Adversarial Attacks on Deep Neural Network-based Power System Event Classification Models

Yuanbin Cheng, Koji Yamashita and Nanpeng Yu
Department of Electrical and Computer Engineering
University of California, Riverside
Riverside, California 92507 USA
ychen871@ucr.edu, kyamashi@ucr.edu, nyu@ece.ucr.edu

***Abstract*—Online event classification is crucial to enhancing the reliability of the power transmission system. With the recent success of deep learning based methods in various domains such as computer vision and natural language processing, researchers started adopting these techniques to solve the power system event identification problem and achieved excellent results. However, the previous work does not consider the vulnerability of deep learning models to adversarial attacks, which could potentially make them unreliable in real world applications. In this paper, we adopt several adversarial attack mechanisms by adding tailored noise signal to the input Phasor Measurement Units (PMU) time series and make the deep learning model misclassify the power system event. Our results reveal that current state-of-the-art deep learning based power system event classifiers are extremely vulnerable to adversarial attacks, which may jeopardize the reliability of the power transmission system.**

***Index Terms*—Adversarial attacks, deep learning, event identification, phasor measurement unit.**

I. Introduction

Phasor Measurement Units (PMUs) have been increasingly deployed in many countries mainly to enhance situational awareness in the bulk power system. The high sampling rate of PMUs enabled the development of data-driven power system event detection and classification algorithms. The provision of large-scale real-world PMU dataset and event labels by the U.S. Department of Energy has advanced the field of deep learning-based power system event identification and classification. A deep convolutional neural network (CNN) with information loading and graph signal processing (GSP)-based sorting algorithms is proposed to classify power system events using PMU data [1]. Different variations of CNN-based models are designed to classify power system events, for instance, [2] designs the spatial pyramid pooling (SPP)-aided CNN, and [3] establishes a hierarchical CNN.

Although deep neural network-based classifiers achieved great success in many fields, they were shown to be vulnerable to adversarial attacks [4], i.e., small modifications to the input data may lead to misclassification [5]–[9]. Most of adversarial attacks on neural networks literature focused on image recognition tasks. The existing work studied how to create a modified/adversarial image, which will be misclassified by the neural network. A fast gradient-based attack [5] was developed as an alternative to expensive optimization techniques [4], where the authors hypothesized that the presence of such adversarial examples are due to the linearity for deep learning models. This adversarial attack was extended by a more costly iterative procedure [6]. Another iterative method to compute a minimal norm adversarial perturbation for a given image was proposed [7], where the authors also introduced a metric to quantify the robustness of the classifiers.

While adversarial attack has been extensively researched in the image recognition field, it has not been well studied in the sub-field of deep learning-based power grid event identification. Cyberattacks against the power grid such as false data injection attacks on the state estimation, protection, and control sub-modules of the power system have received a lot of attention from researchers. However, little attention has been paid to attacks against deep neural network-based monitoring tools in the bulk power systems.

In this paper, we leverage adversarial attack schemes that are effective on image datasets to modify the PMU data and try to quantify the robustness/vulnerability of the state-of-the-art deep learning-based power system event classification models. We also perform a large-scale case study using the real-world PMU data during power system events in the Eastern and Western Interconnections of the U.S. transmission grid to demonstrate that the deep learning-based event classification models are prone to adversarial attacks.

Adversarial attacks can be divided into white-box attacks and black-box attacks based on whether the deep learning model and dataset are available to the attacker [10]. In this initial study on the topic of adversarial attacks against power system event classification models, we explore white-box attacks and plan to investigate black-box attacks in the future.

The main contributions of this paper are:

- We define and formalize the adversarial attacks on deep

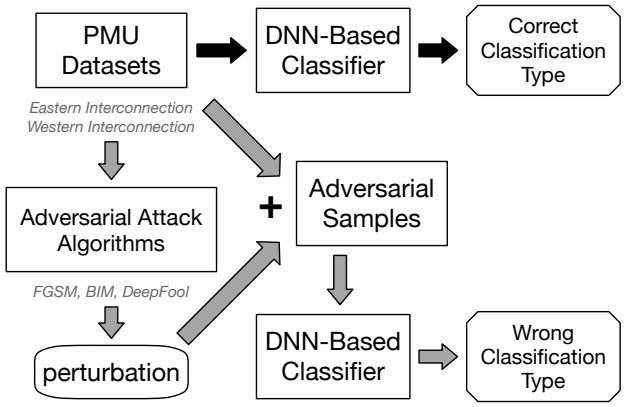

Fig. 1. Adversarial attacks on power system event classifiers with PMU data.

neural network-based power system event classification models that use PMU dataset.

