monitoring, electricity market, and spinning reserve.