- We demonstrate that the state-of-the-art deep learning-based power system event classification algorithms are extremely vulnerable to adversarial attacks.
- We quantify the robustness of different deep learning-based event classification models by calculating the minimum false data injection level to fool them, which heavily depends on the original and target event class/label.

The rest of this paper is organized as follows: Section II introduces the overall framework of adversarial attack on power system event classification models and the key notations. Section III presents the technical methods of three adversarial attack algorithms. Section IV quantifies the effectiveness of the adversarial attacks using a large-scale power system event datasets. Section V concludes the paper.

II. Overall Framework and Key Notations

We first present the overall framework of adversarial attack on the deep learning-based event classification models in the power system. Then, we provide the notations for the power system event classification models and the adversarial attacks.

A. Overall Framework

The overall framework of our proposed method is illustrated in Fig. 1. Suppose we have a trained deep neural network-based classifier that correctly identifies the event type of most PMU data samples. The adversarial attack algorithm has the ability to create a small perturbation from the PMU data sample. By adding the subtle and imperceptible perturbation to the original PMU data sample, the well-trained deep neural network based classifiers will misclassify many power system events.

B. Key Notations

The key notations are summarized below for ease of understanding of the technical methods:

Notation 1: A PMU time series sample $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_T}]$ is an ordered set containing PMU measurements at time steps $t = 1, ...T$. $\mathbf{x_t}$ consists of measurements of active power (P), reactive power (Q), voltage magnitude (V), and frequency (F) from multiple PMUs at time step t. The sampling rate of the PMUs is 30 Hz, and each data sample has a 12-second (360 timestamps) window.

Notation 2: A power system events dataset, $\mathbf{D} = \{(\mathbf{X_1}, \mathbf{Y_1}), (\mathbf{X_2}, \mathbf{Y_2}), ..., (\mathbf{X_N}, \mathbf{Y_N})\}$, is a set consists of the PMU time series sample and event label pair. $\mathbf{X_i}$ is the ith PMU time series sample and $\mathbf{Y_i}$ is the corresponding one-hot encoded event class vector. There are four different classes: normal behavior (non-event), voltage-related event, frequency-related event, and oscillation event.

Notation 3: $f(\cdot)$ denotes a deep learning-based event classification model. It maps the input $\mathbf{X_i}$ to a vector $\hat{Y}$, which reflects the the confidence of the corresponding class.

Notation 4: $J_f(\cdot, \cdot)$ denotes the cross-entropy loss function of the deep neural network-based classification model f.

Notation 5: X' represents the adversarial example comprised of a perturbation and X (the original data sample).

III. Technical Methods

First, we describe the design of a state-of-the-art CNN-based power system event classifier. Then, we present three adversarial attacks that are used to generate adversarial PMU time-series examples to fool the above-mentioned classifier.

A. CNN-based Power System Event Classifier

The CNN-based neural network classifier used in this paper was proposed in [1], and it combined the residual network, the graph signal processing (GSP)-based PMU sorting algorithm, and the information loading-based regularization techniques.

1) Residual Network: This classifier estimates the event type, taking in the pre-processed streaming PMU data. The architecture of the classifier is a CNN, called ResNet-50 [11].

The ResNet-50 may be split into an encoder and an estimator in terms of functions. The encoder is a CNN that primarily consists of the input layer, a max-pooling layer, a series of different convolutional building blocks, and a global average pooling layer.

Mathematically, a building block can be expressed as: $Y_i = g(U_i, \theta_\mathbf{i}) + U_i$, where U_i and Y_i are the input and output of the i-th block, respectively. $\theta_\mathbf{i}$ is the parameter vector for the i-th block. $g(\cdot)$ is the nonlinear activation function.

The estimator is the last layer of the ResNet-50, which is a fully connected layer with outputs normalized by the softmax function. The classifier is trained using the categorical cross-entropy loss function. The training is performed with stochastic gradient descent and the Adam optimizer [12].

2) GSP-based PMU Sorting: Convolutional layers use convolution filters to process local information. They function perfectly, especially when the local patches of data are highly correlated. The GSP-based sorting algorithm rearranged the sequence of PMUs, where the PMUs with high correlations will be placed close to each other [1].

The detailed derivation of the GSP-based sorting algorithm can be found in [1]. We only show the procedure of this algorithm in this paper. The GSP-based PMU sorting algorithm can be summarized in the following four steps:

(1) Derive the Pearson correlation coefficients between PMUs' measurements in the interconnection.

(2) Obtain the graph Laplacian matrix L.

(3) Execute eigenvalue decomposition on L.

(4) Sort PMUs according to eigenvector that corresponds to the second smallest eigenvalue of the L.

3) Information Loading-based Regularization: The information loading-based regularization technique is motivated by the recent theory related to the information losses through the neural classifier [13]. The information loading-based regularization can adjust the information quantity between the input and the hidden representation of a deep neural network. Like the GSP-based PMU sorting, we do not include this technique's detailed problem formulation and derivation in this paper. Interested readers can find them in [1].

The information loading-based regularization technique designs an estimator for estimating the mutual information between the input and last hidden layers. Moreover, the estimated mutual information is added as a weighted penalty term to the original cross-entropy loss function to control the information compression of the classifier. In this paper, we trained classifiers with the same parameter settings in [1].

B. Three Adversarial Attack Algorithms

This subsection provides the details of the three different adversarial attack algorithms that were adopted in this paper to fool the power system event classifiers.

1) Fast Gradient Sign Method: The fast gradient sign method (FGSM) was proposed in [5] to generate adversarial images that fooled the well-known GoogLeNet model. The FGSM attack updates the sample along the direction of the gradient's sign. The perturbation process is expressed as:

$$\eta = \epsilon \cdot sign(\nabla_x J(X, Y_{true})), \tag{1}$$

where ϵ denotes the magnitude of the perturbation (a hyperparameter). The adversarial event sample, X', is easily generated with $X' = X + \eta$. The gradient is efficiently computed using back-propagation.

Targeted Fast Gradient Sign Method: Instead of making the model misclassify the given sample, an extension of the FGSM approach demonstrated by [9], shows that the algorithm was able to specify a class as the attack target. This enhanced method is called the "targeted FGSM". The adversarial example is crafted using the following equation:

$$X' = X - \epsilon \cdot sign(\nabla_x J(X, Y_{target})) \tag{2}$$

2) Basic Iterative Method: The basic iterative method (BIM) [6] evolves the FGSM by changing the one-step update to the multiple small-step updates. Moreover, it clips the accumulated perturbations after each step to limit its l_∞. The result of the image dataset shows that this iterative perturbation is more negligible compared to FGSM.

Algorithm 1 shows the procedure of this iterative attack which requires three hyper-parameters: (1) the number of iterations, I; (2) the amount of maximum perturbation, η; and (3) the per step small perturbation, α. We set I = 100, η = 0.1 and $\alpha = 0.005$ in our experiment.

Algorithm 1 Basic Iterative Method (BIM)

Parameter: I, η, α
Input: event sample X, label Y_{true}
Output: perturbed event sample X'

X' = X
for i = 1 to I **do**
  untargeted: $\eta = \epsilon \cdot sign(\nabla_x J(X, Y_{true}))$;
  targeted: $\eta = \epsilon \cdot sign(\nabla_x J(X, Y_{target}))$;
  untargeted:$X' = X' + \eta$;
  targeted: $X' = X' - \eta$;
  $X' = min(X + \epsilon, max(X - \epsilon, X'))$;
end for

Like the targeted FGSM, as shown in Algorithm 1, BIM can also be extended to the targeted BIM by changing the calculation of the perturbation, η, and the adversarial example, X', for each iteration.

3) DeepFool: The DeepFool [7] is a cutting-edge technique that can effectively deteriorate deep neural network-based image classification performance. The DeepFool can efficiently estimate the minimal norm adversarial perturbation that makes the trained classifier misclassify the given sample.

DeepFool is an iterative method that accumulates the perturbation until it successfully makes the classifier misclassify. Each step of the DeepFool contains two procedures. First, it looks for the closest decision boundary from all classes except the label. Then, it updates the sample by orthogonally projecting to that closest decision boundary. Due to space limitations, this paper does not include the implementation details of DeepFool, which can be found in [7]. Note that the Deepfool algorithm does not need any parameters.

IV. Numerical Study

In this section, we validate the three adversarial attack algorithms with the myriad real-world PMU datasets in the Eastern and Western Interconnections of the United States. Also, we quantify the robustness of the trained event classifiers, and we clarify the impact of the GSP-based sorting algorithm and information loading techniques on the robustness.

First, we briefly describe the PMU data used in this work. Then, we evaluate the untargeted and targeted performance of the three adversarial attacks and analyze the impact of the GSP-based sorting algorithm and information loading techniques. Finally, we quantify the robustness of neural network-based power system event classifiers.

A. Data Source

Each dataset contains the PMU measurements from a separate transmission network with 179 valid PMUs in the Eastern interconnection and 41 valid PMUs in the Western interconnection of the U.S. The raw measurement data contains the sequence of the voltage phasor, current phasor, and frequency. We follow the same procedure in Section III-F of [14] to

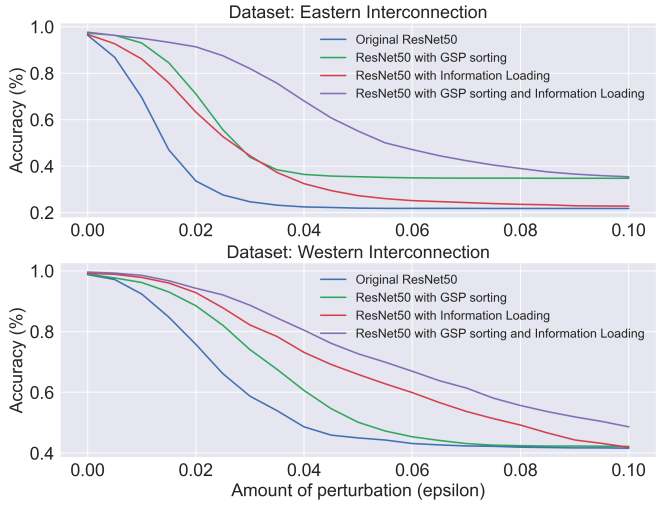

Fig. 2. Power system event detection accuracy with respect to the amount of perturbation for FGSM attacks on Eastern and Western Interconnection.

clean raw data and convert this cleaned data to the tensor that consists of active power P, reactive power Q, voltage magnitude $|V|$, and frequency F. The pre-processing data pipeline includes the bad PMU removal using the PMU status flag or outlier thresholds, the missing data replacement, and the real and reactive power calculation. The event labels were generated from the electric utility and network operators' event log. There are a total of 1,147 (1,204) labeled PMU data samples in the dataset, and are separated into four different types: 825 (625) line events (voltage events), 84 (333) generator events (frequency events), 118 (147) oscillation events and 120 (99) normal system operation behavior, for the Eastern (Western) Interconnection. Each sample in the dataset is a 12-second window. The sampling frequency of the PMUs is 30 Hz. Therefore, the shape of each PMU data sample in Eastern (Western) Interconnection dataset is [360; 179; 4] ([360; 41; 4]). These three dimensions correspond to the timestamp (360), the number of PMUs 179 (41), and the four measurement channels: P, Q, $|V|$, and F.

The datasets are divided into training datasets (80% of the total samples) and testing datasets (20% of the total samples). The event classifier is trained on a machine with four RTX 2080 Ti GPU, the batch size of 16, and 200 training epochs.

B. Adversarial Attack Performance

1) Fast Gradient Sign Method (FGSM): Figure 2 shows the change of the accuracy in the testing dataset with the increasing amount of the perturbation ϵ added to every sample. All four of these classifiers are extremely vulnerable to the adversarial attack. When the perturbation fraction of ϵ increases from 0 to 0.1, the accuracy of system event classification algorithms plummet below 50% for all these four classifiers.

We also compare the performance between the classifiers with and without the techniques used in [1]: GSP-based PMU sorting algorithm and information loading-based regularization. It is observed that both the GSP-based PMU sorting algorithm and information loading-based regularization contribute to increasing the resistance to adversarial attacks although this increase is quite limited.

TABLE I
AVERAGE PERTURBATION l_2 NORM FOR DIFFERENT METHOD WITH 50% OF MISCLASSIFICATION ON EASTERN/WESTERN INTERCONNECTION

Classifier	FGSM	BIM	DeepFool
ResNet50 (RN50)	5.95/5.92	5.37/4.93	2.69/3.05
RN50-GSP	10.61/7.89	8.96/6.59	4.97/3.96
RN50-Info	13.76/15.06	6.91/7.94	3.54/4.83
RN50-GSP-Info	27.34/17.95	13.24/8.26	6.86/5.17

2) Average l_2 Norm for Three Adversarial Attack Algorithms: Table I shows the necessary amount of the perturbation of three adversarial attack algorithms that makes the power system event classifier misclassify 50% of the PMU samples in the Eastern/Western Interconnection. The average l_2 norm is used as the scalar to quantify the amount of perturbation.

Although the classifiers are highly vulnerable to the FGSM algorithm, as shown in this table, the BIM and DeepFool algorithms can fool the classifier with even smaller perturbations. Figure 3 displays an example of the perturbation generated by DeepFool algorithm that fools the RN50-GSP-Info classifier from identifying a normal behavior to a generator event. It is evident that the injected perturbation is imperceptible. Table I also reveals that the ResNet-50 combined with GSP-based PMU sorting and information loading-based regularization has the most effective resistance to adversarial attacks, which justifies that this classifier shows the highest distinctive competence for power system events.

3) Targeted Adversarial Attack: As described in Section III-B, the FGSM and BIM algorithms can be enhanced so that we can freely/actively specify the target class.

TABLE II
AVERAGE l_2 PERTURBATION TO FOOL CLASSIFIER FROM THE LABEL TO THE TARGET CLASS BY TARGET BIM ON WESTERN INTERCONNECTION

Label \ Target	**Normal**	**Voltage**	**Frequency**	**Oscillation**
Normal		3.93	3.67	3.16
Voltage	9.05		6.13	7.85
Frequency	5.18	4.94		5.91
Oscillation	3.20	4.29	4.376	

We adopt the targeted BIM algorithm and derive the average l_2 of the perturbation to fool a sample from the classified label to the target class, as shown in Table II. The classifier used in this experiment is RN50-GSP-Info. This table demonstrates that the voltage event is the hardest to be fooled compared to other types, which is in line with the domain expert's intuition because only the voltage event data generally include single or more significant impulse responses.

C. Power System Event Classifier Robustness Quantification

The robustness of the classifier, f, at point X may be defined as the norm of the minimal perturbation, r, that can fool the classifier [7]. The robustness ρ of the classifier,

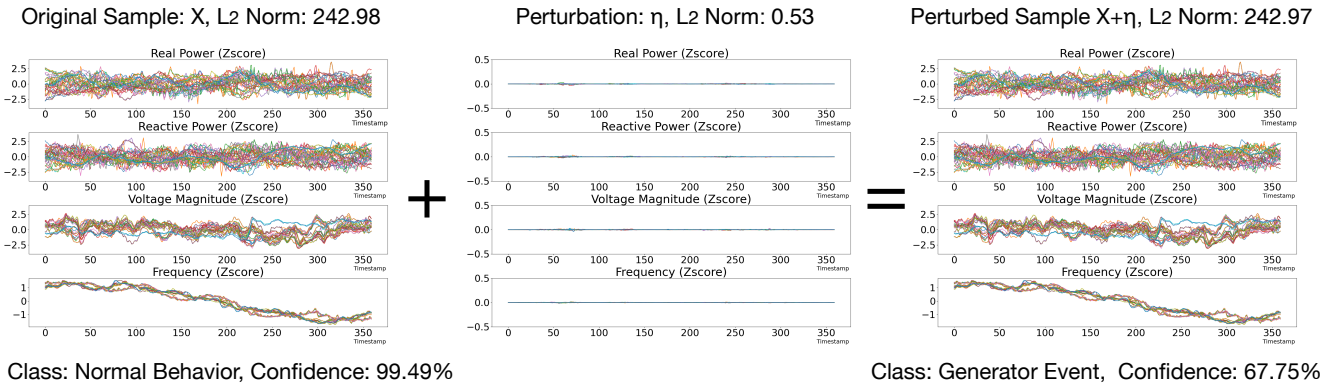

Fig. 3. Example of a tiny perturbation computed by DeepFool that make the model misclassify from normal behavior to generator event.

f, is expressed as the point robustness expectation over the distribution of data:

$$\rho = E_X \frac{\Delta(X, f)}{||X||_2} \tag{3}$$

The minimal normal perturbation $\Delta(X, f)$ to fool the classifier can be estimated via the DeepFool algorithm in (5).

$$\Delta(X, f) = min_r ||r||_2, f(X + r) \neq f(X) \tag{4}$$

TABLE III
CLASSIFIER'S ROBUSTNESS AGAINST DEEPFOOL PERTURBATION

Classificer	Robustness (East)	Robustness (West)
ResNet50 (RN50)	0.0099	0.0221
RN50-GSP	0.0137	0.0249
RN50-Info	0.0194	0.0347
RN50-GSP-Info	0.0245	0.0385

Table III quantifies the robustness of four trained classifiers for Eastern and Western interconnection data. It can be seen that both GSP-based sorting and information loading-based regularization increase classifiers' robustness. Nevertheless, the robustness of all four of these classifiers indicates that the perturbation with the l_2 norm, less than 4% of the sample, will make the classifiers completely useless. Therefore, significant development is needed to improve the robustness of the state-of-the-art neural network-based power system event classifiers.

V. CONCLUSION

This paper adopts three adversarial attacks to fool power system event classifiers and evaluates them on the large-scale real-world Phasor Measurement Units (PMU) dataset. We showed that adding small perturbation signals to the PMU dataset could significantly degrade the performance of the state-of-the-art power system event classifiers. Our result manifests the current power system event classifiers' vulnerability and reveals the necessity of future research to design more robust deep learning models to classify power system events.

VI. ACKNOWLEDGMENTS

This material is based upon work supported by the Department of Energy under Award Number DE-0000916.

REFERENCES

[1] J. Shi, B. Foggo, and N. Yu, "Power system event identification based on deep neural network with information loading," *IEEE Trans. Power Syst.*, vol. 36, no. 6, pp. 5622–5632, 2021.

[2] Y. Yuan, Y. Guo, K. Dehghanpour, Z. Wang, and Y. Wang, "Learning-based real-time event identification using rich real PMU data," *IEEE Trans. Power Syst.*, vol. 36, no. 6, pp. 5044–5055, 2021.

[3] M. Pavlovski, M. Alqudah, T. Dokic, A. A. Hai, M. Kezunovic, and Z. Obradovic, "Hierarchical convolutional neural networks for event classification on PMU measurements," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16*, 2014.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9*, 2015.

[6] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26*, 2017.

[7] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2574–2582, 2016.

[8] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22*, pp. 1625–1634.

[9] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26*, 2017.

[10] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: A survey," *CoRR*, vol. abs/1810.00069, 2018. [Online]. Available: http://arxiv.org/abs/1810.00069

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30*, pp. 770–778.

[12] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9*, 2015.

[13] B. Foggo, N. Yu, J. Shi, and Y. Gao, "Information losses in neural classifiers from sampling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 10, pp. 4073–4083, 2019.

[14] Y. Cheng, N. Yu, B. Foggo, and K. Yamashita, "Online power system event detection via bidirectional generative adversarial networks," *IEEE Trans. Power Syst.*, 2022.